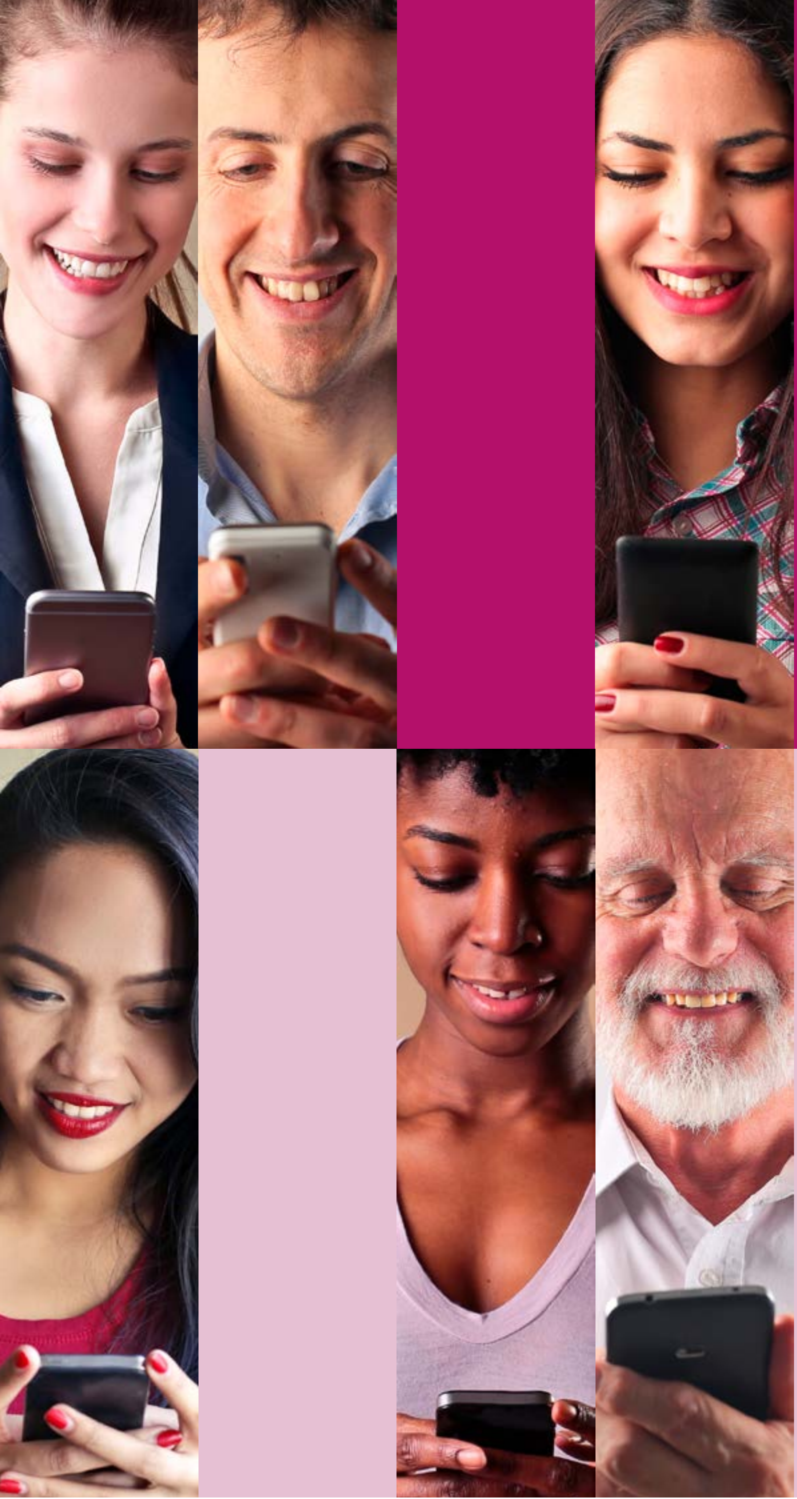


MODELS OF GOVERNANCE OF ONLINE HATE SPEECH

On the emergence of collaborative governance and the challenges of giving redress to targets of online hate speech within a human rights framework in Europe

Alexander Brown

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

Models of Governance of Online Hate Speech

On the emergence of collaborative governance and the challenges of giving redress to targets of online hate speech within a human rights framework in Europe

Alexander Brown

Council of Europe

Cover and layout: Documents and Publications Production Department (SPDP), Council of Europe
Photo: Shutterstock

Online published by the Council of Europe

Table of Contents

EXECUTIVE SUMMARY

For some users of the Internet, even if it is a minority of people, it would not be an exaggeration to speak of the current Internet epoch as being the Internet of Hate. Partly reflecting the scale and seriousness of the problem of online hate speech, the past three years or so has seen several innovations in governance tools for online hate speech across Europe. New governance tools have been proposed and developed, very often through collaboration, by national governments, intergovernmental organisations (such as the Council of Europe and the European Commission), Internet platforms and civil society organisations. Some of these tools are in their infancy, others are yet to be implemented, still more are in the design, planning and final approval stages. These tools must operate within a human rights framework, which for Council of Europe member states is set by the European Convention of Human Rights, the European Court case law, the additional protocol to the Cybercrime convention and other Council of Europe standards including, those of dedicated monitoring bodies, such as the European Commission against Racism and Intolerance and its General Policy Recommendations Nr. 6 and 15.

This six month study (June to December 2019, and updated in April 2020) was commissioned by the Council of Europe, Directorate General of Democracy. It seeks to map and explain but also to critically evaluate these emerging innovations. The study in its review covered among others:

- NetzDG Act in Germany [see sections I.C(ii), I.D, I.F, I.G, IV.A, IV.C, VI.D(iv), VII.C(iii), IX.1, IX.2];
- Avia Bill in France (Draft Law to Fight Against Hate Content on the Internet) [see sections I.C(ii), I.D, I.G, IV.C, IV.G, VII.C(i), VII.C(iii), IX.1];
- Bill on the prevention of undesirable behaviour on social networks in Croatia [see sections IV.C];
- Online Harms White Paper in the UK [see sections I.C(ii), I.G, IV.E];
- Agreement establishing a working procedure between trusted flaggers, a special public prosecutor for digital crimes and Internet platforms forthcoming in a member state of the European Union (anonymous) [see sections I.G, IV.F];
- Proposal for lesser sentences for persons convicted of hate speech offences if the criminal offences were committed on social media in Spain [see section IV.H];
- European Commission's Code of Conduct on Countering Illegal Hate Speech Online [see sections I.B(iv), I.E, IV.B, V.A, V.B(i), IX.6].

For their part, some Internet platforms developed their own governance tools for tackling hate speech (and other forms of harmful content) on their platforms, services, websites and products. First, the vast majority of Internet platforms publish "community standards" or "content policies" that prohibit users from posting or sharing "hate speech" content. Some examples of the relevant definitions of hate speech [see section I.E.]. Second, some Internet platforms have introduced forms of oversight to check and guide how they undertake moderation of online hate speech which this study also addresses, for example:

- Twitter's Trust and Safety Council which was announced in 2016 [see section III.C];
- Facebook's Oversight Board [see sections I.B(iv), I.F, III.C, III.D, V.B(ii), VII.C(ii), IX.2, IX.7].

The study, reflecting on these and other examples, identifies three main levels of governance for online hate speech: the moderation level, the oversight level, and the regulatory level [see section I.C(i)]. It also maps more than 20 different model types of governance tools split across the three main levels, as well as numerous subtypes or variants of these main model types [see sections

II, III, IV]. The study also uncovers the goals, aims, values and expectations of governmental agencies, Internet platforms, civil society organisations and the general public when it comes to the governance of online hate speech [see section VI].

In addition to this, the study identifies 30 separate indicators or measures that could be used by monitoring bodies or other stakeholder organisations to assess the success or progress of different governance tools for online hate speech [see section VIII.A]. The study also highlights 6 factors that are important when monitoring bodies or other stakeholder organisations are selecting which indicators or measures to use [see section VIII.C].

Furthermore, the study examines in detail several important themes or policy issues that cut across recent developments in the story of governance of online hate speech in Europe. These include: the standardization or "common standards" agenda for governance of online hate speech [see sections I.A(ii), I.B(iii), I.D, IX.1]; how best to deal with grey area cases of online hate speech within governance structures [see sections I.E, I.F, II.F, III.D, III.E, IV.C(i), VIII.A, IX.2]; the benefits and challenges of collaborative approaches to governance of online hate speech [see sections V, IX.4]; the need for a victim-sensitive approach to governance of online hate speech within a human rights framework [see sections VII, IX.7].

The terms and scope of the current study testifies to an emphasis by the Council of Europe on "redress" and "victim-sensitivity". Redress is the idea that a core part of the function of governance tools for online hate speech is to provide a means or mechanism for individuals or groups who are targeted or adversely affected by online hate speech to report content, appeal decisions, assert grievances, lodge complaints, seek administrative, civil or criminal remedies, or in some other way claim or pursue resolution or rectification. Victim-sensitivity in turn is about the design and implementation of governance tools for online hate speech being sensitive to the needs and experiences of victims

Finally, the study also draws a series of conclusions and makes a large number of practical recommendations covering ten key areas of the governance of online hate speech:

1. Standardization agenda [see section IX.1]
2. Grey area cases [see section IX.2]
3. Public opinion [see section IX.3]
4. Collaboration [see section IX.4]
5. Mitigating the incentive to over-remove hate speech content [see section IX.5]
6. Monitoring voluntary codes of conduct [see section IX.6]
7. A victim-sensitive approach [see section IX.7]
8. Proactive use of text extraction and machine learning tools [see section IX.8]
9. Indicators of success in the governance of online hate speech [see section IX.9]
10. Equitable sharing in the governance of online hate speech [see section IX.10]

I. CONTEXT

For some people hate speech casts a pall over their lived experience of the Internet: the bits of cyberspace they inhabit, the posts they see and the messages they receive. For these Internet users, even if it is a minority of people, it would not be an exaggeration to speak of the current Internet epoch—following the Internet of Content, the Internet of Services, the Internet of People and the Internet of Things—as being the Internet of Hate.

Evidence of online hate speech comes from numerous different social, institutional, legal and academic sources not least from the extensive country monitoring reports of the European Commission against Racism and Intolerance (ECRI), summarised in Bakalis (2015). This particular study, however, is focused less on summarising the evidence of online hate speech and more on emerging innovations in the governance of online hate speech across Europe. While considering these innovations the study also highlights the need for the governance of online hate speech to operate within a human rights framework.

A. Impetus for the study

This study was commissioned by the Council of Europe, Directorate General of Democracy, which supports the Council of Europe in fields which are vital for the sustainability of democracy, including but not limited to ensuring respect for human dignity without discrimination on the basis of human rights standards.

In particular, the study seeks to map and explain but also to critically evaluate emerging innovations in the governance of online hate speech across Europe with a particular focus on the instruments and tools used by Internet platforms, civil society organisations, and governmental authorities—increasingly in collaboration with each other as multi-stakeholders—in the moderation, oversight and regulation of online hate speech. Particular attention is also placed on the role of redress and victim-sensitivity within a human rights framework.

The impetus for both the emerging innovations and the need to study them can be partly understood in terms of two basic drives: a widespread demand for action against online hate speech and a standardization agenda for the governance of online hate speech.

(i) Demand for action

On 29 November, 2018 the European Commissioner for Security Union, Julian King, gave a speech in which he put Internet platforms on notice about the removal of online hate speech (amongst other things). He stated:

> We can communicate, share information, shop, bank, conduct business, all with an ease and at a volume we have never seen before. But our cyber world also has this darker flip side; a proliferation of illegal content, of hate speech, of malign attempts to sway the way we think; to use our own online lives against us.
>
> These two faces of the digital age are inevitable, given how interconnected our daily lives have become. We need to be honest about the risks, and we need to be ready to act. We cannot afford the internet to be a Wild West where anything goes.

> I haven't used the word 'regulation' yet—but it is important to emphasise that the EU's approach to the digital world is not about attempting to limit its possibilities. It is about preserving trust, making sure that the rule of law applies online as elsewhere.
>
> […] We have seen a widespread demand for action—from the public, from the European Parliament and from Member States. We have reached something of a watershed—one where platforms need to shape up and begin acting responsibly in these different areas, or we will be obliged to legislate to see that they do.[1]

Arguably at the national level the watershed moment arrived on 1 January 2018 when Germany's Network Enforcement Act (NetzDG) came into effect. NetzDG requires "social networks" to remove (delete/block) content that is "manifestly unlawful" within 24 hours of receipt of a complaint, report or flag about the content, and to remove content that is "unlawful" within 7 days. It also creates administrative offences for social networks failing to set up procedures in conformity with the Act's general requirements including but not limited to content removal. (Procedural responsibilities also include a duty to publish "transparency reports" every 6 months, for example.) Fines can be levied for these administrative offences in the millions of Euros. Under the Act, Internet platforms have a responsibility to remove manifestly illegal content when it is reported by bodies or users whose place of residence is located in Germany, even if the content was uploaded, posted or shared by a person in a third country.[2]

In July 2019, for example, Germany's Federal Office of Justice (BfJ) fined Facebook 2 million Euros, among other things, because its NetzDG reporting form was too difficult to find.(ECRI 2020: para. 76)

The NetzDG Act is in one sense Germany taking out an insurance policy against online hate speech: namely, not simply leaving it to trust that Internet platforms will undertake thorough, effective and timely removal of unlawful hate speech content—whether directly via Internet platforms' legal compliance teams or indirectly through their content moderation practices—but making it a legal requirement, enforced or backed up with fines, for Internet platforms to remove unlawful hate speech. Indeed, in its 6th monitoring report on Germany, ECRI makes the following observation about NetzDG's positive effects:

> During the country visit, ECRI was informed about the positive effects of the NEA: the large social network providers have invested considerable resources in applying the law in an efficient manner. Many stakeholders confirmed that the most serious and open forms of hate speech have disappeared from the large platforms and thus do not any more reach the big number of their users. (ECRI 2020: para. 52)

Now the scope of this study is much broader than NetzDG and Facebook. Indeed, one of the key aims of the study is to uncover and highlight the importance of the wide diversity of both governance models and Internet platforms in Europe. Nevertheless, the point of starting with this specific example, and the quote from Julian King, is to illustrate the particular area or terrain of Internet governance (IG) that the study seeks to map. In general, Internet governance (IG)

[1] Speech available at: https://ec.europa.eu/commission/commissioners/2014-2019/king/announcements/financial-times-future-news-impact-social-media-abuse-european-society_en [last accessed 6 October, 2019].

[2] Interestingly, taking a nuanced approach to jurisdictional issues has a precedent in criminal law in Germany. For example, German courts have applied hate speech laws to creators of hate speech content uploaded to websites located in a third country (Alkiviadou 2016: 223-4).

comprises all of the processes of governing the Internet as a whole, whether through institutions, laws, protocols, guidelines, codes of practice, etc. But, as shall be explained in section I.C below, this study focuses on governance tools for online hate speech specifically.

It is worth highlighting at this stage that in its monitoring reports, the Council of Europe, specifically ECRI, has sometimes emphasised the need for governmental authorities to deploy, more rigorously, extant governance measures, such as criminal hate speech laws, in tackling the problem of online hate speech. For instance, in its 5th monitoring cycle report on Spain, ECRI makes the following recommendation:

> ECRI recommends that the Spanish authorities use their regulatory powers with regard to Internet and social media providers, reinforce the civil and administrative law protection against cyber hate speech and continue focusing on criminal investigation of cyber hate speech (ECRI 2018: para. 55).

ECRI has also, both in GPR No. 15 and in its country reports (e.g. ECRI 2017: para. 45), made general recommendations about tackling (online) hate speech that touch on several different areas of governance. For example:

> 7. use regulatory powers with respect to the media (including internet providers, online intermediaries and social media), to promote action to combat the use of hate speech and to challenge its acceptability, while ensuring that such action does not violate the right to freedom of expression and opinion, and accordingly:
>
>> a. ensure effective use is made of any existing powers suitable for this purpose, while not disregarding self-regulatory mechanisms;
>>
>> b. encourage the adoption and use of appropriate codes of conduct and/or conditions of use with respect to hate speech, as well as of effective reporting channels;
>>
>> c. encourage the monitoring and condemnation of the use and dissemination of hate speech;
>>
>> d. encourage the use, if necessary, of content restrictions, word filtering bots and other such techniques;
>>
>> e. encourage appropriate training for editors, journalists and others working in media organisations as to the nature of hate speech, the ways in which its use can be challenged;
>>
>> f. promote and assist the establishment of complaints mechanisms; and
>>
>> g. encourage media professionals to foster ethical journalism;
>
> 8. clarify the scope and applicability of responsibility under civil and administrative law for the use of hate speech which is intended or can reasonably be expected to incite acts of violence, intimidation, hostility or discrimination against those who are targeted by it while respecting the right to freedom of expression and opinion, and accordingly:

> a. determine the particular responsibilities of authors of hate speech, internet service providers, web fora and hosts, online intermediaries, social media platforms, online intermediaries, moderators of blogs and others performing similar roles;
>
> b. ensure the availability of a power, subject to judicial authorisation or approval, to require the deletion of hate speech from web-accessible material and to block sites using hate speech;
>
> c. ensure the availability of a power, subject to judicial authorisation or approval, to require media publishers (including internet providers, online intermediaries and social media platforms) to publish an acknowledgement that something they published constituted hate speech;
>
> d. ensure the availability of a power, subject to judicial authorisation or approval, to enjoin the dissemination of hate speech and to compel the disclosure of the identity of those using it;
>
> e. provide standing for those targeted by hate speech, equality bodies, national human rights institutions and interested non-governmental organisations to bring proceedings that seek to delete hate speech, to require an acknowledgement that it was published or to enjoin its dissemination and to compel the disclosure of the identity of those using it; and
>
> f. provide appropriate training for and facilitate exchange of good practices between judges lawyers and officials who deal with cases involving hate speech.[3]

Returning to Germany, however, it deserves mentions that the NetzDG Act did not emerge out of thin air. Back in March 2017 Germany's Federal Justice and Consumer Protection Minister, Heiko Maas, issued his own warning statement to Internet platforms operating in Germany. Citing a government-funded report that suggested a decline in the performance of Facebook in deleting or blocking unlawful content reported by users, including unlawful hate speech content, he declared:

> Therefore, it is now clear that we must further increase the pressure on social networks. We need legal regulations to make companies even more obligated to eradicate criminal offenses. (cited in Lomas 2017)

For its part, Facebook responded by publicly reaffirming its commitment to removing unlawful hate speech content as per its obligations under NetzDG:

> We have clear rules against hate speech and work hard to keep it off our platform. We are committed to working with the government and our partners to address this societal issue. By the end of the year over 700 people will be working on content review for Facebook in Berlin. (ibid)

This exchange raises several important questions. Is imposing fines on Internet platforms for a pattern of failure to remove unlawful hate speech content an appropriate and proportionate

[3] CRI(2016)15, Strasbourg, 8 December 2015. Available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 [last accessed 7 October, 2019].

means of pursuing regulatory goals? Are there other similarly effective but less restrictive means available? What are the correct performance measures for assessing the progress of Internet platforms in tackling online hate speech? If the measures are based on performance in removing unlawful online hate speech, how can the measures be accurate if the content in question has yet to be judged unlawful in a court of law? More generally, what is the point or purpose of the governance of online hate speech? Should it be to force Internet platforms to achieve a predetermined percentage of content that should be removed, that is, an ideal removal rate? Or instead should it be to encourage companies to invest proper resources and put fair procedures in place for moderating content irrespective of any ideal removal-rate? In other words, should performance measures be outcome-based or process-based? And, should performance measures include an assessment of the degree to which both content moderation and oversight of moderation are sensitive to the particular experiences and needs of targets of online hate speech ("victims")? These are some of the questions this study will investigate.

Yet further questions come to the fore in relation to "grey area" or "difficult" cases. These are cases where it is unclear whether or not a given piece of online content (a) is hate speech, (b) contravenes the Internet platform's own rules on permissible content (i.e. community standards or content policies on hate speech), and (c) is unlawful or illegal based on local hate speech laws (where such laws exist). Which governance models and tools for online hate speech are best equipped and most appropriate for dealing with grey area cases and why?

Meanwhile other key features of the Internet, namely, content, services, people, and things, have certainly not lessened in value and importance in recent years. They must also be protected, and invariably this means promoting and protecting the human right to freedom of expression online. And this too raises questions to be examined in this study. For example, what are suitable models and standards of governance for Internet platforms who remove large amounts of *lawful* hate speech content (both from public content areas and from "closed groups") because it is in breach of their "community standards" or "content policies"? Should Internet regulators make it a legal responsibility for Internet platforms to undertake moderation and oversight within international human rights standards including the right to freedom of expression? And how should any such legal responsibility be enforced?

Interestingly, efforts by governmental agencies and the courts not only in Germany but across many other European countries both inside and outside of the European Union, to "persuade" Internet platforms to suppress online hate speech date back at least to the beginning of the millennium (Frydman and Rorive 2002; Eberwine 2004; Banks 2011).

Arguably what has happened over the past two to three years, however, is that as well as intergovernmental organisations some national governments within Europe have become more resolute in their pursuit of this policy agenda, and have developed, and in some cases rolled out, sophisticated instruments to achieve it. Examples can be found in the following policies or policy proposals:

- NetzDG Act in Germany [see sections I.C(ii), I.D, I.F, I.G, IV.A, IV.C, VI.D(iv), VII.C(iii), IX.1, IX.2];
- Avia Bill in France (Draft Law to Fight Against Hate Content on the Internet) [see sections I.C(ii), I.D, I.G, IV.C, IV.G, VII.C(i), VII.C(iii), IX.1];
- Bill on the prevention of undesirable behaviour on social networks in Croatia [see sections IV.C];
- Online Harms White Paper in the UK [see sections I.C(ii), I.G, IV.E];

- Agreement establishing a working procedure between trusted flaggers, a special public prosecutor for digital crimes and Internet platforms forthcoming in a member state of the European Union (anonymous) [see sections I.G, IV.F];
- Proposal for lesser sentences for persons convicted of hate speech offences if the criminal offences were committed on social media in Spain [see section IV.H];
- European Commission's Code of Conduct on Countering Illegal Hate Speech Online [see sections I.B(iv), I.E, IV.B, V.A, V.B(i), IX.6].

For their part, some Internet platforms have taken "the bull by the horns" and developed their own governance tools for tackling hate speech (and other forms of harmful content) on their platforms, services, websites and products. First, the vast majority of Internet platforms publish "community standards" or "content policies" that prohibit users from posting or sharing "hate speech" content. Some examples of the relevant definitions of hate speech are listed in section I.E below.

Second, some Internet platforms have introduced forms of oversight to check and guide how they undertake moderation of online hate speech. For example:

- Twitter's Trust and Safety Council[4] which was announced in 2016 [see section III.C];
- Facebook's Oversight Board[5] [see sections I.B(iv), I.F, III.C, III.D, V.B(ii), VII.C(ii), IX.2, IX.7].

Such oversight of moderation is widely thought to be especially useful and appropriate for use in grey area cases, or what Facebook dubs "difficult" cases,[6] of online hate speech.

Third, most Internet platforms have "terms of service" (to which users must agree in order to use the platform, service, website or product), which state that users may not post or share "unlawful" or "illegal" content. Some examples are list in section I.E below. Typically this particular kind of "terms of service" is articulated in an unqualified way and so covers all forms of unlawful or illegal content including therefore unlawful or illegal *hate speech*.

Importantly, these Internet platforms typically also employ in-house "legal compliance" teams—and sometimes seek advice from external legal counsel—that monitor content and respond to reports, referrals or flags relating to content suspected of being illegal or unlawful and therefore breaching the Internet platform's terms of service on illegal or unlawful content.[7]

This emerging picture of Internet governance for online hate speech is complex and diverse. It testifies to the fact that key stakeholders are minded to act but also to variations in progress and approach across national governments and different Internet platforms.

As illustrated at the start of the section, both the European Commission and the Council of Europe wish to promote the idea that, when it comes to both Internet platforms and national

[4] Information available at: https://blog.twitter.com/en_us/a/2016/announcing-the-twitter-trust-safety-council.html [last accessed 6 October 2019].

[5] Facebook, Oversight Board Charter, September, 2019. Available at: https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf [last accessed 17 December, 2019].

[6] Facebook, Oversight Board Bylaws, January, 2020. Available at: https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf [last accessed 15 April, 2020].

[7] 1st consultative meeting, London, 17-18 October, 2019.

governments, their doing nothing to combat online hate speech cannot be defensible. This idea is shared by leading political figures working on these issues. For example, Laetitia Avia, a member of the French National Assembly representing La République En Marche!, is currently navigating the new Bill on tackling online hate speech, the so-called Avia Bill, through the French Senate. In her words, "we have to say that [keeping] the status quo is not an option".[8]

Furthermore, the terms and scope of the current study, commissioned by the Council of Europe, testifies to an emphasis by the Council on "redress" and "victim-sensitivity". Redress is the idea that a core part of the function of governance tools for online hate speech is to provide a means or mechanism for individuals or groups who are targeted or adversely affected by online hate speech to report content, appeal decisions, assert grievances, lodge complaints, seek administrative, civil or criminal remedies, or in some other way claim or pursue resolution or rectification. Victim-sensitivity in turn is about the design and implementation of governance tools for online hate speech being sensitive to the needs and experiences of victims.

(ii) Standardization agenda

As well as a demand for action, the current direction of travel seems to be towards creating a common set of standards or digital rulebook across Europe. For example, in 2019 the Committee of Ministers of the Council of Europe established a new interdisciplinary Committee of Experts mandated to draft a Committee of Ministers recommendation on a comprehensive approach to combating hate speech within a human rights framework, which among other will cover online forms of communication.

Similarly, the new President of the European Commission, Ursula von der Leyen, stated in her candidacy document the intention to replace the e-Commerce Directive (2000/31/EC) with "[a] new Digital Services Act [that] will upgrade our liability and safety rules for digital platforms, services and products" (Leyen 2019: 13). In particular: "We should develop a joint approach and common standards to tackle issues such as disinformation and online hate messages" (ibid: 21). It has been reported that very similar ideas are being discussed and promoted by the Digital Single Market (DSM) strategic group within the Directorate General for Communications Networks, Content and Technology (DG CONNECT) within the European Commission (Avram 2019). Moreover, the general policy idea that there should be "common efforts at European level to tackle the phenomenon [of online hate speech]" is also supported by some equality boards who currently operate as monitoring bodies under the European Commission's Code of Conduct monitoring system.[9]

This study seeks to make a contribution to current discussions across Europe about what forms "a joint approach and common standards" could and should take and why. Of course, in practice much will depend on ongoing discussions among members of the Council of Ministers as they formulate a new resolution on hate speech, and on consensus-building within the European Commission as it draws up a comprehensive Digital Services Act that is likely to cover the responsibilities of Internet platforms with regards to removing illegal hate speech.

It is also worth noting that the standardization agenda is not solely about agreeing common standards across European countries it is also about ensuring that the same standards are

[8] Comments during session on Tackling Hate Speech Online, IGF, Berlin, 27 November, 2019.

[9] Trusted flaggers and monitoring bodies questionnaire response 4, 12 December, 2019 [Anonymised].

applied to all Internet platforms within the sector. Thus, an unnamed European Commission official is reported as justifying the need for standardization as follows: “We need to have legal certainty so all operators know the rules of the game” (Khan and Murgia 2019). The chief executives of some Internet platforms are also calling for standardization in governance across Internet platforms. Consider the words of Mark Zuckerberg from March 2019:

> Internet companies should be accountable for enforcing standards on harmful content. It’s impossible to remove all harmful content from the Internet, but when people use dozens of different sharing services—all with their own policies and processes—we need a more standardized approach.
>
> One idea is for third-party bodies to set standards governing the distribution of harmful content and to measure companies against those standards. Regulation could set baselines for what’s prohibited and require companies to build systems for keeping harmful content to a bare minimum. (Zuckerberg 2019)

The reason why Internet platforms like Facebook would call for common standards across all Internet platforms is not hard to fathom. At present Facebook and other mainstream Internet platforms that impose reasonably strict community standards or content policies on hate speech and that devote not insignificant time and resources into moderation and oversight risk losing some users to other, smaller Internet platforms that impose less strict standards or that devote much less time and resources to moderation and oversight.[10]

However, standardization of governance across Internet platforms might also pose a threat to the fact of diversity and pluralism within the sector—something to which this study will return repeatedly.

Moreover, Zuckerberg’s reference to “third-party bodies to set standards” raises some obvious and difficult questions. Which bodies? What standards should they be enforcing? And should the standards be settled at the national or intergovernmental level? And so, we come back to the intentions of the Council of Europe’s Committee of Ministers in commissioning an Expert Committee in 2019 to draft new recommendations in this area, as well as Ursula von der Leyen’s call for “common standards” across Europe. Once again it is not hard to think of reasons why some Internet platforms might favour common standards within Europe. If there is a patchwork of different regulatory models and standards—which develop in a piecemeal, unpredictable and sometimes fast evolving fashion—it can make it more challenging for Internet platforms to operate in Europe. Internet platforms are likely to find it more straightforward to achieve regulatory compliance, and to develop internal policies and procedures to that end, with European-wide standards rather than developing different responses to different regulatory standards within each individual European country.[11]

That being said, country context is a key factor in the design and implementation of regulatory institutions, systems and rules [see section I.(D)(vi)]. Therefore, common standards for the regulation of online hate speech across Europe need not mean identical regulatory models or tools [see section I(D)(vii)].

Much will also depend on how the jurisdictional authority of national regulators will be determined. For instance, national regulators could have jurisdictional authority over (i) Internet platform

[10] 2nd consultative meeting, Berlin, 26 November, 2019.

[11] Ibid.

content that was created or produced in their own country, (ii) Internet platform content that is accessed by users in their own country, or (iii) content posted or shared on an Internet platform that has its legal home in their own country. If the jurisdictional authority of national regulators covers (iii), then this could create a perverse situation in which a national regulator in Ireland, say, could find itself imposing fines on an Internet platform legally registered in Ireland for a pattern of failure to remove illegal content based on Irish hate speech laws, even though much of the content was created and accessed by users in other European countries operating under different hate speech laws.[12] For this, and other reasons to be discussed in this study, the degree of definitional harmonisation on what counts as "hate speech" is an important issue [see section I.E]. And the degree of definitional harmonisation is an important issue not merely across different Internet platforms but also between Internet platforms and national hate speech laws in each country, and also across different national hate speech laws found within Europe.[13]

Finally, national governments and intergovernmental organisations in Europe should be mindful of the fact that when they develop governance regimes for online hate speech what they do is seen by other countries in other global regions, and may become, whether intentionally or unintentionally, a model for others to follow. Thus, in developing the new Digital Services Act the European Union may become or be perceived as "a shining city upon a hill", in the sense that what it does will be observed with keen interest across Asia, the Americas and Africa.

Of course, what the European Union does next in tackling online hate speech, and the success or failure of what it does, will also be closely monitored by all European countries that are not members of the European Union including not least Russia and the United Kingdom. It goes without saying that perceptions of success or failure will depend on which countries and which organisations within those countries are doing the judging and against what indicators. Therefore, another aim of this study is to develop a set of indicators or measures that could be used by monitoring bodies or other stakeholder organisations to assess the success or progress of different governance tools for online hate speech. Among the factors that are important when governmental authorities, monitoring bodies or other stakeholder organisations are selecting which indicators or measures to use will be the degree of support a given indicator commands in any given country context and across different country contexts.

That being said, in some parts of the rest of the world governmental authorities have already implemented regulatory instruments applicable to online hate speech. For example, in China Articles 47 and 48 of the Cybersecurity Law 2016—a law which predates the NetzDG Act in Germany—impose responsibilities on Internet service providers and software platforms to "stop transmission" of illegal or administratively prohibited content, and Article 68 sanctions the Cyberspace Administration (China's Internet regulator) to impose fines on Internet companies for violations of these responsibilities. In September 2017, for instance, the Cyberspace Administration's Beijing and Guangdong offices imposed maximum fines on Tencent, Baidu, and Sina Weibo for breaches of Article 47 by hosting "information of violence and terror, false rumors, pornography, and other information that jeopardizes national security, public safety, and social order" (Gao 2017). This regulatory intervention also predates Germany's Federal Office of Justice (BfJ) imposition of a 2 million Euros fine on Facebook in July 2019 among other things, because its NetzDG reporting form was too difficult to find.

[12] Interview with member of German government, Federal Ministry of Justice and Consumer Protection, 29 November, 2019.

[13] For a comparison of similarities and differences in national hate speech laws across Europe, see Brown (2015: ch. 2), Alkiviadou (2016), and Brown and Sinclair (2019: ch. 2).

B. Wider background to the study

Many of the above developments in the emerging story of the governance of online hate speech within Europe have been called for, welcomed and in some cases precipitated by key texts at the intergovernmental level, including codes, directives, communications, and recommendations. These key texts include, for example, Directive 2000/31/EC of the European Parliament and of the Council on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market ("Directive on Electronic Commerce") of 8 June 2000,[14] ECRI's General Policy Recommendation No. 15 on Combating Hate Speech of 8 December 2015 (and Explanatory Memorandum),[15] the European Commission's Code of Conduct on Countering Illegal Hate Speech Online of May 2016,[16] the European Commission's Communication on Tackling Illegal Content Online of 28 September 2017,[17] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Role and Responsibilities of Internet Intermediaries of March 2018,[18] and the Revised EU Audiovisual Media Services Directive ("AVMS Directive") of 14 November 2018.[19]

Within these key texts can also be found several major themes that provide an important part of the wider background to this study: namely, giving remedies to persons or groups targeted or impacted by online hate speech whilst operating within international human rights frameworks, cementing in law the responsibilities of Internet platforms, taking notice of diversity among Internet platforms, the emergence of new forms of cooperation and collaboration between governments, Internet platforms and civil society organisation, and the inevitability of trade-offs between important concerns, interests and rights at stake in the issue of governance.

(i) Emphasis on the importance of international human rights frameworks

The first major theme is the idea that there is an urgent need to tackle the problem of online hate speech as a matter of protecting human rights but that this must be done using governance tools which themselves operate within the constraints of international human rights frameworks that also include the right to freedom of expression. For example:

[14] Directive 2000/31/EC of the European Parliament and of the Council on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market ("Directive on Electronic Commerce") of 8 June 2000. Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031 [last accessed 7 October, 2019].

[15] CRI(2016)15, Strasbourg, 8 December 2015. Available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 [last accessed 7 October, 2019].

[16] Available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en [last accessed 11 December 2019].

[17] Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms, COM(2017)555, Brussels, 28 September, 2017. Available at: https://ec.europa.eu/digital-single-market/en/news/communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms [last accessed 11 December 2019].

[18] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries of March 2018. Available at: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 [last accessed 5 October, 2019].

[19] Revised EU Audiovisual Media Services ("AVMS") Directive, Directive (EU) 2018/1808 of the European Parliament and of the Council. Available at: https://eur-lex.europa.eu/eli/dir/2018/1808/oj [last accessed 5 October, 2019].

> The IT Companies […] share the European Commission's and EU Member States' commitment to tackle illegal hate speech online. […] The IT Companies and the European Commission also stress the need to defend the right to freedom of expression.[20]
>
> [The] spread of illegal content that can be uploaded and therefore accessed online raises serious concerns that need forceful and effective replies. What is illegal offline is also illegal online. […] The European Union has responded to these concerns through a certain number of measures. However, addressing the detection and removal of illegal content online represents an urgent challenge for the digital society today.[21]
>
> By enhancing the public's ability to seek, receive and impart information without interference and regardless of frontiers, the internet plays a particularly important role with respect to the right to freedom of expression. It also enables the exercise of other rights protected by the Convention and its protocols, such as the right to freedom of assembly and association and the right to education, and it enables access to knowledge and culture, as well as participation in public and political debate and in democratic governance. […] However, the internet […] has spurred the spread of certain forms of harassment, hatred and incitement to violence, in particular on the basis of gender, race and religion, which remain underreported and are rarely remedied or prosecuted.[22]

These aspects of cross-European thinking about the interplay of human rights when it comes to combating hate speech reflect previous key documents on hate speech in general, such as Principle 2 of Recommendation No. R (97) 20 of the Committee of Ministers to Member States of the Council of Europe on "Hate Speech" of 30 October, 1997: "The governments of the member states should establish or maintain a sound legal framework consisting of civil, criminal and administrative law provisions on hate speech which enable administrative and judicial authorities to reconcile in each case respect for freedom of expression with respect for human dignity and the protection of the reputation or the rights of others."[23]

More focus is needed, however, in thinking about exactly which human rights are at stake in the area of hate speech regulation including governance tools for online hate speech content. There is a growing body of European Court of Human Rights (ECtHR) cases that deal with hate speech laws, and typically these cases make reference to laws that can be "necessary in a democratic society", as per Art. 10(2) of ECHR, including based on issues of personal safety, social cohesion, and protecting the institutions of a functioning democracy. Consider *Jersild v. Denmark*,[24] *Aksu v. Turkey*[25] and *Vejdeland and Others v. Sweden*,[26] in which the ECtHR upheld

[20] European Commission's Code of Conduct on Countering Illegal Hate Speech Online of May 2016.

[21] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online.

[22] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries of March 2018.

[23] Available at: https://rm.coe.int/1680505d5b [last accessed 10 December, 2019].

[24] No. 15890/89 (ECtHR, Strasbourg, 23 September, 1994) (relating to the conviction of a Danish journalist for aiding and abetting hate speech offences committed on national TV).

[25] Nos. 4149/04 and 41029/04 (ECtHR, Strasbourg, 15 March, 2012) (relating to decisions taken by domestic courts in Turkey not to uphold complaints made against the creators of dictionaries to be used for educational purposes that included a range of words the definitions of which expressed negative stereotypes of gypsies).

[26] No. 1813/07 (ECtHR, 9 February, Strasbourg, 2012) (relating to the conviction of four members of an organization called National Youth under Ch. 16, s. 8 of the Swedish Criminal Code, for distributing leaflets containing homophobic statements within a secondary school).

convictions under local hate speech laws. On the other hand, in *Erbakan v. Turkey*,[27] *Perinçek v. Switzerland*[28] and *Stomakhin v. Russia*[29] the ECtHR decided that the convictions of political figures for hate speech offences by domestic courts had violated these political figures' human right to freedom of expression under Art. 10(1) of the ECHR.

The academic literature on hate speech laws deepens and extends the analysis of the relevant considerations, pointing to such concerns as mental and emotional health and well-being, security, autonomy, the public good of equal access to information, dignity, recognition, intercultural dialogue, democratic legitimacy, and so on (Brown 2015).

More recently, the ECtHR has drawn a connection between the issue of regulating hate speech and the weighing up of different human rights, including the human right to freedom of expression (Art. 10), the human right not to be discriminated against (Art. 14) and the human right to a private and family life (Art. 8). Consider *Delfi AS v. Estonia*,[30] *R.B. v. Hungary*,[31] *Király et al. v. Hungary*,[32] and *Alković v. Montenegro*.[33]

In addition, more attention needs to be devoted to asking what analytical model should be used in factoring these different human rights into governance tools. For example, ECRI's General Policy Recommendation No. 15 poses the following dilemma that is highly relevant to the governance of online hate speech: "Aware of the grave dangers posed by hate speech for the cohesion of a democratic society, the protection of human rights and the rule of law but conscious of the need to ensure that restrictions on hate speech are not misused to silence minorities and to suppress criticism of official policies, political opposition or religious beliefs".[34] But more clarification is needed on what the right model is for bringing these very valid but also very diverse considerations "under one roof" within any particular governance tool for online hate speech. Is this a matter of "balancing" different human rights, a case of establishing a clear "hierarchy" of human rights (i.e. the "pre-eminence of one right over the other" (Bychawska-Siniarska 2017: 11)), or instead reaching some sort of "principled compromise" (Brown 2015: ch. 10)?

(ii) Growing demands on Internet platforms to do their bit in tackling online hate speech

A second major theme is the growing societal and governmental expectation that Internet platforms have a "special responsibility", and should be "doing more", to tackle online hate speech posted or shared on their platforms, websites and services. In relation to online hate speech in general, including lawful or legal content, ECRI's General Policy Recommendation No. 15 on Combating Hate Speech (and Explanatory Memorandum) highlights the need for better "self-regulation" among Internet platforms.[35] Member states are to "encourage" Internet platforms

[27] No. 59405/00 (ECtHR, Strasbourg, 6 July, 2006).

[28] No. 27510/08 (ECtHR, Strasbourg, 17 December, 2013).

[29] No. 52273/07 (ECtHR, Strasbourg, 9 May, 2018).

[30] No. 64569/09 (ECtHR, Strasbourg, 16 June 2015), at para. 138.

[31] No. 64602/12 (ECtHR, Strasbourg, 12 April 2016), at paras. 78 and 81-84.

[32] No. 10851/13 (ECtHR, Strasbourg, 17 January 2017), at paras. 61-82.

[33] No. 66895/10 (ECtHR, Strasbourg, 5 December 2017), at paras. 63-73.

[34] Available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 [last accessed 7 October, 2019].

[35] CRI(2016)15, Strasbourg, 8 December 2015, Recommendation 7. See also Explanatory Memorandum, paras. 130-144.

to develop and adopt for themselves their own industry-based "codes of conduct", "monitoring [systems]", and "complaints mechanism" for online hate speech.[36]

When it comes to illegal or unlawful hate speech content, however, ECRI recommends that member states seek to "clarify the scope and applicability of responsibility under civil and administrative law for the use of hate speech".[37] This clarification of responsibilities can include, for example, the imposition of legal or administrative sanctions (e.g. fines) on Internet platforms that demonstrate a pattern of "failure to comply with regulatory requirements".[38] Indeed, ECRI also notes that in *Delfi AS v. Estonia*[39] the European Court of Human Rights (ECtHR) "considered the right to freedom of expression not to have been violated where a company was found liable to those targeted by hate speech posted on its internet news portal."[40]

In a similar vein, the Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 ("Directive on electronic commerce") clarifies that any limitation on the legal or administrative liability of Internet intermediaries for illegal online content should be conditional on their acting "expeditiously" to remove or disable access to illegal content, including illegal hate speech content, upon obtaining actual knowledge or awareness of that content.[41]

Arguably at the heart of recent key texts at the European intergovernmental level on this issue sits societal norms about an equitable sharing of responsibility and an equitable division of labour across and within the governance of online hate speech between governmental agencies, civil society organisations and Internet platforms themselves. For example:

> Those online platforms which mediate access to content for most internet users carry a significant societal responsibility in terms of protecting users and society at large and preventing criminals and other persons involved in infringing activities online from exploiting their services. The open digital spaces they provide must not become breeding grounds for, for instance, terror, illegal hate speech, child abuse or trafficking of human beings, or spaces that escape the rule of law. Clearly, the spreading of illegal content online can undermine citizens' trust and confidence in the digital environment, but it could also threaten the further economic development of platform ecosystems and the Digital Single Market. Online platforms should decisively step up their actions to address this problem, as part of the responsibility which flows from their central role in society.[42]
>
> Video-sharing platform services provide audiovisual content which is increasingly accessed by the general public, in particular by young people. This is also true with regard to social media services, which have become an important medium to share information and to entertain and educate, including by providing access to programmes and user-generated videos. Those social media services need to be included in the scope of Directive 2010/13/EU because they compete for the same audiences and

[36] Ibid.

[37] Ibid., Recommendation 8. See also Explanatory Memorandum, paras. 145-155.

[38] Ibid., Explanatory Memorandum, para. 151.

[39] No. 64569/09 (ECtHR, Strasbourg, 16 June 2015).

[40] CRI(2016)15, Strasbourg, 8 December 2015, Explanatory Memorandum, para. 150.

[41] Consider, for example, Art. 14 of Directive 2000/31/EC of the European Parliament and of the Council on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market ("Directive on Electronic Commerce") of 8 June 2000.

[42] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online.

> revenues as audiovisual media services. Furthermore, they also have a considerable impact in that they facilitate the possibility for users to shape and influence the opinions of other users. Therefore, in order to protect minors from harmful content and all citizens from incitement to hatred, violence and terrorism, those services should be covered by Directive 2010/13/EU to the extent that they meet the definition of a video-sharing platform service.[43]

Arguably, however, it is not the mere fact of the spreading of illegal content online that can "undermine citizens' trust and confidence in the digital environment". Rather, it is the perceived failure of governmental agencies and Internet platforms alike to take effective and necessary steps to prevent such content that undermines trust and confidence.

Of course, it is a ubiquitous feature of public policy that public perceptions of the success or diligence of public bodies, civil organisations and private enterprises in tackling given problems are not always accurate, and can tend to err on the side of exaggerating the lack of success or diligence. This is partly caused by a more general lack of trust that government will "do the right thing". It is also precipitated in certain areas of public policy by wildly divergent perceptions among the public about what the "problem" is that public bodies, civil organisations and private enterprises are supposed to be dealing with.

This last point is especially true in the case of online hate speech, where public understandings of the "the problem of hate speech" are radically divergent and deeply rooted in political disagreements, culture wars and ideological schisms (Brown and Sinclair 2019: ch. 1). In many European countries (and elsewhere) the term "hate speech" has become an ideological football, used by political opponents to accuse each other either of using objectionable speech or of using accusations of objectionable speech to close down debate and silence criticism.

Predictably, then, for some people the alleged failure of Internet platforms and governmental agencies to take effective and necessary steps to tackle online hate speech is a matter of under-removal of content. But for other people this failure is essentially one of over-removal of content. In former Eastern Block countries like Hungary, for example, there is a historic and cultural mistrust of governmental agencies *qua* agents of censorship that has translated into particular fears about online content moderation and the removal of hate speech whether by governmental agencies or Internet platforms (see Brown and Sinclair 2019: 85-89).

Be that as it may, surely part of the so-called tech-backlash against Internet platforms is that when it comes to the policy goal of tackling online hate speech some governments, and no doubt some people, believe that Internet platforms are not doing their fair share.

Whether this particular accusation is reasonable, however, depends on what norms of fairness are being relied on. First, people might have the thought that it is fair to expect Internet platforms to "do more" to help clean up the Internet insofar as Internet platforms are responsible for creating the services that facilitate and perhaps even encourage the flow of online hate speech in the first place (see Citron 2014; Delgado and Stefancic 2014; Cohen-Almagor 2015; Brown 2017e, 2018a). This speaks to the logic of backward-looking responsibility or blame. Then again, Internet companies could reply that they did not help to invent hate speech or even the idea of utilising new technologies to assist in the large-scale public dissemination of hate speech[44]—both predate

[43] Art. 4 of the Revised EU Audiovisual Media Services ("AVMS") Directive.

[44] For example: "Hate speakers have always used the latest technologies to spread their messages to as many people as possible, as cheaply as possible and as anonymously as possible. Printed leaflets, mail shots, automated

the Internet. And they could protest that they are no more responsible for hate speech content than postal services are responsible for poison pen letters. This debate about who is responsible for causing online hate speech inevitably partly reduces to legalistic questions about the nature of what Internet platforms do. Are they conduits, facilitators, curators, editors, secondary publishers, or publishers of content for the purposes of establishing responsibility and even liability?

Second, the public might think it fair for Internet platforms to do more to clean up the Internet due to the perception that they are commercial enterprises profiting from platforms, services and websites that, whether by design or not, enable the social spread of hate speech that causes misery to its targets.[45] The suggestion here is that Internet platforms have business models that rely on advertising revenues that depend on user growth and engagement, which in turn depends on algorithms that feed users ever-more of the content they like, which for some users is hate speech content. Indeed, the perception might be that insofar as Internet platforms do more than simply allow hate speech content to be posted or shared but also use algorithms to accelerate, amplify or promote virality of such content, then they must be held responsible for the harmfulness of the content. In other words, the idea is that by accelerating content they are acting not like hosts but like editors. Some platforms might also end up incentivising professional content creators to post or share hate speech in circumstances where this represents good "click bait"—more clicks equals more revenue for content creators. However, the big Internet platforms themselves insist that this is not their business model and that hosting hate speech content reduces not increases their advertising revenues.[46]

Against the charge that they are profiting, directly or indirectly, off the back of the misery caused by online hate speech, Internet platforms could make the further point that their services also provide useful tools for victims of cyberhate, or indeed offline hate speech. For example, they provide opportunities for targets to identify, discuss, come together to console, mutually support, campaign, educate and, importantly, counter-speak against hate speech.

Arguably a more fruitful and consensus-building thought is that it is fair to expect Internet platforms to "do better" in tackling hate speech content based on the mere fact of their capacity to do so (see also Citron 2014). Understood in this way, to speak of Internet platforms doing their fair share is less about "causal" or "backward-looking" responsibility in the sense of "You made the mess so you should clean it up!", but "remedial responsibility" in the sense of "There is a problem that needs fixing and you are well placed to help fix it".[47] The general principle here is that when faced with a shared problem those agents with greater capacity to help tackle it should as a matter of fairness bear a greater and more urgent responsibility to do so.[48]

These different ideas of fairness all point to the same conclusion. It is that assessing Internet platforms' fair share of the practical burden of, and legal responsibility for, tackling online hate

telephone messages—these were just some of the technologies used by white supremacists and anti-Semites in the twentieth century." (Brown and Sinclair 2019: 21).

[45] Interestingly, the European Court of Human Rights has pointed to the question of whether or not Internet platforms profit from enabling content to be shared or posted as being potentially relevant to determining any legal liability that Internet platforms might have for failing to remove illegal content. Consider Magyar Tartalomszolgáltatók Egyesülete and Index.Hu Zrt v. Hungary, No. 22947/13 (ECtHR, Strasbourg, 2 February, 2016), at para. 64, and Pihl v. Sweden, No. 74742/14 (ECtHR, Strasbourg, 9 March 2017), at p. 31.

[46] 1st consultative meeting, London, 17-18 October, 2019. 2nd consultative meeting, Berlin, 26 November, 2019.

[47] For a philosophical analysis of the distinction between these two broad kinds of responsibility, see Miller (2007: ch. 4) and Brown (2009).

[48] For an articulation and defence of this general principle of responsibility assignment, see Wenar (2007).

speech cannot be done using a crude notion of strict equality but rather must rely on the more nuanced concept of equity or fair proportion. An equitable share of burden and responsibility is one that is in fair proportion to Internet platforms' role in facilitating or profiting from online hate speech and/or in fair proportion to Internet platforms' capacity in tackling online hate speech.

It is also important to recognise at this juncture that there is more than one fruitful way to "tackle" online hate speech. Governments across Europe have tended to think in terms of the extent to which Internet platforms meet certain norms and standards on the removal of content, processes for the removal of content, and transparency about content removal and processes. As a result, Internet platforms have often been accused of failing to meet expectations. For example, during a session on tackling illegal hate speech at IGF2019 in Berlin, Internet platforms including Facebook, Twitter and YouTube were accused by some delegates (rightly or wrongly) of failing to provide annual statistics on how many reports of illegal hate speech content they have received and what proportion of those reports they have removed within specified time frames, and, equally importantly, were also accused by some delegates (rightly or wrongly) of failing to give detailed information to the public on the standards, protocols and texts used in the training of their human moderators.[49] But the prior question is whether these are reasonable expectations to have in the first place. Internet platforms also have responsibilities to protect users' data, commercially sensitive information and intellectual property, and so arguably the setting of norms and standards needs to be collaborative.

More importantly, norms and standards on the removal of content, processes for the removal of content, and transparency about content removal and processes are certainly not the beginning and end of the story when it comes to effective governance of online hate speech. For one thing, recent research has suggested that the mere act of Internet platforms publicising and making clearer to users the existence and meaning of their community standards or content policies on hate speech can have a significant impact in terms of lowering the rate of hate speech being shared or posted on those Internet platforms (see Benesch and Matias 2018).

For another thing, Internet platforms are increasingly looking to "content management" tools, such as reducing distribution, making content ineligible for recommendation or sponsorship, preventing people from sharing or liking content, etc., as means of reducing access to potentially harmful content other than via content removal or take down. Content management tools are the obverse of content acceleration, amplification and virality.

Therefore, regulatory interventions like imposing legal responsibilities or duties of care, for example, should encompass not merely expectations about content removal, processes, and transparency but also about Internet platforms doing more to publicise and clarify their community standards on hate speech and to reduce access to hate speech content, especially in grey area or difficult cases.

Accordingly, imposing a wider range of expectations and responsibilities on Internet platforms might be not merely fairer, especially in grey area cases, but might also increase the capacity and chances of Internet platforms actually meeting those responsibilities.

Equally importantly, the assertion that Internet platforms should be doing their fair share in tackling online hate speech cannot be understand properly unless it is set against the converse assertion that has been made by the chief executives of some Internet platforms in recent years,

[49] Session on tackling online hate speech, IGF, Berlin, 27 November, 2019.

namely, that governmental agencies must also do their fair share. Responsibility is a two-way street, in other words. Consider these remarks by Mark Zuckerberg in March 2019:

> I believe we need a more active role for governments and regulators. By updating the rules for the Internet, we can preserve what's best about it—the freedom for people to express themselves and for entrepreneurs to build new things—while also protecting society from broader harms. (Zuckerberg 2019)

This is a timely reminder that governments and regulators have their part to play in tackling online hate speech, such as by taking action to better regulate Internet platforms.

Finally, the idea that Internet platforms along with governmental agencies must do their fair share in tackling online hate speech must be discussed in the context of a much wider set of responsibilities, including the responsibility to operate within human rights frameworks and to protect free speech. If the expectations or responsibilities that the public and governmental agencies place on Internet platforms to remove hate speech become too strong and overly sweeping, there is a risk of incentivising the over-removal of content, at the cost of free speech.

Therefore, in circumstances where governmental agencies impose legal responsibilities on Internet platforms to remove illegal hate speech content within specified time frames and then seek to levy fines on Internet platforms for patterns of failure, and in circumstances where courts issue Internet platforms with judicial notice and take down orders for particular bits of content, it would be short-sighted and unfair to expect Internet platforms to always accept these fines and accede to these orders automatically. A truly responsible Internet platform is one that, on occasion and where appropriate, is willing to defend in the courts its decisions not to remove content, on the grounds of promoting and protecting the human right to freedom of expression.

Of course, an Internet platform operating in a country where an authoritarian regime might ban its operations if the platform in any way challenges the demands made by governmental agencies to remove content has a dilemma. It must balance its aim to provide access to its services to the majority of users in that country against the principle of defending free speech against authoritarian regimes. An Internet platform that errs on the side of staying in the country to provide its services to the majority of users should not necessarily be condemned. Then again, when an Internet platform is operating in a mature democracy with free and open elections, the rule of law and checks and balances on the exercise of power by the ruling government, and in circumstances where there is little or no prospect of the platform being banned from the country simply for seeking to challenge in courts certain fines or orders, then arguably it is reasonable and fair to expect the platform to take a stand in some instances.

(iii) Recognition of the diversity of Internet platforms

A third major theme is awareness of the diversity of Internet platforms: the different functionalities they offer, the different positions they occupy within the sector, and the different practices they have in terms of tackling online hate speech content.

> The term "media and the Internet" is one that embraces many forms of communication with vastly different characteristics and impact. Thus, it covers print media (such as newspapers, journals and books, as well as pamphlets, leaflets and posters) but also audiovisual and electronic media (such as radio, television, digital recordings of sound and image, web sites, apps, emails and a vast array of social media and video games) and undoubtedly other forms of communication that may yet be developed. Moreover, some things spoken, published or otherwise communicated will be truly individual initiatives, while others will be the product of substantial business enterprises. Some such communications will be subject to varying forms of editorial control but others will appear without being reviewed by anyone other than their originator and indeed appear without the prior knowledge of the person providing the particular means of communication. In many instances the author of a communication will be identifiable but in others he or she can remain anonymous. Some communications will reach an audience almost instantaneously but others will depend on the willingness to listen, read or otherwise access what is being communicated. Some will be widely disseminated and/or enduring but others will be barely noticed and/or fleeting in their existence. All these differences need to be taken into account when determining the scope of regulatory action and self-regulation, as well as whether expectations as to what they can achieve are realistic.[50]
>
> A wide, diverse and rapidly evolving range of players, commonly referred to as "internet platforms", facilitate interactions on the internet between natural and legal persons by offering and performing a variety of functions and services. […] A variety of network effects and mergers have led to the existence of fewer, larger entities that dominate the market in a manner that may jeopardise the opportunities for smaller platforms or start-ups and places them in positions of influence or even control of principal modes of public communication. The power of such platforms as protagonists of online expression makes it imperative to clarify their role and impact on human rights, as well as their corresponding duties and responsibilities, including as regards the risk of misuse by criminals of the platforms' services and infrastructure.[51]
>
> The IT Companies underline that the present code of conduct is aimed at guiding their own activities as well as sharing best practices with other internet companies, platforms and social media operators.[52]

The fact that not all Internet platforms are alike raises a challenge for the governance of online hate speech, especially at the regulatory level. There is often a tendency for governments to roll

[50] ECRI, General Policy Recommendation No. 15 on Combating Hate Speech, CRI(2016)15, Strasbourg, 8 December 2015, Explanatory Memorandum, para. 131.

[51] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries.

[52] European Commission's Code of Conduct on Countering Illegal Hate Speech Online of May 2016. Available at: http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf [last accessed 5 October 2019].

out regulatory instruments that apply to large or general categories of businesses or organisations and that leave limited discretion for treating different businesses or organisations differently. To explain, in the study of public policy it is common to distinguish between administrative policies (secondary legislation, general rules, regulations, and policy statements), on the one hand, and administrative measures (particular administrative orders, decisions, and adjudications that relate to a single agent or small number of identifiable agents), on the other hand. Applying this distinction to the governance of online hate speech, administrative policies might include a regulatory regime according to which Internet platforms have a legal responsibility to remove illegal hate speech content enforced with fines and/or a wider code of practice for handling hate speech content that includes duties of due process and transparency. By contrast, administrative measures concern the application to particular cases, such as a decision by a regulator, government ministry or administrative court to levy a fine on a particular Internet platform or to find that a particular Internet platform has failed to live up to a code of practice in certain respects. Importantly, if an Internet platform has a certain type of function, mission, or business model, or distinctive set of values or guiding principles, that it thinks sets it apart from other sorts of Internet platforms, the recognition of the difference needs to be embedded in the administrative policy itself as a legal category (e.g. exemptions, exceptions). This is because regulators, government ministries or administrative courts may not be in a position to unilaterally change or adapt the policy on a case by case basis. Rather, they may only have the power to apply the policy as written, that is, to take administrative measures.

The above-quoted passages also make clear that different regulatory interventions potentially can have different effects on larger and smaller Internet platforms. And so there is a competition policy dimension to the choice of governance tools for online hate speech, such as in circumstances where there is a high probability that certain regulations might enable “big” companies to get even bigger—potentially at the expenses of “small” companies—because they can better cope with the resource burdens imposed on them by complying with the regulations.

Arguably some national governments and intergovernmental organisations have failed to recognise or have paid insufficient attention to the radical implications that the aforementioned facts of pluralism and diversity within the Internet platform sector might have for the development and adoption of governance tools for online hate speech. This point shall be picked up again in section I.D below, and in various places in the remainder of the study including the recommendations in section IX.1

(iv) Promotion of collaborative governance of online hate speech

A fourth major theme is the promotion of cooperation and collaboration between governmental agencies, Internet platforms and civil society organisations in tackling online hate speech. For example:

> The IT Companies support the European Commission and EU Member States in the effort to respond to the challenge of ensuring that online platforms do not offer opportunities for illegal online hate speech to spread virally.[53]

> Online platforms should […] cooperate closely with law enforcement and other competent authorities where appropriate, notably by ensuring that they can be rapidly

[53] European Commission’s Code of Conduct on Countering Illegal Hate Speech Online of May 2016.

> and effectively contacted for requests to remove illegal content expeditiously and also in order to, where appropriate, alert law enforcement to signs of online criminal activity. To avoid duplication of effort and notices and thus reduce the efficiency and effectiveness of the removal process, law enforcement and other competent authorities should also make every effort to cooperate with one another in the definition of effective digital interfaces which facilitate the fast and reliable submission of notification and to ensure efficient identification and reporting of illegal content. Establishing points of contact by platforms and authorities is key for the proper functioning of such cooperation.[54]

> Platforms should seek to collaborate and negotiate with consumer associations, human rights advocates and other organisations representing the interests of users and affected parties, as well as with data protection authorities before adopting and modifying their policies.[55]

The past 36 months or so has seen acceleration within the European context of innovation and take-up of collaborative forms of governance of online hate speech—and, of course, governance of other problematic kinds of content as well.

In the context of the "tech backlash"—including growing popular and political pressure to regulate Internet platforms to tackle forms of harmful speech—national governments, intergovernmental organisations, Internet platforms, and civil society organisations (e.g. trusted flaggers) across Europe are increasingly working together to devise and deliver collaborative governance of online hate speech. Governance tools typically involve multiple partners working together, albeit there is tremendous diversity in the nature and extent of that collaboration.

Collaboration has been promoted for various reasons, to be explored in section V of this study, including but not limited to sharing expertise and resources in handling large numbers of reports or flags. This collaboration has been thought especially appropriate and beneficial in grey area or difficult cases.

Internet platforms are increasingly engaged in collaborative partnerships with academics, civil society organisations, NGOs and other stakeholders to deliver more robust styles of content moderation and systems of oversight of moderation, so as to diversify responsibility for deciding what content appears on the Internet. As Facebook's Brent Harris put it, speaking at the launch event of its 2019 report *Global Feedback & Input on the Facebook Oversight Board for Content Decisions*, "we don't feel we should hold that responsibility alone".[56]

And different national governments across Europe are pursuing different sorts of partnerships with Internet platforms and civil society organisation to support better systems of content moderation, oversight of moderation and regulation. Examples from Croatia, France, Germany, the UK, and a member state of the European Union (anonymous) will be discussed in this study.

[54] Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online.

[55] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries.

[56] Brent Harris, Facebook, comments at invited event, "Who should regulate free speech online?", Chatham House, London, 27 June, 2019.

Moreover, many different Internet platforms have now joined up to the European Commission's Code of Conduct on Countering Illegal Hate Speech Online, whether as founding members of the forum or as later signatories.[57]

Many Internet platforms are keeping a close eye on what other platforms are doing in the areas of content moderation and oversight of content moderation for online hate speech (and other types of content), without necessarily "rushing to judgment". YouTube, Twitter, Snapchat do not have oversight boards but are watching with interest what happens at Facebook, for example.

Content moderation and oversight of moderation for online hate speech are not necessarily areas of corporate endeavour where inter-company competition and the fight for customers precludes sharing of information and closer cooperation in the establishment of best practice. The content moderation style and systems of oversight of moderation adopted by Internet platforms may factor into users' decisions about whether and how much to use different platforms in some small way, but this is very unlikely to be a substantial deciding factor.[58]

So these are areas where potentially Internet platforms could be driven by a wish to "do better"—for their users and society at large—in terms of preventing harm to, and also giving redress to, targets of online hate speech whilst at the same time respecting the human right to freedom of expression, and by a willingness to collaborate with each other in this endeavour.

(v) Trade-offs at the heart of the governance of online hate speech

A fifth major theme is that getting the governance of online hate speech right necessarily and inescapably involves making trade-offs between interests, rights and values of equivalent importance. For example:

> States should regularly consult with all relevant stakeholders with a view to ensuring that an appropriate balance is struck between the public interest, the interests of the users and affected parties, and the interests of the platform.
>
> […]
>
> This positive obligation to ensure the exercise and enjoyment of rights and freedoms includes, due to the horizontal effects of human rights, the protection of individuals from actions of private parties by ensuring compliance with relevant legislative and regulatory frameworks. Moreover, due process guarantees are indispensible, and access to effective remedies should be facilitated vis-à-vis both States and platforms with respect to the services in question.
>
> […]
>
> When restricting access to content in line with their own content-restriction policies, platforms should do so in a transparent and non-discriminatory manner. Any restriction of content should be carried out using the least restrictive technical means and should

[57] Information available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en [last accessed 11 December 2019].

[58] 1st consultative meeting, London, 17-18 October, 2019. 2nd consultative meeting, Berlin, 26 November, 2019.

be limited in scope and duration to what is strictly necessary to avoid the collateral restriction or removal of legal content.[59]

As already noted, concerns have been raised by both users and civil liberties organisations, among others, about potential weaknesses in procedural fairness or due process within many emerging governance tools for online hate speech. But almost inevitably there will be some trade-off between quantity of moderation and quality of moderation, for example. If there is a need to get through extremely large quantities of content moderation there is without doubt going to be downwards pressure on standards of due process. An Internet platform's size and resource capacity will also constrain the styles of moderation it can adopt, such as the mix of human moderation and automated or machine learning moderation.

From a more abstract or philosophical perspective, this is, in one regard, a trade-off between the public interest and the individual interest. It is in the individual's interest to have a high degree of due process applied during the moderation, oversight or regulation of his or her personal case, but at the same time it is also in the public interest to achieve potentially large quantities of moderation, oversight and regulation so as to achieve policy influence. No doubt as a case progresses up the chain from the moderation level to the oversight level and on to the regulatory level, arguably there is a need for increasingly high standards of due process. But there will be a trade-off to be had at all levels, especially at the moderation level.

This is, of course, not the only trade-off. Another to be explored in this study—and one that has hitherto been largely ignored in the debate—is the trade-off between procedural fairness and sensitivity to the experiences and needs of "victims", that is, persons targeted or adversely affected by online hate speech, especially those who have or are considering reporting it.

When governmental agencies and Internet platforms strengthen and enforce procedural fairness or due process considerations, such as by placing a burden of proof on persons reporting online hate speech, testing the credibility of the person doing the reporting, imposing legal sanctions on persons found to have submitted fake reports, for example, this can create a psychological or emotional barrier to genuine victims of online hate speech reporting it.

To give one illustration, whereas procedural fairness or due process considerations might suggest that people who report online hate speech should be required to provide maximum information about themselves, the content they are reporting, why they are reporting the content, and the impact of the content upon themselves, victim-sensitivity may suggest not requiring them to do things as part of the reporting process that might risk retraumatising them.

Likewise, whereas strict fairness might imply that "malicious" reporting of online hate speech content should be made a criminal offence—based on the logic that if posting or failing to remove unlawful online hate speech content can put persons or organisations in legal jeopardy then so should false accusations of posting or failing to remove unlawful online hate speech content—victim-sensitivity could point in the direction of not creating criminal offences that might have the unintended consequence of dissuading genuine victims from making reports.

Original proposals for a victim-sensitive approach to the governance of online hate speech at the moderation, oversight and regulatory levels are set out in section VII of this study. This section also addresses potential challenges to the very idea of a "victim-sensitive" approach.

[59] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries.

C. What's the point of Internet governance for online hate speech?

In general terms, Internet governance (IG) "refers to the rules, policies, standards and practices that coordinate and shape global cyberspace".[60] In the words of Roxana Radu, "[h]undreds of governance instruments are at work to regulate the digital aspects of our lives, from connectivity to online behaviour on social networks" (Radu 2019: 1).

Looking at the current picture across Europe, "[t]he regulatory framework governing the services provided by or through platforms is diverse, multi-tiered and continuously evolving."[61] This is perhaps to be expected given that "[s]tates are confronted with the complex challenge of regulating an environment in which private parties fulfil a crucial role in providing services with significant public service value."[62]

However, this study is exclusively concerned with governance tools that are used specifically to tackle online hate speech post or shared on Internet platforms. For the purposes of this study, in other words, the phrase "governance tools for online hate speech" refers to particular packages or assemblages of norms, principles, mechanisms, systems, structures, laws, rules, regulations and adjudication procedures that are, or can be, used by governmental agencies, Internet platforms, independent supervisory councils, oversight boards, civil society organisations, etc. either separately or collaboratively, to tackle online hate speech.

(i) Three levels of governance: The moderation level, the oversight level and the regulatory level

There are several important dimensions to the question "What's the point of governance tools for online hate speech?" The first is that governance of online hate speech can operate at three levels: the moderation level, the oversight level and the regulatory level. The moderation level is where Internet platforms, or other individuals (e.g. volunteer moderators) or organisations (e.g. subcontractors), engage in the moderation of hate speech. This means assessing bits of content against Internet platforms' "community standards" or "content policies" and deciding whether to leave up, remove, add warning labels or take some other action or non-action in relation to the content. Bits of content come through to moderation by Internet platforms or subcontractors in a variety of different ways, including via user reports, flags from trusted flagger organisations, and by Internet platforms proactively identifying content using their own text extraction and machine learning tools or algorithms. Moderation decisions themselves can also be done by humans, machine learning tools or algorithms, or a combination of both.

Typically speaking, governance of online hate speech at the moderation level is concerned with content deemed to be impermissible based on Internet platforms' own community standards or content policies on hate speech. That being said, content deemed to be impermissible hate speech in this sense might also happen to be lawful or unlawful content depending on local hate speech laws.[63]

[60] Definition provided by the Internet governance project at the School of Public Policy, Georgia Tech. Available at: https://www.internetgovernance.org/what-is-internet-governance/ [last accessed 5 October 2019].

[61] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries.

[62] Ibid.

[63] In other words, normal processes of content moderation, whilst ostensibly aimed at identifiable content deemed to be impermissible based on Internet platforms' own community standards or content policies on hate speech,

The oversight level is where not only content moderation decisions but also content moderation policies and moderation guidelines, processes and procedures established by Internet platforms are subject to scrutiny and checks. This can mean external scrutiny but is not exhausted by external scrutiny. Oversight could include internal oversight processes such as an internal appeals process, for example. What matters is that there is some critical reflection or "double checking" on what is being done at the moderation level, that is to say, an evaluative assessment of content moderation that is separate from the performance of moderation itself.

A range of stakeholders can be involved in oversight, including but not limited to intergovernmental organisations, Internet platforms themselves, volunteer users, equality boards, fully independent dispute resolution services, independent supervisory councils, steering committees or oversight boards. Indeed, oversight boards in turn may be composed of members of civil society organisations, legal experts, academics, policy specialists, or members of NGOs or minority rights organisations. Oversight tends to encompass oversight of content moderation, but in theory it could also cover Internet platforms' legal compliance policies, practices and decisions. The study will return to legal compliance below. However, in that case it is more typical to classify this as regulatory governance instead.

The regulatory level is typically where national governments or governmental agencies become involved in Internet governance. To give one example, regulatory governance can be a matter of creating legislation that imposes legal responsibilities on Internet platforms, establishing statutory duties of care or creating codes of practice, and also levying fines or imposing other sanctions on Internet platforms that fail to meet the relevant legal responsibilities, duties of care or codes of practice. To take another example, regulatory governance can also involve the police or public prosecutors notifying Internet platforms about content they deem to be unlawful hate speech. Regulatory governance might also encompass governments creating bespoke criminal offences, reforming sentencing guidelines or establishing special police units—all with a view to identifying and prosecuting creators or authors of online hate speech directly.

However, Internet platforms can also be involved in regulatory governance, such as by adopting terms of service banning users from posting or sharing unlawful or illegal hate speech, and by including within their staffing and management structures "legal compliance teams" who focus on the removal of such content. Internet platforms' legal compliance teams often remove content in response to administrative notifications from the police or public prosecutors, or judicial "notice and take down" orders from courts. Nevertheless, take down decisions can also be based on identification of illegal content by legal compliance teams themselves, whether manually or using automated machine learning tools or algorithms, or based on referrals of more extreme or serious cases from Internet platforms' moderation teams. Moreover, take down decisions can also result from identification of potentially illegal content based on flags from trusted flagger organisations or reports from users using "legal report forms".

As evident in the examples listed in the previous few paragraphs, regulatory governance can be focused on tackling illegal or unlawful hate speech. Nevertheless, some forms of regulatory governance, including legal responsibilities, duties of care or codes of practice, can also speak to how Internet platforms deal with content that is impermissible, whether it happens to be lawful or unlawful. And so, governmental agencies, such as Internet regulators or government departments, can impose rules or regulations that require Internet platforms not to over-remove

might also end up, whether by design or not, removing some content that also happens to be unlawful or illegal based on local hate speech laws, as well as, of course, some content that is lawful or legal in the local context.

lawful hate speech content, and to undertake content moderation in accordance with adequate standards of due process and transparency. In that scenario, regulatory governance is concerned with what can, and should, happen at the moderation and oversight levels with respect to online hate speech content that raises questions of illegality or unlawfulness.

(ii) Outcome-oriented versus process-oriented governance tools: The NetzDG Act in Germany and the Avia Bill in France as case studies

A second important dimension to the question "What's the point of governance tools for online hate speech?" is the difference between outcome-oriented governance tools, on the one hand, and process-oriented governance tools, on the other hand. Outcome-oriented governance tools have a fixed idea of the governance outcomes they seek. Consider a regulatory system that formalises a legal responsibility on the part of Internet platforms to remove unlawful hate speech content within a specified time frame and that gives governmental authorities the power to apply to the courts for permission to impose administrative fines on Internet platforms for a pattern of failure to remove unlawful hate speech content within that time frame. The desired outcome of the regulatory intervention is that Internet platforms remove unlawful hate speech content within the time frame. This is, in one sense, an outcome-oriented governance tool.

Of course, this is an idealised model of what a governance tool could look like and real world governance regimes and instruments are often complex and hybrid in nature and do not fall neatly into idealised models or categories. For example, most of the particular provisions found in the NetzDG Act and the Avia Bill are more process than outcome oriented, in the sense that they formalise legal responsibilities on the part of Internet platforms to adopt certain processes or procedures rather than attempting to promote or hold platforms accountable for achieving particular outcomes per se. One such process or procedure might be that Internet platforms should follow adequate standards of due process and transparency in how they undertake the moderation of hate speech content, for instance. However, arguably at least some parts of the NetzDG Act and the Avia Bill are in one sense also outcome oriented. For example, both these regimes formalise a responsibility on the part of Internet platforms to adopt a pattern of good behaviour in which they remove manifestly or clearly illegal hate speech content within 24 hours. Internet platforms would be fined not for individual or specific instances of a failure to remove bits of content within 24 hours but instead they would be fined for patterns of failure to remove such content within 24 hours. Whilst in one sense these responsibilities concern processes, clearly the desired governance outcome is that manifestly or clearly illegal hate speech content is removed by the platforms within 24 hours. So in that very narrow and specific sense this particular part of these governance regimes is also outcome oriented.[64]

[64] Some people might insist that the imposition of fines should be intended and used simply as a last resort, and that in one sense the use of fines could be seen as a failure of a regulatory process that is supposed to be a constructive and collaborative enterprise between the government and Internet platforms. Comments made by Serge Abiteboul at a public lecture event titled "How social media could be regulated—Is France a role model?" hosted by Stiftung Neue Verantwortung, Berlin, 27 November, 2019. However, the fact is that the Avia Bill does formalise the legal responsibility on Internet platforms to remove clearly illegal content and does confer powers on governmental agencies to seek to impose fines for a pattern failure to remove. Arguably it does both of things precisely because governmental agencies cannot simply take it on trust that Internet platforms will comply with their responsibilities in the absence of sanctions, even if regulation is supposed to be a collaborative enterprise. In that sense fines are at the very least a sort of insurance policy aimed at achieving the desired regulatory outcome, namely, that Internet platforms remove clearly illegal content within specified time frames.

At any rate, the practical philosophy behind a governance tool (as modelled) which is oriented toward the desired outcome that Internet platforms remove unlawful hate speech content within a specified time frame can be put like this. Suppose one believes that any organisation or body, public or private, that has control over a space used by people, such as a park, a restaurant, a bus, a nightclub, or an Internet platform, has a responsibility to ensure that illegal activity is not allowed within that space.[65] It is a further question how fulfilment of that responsibility should be measured. An outcome-oriented governance tool measures the degree to which the responsibility has been met in terms of outcomes. For example, when it comes to the responsibility of bus companies to combat the problem of people putting up posters soliciting paid sex, an outcome-oriented measure might be what percentage of posters in bus shelters soliciting paid sex are removed by bus companies within a given time frame, say, within 24 hours after notification, perhaps judged over a period of 6 or 12 months. Likewise, when it comes to the responsibility of Internet platforms to combat the problem of people posting illegal hate speech content, an outcome-oriented measure might be what percentage of illegal hate speech content posted on Internet platforms is removed by the platforms within a given time frame, say, within 24 hours after notification, judged over a period of 6 or 12 months.

However, one potential problem with this particular outcome-oriented approach is that it puts pressure on Internet companies to remove content prior to any court finding that given bits of content actually *are* unlawful hate speech. Whilst the general legal responsibility might seem appropriate, as applied to particular cases of content it involves Internet platforms making a sort of pre-judgment of the correct outcome prior to a court of law ever looking at the cases in question.[66] It is one thing to ask bus companies to remove posters soliciting paid sex when it is typically clear-cut whether a poster is or is not soliciting paid sex; it is quite another thing to ask Internet platforms to remove unlawful hate speech content when there is a large grey area of cases where it is not clear-cut whether the content is or is not unlawful hate speech.

In short, this particular outcome-oriented approach underestimates the need for "judicial control": "the requirement that any exercise of such powers be subject to judicial authorisation or approval".[67] This requirement is, in the words of ECRI, "a reflection of the fundamental importance of the courts being able to exercise a supervisory role and thereby provide a safeguard against the possibility of any unjustified interference with the right to freedom of expression".[68]

Of course, one response to this problem is to only require Internet platforms to remove "manifestly" or "clearly" unlawful content within 24 hours, and to give them 7 days to remove merely potentially unlawful hate speech content (e.g. Germany's NetzDG Act), or else to not impose any responsibility to remove merely potentially unlawful hate speech content (e.g. France's Avia Bill). However, this move may severely limit the scope of the legislation, potentially rendering it applicable to only a relatively small percentage of all online hate speech content. The public may deem this to be watering down the law in an unacceptable way.

[65] Interview with Laëtitia Avia, 28 October, 2019.

[66] Note, this tendency to pre-judgment in given cases is partly mitigated if governmental authorities cannot actually levy the fine until they have obtained a warrant or permission to do so from an administrative court, say, and if the court cannot grant permission until it has heard any objections an Internet platform might make. But even so, the fact remains that the Internet platform has a legal responsibility to remove content concerning which *it* not the courts makes a judgment as to illegality. Furthermore, there is another issue of due process if court decisions to reject objections cannot be themselves legally challenged. See section IV.C(i) below.

[67] ECRI, General Policy Recommendation No. 15 on Combating Hate Speech, CRI(2016)15, Strasbourg, 8 December 2015, Explanatory Memorandum, para. 153.

[68] Ibid.

More generally, governments and intergovernmental organisations might take the view that there is nothing “strange” in requiring commercial enterprises—which are legal entities in the sense that they have both legal rights and legal obligations—to take necessary steps to ensure that laws are complied with and, where necessary, in imposing fines for failures in legal compliance. For example, commercial enterprises have an obligation to ensure that their own recruitment and promotion policies and practices do not violate anti-discrimination laws.[69] Then again, some civil liberties organisations deem that imposing a legal responsibility on Internet platforms to remove unlawful hate speech is, in effect, outsourcing quasi-judicial decisions concerning third party content to Internet platforms (Article 19 2017; GNI 2017). They take the view that it is one thing to require commercial enterprises to ensure that their own recruitment and promotion policies and practices comply with anti-discrimination laws; it is quite another thing to require Internet platforms to make determinations as to whether the content posted by third parties on their platforms is unlawful, especially given the intricacies of applying legal definitions of hate speech that are open to interpretation and highly context dependent.

By contrast, a process-oriented approach emphasises that Internet platforms have a responsibility to handle suspected cases of hate speech in a “good faith” or “responsible” fashion, irrespective of whether the final decision in any given case is to remove or not to remove. Such an approach is highly applicable to how Internet platforms handle potentially unlawful hate speech content for the reasons stated above (judgments prior to court rulings). But it is arguably also fitting where Internet platforms are routinely removing probably lawful hate speech content, where there is not an agreed outcome against which they can be judged.

Adopting a process-oriented approach suggests, for example, moving towards duties to take “reasonable steps” to combat the problem. For example, if bus companies have a responsibility to combat the problem of people putting up in bus shelters posters soliciting paid sex, then a process-oriented measure of progress in fulfilling this responsibility might ask about how many employees they hire to remove the posters and what sort of training employees receive in identifying illegal posters. Likewise, if Internet platforms have a responsibility to combat the problem of people posting illegal hate speech content, then a process-oriented measure of progress in fulfilling this responsibility might ask about how many teams of employees they hire to remove illegal hate speech content and what sort of training employees receive in determining what is illegal hate speech. The process-oriented approach puts the emphasis on companies taking reasonable steps to remove illegal hate speech content, in other words.

Importantly, the process-oriented approach also points in the direction of not imposing fines on Internet platforms for alleged patterns of failure to remove illegal hate speech content. A radical way to do this would be simply refraining from enacting, or repealing, any Internet laws that as well as imposing a legal responsibility on Internet platforms to remove unlawful hate speech content within specified times frames also impose fines for a pattern of failure to discharge this responsibility.

Another way forward is for governments to maintain these legal responsibilities and associated fines regime but provide exemptions to Internet platforms that are granted a “responsible platform” status because, for example, they make a “good faith” effort to tackle the problem of users posting illegal hate speech content. In addition, governments could institute leniency programmes involving reductions in fines for Internet platforms that fully cooperate or “come

[69] Interview with European Commission, 11 October, 2019.

clean" about the real numbers of reports of illegal hate speech content they receive and act on each year. These are original proposals from this study. See sections IV.C(iii) and IV.C(iv) below.

Yet another approach would be for governments to establish a "duty of care" or "code or practice" for how Internet platforms should handle suspected cases of hate speech, and to place a focus on process-oriented obligations such as relating to due process and transparency, for instance. In the words of the civil society organisation Carnegie UK Trust responding to the UK government's Online Harms White Paper: "In the codes, the Government chose to elaborate there is undue emphasis on notice and take down processes with the unfortunate consequence that the Government appears to prioritise these over the safety by design features inherent in a systemic statutory duty of care" (Carnegie UK Trust 2019).

(iii) Understanding the purpose or function of governance tools for online hate speech

A third key feature of the question "What's the point of governance tools for online hate speech?" is to understand the purpose or function of such tools. For example, part of the logic of governments embracing outcome-oriented governance tools for online hate speech, such as fines for under-removal of unlawful online hate speech content, is as a response to the use of outcome-based performance indicators to assess the progress of Internet platforms. On this model, organisations (e.g. governmental agencies, intergovernmental organisations, civil society organisations) assess the performance of Internet platforms by taking a sample of cases that they, the monitoring organisations, themselves deem to be unlawful hate speech and examining how swiftly, if at all, the content was removed. Based on this they come up with a percentage of how much content was removed within 24 hours, say, over the monitoring period. If the figure is 30 percent, for example, they typically judge this poor performance but if the figure is 80 percent they typically deem this excellent performance (Lomas 2017).

But these assessments involve content that has yet to be established as actually unlawful in a court of law. *If* (see above) it is problematic to outsource quasi-judicial decisions to Internet companies—that is, to place a legal responsibility on Internet platforms to remove unlawful content without any prior legal proceedings, based on high standards of due process, to determine that particular bits of content are actually unlawful—then arguably it is equally problematic to outsource quasi-judicial assessments to monitoring organisations—that is, to allow monitoring organisations to pass judgment on Internet platforms for having performed well or poorly in removing unlawful hate speech content without recourse to any legal proceedings to determine that particular bits of content were actually unlawful.

Some people might think that a fairer outcome-based performance measure would be the rate of removal of reported or flagged hate speech content *of any kind* (whether illegal or legal). But then it becomes difficult to set an ideal removal rate that is non-arbitrary. After all, the rate of removal for each Internet platform will be sensitive to the type of users, the type of content, its reporting and flagging procedures, the amount of reports or flags it receives, the resources it has for moderation, and the relative stringency of its community standard or content policy on hate speech. Thus, suppose a large, mainstream Internet platform removes around two-thirds of all content reported or flagged as being in contravention of its community standard or content policy on hate speech. Is that good or bad performance? That depends on all the contingencies listed above. Would the large, mainstream Internet platform's two-thirds removal rate constitute a better or worse performance compared to a smaller, niche Internet platform with different users, different content, different mission, different amount of resources, and a different content policy on hate speech, that receives relatively few reports and only removes half of them? Once again it is hard to say.

It is also difficult to come up with an ideal removal rate for hate speech content without some agreed standard of what hate speech is or, to be more precise, what sort of hate speech is impermissible hate speech. So long there is definitional divergence due to the fact that Internet platforms have different users, serve different purposes, have different missions, and different corporate values, for instance, there is lacking a non-arbitrary standard of impermissible hate speech that could be used as the basis for a performance indicator like an ideal removal rate. The only standards that might be universally applied are legal standards. But then the previous problem re-emerges. Therefore, progress towards a non-arbitrary ideal removal rate depends on Internet platforms harmonising their definitions of hate speech, so as to enable like-for-like comparisons. But this, of course, will require no merely coordination, goodwill and cooperation but also a potential threat to diversity and pluralism within the Internet platform sector.

Reflecting on all this, it is important to understand that the purpose or function of governance tools for online hate speech goes well beyond responding to certain sorts of outcome-based performance indicators. So what is the purpose or function of governance tools for online hate speech? There are perhaps as many different functions as there are types of governance tools for online hate speech in the first place. Some possible functions include:

- give meaningful redress to targets of online hate speech;
- address the problem of the under-removal of illegal hate speech content;
- fill a vacuum left by the inability or failure of the criminal justice system to properly enforce hate speech laws in online environments;
- counteract the underreporting of online hate speech;
- referee conflicting preferences and demands about online hate speech among users and within society at large;
- enable the removal or management of online hate speech content whilst at the same time preserving values of content neutrality and protecting freedom of expression;
- address the problem of the over-removal of legal hate speech content;
- ensure that Internet platforms operate in accordance with international human rights standards when engaged in content moderation and legal compliance;
- determine the right thing to do with grey area cases of online hate speech;
- formalise and promote collaboration among governmental agencies, Internet platforms and civil society organisations in combating online hate speech;
- establish a model of best practice for other Internet platforms to join over time;
- solve a coordination problem among stakeholders who share the policy goal of combating online hate speech;
- provide a pathway to legitimacy for how Internet platforms deal with online hate speech;
- create a common framework that promotes definitional harmonisation and policy convergence in combating online hate speech;
- provide a workable framework for removing or managing online hate speech content consistent with the business models and corporate values of Internet platforms;
- establish different sets of rules for removing or managing online hate speech content that can be suitable for different Internet platforms, large and small, and which thereby promote rather hinder fair competition and diversity in the sector.

One thing that is immediately obvious about these possible functions is that many of them cut in different directions, meaning that in practice they are likely to support the use of different governance tools for online hate speech. In practical terms this means that no single governance tool is likely to embody or fulfil all the functions at once or perfectly, and that to embody or fulfil a range of functions will require availing of a range or plurality of governance tools. But then of course the challenge of assembling a complex and divergent range of governance tools is to establish coherence and mutual reinforcement between the tools.

D. On the prima facie need for pluralism, disaggregation and integration within the governance of online hate speech

A comprehensive governance regime for tackling online hate speech will almost certainly need to integrate different types of governance tools from each of the three levels of Internet governance: that is, tools from the moderation level, tools from the oversight level, and tools from the regulatory level. This alone means that an integrative approach is necessary.

There is also a prima facie need for pluralism and disaggregation *within* each of the three levels of Internet governance. Consider the regulatory level, for instance. Hitherto there has been pluralism in the regulatory models pursued across different countries. (Contrast the NetzDG Act in Germany, which imposes both legal responsibility and liability to fines on Internet platforms, with the Communications Decency Act in the United States, which provides immunity from liability to Internet platforms but at the same time a kind of moral responsibility not to abuse that immunity). And there has been pluralism in the regulatory regimes applied to different kinds of harmful speech within the same countries. This includes different regulatory regimes for extremist speech, child abuse images, hate speech, pornography, and copyright material, for instance (e.g. in France the LOPPSI 2 Bill dealing with child pornography online and the Avia Bill dealing with cyberhate). But this has not been match by pluralism in the regulation of the same kinds of speech across different kinds of Internet platforms within the same countries.

In other words, regulatory models for the same kinds of speech (e.g. hate speech) tend to be applied at the national level to all Internet platforms operating in that country in the same way, irrespective of relevant differences between those platforms.

Differences in the size of Internet platforms are a notable and important exception to this general trend towards aggregation—for instance, some of the requirements in NetzDG only apply to "[p]roviders of social networks which receive more than 100 complaints per calendar year about unlawful content".

Nonetheless, differences between platforms in the functionalities provided to users, in their moderation, compliance and oversight practices, and in their corporate values, mission and business models tend to be ignored. This aggregated, or one-size-fits-all approach to Internet regulation may make life easier from public administration and public policy perspectives, and perhaps also from a justiciability point of view where the imposition of fines and administrative law is concerned, but it risks creating crude and unsuitable Internet regulation.

The current point is underscored in the Committee of Ministers to Member States of the Council of Europe Recommendation on the Role and Responsibilities of Internet Platforms of March 2018: "Owing to the multiple roles platforms play, their corresponding duties and responsibilities and their protection under law should be determined with respect to the specific services and functions that are performed."[70]

Aggregated or one-size-fits-all approaches potentially ignore several highly relevant forms of diversity and difference that merit more public attention and more careful recognition among governments and intergovernmental organisations.

[70] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries.

(i) Public content areas versus closed groups

First, the Internet is many things and characterisable in many ways. But among its standout characteristics is its biodiversity—both in terms of the diversity of its users but also the diversity of its different spaces or ecosystems. In *Packingham v. North Carolina*,[71] Justice Kennedy of the US Supreme Court declared:

> While in the past there may have been difficulty in identifying the most important places (in a spatial sense) for the exchange of views, today the answer is clear. It is cyberspace—the "vast democratic forums of the Internet" in general […], and social media in particular.[72]

While it is hard to dissent from the idea that the Internet can be a host for places and spaces that are important for "the exchange of views" and other important free speech activities, Justice Kennedy's characterisation testifies to an unfortunate tendency to lump parts of the Internet together. Contrary to the above opinion, in reality cyberspace is not one amorphous space "but a complex amalgam of public, private, and mixed spaces, each with their own characteristic speaker intentions and sociolinguistic conventions" (Brown 2019a: 215).

To take one example, consider the difference between public content areas and closed groups within Internet platforms. No doubt an argument could be made for large, mainstream Internet platforms applying their community standards or content policies on hate speech in the same way across all spaces on their websites, services or platforms, including public content areas and closed groups alike. But there is surely at least a debate to be had about whether specialist or artisanal platforms like Reddit, for example, should be allowed to give their communities or user groups, at the subplatform level, greater leeway in coming up with their own content policies and moderation practices. This would be a way of recognising the special purpose or the unique selling point of some spaces on the Internet, including spaces that exist inside the frameworks provided by Internet platforms: namely, to enable closed groups to moderate their own content in accordance with rules they devise themselves.

That being said, it may be that if a focus is placed on the victim's needs and interests, then a victim-sensitive approach to moderation would require the application of community standards or content policies relating to obvious and severe instances of hate speech irrespective of whether the content appears in a public content area or a closed group [see section VII.C(i)].

[71] 137 S. Ct. 1730, 1737–38 (2017), at 1735.
[72] Ibid., at 1735.

(ii) Different types of Internet platforms

This leads directly to a second, highly relevant form of diversity, namely, diversity among the types of Internet platforms where user-generated hate speech content might be posted, shared hosted or transmitted. For example:

- social networking platforms (e.g. Facebook, Tagged, Gab);
- microblogging platforms (e.g. Twitter, Tumbler);
- blogging platforms (e.g. Medium, LiveJournal);
- Internet messaging platforms (e.g. WhatsApp, Messenger, Snapchat, Slack, Discord);
- discussion forum and bulletin board platforms (e.g. Google Groups, Digital Spy, Tianya Club, 4chan, 8chan, ATRL, Forum.hr, Gaia Online);
- social news aggregation and rating platforms (e.g. Reddit, Voat, Slashdot);
- video sharing platforms (e.g. YouTube, Instagram, Vimeo, Flickr, Dailymotion, TicTok).

This diversity among Internet platforms has given rise to, or necessitated, calls within the industry for greater diversity in approaches to Internet governance. As Robyn Caplan (2018) correctly observes:

> Representatives [of Internet platforms] frequently […] note that regulations that do not consider differences between platforms threaten to "lump all the technology together in ways that do not make good sense … and fail to recognize that users will have very different purposes for accessing information on different types of platforms." The argument that legislation could be overly broad and unintentionally limit an industry is a familiar complaint from private companies worried about regulation. (Caplan 2018, 8-9)

Caplan offers one framework for disaggregation by drawing a potentially regulation-relevant distinction between three different approaches to moderation adopted by different Internet platforms befitting their different "missions, business models, and size of team": "artisanal", "community-reliant" and "industrial".

> Smaller-scale operations most often emphasize a hands-on approach to content moderation. Alex Feerst, head of legal for the blogging platform Medium, referred to their approach as "artisanal," or (being tongue-in-cheek) as "small-batch," to note that despite their more than 80 million users, their moderation approach is still done manually, "by human beings." […] Artisanal approaches are […] limited in their use of automation. (17)
>
> […]
>
> These companies also seemed to place an emphasis on learning from each case, developing rules more slowly over time, with little worry they would need to construct a black-and-white rule to be deployed by an algorithm (largely because they lack the requisite financial and technical resources, including enough data to train an algorithmic model). Because of this, their rules tend to be opaque and less consistent, leading to concerns about transparency and fairness in their application. (19)

> Community-reliant organizations are platform companies that have created structures for large groups of volunteer users to implement and add to the overarching policy decisions of a small team employed by the company. Because the users are doing a significant portion of the actual moderation, these organizations cannot be neatly understood by team size. (20)
>
> Public interest in content moderation has typically focused on a small number of larger companies—mainly Facebook and Google (YouTube primarily)—that have been called "industrial" due to their scale and number of users, the size of their content moderation teams, their operationalizing of rules, and the separation between policy and enforcement at their companies. These companies tend to have more resources and are continuing to add employees in content moderation rapidly. (23)

(iii) Different kinds of content

A third, but related, fact about diversity within the sector that might also point in the direction of imposing less restrictive governance regimes on some Internet platforms has to do with the kind of content being hosted or shared. There is arguably an important difference between an Internet platform that contains almost exclusively journalistic content, or content published by professional publishers such as newspapers and magazines, based on their own professional codes of practice, and an Internet platform that contains almost exclusively content produced by ordinary members of the public, based simply on their own sense of propriety or lack thereof. Public expectations about the appropriate level of restrictiveness of governance regimes may be very different for the former kind of platform than for the latter kind of platform.

Then again, disaggregating the governance of online hate speech by holding different Internet platforms to different standards is not without its dangers. One is that it would give would-be hate speakers an opportunity to sail around looking for safe harbours, that is, services and platforms operating under less restrictive governance regimes. Then again, if a platform does contain almost exclusively journalistic content or content published by professional publishers such as newspapers and magazines, then would-be hate speakers would find it difficult to become publishers on that platform. They would be unlikely to be granted "authorised publisher" status because the platform would be careful about its corporate relationships.

(iv) Differential reputational damage to Internet platforms

A fourth relevant dimension of diversity and pluralism within the sector is the fact that Internet platforms may differ in the reputational damage done to them by being perceived to be either under-removing or over-removing hate speech content.

This may be partly about the size of the company, type of users and content on the platform, and rate of reporting. Caplan, for example, hypothesises that when it comes to "artisanal" Internet platforms "[t]hough they had fewer resources, they also had fewer reports, and arguably, lower stakes, reputationally and financially, if they failed to make a good decision" (Caplan 2018: 19). The idea seems to be that "getting it wrong" matters more the bigger the scale.

Reputational damage caused by "getting it wrong" on the moderation of hate speech content might also be a function of the corporate values and platform features being sold by the Internet platform. An Internet platform that sells itself as a protector of free speech and as providing "a space" for controversial or offensive speech, might suffer greater damage if it adopts or is compelled to adopt governance tools that err on the side of over-removal, for example. The public may be prepared to forgive many failings in Internet platforms but selling out or rank hypocrisy might not be among them.

Then again, as user expectations about the removal of hate speech content change, so the danger of reputational damage caused by failing to act increases, changing the calculus for Internet platforms. Interestingly, the scholar Kate Klonick (2017) has argued that three of the major online platforms, Facebook, Twitter, and YouTube, "curate user content with an eye to American free speech norms, corporate responsibility, and the economic necessity of creating an environment that reflects the expectations of their users" (Klonick 2017: 1599).

(v) Differing amounts, types and degrees of harmfulness of hate speech

A fifth relevant aspect of diversity is the fact that different Internet platforms, by their nature, may attract more or less online hate speech, and different types of online hate speech, because of the sorts of users they attract. Contrast Snapchat and 8-Chan, for instance. Furthermore, the types of hate speech that different Internet platforms attract may also be more or less harmful.

Harmfulness can be partly a function of whom the targets of hate speech are. Some groups or communities may be subject to more hate speech or to qualitatively more extreme or severe hate speech. Some groups or communities (e.g. Roma) may be especially vulnerable to the negative effects of online hate speech because of their position in society or wider societal factors concerning citizenship states, differentials in power, social exclusion, economic disadvantage and oppression. Persons with intersecting vulnerable identities may also be especially vulnerable (e.g. Muslim women). According to the Institute for Strategic Dialogue (ISD), "[i]t is important to note that those with multi-intersecting identities will experience online abuse differently and in most cases be disproportionately impacted" (ISD 2019).

However, harmfulness will also depend on characteristics, features and functionalities of the Internet platforms themselves (e.g. scale of the audience, anonymity of the speaker, instantaneousness interpersonal messaging, the "piling on" against the target by others, the permanence of the content, the "captive audience" dimension of some places and spaces on the Internet) (see Citron 2014; Delgado and Stefancic 2014; Cohen-Almagor 2015; Saccardo 2016; Brown 2017e, 2018a).

Moreover, when it comes to the harmfulness of online hate speech what is arguably more relevant is not necessarily how long the content remains online, albeit the permanence of content can, and does make a difference in some instances, but the reach of the content. The length of time that content is left up on Internet platforms is at best only a crude proxy for its reach, because reach depends more crucially on how many users engage with the content and whether the platform takes action (manually or automatically) to promote or demote the content, that is, to accelerate or reduce its distribution.[73] That time left up on Internet platforms does not necessarily translate to reach and therefore to harmfulness is ironic given the focus of some regulatory regimes on removing content within 24 hours (e.g. NetzDG).

[73] 1st consultative meeting, London, 17-18 October, 2019.

Harmfulness may also reflect how many users see the content, and which kinds of users see it. The sheer scale of the audience online, for example, can make the sense of victimisation or vilification of being publicly targeted by hate speech that much more intense or profound (see Saccardo 2016; Brown 2018b). According to Paul Giannasi, for example, Greta Thunberg's Twitter account averages 30 hate messages per minute referencing in derogatory terms the fact of her Asperger's syndrome. This peaked to 500 per minute the day she met Barrack Obama.[74] It does not seem outlandish to imagine that when it comes to a sense of victimisation or vilification "numbers count".

Then again, potentially there is also something especially humiliating about being targeted by online hate speech on a social networking platform, say, in front of, virtually, one's friends, family and work colleagues (see Brown 2018b).

Thus, if the stakes are higher for large, industrial-scale Internet platforms that permit mass communication within public content areas, that is, if the risks, dangers and harms of online hate speech might be greater because of size of audience and extent of distress or humiliation caused, then the necessity of more restrictive governance tools might be greater.

Moreover, some Internet platforms allow for more anonymity than others. 4chan is an example of an Internet discussion forum where users can post messages and images anonymously. Yet it is well-understood that anonymity can be a facilitator of user-generated hate speech content. For example:

> It has been suggested that the anonymity of the Internet can provide opportunities for freer speech because people can say what they think without fear that other people will react or respond unfavourably simply because of the colour of their skin, their sexual orientation, or even their gender identity, for instance […]. This cuts both ways, however. For, there is also evidence to suggest that the Internet disinhibits speakers to say things they would not otherwise say, face-to-face […]. There are different strands to this cyber-psychological phenomenon, but one is that anonymity—even perceived anonymity—can embolden people to be more outrageous, obnoxious, or hateful in what they say than would be the case in real life […]. For instance, the perceived anonymity of the Internet may remove fear of being held accountable for cyberhate and may also evince a sense that the normal rules of conduct do not apply; the associated feeling of liberation may drive people to give in to their worst tendencies […]. (Brown 2018a: 298-9)

Likewise, many Internet platforms furnish users with opportunities for both instantaneous messaging and mass communication. Yet these same opportunities can facilitate and even encourage "forms of hate speech that are spontaneous in the sense of being instant responses, gut reactions, unconsidered judgments, off-the-cuff remarks, unfiltered commentary, and first thoughts" (Brown 2018: 304). There may be forms of Internet platform functionality, content moderation and oversight that are valuable precisely because they allow authors, reporters and moderators of hate speech content to come together to reflect in the cold light of day over a number of days about the content, and to seek a consensus.

Together the above points seem to suggest that, other things remaining equal, the lesser the amount of hate speech, the lesser the size of the audience, and the lesser the severity and

[74] Comments by Paul Giannasi, Workshop on the regulation of online hate speech hosted by Equally Ours, London, 18 September, 2019.

harmfulness of the hate speech found on a given Internet platform, the less restrictive the governance regime should be.

(vi) Pluralism of country contexts

A sixth feature of diversity is pluralism of country contexts, including socio-political conditions and the ecosystem of domestic hate speech laws and regulatory frameworks for the Internet. In a country in which hate speech has been identified as a cause of harm to well-being and dignity, and a source of social strife, violence and hate crime, in which there are robust hate speech laws in place, and where media regulators already impose strict regulations against publishing hate speech in newspapers and on television and radio, for example, it might be more fitting to impose a strict governance regime on online hate speech. By contrast, public expectations about how restrictive governance tools for online hate speech ought to be might be very different in a country where hate speech is not generally considered a problem or else is deemed a protected category based on principles of free speech, where there are few if any hate speech laws, and where media regulators in general do not ban hate speech.

An important feature of country context is the degree of what might be called “governance alignment” between the laws and regulations created and enforced by governmental authorities and agencies and the content moderation and oversight of moderation employed by Internet platforms. One aspect of this is the degree of definitional harmonisation (vertically) between Internet platforms’ “community standards” or “content policies” on hate speech (such as they exist), on the one hand, and local hate speech laws, on the other hand.

Another aspect is the degree of procedural harmonisation, that is, whether the sorts of moderation and oversight practices that governmental agencies (e.g. Internet regulators) might require Internet platforms to engage in (e.g. through a duty of care or code of practice) can already be found in the content moderation and oversight practices of the Internet platforms concerned. Where significant governance alignment exists between what governmental agencies expect and what Internet platforms are already doing, it is possible that imposing stiff fines on Internet platforms for failing to achieve certain desired regulatory outcomes (e.g. ideal removal rates) would be disproportionate, redundant and potentially counter-productive.

(vii) Implications of diversity and pluralism for the governance of online hate speech

Pulling all these different strands together, one key hypothesis of the study is that variety in the types of Internet platforms can necessitate pluralism within styles of content moderation (meaning that different styles of moderation befit different types of Internet platforms), that pluralism within styles of moderation makes appropriate pluralism within systems of oversight of moderation (meaning that different systems of oversight of moderation befit different styles of moderation), and that pluralism within systems of oversight of moderation suggests the need for pluralism within regulatory instruments (meaning that different regulatory instruments are more or less appropriate for different systems of oversight of moderation).

There is, of course, one major difficulty with this hypothesis. Even if it is true, it is likely to be unpopular among governments and impractical for the relevant governmental agencies (e.g. government departments, Internet regulators). For one thing, there is the potential resource burden involved in running parallel regulatory regimes for different categories of Internet platforms. Furthermore, governments may prefer a singularity of purpose when approaching Internet regulation, fearing that overly complex parallel regulatory regimes may lead to ambiguity and confusion and ultimately to weaker policy influence and regulatory compliance. There is also the potential political difficulty of explaining and selling highly complex regulatory regimes to the public and getting them through parliaments.

So, in the end achieving greater recognition of facts about diversity among Internet platforms is, in practice, likely to be more a question of building in exceptions, exemptions and leniency programmes under the rubric of the main or universal regulatory regime. Concrete proposals for exceptions, exemptions and leniency programmes are examined in sections IV.C and IX.1

A second key hypothesis, reflecting the fact of pluralism of country contexts [see section I.D(vi)], is that common standards for the regulation of online hate speech in Europe need not mean identical regulatory models or tools.

Suppose a new Digital Services Act at the European level incorporates rules imposing a duty on all member states to uphold a common standard on tackling online hate speech. These rules could, and should, retain three important forms of decentralisation. First, decentralised regulatory authorities, meaning each country establishes its own national regulator or devolves more powers to existing regulators.

Second, a common standard on the responsibility of Internet platforms to remove illegal hate speech content within a specified time frame but with each national regulator applying its own local hate speech laws.

Third, a common standard on the responsibility of Internet platforms to remove illegal hate speech content within a specified time frame but with each national regulator designing and implementing slightly different exceptions, exemptions and leniency programmes under this main rubric.

All of this means that regulators could work in ways that reflect national regulatory contexts, as well as enforcing responsibilities to remove content based on national hate speech laws.

E. Definitional Issues

Another important feature of the current state of play in the governance of online hate speech is the lack of definitional harmonisation across national governments, intergovernmental organisations, Internet platforms and civil society organisations. This is partly because there is not one concept of hate speech, but arguably both a legal concept and an ordinary or popular concept (Brown 2017a). Even focusing on the legal concept, there is a mismatch between the content (the particular forms of prohibited hate speech) and scope (the range of protected characteristics) of most international hate speech instruments and the content and scope of most national hate speech legislation (see Brown and Sinclair 2019: 136-8).

Turning to the ordinary or popular concept, "[i]n ordinary discourse the term 'hate speech' has become an umbrella term, as well as an opaque idiom, with multiple meanings covering a heterogeneous collection of expressive phenomena" (Brown and Sinclair 2019: 16). This heterogeneity is reflected in the following account of hate speech given in ECRI's GPR No. 15:

> Considering that hate speech is to be understood for the purpose of the present General Policy Recommendation as the advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of "race", colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status;
>
> Recognising that hate speech may take the form of the public denial, trivialisation, justification or condonation of crimes of genocide, crimes against humanity or war crimes which have been found by courts to have occurred, and of the glorification of persons convicted for having committed such crimes;
>
> […]
>
> Recognising that the use of hate speech may be intended to incite, or reasonably expected to have the effect of inciting others to commit, acts of violence, intimidation, hostility or discrimination against those who are targeted by it and that this is an especially serious form of such speech;[75]

The above overview of the varieties of hate speech is a starting point, but it can hardly claim, nor does it claim, to be exhaustive. For example, there are forms of speech not necessarily listed above but which nonetheless could qualify as "hate speech", such as if motivated by bias, prejudice and contempt for members of oppressed or vulnerable groups, if intended to stir up animosity, enmity or a lowering of civic standing towards members of such groups, if causing fear, anxiety, intimidation or humiliation among members of such groups, and so on. Consider false rumours, dogwhistles, fake news or deepfakes that target prominent Jewish political or business figures for reasons of, or in order to promote, anti-Semitism.

[75] CRI(2016)15, Strasbourg, 8 December 2015. Available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 [last accessed 7 October, 2019].

That there are so many varieties of hate speech is hardly surprising given the fact that the ordinary or popular concept hate speech serves so many distinct purposes (Brown 2017b), many of which are bound up with intense political disputes (Brown and Sinclair 2019: ch. 1).

Another reason for the lack of definitional harmonisation across governance tools for online hate speech is that governments, Internet platforms and civil society organisations often have different reasons, motives and goals for designing, developing and implementing governance tools for online hate speech in the ways that they do, and these reasons can point in the direction of quite distinct and often mutually inconsistent definitions.

To illustrate this point, many Internet platforms do not include qualifiers like "unlawful" and "illegal" in the definitions of hate speech provided in their "community standards" and "content policies" at the moderation level. For example:

> [Facebook] We define hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We also provide some protections for immigration status. We define "attack" as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation.[76]
>
> [Twitter] You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories. […] You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.[77]
>
> [YouTube] Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity, Nationality, Race, Immigration status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran status.[78]
>
> [Snapchat] Don't post any content that demeans, defames or promotes discrimination or violence on the basis of race, ethnicity, national origin, religion, sexual orientation, gender identity, disability or veteran status.[79]

However, at the regulatory level, governments *do* typically focus on the regulation of "unlawful" or "illegal" content including hate speech. The NetzDG Act is one example. Intergovernmental organisations do likewise. The European Commission's Code of Conduct on Countering Illegal Hate Speech Online, for example, includes an agreement for "IT Companies to have in place

[76] Facebook "community standard" on "hate speech". Available at: https://www.facebook.com/communitystandards/hate_speech [last accessed 7 October 2019].

[77] Twitter, "Hateful conduct policy". Available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [last accessed 7 October 2019].

[78] Youtube, "hate speech policy". Available at: https://support.google.com/youtube/answer/2801939?hl=en-GB [last accessed 7 October 2019].

[79] Snapchat, Community guidelines on hate speech. Available at: https://support.google.com/youtube/answer/2801939?hl=en-GB [last accessed 7 October 2019].

clear and effective processes to review notifications regarding illegal hate speech on their services so they can remove or disable access to such content".[80]

What is more, the majority of Internet platforms have "legal compliance" teams that work to ensure that unlawful or illegal content is removed from their platforms, services, websites or products—a form of self-regulation. They also have "terms of service" that prohibit users from posting or sharing illegal or unlawful content. Since the terms "illegal" or "unlawful" are typically unqualified within these terms of services, they also cover illegal or unlawful hate speech content even though the terms of service do not actually refer to "hate speech". For example:

> Facebook [You may not use our Products to do or share anything […] [t]hat is unlawful […].[81]
>
> [Twitter] We reserve the right to remove Content that violates the User Agreement, including for example […] unlawful conduct […].[82]
>
> [Reddit] Content is prohibited if it [i]s illegal […].[83]
>
> [Snapchat] By using the Services, you agree that: You will not use the Services for any purpose that is illegal […].[84]

Reflecting on the above differences between community standards or content codes on hate speech (the moderation level), on the one hand, and Internet laws, codes of practice and terms of service relating to hate speech (regulatory level), on the other hand, is not merely a semantic exercise. It makes a real difference to governance, and it does so precisely because definitions of hate speech found in local hate speech laws are typically narrower and less inclusive than definitions of hate speech found in Internet platforms' community standards or content policies.

Another important issue is grey area cases. For every bit of content that clearly or manifestly contravenes an Internet platform's community standard or content policy on hate speech, and for every bit of content that clearly or manifestly violates local hate speech laws (if they exist), there will invariably be several other bits of content that fall into a grey area of ambiguity. Grey area or difficult cases are those which could potentially be hate speech as defined by an Internet platform's community standard or content policy on hate speech, or that could potentially be hate speech as defined by local hate speech laws, but it is not obvious.[85]

No doubt all cases of suspected hate speech are subject to some discretion in interpretation and reasonable disagreement as to status. But grey area cases are characteristically subject to a much higher levels of discretion in interpretation and reasonable disagreement. The status of grey area cases might be heavily contested among even highly skilled, well trained and experienced practioners such as professional moderators and legal professionals.

[80] Available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en [last accessed 11 December 2019].

[81] Available at: https://m.facebook.com/legal/terms [last accessed 1 October, 2019].

[82] Available at: https://twitter.com/en/tos [last accessed 1 October, 2019].

[83] Available at: https://www.redditinc.com/policies/content-policy [last accessed 5 October 2019].

[84] Available at: https://www.snap.com/en-GB/terms [last accessed 1 October, 2019].

[85] This study will not discuss categories of speech that are clearly not hate speech but which might also fall into an area of governance ambiguity because, for example, there is widespread but hitherto unresolved public debate on whether or not such speech should be prohibited by Internet platforms under their community standards or content policies and whether or not such speech should be made illegal or unlawful under local laws.

Grey area cases pose a particular challenge for the governance of online hate speech because typically governance tools will be expected to have something to say about both the clear or manifest cases and the grey area cases. However, not all governance tools are equally capable of meeting this expectation. Below this study highlights in several places where particular governance tools are especially well placed to tackle grey area cases.

There is yet another reason why the definition of hate speech matters, namely, that some forms of online hate speech may be more dangerous or harmful than others. So if a definition includes some forms but not others, then the definition might be including more or less dangerous and harmful forms of hate speech. More research is needed on this issue.[86] As the No Hate Speech Movement (2014) observes, hitherto it has been uncommon for organisations tackling online hate speech "to distinguish between […] the worst and most dangerous forms of hate speech from those which may be merely unpleasant, disturbing or displaying racist or intolerant attitudes" (No Hate Speech Movement 2014: 23).

However, such distinctions are not unheard of. For example, within its "community standard" on "hate speech", Facebook "separate[s] attacks into three tiers of severity". One way of reading "severity" is that it speaks to degrees of the extent of harm and/or the risk of harm to victims and society in general of different forms of hate speech. At any rate, the three tiers identified by Facebook are as follows, with Tier 1 being the most severe and Tier 3 the least:

Tier 1

Content targeting a person or group of people (including all subsets except those described as having carried out violent crimes or sexual offences) on the basis of their aforementioned protected characteristic(s) or immigration status with:

- Violent speech or support in written or visual form
- Dehumanising speech or imagery in the form of comparisons, generalisations or unqualified behavioural statements to or about:
  - Insects
  - Animals that are culturally perceived as intellectually or physically inferior
  - Filth, bacteria, disease and faeces
  - Sexual predator
  - Subhumanity
  - Violent and sexual criminals
  - Other criminals (including but not limited to "thieves", "bank robbers" or saying that "all [protected characteristic or quasi-protected characteristic] are 'criminals'")
- Mocking the concept, events or victims of hate crimes, even if no real person is depicted in an image
- Designated dehumanising comparisons, generalisations or unqualified behavioural statements (in written or visual form).

Tier 2

Content targeting a person or group of people on the basis of their protected characteristic(s) with:

[86] Indeed, as Brown (2018a) points out, more research is also needed on whether the harmfulness or risk of harmfulness is of a different quality and/or order of magnitude for online hate speech as compared to offline hate speech.

- Generalisations that state inferiority (in written or visual form) in the following ways:
  - Physical deficiencies are defined as those about:
    - Hygiene, including, but not limited to: filthy, dirty, smelly
    - Physical appearance, including, but not limited to: ugly, hideous
  - Mental deficiencies are defined as those about:
    - Intellectual capacity, including, but not limited to: dumb, stupid, idiots
    - Education, including, but not limited to: illiterate, uneducated
    - Mental health, including, but not limited to: mentally ill, retarded, crazy, insane
  - Moral deficiencies are defined as those about:
    - Culturally perceived negative character trait, including, but not limited to: coward, liar, arrogant, ignorant
    - Derogatory terms related to sexual activity, including, but not limited to: whore, slut, perverts
- Other statements of inferiority, which we define as:
  - Expressions about being less than adequate, including, but not limited to: worthless, useless
  - Expressions about being better/worse than another protected characteristic, including, but not limited to: "I believe that males are superior to females."
  - Expressions about deviating from the norm, including, but not limited to: freaks, abnormal
- Expressions of contempt or their visual equivalent, which we define as:
  - Self-admission to intolerance on the basis of protected characteristics, including, but not limited to: homophobic, islamophobic, racist
  - Expressions that a protected characteristic shouldn't exist
  - Expressions of hate, including, but not limited to: despise, hate
- Expressions of dismissal, including, but not limited to: don't respect, don't like, don't care for
- Expressions of disgust or their visual equivalent, which we define as:
  - Expressions suggesting that the target causes sickness, including, but not limited to: vomit, throw up
  - Expressions of repulsion or distaste, including, but not limited to: vile, disgusting, yuck
- Cursing, such as:
  - Referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole
  - Profane terms or phrases with the intent to insult, including, but not limited to: fuck, bitch, motherfucker
  - Terms or phrases calling for engagement in sexual activity, or contact with genitalia or anus, or with faeces or urine, including, but not limited to: suck my dick, kiss my ass, eat shit

Tier 3

Content targeting a person or group of people on the basis of their protected characteristic(s) with any of the following:

- Calls for segregation

- Explicit exclusion, which includes, but is not limited to, “expel” or “not allowed”.
- Political exclusion defined as denial of right to political participation.
- Economic exclusion defined as denial of access to economic entitlements and limiting participation in the labour market.
- Social exclusion defined as including, but not limited to, denial of opportunity to gain access to spaces (incl. online) and social services.

> We do allow criticism of immigration policies and arguments for restricting those policies.
>
> Content that describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics.[87]

No doubt much more discussion needs to be had about the rationale, benchmark and evidence behind this “tiers of severity” system of classifying the type and seriousness of online hate speech. But at the very least it breaks ground and begins an important and overdue conversation. Clearly there is an important role for governments, intergovernmental organisations, civil societies organisations as well as Internet platforms themselves in thinking about whether “tiers of severity” is the right model for classifying types of online hate speech, and if so, what implications a “tiers of severity” classification system might have for making decisions about the appropriate governance tools for online hate speech.

[87] Facebook Community standards, Hate Speech. Available at: https://www.facebook.com/communitystandards/hate_speech/ [last accessed 16 April, 2020].

F. Governance firewalls: Facebook's Oversight Board as a case study

The previous section showed an important difference in the focus of governance tools. At the moderation and oversight levels the focus tends to be on impermissible hate speech defined by Internet platforms' own community standards or content policies, whereas at the regulatory level the focus shifts to "illegal" or "unlawful" hate speech defined by local hate speech laws. Interestingly, some Internet platforms seek to build a governance firewall between these levels. For example, it is a common practice among Internet platforms to report that they operate management structures aimed at achieving a clear division of labour in the governance of online hate speech between "moderation" and "legal compliance". Speaking to this division of labour, an Internet platform might state publicly that its "content moderation teams" or "community operations teams" handle cases of impermissible hate speech (such as come to their attention through user reporting mechanisms and from trusted flaggers), whereas its "legal compliance teams" handle cases of potentially unlawful or illegal hate speech (such as are brought to their attention by local law enforcement agencies and from trusted flaggers).[88]

But there is an important difference between operating a management structure with this goal in mind and actually achieving it. In reality an Internet platform's moderation teams will deal with a very substantial amount of user reports and trusted flags relating to alleged hate speech content, and, given the character of some local hate speech laws, this inevitably means that each day the moderation teams will be making take down decisions on a very substantial amount of content that also happens to be most probably unlawful hate speech content under local laws. This is true even if the moderation teams are ostensibly applying the platform's own community standards or content policies rather than local hate speech laws.

This is likely to happen because even non-identical definitions of hate speech are likely to extend to or cover some of the very same content. Consider the following hypothetical example involving social media posts in England by a far-right anti-Muslim group that has opened up accounts on Facebook and Twitter: "You simply can't deny that Muslims are terrorists, rapists and paedophiles, and that they deserve only hatred—so I urge you to hate them, and when the world comes together in its hatred of Muslims they had better watch out!". Suppose users report this post to the community operations teams or moderation teams at Facebook and Twitter. It seems highly likely that both teams would deem this to be prohibited by their (different) content policies on hate speech and so they would presumably remove the posts. But at the same time the posts would also likely to be unlawful under Part 3A of the Public Order Act 1986 in England and Wales. So the content moderation teams would be, in effect, also removing probably unlawful hate speech content, even if they did not consult with their legal compliance teams about this particular content.[89]

In other words, if there is a degree of overlap between the substantive definitions of hate speech found in an Internet platform's community standards, on the one hand, and the substantive definitions of hate speech operating within local hate speech laws, on the other hand, then the platform's moderation teams will find themselves, whether by design or not, removing hate speech content that is also illegal, irrespective of the fact that the platform also has legal compliance teams whose primary role is to deal with potentially illegal content.

[88] 1st consultative meeting, London, 17-18 October, 2019.

[89] It is worth noting that incitement to hatred laws are one of the most, if not the most, common kinds of domestic hate speech laws both in Europe and globally (Brown 2015: ch. 2).

Therefore, even though an Internet platform could reasonably say that its moderation teams are only in the business of applying community standards or content policies (including on hate speech) and are not tasked with legal compliance (the job of taking down unlawful content including unlawful hate speech content), it would be unrealistic for an Internet platform to maintain that its moderation teams will never end up taking down content that is both in contravention of its community standards and probably in breach of local hate speech laws.

Interestingly, some Internet platforms also seek to build a governance firewall between the oversight of their content moderation policies and practices on impermissible hate speech (i.e. hate speech that potentially violates their community standards or content policies), on the one hand, and any issues around legal compliance and the removal of potentially unlawful or illegal hate speech, on the other hand. They do this by framing and designing their oversight tools to "stay out" of legal issues.

Facebook, for example, has been very explicit in wanting to keep separate the work of its proposed Oversight Board, including the oversight of its content moderation policies and practices on hate speech, on the one hand, and local laws and government regulations in relation to unlawful online hate speech, on the other hand. In the words of Brent Harris again, "Facebook's oversight board is no substitute for local law".[90] This reiterates the formal position Facebook has taken in all its documentation on the Oversight Board to date. For example:

> On legal questions, for example, many have pressed for issues of local law to be included under the Board's remit. Facebook, meanwhile, has explained that it will not be in scope. (Facebook 2019, 35)
>
> [A]s one Facebook representative stated: The Board, and we're very intentional on this, will not actually be about making those decisions that are legally prohibited country by country, and the reason for that is that we actually cannot confer on the board greater authority than Facebook itself has. We, as a company, [respect] the laws of different countries and different places … this is really a delegation of authority and part of how we're envisioning exercising our responsibility, and we actually can't go beyond those lines. (36)
>
> In limited circumstances where the board's decision on a case could result in criminal liability or regulatory sanctions, the board will not take the case for review.[91]

Although the above passages appear similar, in fact they make two subtly different points. One is a crude or catchall point that the Oversight Board will not hear cases that could implicate issues of the legality or illegality of the relevant content based on local hate speech laws. The second, more nuanced point is that the Oversight Board will not hear cases in "limited circumstances" where its decisions, if acted upon by Facebook, could render Facebook's senior managers liable to criminal liability or could make Facebook as a corporate entity the target of regulatory sanctions. This point is mirrored in Art. 4 of Facebook's Oversight Board Charter: "The board's

[90] Brent Harris, Facebook, comments at invited event, "Who should regulate free speech online?", Chatham House, London, 27 June, 2019.

[91] Facebook, Oversight Board Charter, September. Available at: https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf.

resolution of each case will be binding and Facebook will implement it promptly, unless implementation of a resolution could violate the law."[92]

If the first point is taken at face value, then it could have the potential to make the Oversight Board of limited usefulness in almost any country that has reasonably inclusive or broad hate speech laws. Recall the point made above that in practice specific bits of hate speech content could fall foul of both an Internet platform's community standard or content policy on hate speech and local hate speech laws. Crucially this means that when an Internet platform's content moderation teams remove content because the content breaches the platform's community standard or content policy on hate speech, the moderation teams would be, in effect, also removing probably unlawful hate speech content in almost any country that has reasonably inclusive or broad hate speech laws. But if the Oversight Board will not hear cases that could implicate issues of the legality or illegality of hate speech content, then that renders ineligible for the Board's consideration potentially a vast amount of the Internet platform's ordinary content moderation decisions.

However, it seems likely that Facebook does not, in reality, intend or mean the crude or catchall point. In practice what could happen is that when content is reported by users or trusted flaggers to Facebook, through its moderation channels, as being potentially in breach of its community standard on hate speech, and those cases are only dealt with by its moderation teams (or "community operations teams"), then those cases can indeed be referred to the Oversight Board. Moreover, those cases would remain eligible to be referred to, and heard by, the Oversight Board even if in theory those cases could implicate issues of the legality or illegality of the relevant content based on local hate speech laws. What matters is that the cases are reported to the community operations teams and then referred to the Oversight Board as having to do with Facebook's community standard on hate speech and that the Board comes to a decision on that particular content in relation to the community standard.[93]

That then leaves the second, more nuanced point that Facebook's new Oversight Board will not hear cases in "limited circumstances" where its decisions, if acted upon by Facebook, could render Facebook's senior managers liable to criminal liability or could make Facebook as a corporate entity the target of regulatory sanctions, for example. Critically evaluating this second point is difficult unless and until the Oversight Board is up and running. The truth of the assertion that there will only be limited circumstances where the Oversight Board will not hear cases depends on how narrowly or broadly Facebook interprets the key test "could result in criminal liability or regulatory sanctions" once the Oversight Board is up and running.

If Facebook moves forward with a very narrow or super-cautious reading of this key test, then the effect could be as follows. In countries with non-existent or very limited criminal liability and soft regulatory standards for Internet platforms vis-à-vis a pattern of failure to remove illegal hate speech content, then Facebook would have little grounds to fear its senior managers being held criminally liable or to fear regulatory sanctions for such a pattern of failure. Consequently, in those sorts of countries, such as in the UK under the current status quo, Facebook would probably be willing to refer cases to the Oversight Board, whether they have come through moderation channels or through legal compliance channels within Facebook, because there is little grounds to fear criminal liability or regulatory sanctions.

[92] Facebook, Oversight Board Charter, September, 2019. Available at: https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf.
[93] Interview with Facebook, 22 November 2019.

However, it might be a different story in other countries where there is a more hostile environment for Internet platforms in terms of their facing criminal liability or regulatory sanctions for a pattern of failure to remove illegal hate speech content. In those sorts of countries it is unclear that the building of a governance firewall is appropriate, beneficial or even possible. For example, in Germany the NetzDG Act requires social networks to remove “unlawful” content within 7 days (as opposed to 24 hours for “manifestly” unlawful content). Given the variety and breadth of Germany’s hate speech laws (see Brown 2015, ch. 2; Brown and Sinclair 2019: 81-85), then potentially a great deal of content that Facebook’s moderation teams and legal compliance teams might evaluate for possible removal, and that Facebook or its users might potentially want to refer to the Oversight Board, could be at least potentially unlawful. The NetzDG Act also exposes Internet platforms to regulatory sanctions (fines) for patterns of failure to abide by their legal responsibilities to remove unlawful hate speech content within specified time frames. And so in Germany Facebook might be far less inclined to allow the Oversight Board to hear cases for fear of regulatory sanctions, especially in cases where the content has come through legal compliance channels within Facebook, that is, where it has been reported or flagged to Facebook as being potentially illegal under Germany’s various hate speech laws. This limitation on the cases the Board will hear could drastically reduce the usefulness of the Board in Germany, potentially rendering it otiose in that context.

So the question is this: would Facebook be justified in being so cautious in Germany? Here is one reason to think not. In Germany the NetzDG Act has placed a legal responsibility on Internet platforms to remove illegal hate speech content. Therefore, in one sense the government has already conferred on Internet platforms a certain sort of authority to make determinations about the potential illegality of hate speech content under German laws. In this country at least, for the Oversight Board to hear cases involving potentially unlawful hate speech content is not a case of Facebook illegitimately seizing authority over legal issues but instead responding to the legal reality that the government has already given Internet platforms authority to make quasi-judicial decisions about illegal hate speech content. In other words, in Germany Facebook cannot simply say, “decisions about illegal hate speech content are above our pay grade” because the NetzDG Act says otherwise. Moreover, where is the harm in Facebook seeking a “second opinion” concerning the potential illegality of particular bits of hate speech content under German law, especially in cases where Facebook’s legal compliance teams have deemed the content to be legal but might be in error?

Of course, Facebook could respond by making the argument that when its legal teams, perhaps also with the help of external legal counsel who are experts on local hate speech laws, have carefully considered content, they can be confident in the accuracy of the decisions they reach. Moreover, Facebook has designed the composition of the Oversight Board to include a broad range of perspectives and expertises (a kind of representativeness), such that particular subpanels of the Board dealing with specific cases probably will lack individuals with the necessary legal expertise to decide issues of legality. However, the legality of hate speech content under local laws is notoriously difficult to determine and open to wide-ranging legal interpretation and plenty of reasonable disagreement. So there would seem to be usefulness in second, third and fourth opinions, especially in grey area or difficult cases. In addition, the membership of the Oversight Board is a choice made by Facebook. It could decide instead to fill the Board with legal experts. So arguably it is not a decisive or sufficient response to say “the Oversight Board’s composition means that it lacks legal competency” because this attracts the follow up question: Why has Facebook chosen to compose the Board in that way? Is Facebook right to prioritise making the Board representative over making it legally competent?

Furthermore, there is another way in which making the Oversight Board ineligible to hear cases in certain circumstances, such as cases in Germany that fall under the NetzDG regulatory framework, would seem to be a missed opportunity. After all, under s. 3(2)3.b) of the NetzDG Act, the general requirement for Internet platforms to remove unlawful content within 7 days of receiving a report or complaint does not apply *inter alia* in circumstances where the platform refers the case to a competent independent institution ("institution of regulated self-governance") within 7 days of receipt, and agrees to accept the decision of that institution. Therefore, for Facebook to decide that its Oversight Board will not hear cases that potentially raise issues of illegality under the NetzDG regulatory framework arguably constitutes a missed opportunity on the part of Facebook to refer grey area cases of potentially unlawful hate speech to the Board, in doing so to seek to qualify for the 7 day removal exception set out above.

Then again, Facebook might also point out that NetzDG sets a high standard for the sort of institution that can be accredited as an "institution of regulated self-governance", including requirements or tests concerning its legal competence and financial independence. Since Facebook's Oversight Board is envisaged as having a diverse membership beyond merely legal experts and has a single funding source not multiple sources, then this is not a real missed opportunity after all. Then again, this invites another follow up question: Why not construct the Oversight Board so that it could meet the higher standard?

Presumably the answer is complex. But it might include: (i) Facebook intends the Board to serve globally and so cannot design it with a single country in mind, (ii) the Board emerged from a lengthy and comprehensive process of global public consultation and has credibility precisely because of this and so reengineering the Board to take up a very particular governance role outlined under NetzDG in Germany might undermine the consultation process and credibility, and (iii) Facebook could always contribute to supporting another bespoke institution in Germany that could play the particular regulatory governance role outlined under NetzDG. Indeed, Facebook along with Google has in the past 18 months or so worked to support an application by the German organisation FSM[94] to gain accreditation as an institution of regulated self-governance under NetzDG, albeit the application has yet to be approved.[95]

But there is a general point worth making nonetheless. It is that there is a good reason why Internet platforms should refer cases to competent independent institutions to seek second opinions over and above any decisions their legal compliance teams or external legal counsel might reach. The reason is that such institutions could provide legitimate checks and balances on the exercise of content removal powers by Internet platforms. Indeed, if regimes of regulatory fines could create an unwelcome bias or tendency among Internet platforms to remove suspected illegal hate speech content on a "safety first" approach, then referring grey area or difficult cases to competent independent institutions might help to mitigate that tendency. Putting this another way, the limited circumstances where Facebook takes down content because it is potentially illegal but also fears criminal liability or regulatory sanction for not taking it down are precisely the cases where checks and balances are needed the most—the sorts of checks and balances that oversight boards can deliver.

As such Facebook might consider fundamentally changing the Oversight Board, or at least a country subpanel thereof, so as to satisfy the qualifying conditions for the 7 day removal exception set out in the NetzDG Act. Alternatively, the German government might consider

[94] Information on FSM available at: https://www.fsm.de/en/about-us.

[95] Comments by Sabine Frank (Google Germany) during a session on multistakeholder approaches to tackling online hate speech, IGF, Berlin, 28 November, 2019.

relaxing the high standards for the sort of institution that can be accredited as an "institution of regulated self-governance". Or both parties could work collaboratively to reach a compromise.

Indeed, it is possible that in the future the new Digital Services Act could require member states throughout Europe to impose fines on Internet platforms for a pattern of failure to remove illegal hate speech content and could also mirror NetzDG by providing for a 7 day removal exception where platforms send grey area cases to independent institutions (e.g. oversight boards). In that event, Facebook could no longer treat Germany as a special case, and might need to rethink the nature and function of its Oversight Board for use in Europe as a whole.

Of course, it could be seen as a weakness of a model of regulatory governance that it outsources authority to make quasi-judicial decisions concerning the legality or illegality of content either to Internet platforms or to institutions of regulated self-governance (Article 19 2017; GNI 2017). But from the point of view of the Internet platforms operating under these Internet regulations, they have to do the best they can in the circumstances. And arguably building a governance firewall is not best practice even in these limited circumstances.

It is quite understandable that Internet companies want to avoid being fined for under-removal of unlawful hate speech content, but arguably the answer is not for them to simply stick with the initial decisions reached by their legal teams and external legal counsel. There can be value from the perspectives of both legal certainty and legitimate checks and balances in seeking second opinions. And, just to be clear, this value or importance holds for both initial decisions that content is illegal hate speech and for initial decisions that content is not illegal hate speech.

Indeed, just imagine the situation if Germany or some other country decided to extend its regulatory regime by imposing fines on Internet platforms both for patterns of under-removal of unlawful hate speech content and for patterns of over-removal of lawful hate speech content. This proposal is outlined in section IV.D below. Surely in those circumstances Internet platforms would have yet another good reason to refer cases, especially the grey area cases, to competent independent institutions for a second or third opinion. Such an institution could be extremely helpful to Internet platforms in helping them to tread the virtuous yet hard to discern "middle path" between under-removal and over-removal of content.

In addition to these points, it is also worth reflecting on the fact that if a country decided to adopt a purely process-oriented regulatory regime that simply required Internet companies to handle any suspected cases of illegal hate speech content with fair procedures, with transparency and in recognition of appropriate human rights standards, but with no legal responsibility to remove illegal content within specified time frames, then surely an Internet platform would have much less reason to attempt to build a governance firewall between oversight of moderation and legal compliance in the first place.

Of course, if an Internet platform has reasons relating to its particular user-base, business mission or corporate values to go beyond the scope of local hate speech laws and to remove legal hate speech that nevertheless breaches its community standards or content policies, then arguably in principle it should be able to do so based on the freedom to conduct a business. Then again, even here arguably the platform should still conduct its moderation and oversight keeping in mind process-based regulatory standards like due process and transparency.

Indeed, under the United Nations Guiding Principles on Business and Human Rights,[96] or the Ruggie Principles for short, businesses should identify, prevent, mitigate, and account for, any damage they might potentially cause to human rights, and must avail of procedures for remedying the negative consequences on human rights they cause or contribute to causing. This includes negative consequences on the human right to freedom of expression.

[96] A/HRC/17/31, 21 March, 2011, Annex. Available at: https://www.ohchr.org/documents/publications/GuidingprinciplesBusinesshr_eN.pdf [last accessed 7 October, 2019].

G. Aims, scope and methods of the study

The main aim of this study is to provide a comprehensive but also thorough mapping of varieties and innovations in the governance of online hate speech across Europe. The governance tools set out and discussed in sections II, III and IV (e.g. Regulatory-A) are intended as "ideal types" or models, in the sense that they generalise from actual examples of governance regimes or instruments used by specific governmental authorities, intergovernmental organisations, Internet platforms or civil society organisations.

In some instances the study cites actual governance regimes or instruments as illustrations of the models, but in many cases this has not been possible. Some of the governmental authorities, Internet platforms and civil society organisations who participated in the study requested that the information they provided remain fully or partially anonymous. In some instances this was because these organisations were still in the process of developing or reforming governance regimes or instruments and were not yet in a position to go public with their new systems. In other instances this was due to the fact that the information they provided to the study included some commercially sensitive insights. Nevertheless, wherever possible the study provides footnotes attributing particular information to specific organisations.

Although generalisations, the governance tools or models discussed in this study are in one important sense more focused than many of the specific examples of the governance regimes or instruments from which they generalise. For example, both the NetzDG Act in Germany and the instruments envisaged in the UK government's recent Online Harms White Paper refer to a broader range of content than the tools discussed in this study. The scope and terms of this study are narrower than this. Likewise, the Avia Bill in France sets out various legal responsibilities, not just concerning the removal of clearly illegal hate speech content within 24 hours. It is broader, therefore, than the governance model discussed in section IV.C—Regulatory-C. Thus, any attempt to critically examine, assess and measure the success or progress of specific actual governance regimes or instruments would need to take that instrument in the round, looking at how it deals with various different sorts of "harmful" content and the many legal responsibilities it imposes on Internet platforms, for instance.

It is also important to recognise that for the most part this study focuses on *potential* strengths and weaknesses, rather than concrete strengths and weaknesses based on empirical evidence, information, cases and data concerning specific actual governance regimes and instruments in action over time. This study is inherently prospective and speculative because it is mapping newly emerging governance tools many of which have yet to been initiated or put into action, and many of which have given rise to few, if any, actual legal cases or administrative proceedings. For example, during the period of this study Facebook's Oversight Board published its Charter but did not announce its membership or hear any cases.

In many instances objects of the study were "moving targets". Consider several examples. First, the Avia Bill in France has yet to be finally voted on and underwent several amendments during the period of the study. Second, at the time of writing the NetzDG Act was in the process of potentially being altered in response to an amendment bill which itself had not been finally voted on.[97] Third, an agreement establishing a working procedure between trusted flaggers, a special public prosecutor for digital crimes and Internet platforms forthcoming in a member state of the European Union (anonymous) has yet to be announced publicly and continued to undergo

[97] The amendment bill is available at: https://www.bmjv.de/SharedDocs/Artikel/DE/2020/040120_NetzDG.html [last accessed 16 April, 2020].

revisions during the period of the study. Fourth, the UK government has yet to put forward any concrete legislative plans building on the Online Harms White Paper. Whilst it did publish its Initial Consultation Response to public consultation on the Paper, this occurred after the main research for the study had been completed.[98]

As a consequence further empirical research is needed in the future to test these potential strengths and weakness against actual experience, information, cases and data. Nevertheless, this mapping study is intended to provide a useful analytical framework for this future research. It also provides a current snapshot on the innovations that have emerged in recent years. Finally, it provides an indicative assessment of potential strengths and weaknesses.

This study was completed over a six month period from June to December 2019, and then partially updated in April 2020. All research, except for the YouGov surveys in section VI.D, was carried out by the study consultant, Alexander Brown, but with additional assistance from members of staff within the Council of Europe, Directorate General of Democracy in identifying and contacting participant organisations, and in organising and facilitating two consultative meetings in London and Berlin respectively. The ideas and evidence presented, the arguments presented and the recommendations made in this study are the responsibility of the study consultant, Alexander Brown, and do not necessarily reflect the official policy of the Council of Europe, or of the various participating organisations. The study was written by Alexander Brown, but with some additional insights, suggestions and corrections from members of staff within the Council of Europe, Directorate General of Democracy.

The study consultant along with members of staff within the Council of Europe sent out invitations to engage with the study to governmental authorities, Internet platforms and civil society organisations from across Europe. The aim was to engage with a large and representative sample of organisations. In actuality the study received the most engagement overall from governmental authorities based in Belgium, France, Germany, the UK, and a member state of the European Union (anonymous), from Internet platforms Facebook, Twitter, Snapchat, and Youtube, and from civil society organisations Article 19, Gitanos, INACH, No Hate Speech Movement and UNIA. However, the full list of organisations that participated, consulted or engaged with the study can be found at the end of this report.

When sending requests or invitations to organisations to participate in the study, an attempt was made to achieve: balance between governmental agencies, Internet platforms and civil society organisations; balance among governmental agencies between government departments, regulators, police and public prosecutors; balance among Internet platforms between large or mainstream platforms and smaller more specialised platforms; and balance among civil society organisations between equality boards or associations thereof, trusted flaggers or associations thereof, monitoring bodies or associations thereof, NGOs with a focus on protecting particular groups, and civil liberties organisations with a focus on promoting free speech. An attempt was also made to achieve balance among the European countries in which these agencies, companies and organisations were based or operated.

Organisations were initially identified based on their reputation in this area of human rights, based on their own reports and studies on the issues discussed in the study, existing contacts with the

[98] UK government, Initial Consultation Response to public consultation on the Online Harms White Paper, 12 February, 2020. Available at: https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response [last accessed 16 April, 2020].

Council of Europe (Directorate General of Democracy, especially ECRI and the No Hate Speech Movement), and existing contacts with the study consultant, Alexander Brown.

Finally, the study relied on several methods as follows:

A meta-survey of existing policy position statements, studies, reports, articles, books and commentaries on the governance of online hate speech produced by governmental agencies, Intergovernmental organisations, Internet platforms, civil society organisations, NGOs, equality boards, media journalists and academics. This content is contained in the list of references at the end of the study.

Semi-structured interviews with senior members of staff within approximately 80 percent of the organisations listed as participating, consulting, or engaging with the study.

Questionnaire responses received from approximately 25 percent of the organisations listed as participating, consulting, or engaging with the study.

Personal narrative statements ("your stories") received from individuals on a self-selecting basis. These individuals were contacted on behalf of the study by organisations listed as participating, consulting, or engaging with the study. They were all in that sense "end users" of these organisations' services. The response rate was approximately 2 percent.

Notes taken from consultative meetings on preliminary findings of the study held in London, 17-18 October, 2019 and Berlin, 26 November, 2019 involving approximately 70 percent of the organisations listed as participating, consulting, or engaging with the study.

Yougov public opinion survey methodology (UK): This survey has been conducted using an online interview administered members of the YouGov Plc GB panel of 185,000+ individuals who have agreed to take part in surveys. An email was sent to panellists selected at random from the base sample according to the sample definition, inviting them to take part in the survey and providing a link to the survey. (The sample definition could be "GB adult population" or a subset such as "GB adult females"). YouGov Plc normally achieves a response rate of between 35 percent and 50 percent to surveys however this does vary dependent upon the subject matter, complexity and length of the questionnaire. The responding sample is weighted to the profile of the sample definition to provide a representative reporting sample. The profile is normally derived from census data or, if not available from the census, from industry accepted data. YouGov plc make every effort to provide representative information. All results are based on a sample and are therefore subject to statistical errors normally associated with sample-based information. All figures, unless otherwise stated, are from YouGov Plc. Total sample size was 1,633 adults. Fieldwork was undertaken between 8th - 9th November 2019. The survey was carried out online. The figures have been weighted and are representative of all GB adults (aged 18+).

Yougov public opinion survey methodology (France): All figures, unless otherwise stated, are from YouGov Plc. Total sample size was 1008 adults. Fieldwork was undertaken between 13th - 14th November 2019. The survey was carried out online.

The figures have been weighted and are representative of all adults (aged 18+) in France.

Yougov public opinion survey methodology (Germany): All figures, unless otherwise stated, are from YouGov Plc. Total sample size was 2055 adults. Fieldwork was undertaken between 12th - 14th November 2019. The survey was carried out online. The figures have been weighted and are representative of all adults (aged 18+) in Germany. YouGov is a member of the British Polling Council.

II. FIRST LEVEL OF INTERNET GOVERNANCE: THE MODERATION LEVEL

At the first level of governance of online hate speech, the moderation level, the Internet platform uses governance tools to directly tackle hate speech content that has been or could be posted, published or transmitted on its platform, service, website or product.

A. Professionalised moderation

Under what this study will label governance tool Moderation-A, the Internet platform itself undertakes moderation of hate speech content based on the platform's "community standard" or "content policy" on hate speech (such as it exists). Table 1 sets out potential strengths and weaknesses of Moderation-A, including some of its variants.

Table 1. Moderation A: Professionalised moderation

Tool	Strengths	Weaknesses
Moderation-A: The Internet platform undertakes moderation of hate speech content. Moderation is based on the Internet platform's content policies and moderation guidelines, processes and procedures.	- Internet platforms retain control over content and can undertake moderation based on their corporate values, mission and business model - Potential for Internet platforms to retain control over training and development within moderation systems - Moderation is not subject to government enforcement of moderation practices which places a limit on state control over Internet content	- Potential for moderation to reflect group think within the Internet platform - Moderation may be unrepresentative of the general population both demographically and in terms of norms - Potential threat to free speech if the platform's content policies on free speech are overly broad
Collaboration potential		
Medium (Internet platforms; users; trusted flaggers)		

Tool variants		Strengths	Weaknesses
When does moderation take place	Pre-publication moderation	- Potentially prevents users from being exposed to hate speech content in the first place	- Potentially greater threat to free speech since users will be denied the opportunity to post, share or access content in the first place - Less suited to grey area cases
	Post-publication moderation	- Potentially allows more time for content to be assessed - Suitable for grey area cases	- Post-publication moderation may not prevent users from seeing hate speech content prior to its removal
Responsibility for moderation policymaking and moderation enforcement	Same teams decide content policies and undertake moderation	- Ensures that there is immediate and direct feedback from moderation enforcement to moderation policymaking - Suited to small, artisanal Internet platforms	- Potential that "ease" of moderation enforcement restricts the ambition of moderation policymaking
	Separate teams decide content policies and undertake moderation	- Suited to large, mainstream platforms	- Potentially reduces the extent to which the day-to-day experiences and expertise of moderation teams directly and immediately feed into decisions about content policies

A potential benefit of Moderation-A is that moderation remains under the control of private enterprises and is not subject to government enforcement of moderation policies and practices. This in itself places a limit on state control over Internet content. This may be especially important in countries with authoritarian or illegitimate governments who may seek to use state control over Internet content as a vehicle for censorship and silencing dissenting voices.

Of course, within Moderation-A there will be significant variation between Internet platforms in terms of respect for human rights, due process, transparency, and so on. And so for this reason many civil liberties organisations campaign for Internet platforms to adopt best practices. In the words of the Electronic Frontier Foundation, for example:

> YouTube's moderation of the videos, like many of its content moderation decisions, was faulty in many ways […]. And YouTube is not alone: Facebook, Twitter, and others have made, and will continue to make, wrong decisions to take down content, and we will continue to call them out for it.
>
> But the answer to bad content moderation isn't to empower the government to enforce moderation practices. Rather, the answer, as we told the court, is for users' platforms to adopt moderation frameworks that are consistent with human rights, with clear take down rules, fair and transparent removal processes, and mechanisms for users to appeal take down decisions. Our brief thus concludes with a discussion of the Santa Clara Principles, a set of minimum standards we helped craft for content moderation practices that provide meaningful due process to affected speakers and better ensure that the enforcement of content guidelines is fair, unbiased, proportional, and respectful of users' rights. (Greene 2018)

Notwithstanding the merits of the Santa Clara Principles in themselves, ruling governments in mature democracies with long and consistent track records of the rule of law, democratic elections, checks and balances on the exercise of state power, and so on, may take the different view that governmental agencies (e.g. Internet regulators) not merely have a legitimate right but also have a duty to the public to enforce minimum standards on content moderation, such as through a statutory duty of care or code of practice [see section IV.E].

Indeed, the public might argue that for Internet platforms to retain total control over content moderation is not much better and not much more legitimate than handing the control over to governments. Either way, there is a problem of "absence of consent", in the words of Rebecca Mackinnon (2012). For example: "When we sign up for Web services, social networking platforms, broadband service, or mobile wireless networks, and we click "agree" to the terms of service, we give them false and uninformed consent to operate as they like" (ibid).

Concerns about Internet platforms having the power—potentially without full and informed consent—to remove content may be even greater for pre-publication moderation than for post-publication moderation. These and other concerns are set out in Table 1 above.

Notably, in her study of Internet platforms Caplan (2018) notes an important difference between "artisanal" and "community reliant" forms of moderation, on the one hand, and the sort of "industrial" moderation typical of Internet platforms like Facebook, namely, that in the former case the same (small) teams perform both the development of the content policies and the enforcement of those policies (Caplan 2018: 19), whereas in the latter case there is a "separation between policy and enforcement at their companies" (23). The potential strengths and weaknesses of these variants are also listed in Table 1 above.

Other possible variants of Moderation-A cluster around different ways of selecting content for moderation and around different answers to the question of who or what makes the moderation decisions. The strengths and weaknesses of these variations are set out in Table 2 below.

Table 2. Moderation A: Professionalised moderation (selection of content and decision-maker variants)

Tool			
Moderation-A: (*continued*) The Internet platform undertakes professional moderation of hate speech content using a combination of full-time employees and contract labour			
Tool variants		Strengths	Weaknesses
Selection of content for moderation	Content (i) reported by users	- Remedy for victims - Democratisation of reporting - Gives users "ownership" over content	- Potentially subject to frivolous or malicious reporting - Reliant on users having the technical know-how to report content - Reliant on users feeling safe and secure to report content
	Content (i) reported by users, and (ii) flagged by trusted flaggers, but with greater urgency, priority and prima facie credence given to trusted flaggers	- Potential for greater impartiality among trusted flaggers - A prioritisation system is useful in circumstances of excess workload for human moderators	- Potential lack of transparency over the choice of trusted flaggers - Potential lack of independence and neutrality of trusted flaggers - Potential lack of balance due to absence of "second options" or trusted *un*flaggers - Risk of Internet platforms over reliance on trusted flaggers and not being proactive in identifying content
	Content (i) reported by users, (ii) flagged by trusted flaggers, and (iii) automatically flagged machine learning tools or algorithms	- Capacity to deal with vast amounts of content - Text extraction and machine learning tools or algorithms less subject to bias and human error	- Potential increase in the quantity of content being presented for moderation - Potentially only larger Internet platforms have the wherewithal to develop machine learning tools or algorithms - Limitations in the accuracy of machine learning tools or algorithms (over- and under-flagging) - Reliant on the quality of the "training data" or "benchmark data set" used within the machine learning tools or algorithms
Who or what makes the moderation decisions	Moderation decisions taken by human beings	- Human beings are more sensitive to semantic nuance, linguistic context, slang, wider social context, etc. - Potentially greater public confidence in moderation decisions - Suitable for grey area cases	- Human beings prone to prejudice, unconscious bias and human error - Potential for psychological or emotional harm to human moderators from extended exposure to hateful content if not properly trained and supported - High cost of human moderation - Slower speed of moderation decisions make human moderation less suitable for pre-publication moderation
	Moderation decisions taken automatically by machine learning tools or algorithms	- Potentially capable of much faster decision-making - Capable of meeting the scale of content needing to be moderated - Potentially more suitable for pre-publication moderation	- Public tolerance of errors made by automated moderation tools potentially lower than of errors made by human moderators - Reliant on the quality of the "training data" or "benchmark data set" used within the machine learning tools or algorithms - High upfront cost of recruiting developers with skills to create accurate automated moderation tools - Potentially only larger Internet platforms have the wherewithal to develop automated moderation tools - Less suitable for grey area cases - Potential reduced accountability for specific moderation decisions

In terms of the selection of content to go forward for moderation, one variant involves Internet platforms identifying content from multiple sources, including (i) user reports, (ii) flags from trusted flaggers, and (iii) automated flags from text extraction and machine learning tools or algorithms. To give an indication of the extent to which Internet platforms are already making use of automated content flagging systems, consider the following insights from Sabine Frank of Google Germany. “In Q2 [2019] we have removed 9 million videos—78 percent of the videos have been flagged by machine learning technology and [in] 81 percent of these cases no human view has been on that video before it has been removed.”[99]

In terms of who or what makes the moderation decisions, one variant of Moderation-A involves human moderation, typically using a combination of full-time employees and contract labour depending on the size, resources, business model and corporate values of the Internet platform. Moderation teams or “community operations teams” apply the platform’s “community standard” or “content policy” on hate speech to particular bits of content presented to them for human decision-making. Moderators will receive training, guidelines, procedures and protocols for carrying out content moderation, such as training on examples of content that the Internet platforms has already deemed to be in contravention of its content policy on hate speech.

However, human moderation can pose a risk to the psychological or emotional health and well-being of moderators, including professional moderators. Internet platforms differ in how much support is given to human moderators located at their headquarters, at their country offices and at the offices of their subcontractors. As a result, levels of protections for employees and contract labour involved in human moderation can vary within the tech industry, as with other industries. Moderation-A puts a premium on the corporate values and management of Internet platforms in terms of their attention to, and delivery of, protection of human moderators.

The second variant involves automated moderation based on machine learning tools or algorithms. Whilst it does not raise employee protection issues, as pointed out in a recent report by Ofcom (2019) it does have some important potential weaknesses. For one thing, public tolerance of errors made by automatic moderation tools is potentially lower than of errors made by human moderators. Moreover, automated moderation may be less suited to grey area or difficult cases, where there is greater need for understanding of and sensitivity to semantic nuance, linguistic context, slang, and wider social context.[100] These and other potential strengths and weaknesses of automated moderation are set out in Table 2 above.

[99] Comments by Sabine Frank (Google Germany) during a session on multistakeholder approaches to tackling online hate speech, IGF, Berlin, 28 November, 2019.

[100] 1st consultative meeting, London, 17-18 October, 2019. 2nd consultative meeting, Berlin, 26 November, 2019.

B. Distributed moderation

An alternative to professionalised moderation is for the Internet platform to invite volunteer users to undertake moderation of hate speech content on its behalf. Call this governance tool Moderation-B. Moderation-B is a form of decentralized, peer-to-peer, democratic, user-led moderation or distributed moderation. Strengths and weaknesses of this tool are listed in Table 3 below. Table 3 also contains potential strengths and weaknesses of variants of Moderation-B based on whether volunteer moderators do or do not also have a say over the content policies and moderation guidelines, processes and procedures that provide the moderation framework.

Table 3. Moderation B: Distributed moderation

Tool	Strengths	Weaknesses
Moderation-B: Volunteer users undertake moderation of hate speech content on behalf of the Internet platform (distributed moderation)	- Potential to achieve democratisation of moderation - Reduces the moderation burden on Internet companies - Enables innovation and market disruption among small or start-up platforms and services - Potentially mitigates the negative effects of any oligopolistic tendencies within the Internet industry	- Outsourcing of moderation could create ambiguity over where liability rests for unlawful moderation - Risk of lower levels of training, expertise, professionalism among volunteer moderators - Risk of inconsistent or arbitrary moderation - Moderation will be heavily influenced by the ethos fostered by the platform and the culture among the community of moderators - Moderators may be unrepresentative of the general population both demographically and in terms of norms and ethos - Risk of malicious, ill-informed or discriminatory moderation - Public perception of exploitation of unpaid labour - Risk of harm to the well-being and mental health of volunteers moderators if receive less support than employees or contract labour - Moderation may not prevent users from seeing hate speech content prior to its removal
Collaboration potential		
Medium (Internet platforms; volunteer users)		

Tool variants		Strengths	Weaknesses
Whose content policies	Volunteer users undertake moderation based on the content policies and moderation guidelines, processes and procedures provided by the Internet platform	- Partial democratization of moderation - Internet platform retains control over moderation and can pursue its corporate values, mission and business model	- Moderators may misunderstand or misinterpret both the letter and intention of policies and guidelines - Polices and guidelines potentially fail to reflect the experiences and insights of volunteer moderators
	The Internet platform gives volunteer moderators of groups and communities within the platform or service significant discretion to determine their own content policies and moderation guidelines, processes and procedures	- Potentially increases the motivation and “buy in” of volunteer users - Full democratization of moderation - Suited to smaller, artisanal Internet platforms whose corporate values and business mission emphasises the importance of democratising the Internet	- Internet platform gives up control of moderation - Potentially unsuitable for mainstream Internet websites, platforms and services - Outsourcing moderation could create ambiguity over where liability rests for unlawful moderation - Risks lowering the quality of content policies and moderation guidelines, processes and procedures - Moderators may be unrepresentative of the general population - Risks allowing malicious, ill-informed or discriminatory content

Reddit is an example of a platform that uses distributed moderation. Whilst Reddit has content policies that do not explicitly refer to "hate speech", the policies do prohibit content if it is *inter alia* "illegal", "encourages or incites violence" or "Threatens, harasses, or bullies or encourages others to do so".[101] But due to its utilisation of distributed moderation techniques the day-to-day interpretation and impact of Reddit's content policies largely falls on the shoulders of its community of volunteer moderators. Interestingly, studies of Reddit's distributed moderation model suggest that there is a tendency among its moderators to refrain from removing or "shredding" content given the wider ethos of the platform as being avowedly pro-free speech (Massanari 2017). This ethos feeds into the culture among its community of moderators, where a sort of peer pressure can make it harder for individual moderators to take a stand against particular examples of hate speech content that would certainly be removed by other platforms.

[101] Available at: https://www.redditinc.com/policies/content-policy [last accessed 5 October 2019].

C. Pre-moderation by professional publishers of content

Under what this study labels governance tool Moderation-C, the Internet platform expects the pre-moderation of hate speech content to be undertaken by the professional publishers that use the platform to post text, images and videos to public content areas. Pre-moderation involves the checking of content prior to posting against relevant content policies. Potential strengths and weaknesses of Moderation-C, and variants, are set out in Table 4 below.

Table 4. Moderation-C: Pre-moderation by professional publishers of content

<table>
<tr><td colspan="2">Tool</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td colspan="2">Moderation-C: The Internet platform expects pre-moderation of hate speech content to be undertaken by the professional publishers that use the platform or service to post text, images and videos to public content areas (e.g. newspapers, magazines, other media businesses, agents or PR for celebrities)</td><td>- Reduces the resource burden on the Internet platform
- Filters out hate speech content before users see
- Prior blocking of content may significantly reduce the risks of distress, traumatisation, fear, humiliation, etc.</td><td>- Not suitable where the platform or service is used by ordinary users as well as professional publishers to post public content
- The Internet company still has to undertake its own content moderation so potentially involves inefficient duplicative moderation</td></tr>
<tr><td colspan="4">Collaboration potential</td></tr>
<tr><td colspan="4">Medium to low (Internet platforms; professional publishers)</td></tr>
<tr><td colspan="2">Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td rowspan="2">Whose content policies?</td><td>The Internet platform expects pre-moderation to be based on its own content policies</td><td>- Promotes definitional harmonisation of “hate speech” between the Platform and its professional publishers
- Potentially means the Internet platform need not undertake its own secondary moderation of the content</td><td>- The expectation of pre-moderation could potentially restrict the number of professional publishers who might be interested in using the platform (i.e. commercial opportunities)</td></tr>
<tr><td>The Internet platform allows pre-moderation to be based on the professional publisher’s content policies but expects consultation and dialogue on policies</td><td>- Enables professional publishers to use the platform as a tool whilst maintaining their own moderation standards and “creative control”</td><td>- Risks definitional divergence
- Outsources moderation to professional publishers of content
- Creates ambiguity around the liability of the Internet platform for unlawful content</td></tr>
</table>

Whilst Moderation-C is a simplified and generalised description of one possible moderation tool, there are examples of actual platforms that share at least some features in common with it. Snapchat is one such example. Whilst it is primarily a 1:1 or closed group messaging platform, it does also have public content areas. The vast majority of content in its public content areas comes from professional publishers with whom Snapchat has commercial relationships. These publishers undertake pre-moderation based on their own content policies and professional codes of conduct (where applicable), but also in accordance with Snapchat’s content policies or “community guidelines”, which explicitly prohibit *inter alia* “hate speech”.[102]

Under one possible variant of Moderation-C, because professional publishers already pre-moderate content in accordance with the platform’s standards, the platform does not undertake its own secondary or duplicate pre-moderation of professional publishers’ content. The benefit for the platform is reduced resource burden. The weakness is that this approach might be viewed by the public as a way of passing responsibility onto other organisations.

[102] Available at: https://www.snap.com/en-GB/community-guidelines [last accessed 5 October, 2019].

D. Facilitated user self-moderation

Under Moderation-D, Internet platforms create optional (“opt-in”) functionalities that give users the ability to block certain words, phrases or emojis from appearing on the displays or collections of content they access through the platforms (e.g. “filtering” and “safe search” functionalities). These functionalities can give users the ability, for example, to block racist epithets or slurs from appearing on displays or collection of content they access. Potential strengths and weaknesses of Moderation-D, and its variants, are set out in Table 5 below.

Table 5. Moderation-D: Facilitated user self-moderation

<table>
<tr><td colspan="2">Tool</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td colspan="2">Moderation-D: Internet platforms create opt-in functionalities that give users the ability to block certain words, phrases or emojis from appearing on the displays or collections of content they access through the platforms</td><td>- Gives users control over moderation
- Provides potential “victims” of hate speech a form of empowerment
- Means that the user would not see certain content in the first place rather than seeing it and having to select to hide it, block other users from posting it, or reporting it to the platform
- Prior blocking of content may significantly reduce the risks of distress, traumatisation, fear, humiliation, etc.</td><td>- Relies on users being “computer savvy” or having necessary skills
- Places a time burden on users
- Potentially creates ambiguities over the legal liabilities of platforms
- Does not remove hate speech from the platform in general
- Potentially makes users ignorant about the extent of cyberhate
- Potentially could make users less resilient</td></tr>
<tr><td colspan="4">Collaboration potential</td></tr>
<tr><td colspan="4">Low to Medium (Internet platforms; users)</td></tr>
<tr><td colspan="2">Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td rowspan="2">Which content areas</td><td>Content blocking function enabled for users’ personal content pages, recent activity pages, or timelines, which typically contain content by them, content selected by them, content about them, or content in response to them</td><td>- Arguably important for users to have moderation control over personal content pages since these may be seen as especially personal, sensitive or private spaces</td><td>- Potentially facilitates users occupying “echo chambers”</td></tr>
<tr><td>Content blocking function enabled on users’ general content areas, newsfeeds or front pages, which typically contain any sort of content</td><td>- Enables users to turn general content areas of the Internet platform into “safe spaces” for them to use</td><td>- Potentially could be used by users to turn general content areas into “echo chambers”</td></tr>
</table>

E. Auto-moderation

Under Moderation-E, Internet platforms set up their servers so that content is automatically deleted after a specified period of time, say, 24 hours or 7 days. Potential strengths and weaknesses of Moderation-E, and its variants, are set out in Table 6 below.

Table 6. Moderation-E: Auto-moderation

Tool		Strengths	Weaknesses
Moderation-E: Internet platforms set up their servers so that content is automatically deleted after a specified period of time, say, 24 hours or 7 days		- Potentially saves on moderation costs - Prevents hate speech content from remaining on platforms for lengthy periods - Potentially provides more of a "safe space" for users	- Potentially limits the number of users that will sign up to the platform - Perversely the ephemeral nature of the content could lead some users to be even more disinhibited and may actually increase hate speech content - Does not prevent users from seeing hate speech content prior to deletion
Collaboration potential			
Low (Internet platforms)			
Tool variants		Strengths	Weaknesses
Which content areas	Automatic deletion applies to direct messaging between users	- May be important to apply automatic deletion to direct messaging if hate speech content within direct messaging poses a higher risk of distress, traumatisation, fear, humiliation, etc.	- Might not be the least restrictive option, such as if users can already turn on a function that prevents strangers from sending them direct messages - May reduce the desirability of a platform as a direct messaging tool
	Automatic deletion applies to content posted on users' personal content pages, recent activity pages, or timelines	- May be important to apply automatic deletion to personal content pages if hate speech content within personal content pages poses a higher risk of distress, traumatisation, fear, humiliation, etc.	- Might not be the least restrictive option, such as if users cam already turn on a function that allows them to block strangers or block certain content
	Automatic deletion applies to all content areas	- May be important to apply automatic deletion to all content areas if hate speech content poses similar risk of distress, traumatisation, fear, humiliation, etc.	- Might not be the least restrictive option, such as if moderation can be successfully done within the same period - May significantly reduce the usefulness of a platform as a social networking platform, a microblogging platform, a video sharing platform, etc.

Interestingly, Snapchat has optional ("opt-in") functionality that involves some automatic deletion of direct messages ("snaps", "chats") after specified periods, and also automatic deletion of personal content ("my story") after specified periods.[103] Hitherto automatic deletion functionality has been an important part of Snapchat's unique selling point. More recently, however, Snapchat has allowed users to opt-in to other functionalities that enable direct messages to appear on the platform more permanently (archiving messages, "memories").[104] Presumably this has been done in response to market research about its users' preferences.

However, typically social networking platforms, microblogging platfoms, and video-sharing platforms have not adopted automatic deletion functionality, not even as an opt-in function.

[103] Information available at: https://support.snapchat.com/en-GB/article/when-are-snaps-chats-deleted [last accessed 24 October, 2019].

[104] Ibid.

F. Content management

Moderation-F involves content management whereby Internet platforms take practical and technological steps to reduce access to potential hate speech content, as distinguished from more traditional forms of moderation such as content removal. Steps involved in reducing access might include: reducing content distribution, adding warning labels to content, preventing users from adding comments below the content, disabling likes and dislikes functionality for content, preventing the sharing of content, making the content ineligible for Ads, excluding content from sponsored content, promoted content and recommended content features, and filtering content.[105] Other steps might include preventive measures such as Internet platforms highlighting to users the existence of their community standards or content policies on hate speech, so as to discourage users from posting or sharing hate speech in the first place. Content management tools are the obverse of content acceleration, amplification and virality. Potential strengths and weaknesses of Moderation-E are set out in Table 7 below.

Table 7. Moderation-F: Content management

Tool	Strengths	Weaknesses
Moderation-F: Internet platforms take steps to reduce access to potential hate speech content	- Less restrictive measure whilst still reducing the risks of harmful exposure to hate speech content - Potentially more protective of free speech simply due to the fact of non-removal - Reduced resource burden on Internet platforms - Useful for grey area cases and less severe hate speech content	- Non-removal of content might not eliminate entirely the risks of harmful exposure to hate speech - Potentially more reliant on machine learning tools or algorithms and therefore susceptible to inaccuracies - Symbolism of non-removal may have unintended consequences (e.g. implicit condoning)
Collaboration potential		
Low (Internet platforms; users)		

Arguably this sort of content management is most appropriately applied to (1) grey area cases involving content that is not manifestly or clearly in violation of the relevant community standard or content policy on hate speech and (2) less extreme or severe content that whilst clearly being in violation of the relevant community standard or content policy on hate speech, is not a severe or extreme case in itself.[106] Recall from section I.E Facebook's three-tier approach to defining hate speech.

That being said, it may be that in the future due to the phenomenon of identical or very similar hate speech content appearing across multiple platforms at the same time, due to the sheer volume of content to be managed, and due to limitations in human resources in the sphere of moderation, Internet platforms will become increasingly reliant on content management rather than traditional content moderation for all forms of hate speech content.[107]

However, as outlined in Table 7 above, content management as a governance tool for online hate speech is potentially more reliant on machine learning tools or algorithms, and therefore susceptible to inaccuracies.

[105] 2nd consultative meeting, Berlin, 26 November, 2019. Internet platform questionnaire response 1, 19 November, 2019.
[106] 2nd consultative meeting, Berlin, 26 November, 2019.
[107] Ibid.

Interestingly, Twitter has recently adopted the tool of using warning labels (a type of content management) for tweets that violate its content policy on hate speech when the tweets originate from the accounts of political figures.[108]

For critical discussion of whether this is compatible with a victim-sensitive approach to the governance of online hate speech, see section VII.A below.

[108] See Twitter Rules, About Public-Interest Exceptions on Twitter. Available at: https://help.twitter.com/en/rules-and-policies/public-interest.

III. SECOND LEVEL OF INTERNET GOVERNANCE: THE OVERSIGHT LEVEL

As explained in section I.C, the second level of governance of online hate speech, the oversight level, is where not only the content moderation decisions but also the content policies and moderation guidelines, processes and procedures established by the Internet platform are subject to scrutiny and checks. This too can take numerous different forms.

A. Public consultation on content policies and moderation guidelines, processes and procedures

Oversight-A involves a form of oversight in which Internet platforms critically evaluate their content moderation policies and practices through a process of public consultation, inviting and taking account of the ideas, opinions, interests, values and ultimately social norms of society. It is the least demanding, rigorous and restrictive form of oversight, so much so that it might reasonably be considered a limiting case, that is, at the lowest end of what oversight could be. Potential strengths and weaknesses of Oversight-A, and its variants, are set out in Table 8.

Table 8. Oversight-A: Public consultation

Tool		Strengths	Weaknesses
Oversight-A: Internet platform critically evaluates its content policies and moderation guidelines, processes and procedures concerning hate speech content based on public consultation (e.g. users, trusted flaggers, governmental agencies, civil liberties organisations, minority rights groups or NGOs, and other stakeholders)		- Expands the knowledge and expertise base of moderation - Recognises the significant role in public life of moderation - Provides an avenue for the public to influence moderation - Provides a form of public accountability - Need for public consultation befits the size and reach of (some) Internet platforms and services and the important of quality moderation	- Potentially weak impact or policy influence - Symbolism of the Internet platform not responding fully to the public consultation may have unintended consequences (e.g. appearance of a PR exercise)
Collaboration potential			
Medium (Internet platforms; the public)			
Tool variants		Strengths	Weaknesses
Transparency	Internet platform undertakes the public consultation but does not publish the submissions or its response to the public consultation	- Reduces the risk of creating unmanageable expectations of policy influence or impact - Potentially creates a “safe space” for organisations to make submissions on commercially sensitive issues and for the Internet platform to discuss these issues frankly but in private	- Potential for weak policy influence or impact - Lack of transparency - Gives the appearance of a PR exercise
	Internet platform undertakes the public consultation and publishes the submissions and its response to the public consultation	- Increased transparency - Provides a form of public accountability - Potential for improved credibility in the community - Potential to increase public support for moderation	- More time-consuming and expensive - Risks diluting fundamental corporate mission - Potential to decrease public support for moderation

B. Internal appeals processes

Another form of oversight is provided by Internet platforms' own internal appeals processes, Oversight-B. Under these processes, users who report content, or whose content has been reported, as being in contravention of the Internet platform's community standard or content policy on hate speech, can appeal against the Internet platform's moderation decision to leave up or take down the reported content. The appeals are considered by the Internet platform with reference to the same community standard or content policy on hate speech. Potential strengths and weakness of Oversight-B, and its variants, are set out in Table 9 below.

Table 9. Oversight-B: Internal appeals process

Tool		Strengths	Weaknesses
Oversight-B: Internet platforms' internal appeals processes		- Internet platforms retain control over moderation and oversight - Provides victims of hate speech with redress - Provides creators of content with a right of appeal	- Not independent, albeit there can be separation between moderation teams and appeals teams (oversight) within the Internet platform - Potential for lack of transparency if Internet platforms fail to publish the training, protocols and procedures given to appeals teams
Collaboration potential			
Low (Internet platforms)			
Tool variants		Strengths	Weaknesses
Types of appeal	Appeals against decisions to remove only	- Provides a right of appeal to users that believe their right to freedom of expression has been violated	- Fails to provide redress for users that believe their right not to be subjected to hate speech has been violated - Victim-insensitive
	Appeals against both decisions to remove and decisions not to remove	- Potentially increases the credibility of the Internet platform as being a "fair broker" - Provides a right of appeal to creators of content and redress to victims of hate speech	- Increases workload for the appeals teams
Who handles the appeals?	Appeals handled by the same teams as are responsible for moderation	- Pre-existing familiarity with the case - Less time-consuming - Potentially less expensive - Consistency of decisions - Policy influence	- Potentially a superficial and non-credible form of oversight - Lacking independence and separation of decision-making power - May not meet high standards of due process - Less likely to admit errors and overturn bad decisions and therefore risks duplicating errors - More subject to bias
	Appeals handled by a different team to that responsible for moderation	- Some independence and separation of decision-making power - Increased chance of "bad" moderation decisions being overturned - Greater credibility of oversight	- Potential for inconsistency of decisions - Still risks falling short of high standards of due process - Lacking full independence - Decisions may still lack the higher credibility of a fully independent oversight board

Notably, in 2018 Facebook launched a new piece of oversight functionality on its platform, "appeals for posts that were removed for nudity / sexual activity, hate speech or graphic violence".[109] This year it updated its progress as follows: "We are beginning to provide appeals not just for content that we took action on, but also for content that was reported but not acted

[109] Appeals process announcement available at: https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/ [last accessed 6 October, 2019].

on."[110] The potential strengths and weaknesses of extending the internal appeals process to include appeals against moderation decisions not to remove alleged hate speech content are set out in Table 9 above.

[110] Update on the appeals process rollout available at: https://transparency.facebook.com/community-standards-enforcement/guide#section5 [last accessed 6 October, 2019].

C. General recommendations from an independent supervisory council, steering committee or oversight board

Oversight-C involves an independent supervisory council, steering committee or oversight board making general recommendations (GRs) to the Internet platform about its content policies and moderation guidelines, processes and procedures. Hitherto few of these councils, committees or boards actually exist, but where they do exist or are being created Internet platforms have played a substantial role in the set up. Moreover, these councils, committees or boards tend to provide mono-relational oversight, that is to say, each Internet platform has its own council, committee or board. Examples include Twitter's Trust and Safety Council, which was announced in 2016,[111] and Facebook's more recent Oversight Board.[112] In the future, however, it is possible that such councils, committees or boards might evolve to provide multi-relational oversight, whereby a single council, committee or board provides oversight for multiple platforms. If Facebook's Oversight Board is seen to be successful, for example, then perhaps other Internet platforms may join it. Indeed, its corporate structure allows for just this eventuality.[113] Reasons for other platforms to join Facebook's Oversight Board might include cost saving or economies of scale, or else if the Board earns a reputation or accurate and credible oversight. Potential strengths and weaknesses of Oversight-C are set out in Table 10.

Table 10. Oversight-C: Independent supervisory council, steering committee or oversight board

<table>
<tr><td colspan="2">Tool</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td colspan="2">Oversight-C: Independent supervisory council, steering committee or oversight board makes general recommendations (GRs) to the Internet platform about its content policies and moderation guidelines, processes and procedures</td><td>- Potential for independence oversight on content policies and moderation procedures
- Potential for transparency
- Potential to introduce a broader range of external knowledge and expertise into the oversight of moderation</td><td>- Challenge of ensuring that the policy influence or impact of the council, committee or board is not too weak and not too strong
- Challenge of selecting the composition of the council, committee or board
- Challenge of making the council, committee, or board representative of the general population both demographically and in terms of norms
- Challenge of providing independent funding for the council, committee or board
- Challenge of public perceptions of lack of genuine independence of the council, committee or board</td></tr>
<tr><td colspan="4">Collaboration potential</td></tr>
<tr><td colspan="4">Medium or high (Internet platforms; independent supervisory council, steering committee or oversight board)</td></tr>
<tr><td colspan="2">Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td rowspan="2">Standing of GRs</td><td>GRs are non-binding or advisory only</td><td>- Internet platform retains control over content decisions
- Enables the Internet platform to comply with local laws</td><td>- Weak policy influence or impact
- Potentially gives the impression that the council, committee or board is a PR exercise
- Risks alienating members of the council</td></tr>
<tr><td>GRs are binding</td><td>- Strong policy influence or impact
- Strengthens credibility of the independent supervisory council, steering committee or oversight board</td><td>- Internet platforms loses creative control over content decisions
- Risks undermining the "contract" or bond of trust between Internet platforms and their users
- Outsourcing decisions on content policy, etc. may be incompatible with regimes of liability</td></tr>
</table>

[111] Information available at: https://blog.twitter.com/en_us/a/2016/announcing-the-twitter-trust-safety-council.html [last accessed 6 October 2019].

[112] Information available at: https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf [last accessed 6 October 2019].

[113] Interview with Facebook, 6 August, 2019. Comments by Facebook, 1st consultative meeting, London, 17-18 October, 2019.

D. Referrals of grey area or difficult cases to an independent supervisory council, steering committee or oversight board

Under Oversight-D, the Internet platform refers grey area or difficult cases to an independent supervisory council, steering committee or oversight board. These are cases where it is ambiguous, unclear and open to different interpretations and reasonable disagreement about whether the content falls foul of the Internet platform's content policy on hate speech. Facebook envisages such referrals to its new Oversight Board,[114] that is, cases of uncertainty over whether given bits of content breach its community standard on hate speech.[115] Table 11 below outlines potential strengths and weaknesses of Oversight-D.

Table 11. Oversight D: Referrals of grey area cases

<table>
<tr><td colspan="2">Tool</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td colspan="2">Oversight-D: Internet platform refers grey area or difficult cases to an independent supervisory council, steering committee or oversight board</td><td>- Potential for independence in the oversight of moderation decisions
- Potential for transparency
- Potential to introduce a broader range of external knowledge and expertise into the oversight of moderation
- Provides victims with a means of redress
- Gives creators of content a right of appeal
- Potential for policy influence or impact on moderation policies and guidelines
- Reduces burden on Internet platforms</td><td>- Potentially falls short of the high standards of due process associated with court proceedings
- Challenge of ensuring that the policy influence or impact of the council, committee or board is not too weak and not too strong
- Users may have to refer cases to several different bodies if the same content appears across multiple Internet platforms</td></tr>
<tr><td colspan="4">Collaboration potential</td></tr>
<tr><td colspan="4">Medium or high (Internet platforms; independent supervisory council, steering committee or oversight board)</td></tr>
<tr><td colspan="2">Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td rowspan="2">Case referral power</td><td>Only the Internet platform can refer cases</td><td>- Ensures that the council, committee or board focuses on cases deemed important to the Internet platform
- Decreases the risk of unsound or vexations referrals to the council, committee or board</td><td>- Victims lack means of redress to an independent body
- Creators of content lack access to a right of appeal to an independent body
- Gives the appearance that the council, committee or board lacks independence</td></tr>
<tr><td>Both the Internet platform and users can refer cases</td><td>- Potentially enhances perception that the council, committee or board is independent
- Victims have means of redress to an independent body
- Creators of content have access to a right of appeal to an independent body
- If the council, committee or board retains case selection powers, this mitigates the effects of users making unsound or vexations referrals</td><td>- Increases case selection workload on the council, committee or board
- Potentially creates unrealistic expectations among users who decide to make referrals</td></tr>
</table>

Notable variants of Oversight-D relate to case referral powers, set out in Table 11 above. Arguably the potential benefits of allowing both Internet platforms and users to refer cases to an independent supervisory council, steering committee or oversight board outweigh the weaknesses, especially if members of the council, committee or board are given proper support and time, and if the Internet platform is careful to manage the expectations of users when making referrals, such as providing information on the percentage of referrals actually heard.

[114] Facebook, Oversight Board Charter, September, 2019.

[115] Available at https://www.facebook.com/communitystandards/hate_speech [last accessed 5 October, 2019].

Other variants of Oversight-D concern case selection by the independent supervisory council, steering committee or oversight board, which are set out in Table 12 below. As a way of democratizing the oversight of moderation, the petition system stands out. This involves a mechanism by which cases must automatically be heard by the independent supervisory council, steering committee or oversight board if a threshold or minimum number of users support or "vote" for this (e.g. a user-led rating, "like" or electronic petition system whereby appeals referrals receiving a given rating, number of likes or e-signatures must be heard).

In the report on its public consultation over the Oversight Board, however, Facebook stated this about the issue of case selection through petitions.

> Some suggested opening up case selection to public voting, though others feared that well-organized groups could dominate the process at the expense of stakeholders with less influence and resources. Feedback has generally supported a Board that retains "certiorari power" to select which requests for review that it wishes to hear—in other words, a Board that has the discretion "to control its docket." (Facebook 2019, 23)

Following on from and reflecting this statement, the relevant part of Facebook's Oversight Board Charter states: "The board has the discretion to choose which requests it will review and decide upon."[116]

Table 12. Oversight D: Referrals of grey area cases (case selection variants)

Tool			
Oversight-D (*continued*) Internet platform refers "hard" cases to an independent supervisory council, steering committee or oversight board			
Tool variants		Strengths	Weaknesses
Case selection	Case selection done by the Internet platform	- Focus on cases deemed important to the Internet platform	- Lack of independence due to the independent supervisory council being unable to control its own docket
	Case selection done by the independent supervisory council, steering committee or oversight board acting on its own initiative and expertise	- Greater independence - Potential for the council, committee or board to focus on cases it deems important - Potential greater policy influence	- Potentially decreases the chances that substantive decisions will have policy influence on the Internet platforms
	Case selection done by the independent supervisory council based on public consultation (e.g. publishes list of cases it is considering hearing and invites statements or "amicus briefs" about the merits of hearing cases)	- Partial democratization of oversight - Provides a form of public accountability	- Risk of creating unrealistic expectations among the public that could lead to damage to the reputation of the council, committee or board
	Case selection based on a user petition system: Any case must be heard if user support for its being heard passes a certain threshold based on votes, e-signatures or similar	- Democratization of oversight - Increased credibility of the council, committee or board	- Potential that well-organized groups could dominate the user-petition system at the expense of stakeholders with less influence and resources

Nevertheless, are concerns over the public petition variant reasonable? According to Global Partners Digital:

[116] Facebook, Oversight Board Charter, September, 2019.

> With regard to cases raised by users, we would have concerns over the use of public petitioning as a mechanism. There is a risk that this could lead to the dominance of cases raised by well-organised stakeholders at the expense of less well-resourced individuals and groups. (Global Partners Digital 2019, 3-4)

However, arguably it is equally possible that simply allowing the Oversight Board to select its own cases could lead to "capture" of the Board's case selection decision-making, namely, that case selections could become heavily influenced "by well-organised stakeholders at the expense of less well-resourced individuals and groups" due to the greater capacity of some stakeholders to influence the case selection decisions of the Board.

At any rate, surely case selection is not a binary decision: it would be entirely feasible to pursue a mixed strategy in which the Oversight Board hears a combination of both board-selected and user-selected cases. This would allow at least some cases to be heard through the user petition system. The Oversight Board could control the total numbers of cases being referred in this way by adjusting the threshold at which referral is trigged under the petition system.

There is also the possibility of a hybrid system: the Board could publish a list of cases it is considering hearing and could invite from users (or the public at large) statements or "amicus briefs" setting out the merits of hearing certain cases.[117] This would be a way of collecting diffuse information and experiences concerning online hate speech (its nature and effects) that could be highly relevant to case selection that might otherwise be lost if case selection powers reside solely with the Board. It might also be a way of the Board recognising the particular experiences and needs of targets of online hate speech ("victims"). For more on this, see section VII.C(ii) below. This hybrid system could also help to lend greater credibility and legitimacy to case selection decision-making and ultimately even to the decisions themselves.

Other notable variants of Oversight-D include the scope of referrals both in terms of the type of moderation decisions that can be referred and in terms of whether both moderation decisions and legal compliance decisions can be referred. These are set out in Table 13 below.

[117] Note, similar suggestions have been made by some US legal scholars in connection with democratizing reforms of the Supreme Courts' extensive certiorari power (see Watts 2011).

Table 13. Oversight-D: Referrals of grey area cases (type of moderation decision referrals variants)

Tool			
Oversight-D (*continued*) Internet platform refers grey area cases to an independent supervisory council, steering committee or oversight board			
Tool variants		Strengths	Weaknesses
Scope of referrals in terms of moderation decisions to remove and not to remove	Referrals of decisions to remove content only	- Provides a right of appeal to creators of content	- Fails to provide a means of redress to victims of hate speech
	Referrals of both decisions to remove content and decisions not to remove content	- Provides both a right of appeal to creators of content and a means of redress to victims of hate speech	- Increases workload
Scope of referrals in terms of moderation decisions and legal compliance decisions	Referrals of moderation decisions concerning community standards or content policies only	- Simplifies the task of the council, committee or board - Means that the members of the council, committee or board need not have legal expertise	- Potentially a missed opportunity in countries where Internet platforms have a legal responsibility to make determinations as to lawfulness or have incentives to refer such determinations to independent bodies
	Referrals of both moderation decisions concerning community standards and legal compliance decisions concerning local hate speech laws	- Reflects a willingness on the part of Internet platforms to seek second or third opinions on legal issues - Draws on wider legal expertise - May enhance the credibility of the council, committee or board	- Potentially increases the workload in terms of the complexity of cases - Requires legal expertise - Potentially creates ambiguities over liability for bad decisions - Council, committee or board lacks the high standards of due process associated with court proceedings

As noted in section I.F above, in the case of Facebook it has made clear that its Oversight Board will not look at cases that implicate questions of the lawfulness or unlawfulness of given bits of hate speech content in "limited circumstances" where this could expose senior managers at Facebook to criminal liability or expose Facebook as a corporate entity to regulatory sanctions. However, as also noted in section I.F, it could be argued that this attempt to build a governance firewall between oversight and legal compliance might be a missed opportunity in countries like Germany where Internet regulations require Internet platforms to make quasi-legal determinations as to lawfulness of content and also provide incentives to Internet platforms to refer such determinations to competent independent bodies.

Moreover, it was argued that Internet platforms should refer cases that raise issues of the lawfulness of content to a competent independent body precisely because of the need for second, third and fourth opinions. Given the margin of error in hate speech cases, Internet platforms cannot and should not rely solely on their legal compliance teams and external legal counsel. Of course, if Internet platforms do refer cases raising issues of lawfulness of content to a competent independent body, then strictly speaking this body becomes a regulatory tool and not an oversight tool. Thus, the tool is also listed in section IV.A, Table 19 below.

However, if Internet platforms accept these arguments, an alternative way to proceed is for them to simply outsource all day-to-day legal compliance work to specialist law firms (legal compliance consultants) or to civil society organisations that already operate as trusted flaggers and/or monitoring bodies and that have the necessary legal expertise to undertake this work. See section IV.A, Table 19 below, for potential strengths and weaknesses of this idea.

Still more variants of Oversight-D include the basis on which case decisions are made, as set out in Table 14 below.

Table 14. Oversight-D: Referrals of grey area cases (basis of case decision-making variants)

Tool			
Oversight-D (*continued*) Internet platform refers "hard" cases to an independent supervisory council, steering committee or oversight board			
Tool variants		Strengths	Weaknesses
Case decision-making (substance and procedure)	Case decision-making based on (i) the content policies and corporate values of the Internet platform	- Increases the probability that case decisions will have policy influence	- Potentially lacking high standards of due process - Highly dependent on differing degrees to which Internet platforms' content policies live up to international human rights standards
	Case decision-making based on (i) the content policies and corporate values of the Internet platform and (ii) recognition of quasi-judicial requirements of due process and wider international human rights standards	- Potentially more likely to satisfy higher standards of due process - Potentially more likely to respect international human rights standards	- Potentially lowers the policy influence of case decisions, especially where the oversight board puts more emphasis on (ii) than (i) whereas the Internet platform does the opposite

Finally, Table 15 below sets out variants concerning the standing of the decisions made by the independent supervisory council, steering committee or oversight board.

Table 15. Oversight-D: Referrals of grey area cases (standing of case decisions variants)

Tool			
Oversight-D (*continued*) Internet platform refers "hard" cases to an independent supervisory council, steering committee or oversight board			
Tool variants		Strengths	Weaknesses
Standing of case decisions	Internet platform stipulates that case decisions by the council, committee or board are advisory only and non-binding on the platform	- Internet platform retains control over content and can realise its business mission - Protects the "contract" or bond of trust between the Internet platform and its users	- Risks having weak policy influence - Risks giving the appearance of a superficial or 'fake' right of appeal - Risks merely duplicating the internal appeals process as the platform would have to decide on the internal appeal then make a similar decision about whether or not to accept the result of the external appeal
	Internet platform stipulates that case decisions by the council, committee or board are binding on the platform	- Strong policy influence or impact - Increases transparency and accountability of moderation decisions taken by the platform - Potentially enhances public trust in, and the credibility of, the council, committee or board, as well as the Internet platform itself	- Risks undermining the "contract" or bond of trust between the Internet platform and its users

Interestingly, Art. 4 of Facebook's Oversight Board Charter stipulates that the Oversight Board's decisions will be binding on Facebook, whilst any general recommendations or "policy guidance" offered by the Board would remain advisory only and non-binding:

> The board's resolution of each case will be binding and Facebook will implement it promptly, unless implementation of a resolution could violate the law. In instances where Facebook identifies that identical content with parallel context—which the board has already decided upon—remains on Facebook, it will take action by analyzing whether it is technically and operationally feasible to apply the board's decision to that

> content as well. When a decision includes policy guidance or a policy advisory opinion, Facebook will take further action by analyzing the operational procedures required to implement the guidance, considering it in the formal policy development process of Facebook, and transparently communicating about actions taken as a result.[118]

On one possible reading, this position amounts to Facebook taking a nuanced approach to its Internet governance responsibilities and to its relationship with its users. First, Facebook maintains that it has a general responsibility to "own" its policies on content moderation based on the principle that because Facebook made the platform then it should be responsible for key policies concerning what users can and cannot post or share on the platform; and this general responsibility on the part of Facebook is also something that its users would fully expect when signing up to the platform. But second, Facebook makes an important qualification to its general responsibility based on the principle that whilst it should "own" its policies it is not required, and its users would not expect (and may not want), the platform to make literally "all the calls" concerning how the policies are applied to given cases.[119]

[118] Facebook, Oversight Board Charter, September, 2019. Available at: https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf.

[119] Comments by Facebook, 1st consultative meeting, London, 17-18 October, 2019.

E. Fully independent dispute resolution procedure or mediation process

Oversight-E involves users and Internet platforms who are locked in dispute over content moderation decisions agreeing to avail themselves of a fully independent dispute resolution procedure or mediation process after any internal appeals process has been exhausted. Whilst this study is not aware of any European countries where Oversight-E is currently in operation for online hate speech, there are clear parallels with the Racial Discrimination Act 1975 in Australia which provides for the Australian Human Rights Commission to receive, inquire into and conciliate complaints about unlawful discrimination, including unlawful racial vilification (hate speech). Interestingly, Romanian anti-discrimination legislation provides for easily accessible proceedings in discrimination cases, including a conciliation mechanism operated by the Romanian National Equality Board (ECRI 2019: 13). If in the future this was extended to include cases of online hate speech, then this would also be an example of Oversight-E in action. Indeed, Equinet has argued that the mandate of national equality bodies across Europe should be extended to cover tackling hate speech, offline and online (Equinet 2018: 24). The potential strengths and weaknesses of Oversight-E are expressed in Table 16.

Table 16. Oversight-E: Fully independent dispute resolution procedure or mediation process

<table>
<tr><td colspan="2">Tool</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td colspan="2">Oversight-E: Users and Internet platforms can agree to avail themselves of a fully independent dispute resolution procedure or mediation process after any internal appeals process has been exhausted. This gives an opportunity for creators, reporters (victims) and moderators of online hate speech to revisit the content in the cold light of day and reach a consensus through conciliation.</td><td>- Non-adversarial
- Conducted in private and therefore may be less intimidating for “victims”
- Element of restorative justice
- Increased chance of mutually agreed results
- Improved chance of changing user behaviour
- Well suited to grey area cases</td><td>- Potentially weak policy influence or impact
- Lack of transparency due to the “behind-closed-doors” procedure
- Potentially fruitless exercise leading to greater anger and frustration among users
- Lack of scalability
- Expensive</td></tr>
<tr><td colspan="4">Collaboration potential</td></tr>
<tr><td colspan="4">Medium to high (Internet platforms; users; a fully independent dispute resolution service)</td></tr>
<tr><td colspan="2">Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td rowspan="2">What happens in the event that the dispute resolution procedure or mediation process fails to produce a mutually agreed outcome?</td><td>The default position is that the original content moderation and appeal decision by the Internet platform stands</td><td>- Internet platforms retain control over content thus enabling them to fully realise their corporate values and business mission
- Avoids ambiguity over legal liability for mutually agreed outcomes</td><td>- Limited incentive on the part of the Internet platform to back down and so potentially weak policy influence or impact</td></tr>
<tr><td>The default position is that the user-complainant has recourse to some other external appeal process</td><td>- Prevents Internet platforms from stonewalling the user-complainant
- Gives Internet platforms an incentive to make good faith efforts at compromise during the mediation process</td><td>- Potential ambiguity over legal liability for mutually agreed outcomes
- If users are likely to end up making an external appeal to a super-complaints body (e.g. regulatory), this could render the despite resolution procedure redundant</td></tr>
</table>

F. User rating system

Oversight-F involves a user rating system for the evaluation of content moderation decisions taken by the Internet platform. Specifically, volunteer users are given information relating to moderation decisions and are invited to rate those decisions based on ratings, likes or similar measures of support. Potential strengths and weaknesses appear in Table 17 below.

Table 17. Oversight F: User rating system

Tool		Strengths	Weaknesses
Oversight-F: User rating system for the assessment and evaluation of moderation decisions by the Internet platform		- Potentially increases the credibility and/or user satisfaction with moderation decisions - Potentially improves user knowledge of and engagement with moderation decision-making - Potentially promotes greater self-reflection among users about their own online conduct	- Risk of turning oversight of moderation into a mere popularity contest - Arguably does not constitute a genuine means of redress for victims of hate speech or a genuine right of appeal for creators of content - Could lead to inconsistent oversight over time - Risk of malicious, ill-informed or discriminatory ratings - Potentially creates ambiguity over the legal liability of Internet platforms for moderation decisions - Potentially protects or leads to moderation that falls short of international human rights standards
Collaboration potential			
Low (Internet platforms; volunteer users)			
Tool variants		Strengths	Weaknesses
Standing of the user ratings	The user ratings are not made public and are non-binding	- Internet platform retains control over content and can realise its business mission	- Potentially low policy influence or impact - Lack of transparency
	The user ratings are made public but are non-binding	- Transparency - Potential for some policy influence or impact due to reputational damage of failure to reflect ratings	- Potentially low policy influence or impact - Internet companies may suffer reputational damage
	The user ratings are made public and are binding, meaning that a moderation decision stands only if user support for it passes a certain threshold	- Distributed oversight, meaning highly responsive to user views - High policy influence or impact - Representativeness - Suited to smaller, artisanal Internet platforms	- In some circumstances user ratings could prompt or lead to further instances of online hate speech - The Internet platforms relinquishes control of moderation decisions which may impeded pursuit of its business mission - Not suited to large, mainstream Internet platforms

One notable danger of Oversight-G is that in some circumstances it could actually lead to further instances of online hate speech. For example, hate speakers could feel emboldened, legitimised and normalised by low user ratings given to content moderation decisions not to remove hate speech content.

IV. THIRD LEVEL OF INTERNET GOVERNANCE: THE REGULATORY LEVEL

At the third level of the governance of online hate speech, the regulatory level, the emphasis shifts clearly to combating unlawful or illegal hate speech content posted or shared on Internet platforms. Typically at the regulatory level governmental agencies may intervene to compel Internet platforms to remove unlawful or illegal hate speech, or else impose duties of care or codes of practice that enshrine and promote certain desired procedural values in moderation and/or oversight of moderation such as due process or transparency. Nevertheless, governmental agencies do not have a monopoly on regulation. Even self-regulation is a form of regulation, such as when Internet platforms themselves work towards the aim of removing unlawful or illegal hate speech content through so-called legal compliance measures.

A. Legal compliance

Under Regulatory-A, Internet platforms adopt “terms of service” which state that users may not post or share “unlawful” or “illegal” content, including unlawful or illegal hate speech content, and, furthermore, platforms employ legal compliance teams that monitor content, or that respond to legal reporting forms, notices, reports or referrals relating to content that is suspected of being illegal based on local laws, and therefore in breach of the aforementioned terms of service on illegal content. Legal compliance is not the same as moderation because legal compliance concerns local laws and the Internet platform’s terms of service on illegal content in general including but not limited to illegal hate speech content, whereas moderation concerns the platform’s community standards or content policies including but not limited to standards or policies on hate speech. Nevertheless, legal compliance could be thought of as form of “regulatory moderation” or “content regulation” due to the fact that it involves the Internet platform removing content because it is deemed to be unlawful or illegal content.[120] The potential strengths and weaknesses of Regulatory-A are set out in Table 18 below.

Table 18. Regulatory-A: Legal compliance

Tool	Strengths	Weaknesses
Regulatory-A: Internet platforms adopt “terms of service” which prohibit users from posting or sharing “unlawful” or “illegal” content, and have “legal compliance” teams that remove content deemed to be in breach of these terms of service	- Internet platforms take responsibility for tackling the problem of illegal hate speech - Arguably represents an equitable sharing of the burden of removing illegal hate speech	- Dependent on Internet platforms’ legal compliance teams being properly resourced, competent and effective - May not be achievable for smaller, artisanal Internet platforms - Challenge of hiring staff with appropriate legal competence in local hate speech laws - Internet platforms may also have to pay for external legal counsel or legal compliance consultants due to nuances of local hate speech and the complexities of local hate speech laws
Collaboration potential		
Low to Medium (Internet platforms; law enforcement agencies; civil society organisations)		

Typically users must click to agree the Internet platform’s terms of service before they can use the platform, service, website or product. Examples of terms of service that prohibit users from posting or sharing unlawful or illegal content are listed in section I.E above. Typically this

[120] 1st consultative meeting, London, 17-18 October, 2019.

particular kind of terms of service is articulated in an unqualified way and so covers all forms of unlawful or illegal content including therefore unlawful or illegal hate speech content.

Importantly, legal compliance teams can receive notifications of potentially illegal or unlawful hate speech content from a range of stakeholders, including from local law enforcement agencies (e.g. police, public prosecutors, courts, including both informal or administrative notifications and more formal or judicial "notice and take down" orders), from trusted flagger organisations, directly from users through legal reporting forms, and also from their own content moderation teams and even from their own machine learning tools or algorithms.[121]

One question that arises from Regulatory-A is how responsive Internet platforms' legal compliance teams should be to governmental agencies, trusted flaggers and users respectively. Should they give greater credence and urgency to administrative requests from public prosecutors than to legal report forms filled out by users, for instance?[122]

In any event, this study suggests that governmental authorities, following recommendations of intergovernmental organisations, ought to place legal responsibilities on Internet platforms' to ensure that their legal compliance teams should be proactive in identifying unlawful or illegal hate speech content, such as by more closely monitoring accounts after one instance of illegal content has been identified on those account, or by using text extraction and machine learning tools or algorithms to search for illegal content.

Furthermore, the accuracy of these automated systems for flagging unlawful content will be heavily reliant on the quality of the "training data" or "benchmark data set" used in the programming of the machine learning tools or algorithms. This in turn will depend on legal compliance teams furnishing programmers with cases that the legal compliance teams have reach definitive judgments about. For even greater accuracy, this study recommends that legal teams should supply programmers with sample relevant court decisions (summaries) in the relevant countries (i.e. cases where local hate speech laws have been applied by the courts to bits of content).

Another important issue is the extent to which Internet platforms' legal compliance teams "work with" other stakeholders in making substantive legal compliance decisions in particular cases. The potential strengths and weaknesses of two variants are listed in Table 19 below.

[121] Ibid.

[122] 1st consultative meeting, London, 17-18 October, 2019.

Table 19. Regulatory-A: Legal compliance (consultation, case referral and outsourcing variants)

Tool			
Regulatory-A (*continued*) Internet platforms adopt “terms of service” which prohibit users from posting or sharing “unlawful” or “illegal” content, and have “legal compliance” teams that remove content deemed to be in breach of these terms of service			
Tool variants		Strengths	Weaknesses
Consultation, case referral, and outsourcing	Legal compliance teams make decisions based on consultation with local law enforcement agencies and local experts including civil society organisations that already operate as trusted flaggers and/or monitoring bodies	- The nuances of local hate speech and complexity of local hate speech laws can make it difficult for Internet platforms’ legal compliance teams to handle grey area cases thus creating a need for further consultation - Internet platforms’ legal compliance teams make final decisions	- Further consultation with local law enforcement agencies could make Internet platforms less likely to challenge administrative take down requests
	Whilst retaining overall responsibility for most legal compliance decisions, Internet platforms’ legal compliance teams refer difficult cases to a competent independent body and agree to abide by its decisions	- Suitable for handling difficult cases - Draws on a wider body of expertise - Enables Internet platforms to take advantage of exceptions to legal responsibilities provided in some countries (e.g. the NetzDG Act in Germany)	- Potentially undermines the independence of the Internet platform’s legal compliance decisions - Potentially creates ambiguity around where legal liability rests for failures in legal compliance
	Internet platforms outsource day-to-day legal compliance work to specialist law firms (legal compliance consultants) or to civil society organisations that already operate as trusted flaggers and/or monitoring bodies	- Suitable for smaller, artisanal Internet platforms that find it difficult to handle the volume of legal compliance work - Outsourcing day-to-day legal compliance work may achieve efficiency savings even for larger Internet platforms	- Potentially creates ambiguity around where legal liability rests for failures in legal compliance - Potential conflict of interest on the part of the relevant civil society organisations - Potential for civil society organisations to lose credibility as independent monitoring bodies

In terms of the idea of greater consultation, this could take different forms. For example, in its recent report on hate speech policy within the EU, INACH (2019) has recommended that its members—note that its members include civil society organisations, NGOs and other bodies who are trusted flaggers and members who are monitoring bodies under the European Commission’s Code of Conduct monitoring system—should play a more active role in training Internet platforms’ content moderation teams on local hate speech laws.

> Since the start of the monitoring exercises, and even beforehand to become trusted flaggers, NGOs have received a lot of training and instructions from social media companies on how to use their reporting systems and lately on what they remove and what not. These exchanges of knowledge were useful, however, INACH thinks that there is a very strong argument for NGOs to train the moderators of social media. People who work for our members are experts in their fields and they have tremendous knowledge on hate speech and the laws that regulate it on national and EU levels. Yet, social media companies never allow them to train their moderators or to have discussions with them. (INACH 2019: 14)

Internet platforms might respond that decisions about removal of content reported as potentially illegal hate speech are not taken by moderation teams but by legal compliance teams. However, within some Internet platforms even reports of illegal hate speech content are initially sent to moderation teams (or "community operations teams") to determine if the content contravenes the platform's own community standard on hate speech before it is sent to dedicated and specially trained and competent legal compliance teams.[123] And so there might be value in training even moderation teams on local hate speech laws. Furthermore, arguably INACH's recommendation could, and should, be extended to cover training for legal compliance teams as well, especially in countries where local hate speech laws are complex and/or where the application of local hate speech laws to actual bits of content is very difficult due to semantic nuances, linguistic context, slang, and wider social context. Of course, Internet platforms might already make use of external legal counsel in particular countries. But it could be that some civil society organisations have significant experience and expertise to offer, and might potentially offer greater levels of representativeness, independence and impartiality.

The alternative would be for Internet platforms to outsource day-to-day legal compliance work to specialist law firms, that is, legal compliance consultants, or to civil society organisations with local expertise, such existing trusted flagger organisations and/or monitoring bodies. The strengths and weaknesses of that alternative are set out in Table 19 above. This might be a good option for smaller, artisanal platforms. However, one drawback is a potential conflict of interest and loss of credibility on the part of civil society organisations or NGOs if they are intending to operate not only as trusted flaggers and monitoring bodies but also as organisations doing day-to-day legal compliance work for the same Internet platforms.[124]

[123] Interview with Facebook, 22 November 2019.
[124] Interview with INACH, 2 December 2019.

B. Voluntary code of practice

Under Regulatory-B, Internet platforms sign up to a voluntary code of practice (or "code of conduct") on combating unlawful or illegal hate speech, some of the potential strengths and weaknesses of which are set out in Table 20 below.

Table 20. Moderation-B: Voluntary code of practice

Tool	Strengths	Weaknesses
Regulatory-B: Internet platforms sign up to a voluntary code of practice (or "code of conduct") on combating unlawful or illegal hate speech content, including an agreement on periodic monitoring exercises	- Promotes forms of collaboration (among Internet platforms, monitoring bodies, and intergovernmental organisations) - Flow of knowledge and expertise from monitoring bodies to Internet platforms, especial relating to contextual or local circumstances and forms of hate speech - Voluntary codes of conduct potentially among the least restrictive forms of governance partly due to their voluntariness - Potential to build trust between stakeholders because the code is voluntary and non-judicial - Reputational benefits of membership (or disbenefits of non-membership) could lead to a "virtuous cycle" in which pressure is placed on more Internet platforms to join - As more Internet platforms join the code could potentially become a "best practice" standard across the sector - Potential definitional harmonisation over the term "hate speech" between domestic and international law and Internet platforms' content policies	- Potentially weak policy influence - Enforcement is based on monitoring exercises, "naming and shaming" and reputational damage - Voluntary code potentially has weak policy influence as Internet platforms can always withdraw and because enforcement is relatively unrestrictive - Risk that the voluntary code of practice will include relatively undemanding responsibilities compared to other governance tools - Potential for lack of public consultation on the content of the code with civil liberties organisations, minority rights groups or NGOs, and other stakeholders
Collaboration potential		
Medium to high (Internet platforms; intergovernmental organisation; national governments)		

An example of Regulatory-B is the European Commission's Code of Conduct on Countering Illegal Hate Speech Online of May 2016.[125] Results of the fourth monitoring exercise, conducted in Fall 2018, on the implementation of the Code were published in February 2019. It found that "[o]n average, IT companies are removing 72 percent of the illegal hate speech notified to them" (European Commission 2019: 1).

One major challenge with Regulatory-B is to ensure that the authorship or creation of any voluntary code of practice is a legitimate and inclusive process. One variant is for an intergovernmental organisation to take the lead in creating it. Another is for the Internet platforms themselves to take the lead in its creation. Potential strengths and weaknesses of these variants (i.e. respective roles of "creator" and "consultee") are outlined in Table 21 below.

[125] Basic information about the Code including current signatories is available at: https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300 [last accessed 9 October, 2019].

Table 21. Moderation-B: Voluntary code of practice (genesis of the code variants)

Tool			
Regulatory-B (*continued*) Internet platforms sign up to a voluntary code of practice (or "code of conduct") on combating unlawful hate speech content, including an agreement on periodic monitoring exercises			
Tool variants		Strengths	Weaknesses
Genesis of the code	Code created by an intergovernmental organisation and national governments in collaboration with Internet platforms	- Independence - Potential to be more attuned to international norms on combating unlawful hate speech - Definitional harmonisation with international hate speech instruments and national laws	- If the impetus for the code comes from an international organisation rather than from Internet platforms themselves, this could limit engagement and "buy in" among Internet platforms in the long term - Risk that the code will reflect the perspectives and aims of the international organisation more than the viewpoints and business missions of Internet platforms - Potentially pays insufficient regard to wider thinking (if there is lack of public consultation)
	Code created by Internet platforms in collaboration with an intergovernmental organisation and national governments	- Internet platforms potentially likely to engage more fully and over the longer-term with codes of conduct concerning which they feel some "ownership"	- Potential that the code would focus exclusively on Internet platforms' community standards on hate speech and ignore legal definitions and standards - Potential to be less demanding in its requirements - May reflect silo thinking among Internet platforms - Potentially pays insufficient regard to wider thinking if there is lack of public consultation

A more general concern relating to either of the variants outlined in Table 21 is about lack of wider public consultation. For example, in its response to the creation of the European Commission's Code of Conduct, Article 19 noted its concern that "there was apparently no involvement of [Civil Society Organisations] CSOs defending freedom of expression in the drafting of the Code of Conduct" (Article 19 2016: 16).

For its part, the European Commission maintains that civil liberties organisations (e.g. Centre for Democracy & Technology, Access Now) were included in the wider consultation and were given ample opportunity to participate in consultative meetings and to make submissions.[126] In the case of civil liberties organisations that decided to withdraw from the consultation process (e.g. Access Now), the European Commission observes that it has over time continued to receive and respond to submissions and have informal contacts with such organisations.[127]

Another major challenge faced by Regulatory-B, as outlined in Table 22 below, is how to monitor compliance with any voluntary code of conduct by Internet platforms who sign up to it.

[126] Interview with the European Commission, DG Justice, 11 October 2019.
[127] Ibid.

Table 22. Moderation-B: Voluntary code of practice (monitoring variants)

Tool			
Regulatory-B (*continued*) Internet platforms sign up to a voluntary code of practice (or "code of conduct") on combating unlawful hate speech content, including an agreement on periodic monitoring exercises			
Tool variants		Strengths	Weaknesses
Monitoring progress	Internet platforms agree for an international organisation to appoint qualified third party non-governmental bodies to undertake periodic monitoring exercises to measure progress in honouring the code, where the monitoring period is short, monitoring dates predetermined and monitoring dates made known in advance, directly or indirectly, to the Internet platforms being monitored	- Independent assessment of progress in honouring the code of practice - Administratively easier to manage	- Risk that Internet companies will game the monitoring process by putting additional resources into content moderation during known monitoring periods
	As above but the monitoring period is long, monitoring dates not predetermined and monitoring dates not made known in advance to the Internet platforms being monitored	- Monitoring exercise more likely to achieve a "true" and "accurate" picture of progress	- Risks Internet platforms not signing up to the code

Under one variant, Internet platforms agree that an International organisation will appoint qualified third party non-governmental bodies to undertake periodic monitoring exercises to measure progress in honouring the code, where the monitoring period is short, monitoring dates predetermined and monitoring dates made known in advance, directly or indirectly, to the Internet platforms being monitored.

However, this opens up the possibility that Internet companies could game the monitoring process by putting additional resources (e.g. management focus, staff redistribution, overtime, use of contract labour) into moderation during known monitoring periods.[128]

Although not prohibited by the terms, or letter, of the monitoring agreement or protocol, such stratagems—or "logic of consequences", that is, forms of rational self-interested behaviour in pursuit of desired outcomes—on the part of Internet platforms undermine the capacity of the independent monitoring bodies to establish a true and accurate picture, and so arguably do not live up to the spirit of the monitoring agreement.

This challenge is certainly not unique to the governance of hate speech posted or shared on Internet platforms, of course. It is a problem discussed in depth within the academic literature on public sector performance management systems in nearly every area of public policy, for example (Pollitt 2013). But the fact that there is a rationale or logic of consequences for any sort of organisation to game performance management systems or governance mechanisms does not make this practice desirable or something inevitable and immutable.

No doubt some Internet platforms will refrain from gaming any monitoring system to which they have voluntarily agreed because they will pursue a "logic of appropriateness", that is, forms of behaviour oriented towards compliance with what social or community norms deem to be the right thing to do. In the case of a high profile voluntary code of practice, the relevant social or community norm would be to cooperate with both the letter and spirit of monitoring. But it is unlikely that all Internet platforms will demonstrate this logic.

[128] Interview with trusted flagger, 1 October, 2019. Interview with trusted flagger, 3 October, 2019.

However, rather than depend on contingent facts about the different logics pursued by Internet platforms, one way to make it more difficult for any Internet platforms to game the monitoring process, as set out in Table 22 above, would be to adopt longer formal monitoring periods, to allow flexibility or randomness in the actual monitoring dates, and to not make the actual monitoring dates known in advance to the Internet platforms being monitored.

Therefore, this study recommends that intergovernmental organisations in charge of monitoring systems (e.g. the European Commission and the Code of Conduct monitoring system) should give monitoring bodies the discretion to undertake the monitoring exercise within a continuous 12 month formal monitoring period, and to undertake actual monitoring on selected dates of their own choosing, at random, and without any form of notification, direct or indirect, as to the actual dates of monitoring given to the Internet platforms being monitored, for example.

This need not be something that Internet platforms eye with suspicion. By analogy, businesses in the retail sector often hire the services of “mystery shoppers” to help them get a better grasp of customer experiences and how well the business is doing in meeting its customer service goals. Likewise, Internet platforms might come to welcome the use of “mystery flaggers” during a 12 month formal monitoring period to help monitoring bodies obtain a true and accurate picture of how well Internet platforms are complying with the code of practice (i.e. progress towards the moderation goals they have agreed to). Internet platforms should welcome any tool that can result in a better assessment of how far they have come in raising the quality and quantity of their moderation of online hate speech content and how far they have left to travel.

Note, the monitoring system (or “common methodology”) agreed on 5 October 2016 by Internet platforms with the European Commission’s sub-group on combating illegal online hate speech pursuant to the Code of Conduct on Countering Illegal Hate Speech Online of May 2016[129] sits in a fuzzy zone between the two variants outlined in Table 22 above. On the one hand, there is no requirement as part of that agreement for the European Commission or the monitoring bodies to predetermine the actual days on which monitoring occurs, and no requirement to make the monitoring period know in advance to the Internet platforms. However, (i) the agreement indicates that monitoring will usually last between 6-8 weeks, (ii) each of the reports on the four monitoring cycles to date have made clear the exact dates that were used for monitoring in the relevant cycle (European Commission 2016b, 2017, 2018, 2019), (iii) the dates for monitoring periods for the third and fourth cycles were extremely similar (European Commission 2018, 2019) and (iv) reports have often been published around the same time (European Commission 2016b, 2017, 2018, 2019). Together this has created a situation in which potentially Internet platforms can predict with a reasonable degree of accuracy when monitoring might occur. And this creates a risk of “gaming”.

More importantly, there is evidence that under the European Commission’s Code of Conduct monitoring system, the monitoring period has been made known, directly or indirectly, in advance to the Internet platforms being monitored. The evidence for this is varied and includes: (v) Internet platforms have typically organised training sessions for monitoring bodies about the platforms’ procedures and policies, reporting technologies and other matters, timed precisely to take place just ahead of, or during, the monitoring period,[130] (vi) Internet platforms have been aware when monitoring bodies have submitted notifications in quantities and in ways that indicate that the

[129] Minutes of meeting of the European Commission’s sub-group on countering illegal hate speech online concerning the monitoring process and methodology, 5 October, 2019.

[130] 1st consultative meeting, London, 17-18 October, 2019.

monitoring period is in play,[131] and (vii) some monitoring bodies have received verbal and email indications from some Internet platforms that clearly suggest that those Internet platforms have come to know in advance the monitoring period.

Consider as evidence for point (vii) the following statement from a monitoring body about email exchanges it has had with Internet platforms concerning monitoring periods.

> Yesterday we received an email from [a major Internet platform] about the updated guidelines for reporting content, "in view of the upcoming EU Code of Conduct 5th monitoring exercise". It is not the first time that we know that the companies are informed about the monitoring exercises, but it seems to me that now they are even more straightforward and blunt about it, while before they would only mention it orally and in the form of a question, so as to hide that they really knew already.[132]

Importantly, the above points echo experiences, insights and proposals already published in a study on the European Commission's Code of Conduct monitoring cycles by the civil society organisation International Network Against Cyber Hate (INACH):

> The main issue was that a bias was put in place mainly due to the fact that social media companies were informed about many details of the exercise in advance, such as when it was going to take place and who was involved. A more restrictive approach should therefore be taken during the future monitoring exercises. It would, for instance, be advised not to inform social media beforehand about the timeframe of the exercise. Moreover, the ongoing exercise should be masked in some way, making the companies unaware that it is ongoing. This could be done by using low level monitoring for a longer period of time or by carrying out the exercise in a rolling or snowball manner, where NGOs do not start it at the same time and end it at the same time, but spread it out for a longer period and start one after another. The idea to make the ratio of flagging content as normal users and not as trusted flaggers bigger is also one that should be thoroughly considered. This could be done by using anonymous or fake accounts. Using these methods the elevated activity could be a bit more hidden. Moreover, these measures could ensure less bias during future exercises and thus their results would be even more representative. (INACH 2017: 5-6)

Reflecting on the above, this study recommends that reforms are made to the monitoring system to ensure that Internet platforms are not made aware—deliberately or structurally—of the period of monitoring, such as by extending the monitoring period to 12 months of the year.

[131] Ibid.

[132] Correspondence with a monitoring body that is part of the European Commission's Code of Conduct monitoring system, 22 October, 2019 [Anonymised].

C. Legal responsibility to remove unlawful hate speech enforced with fines

Regulatory-C involves using legislation to impose a legal responsibility on Internet platforms to undertake removal of unlawful or illegal hate speech content, and to grant powers to governmental authorities to impose fines, or apply to courts for permission to impose fines, on Internet platforms for a pattern of failure to comply with this legal responsibility. To be more precise, Regulatory-C involves placing a legal responsibility on Internet platforms to remove unlawful or illegal hate speech content within specified time frames—for example, "manifestly" or "clearly" unlawful or illegal content within 24 hours of being reported and/or merely potentially unlawful or illegal content within 7 days—with liability to fines for a pattern of failure to remove illegal hate speech content within the specified time frames. Importantly, the legal responsibility does not require or depend on a specific court order or judicial ruling as to the unlawfulness of any given piece of content. Furthermore, one-off failures or failures taken in isolation are not subject to fines. Rather, it is a pattern of failure over time that can attract fines. Potential strengths and weaknesses of Regulatory-C are set out in Table 23 below.

Table 23. Regulatory-C: Legal responsibility to remove unlawful hate speech enforced with fines

Tool	Strengths	Weaknesses
Regulatory-C: Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified period (e.g. "manifestly" unlawful within 24 hours, otherwise unlawful within 7 days) with liability to fines for a pattern of failure to comply with that responsibility.	- Potentially medium to strong impact and policy influence - Reduces significantly the resource burden on the courts and wider criminal justice system in seeking the removal of unlawful hate speech content on a case by case basis - Provides clarity over where the responsibility to remove unlawful hate speech content rests	- Risk that some Internet platforms could treat this simply as a "business tax" rather than a reason to change policies and practices - Outsourcing quasi-judicial powers to Internet platforms that potentially lack high standards of due process - Results in the removal of hate speech content that has not actually been tested in a court of law as being unlawful or illegal - Incentivizes Internet platforms to engage in over-removal of content just to be "on the safe side" from a legal perspective - The most severe fines might be considered a disproportionate response to a pattern of failure to remove hate speech content - The choice of threshold at which a pattern of failure to remove hate speech content attracts fines may seem arbitrary or may be perceived as "regulatory capture" by industry interests - If any "grace period" is too short, the fine regime could potentially give Internet platforms insufficient breathing space to gradually reform and change their policies and practices over time - Fines could render the business models of some Internet platforms unviable and so could damage competition in the digital economy - Potentially gives insufficient protection for journalistic content - Potentially ignores relevant differences between legal compliance practices across Internet platforms - Resource burden and technical challenges of gathering accurate information on whether Internet platforms are in fact failing to remove unlawful hate speech within the specified time frames
Collaboration potential		
Low (Government; legislature; Internet platforms)		

This type of governance tool can be illustrated by *some* parts of the NetzDG Act in Germany, the Avia Bill in France and a proposed new law on the prevention of undesirable behaviour on social networks in Croatia, albeit these real world regulatory regimes (or proposed regimes in the case of the Avia Bill and the proposed new law in Croatia) incorporate several important features and dimensions in addition to, and different from, the Regulatory-C model.

As set out in Table 23 above, there are several potential problems with Regulatory-C in its basic form. Some of the most serious potential problems are explained in more detail below. First, the legal responsibility on Internet platforms to remove unlawful hate speech content does not require or depend on a specific court order or judicial ruling as to the unlawfulness of any specific bit of content. As Article 19 puts it in relation to NetzDG in Germany, “content is still removed without a determination of legality by a judicial body” (Article 19 2017: 21). It is potentially problematic to impose on Internet platforms a legal responsibility to remove unlawful hate speech content whilst simultaneously not requiring them to submit to or seek any prior court order or judicial ruling as to the unlawfulness of specific bits of content.

From the perspective of some civil liberties organisations, this legal responsibility is problematic because it effectively outsources quasi-judicial or criminal justice powers to Internet platforms even though Internet platforms characteristically lack the capacity and expertise to achieve the same high levels of due process as can be found in court proceedings (Article 19 2017; GNI 2017). More generally, Internet platforms lack the necessary authority, mandate or status to make legal adjudications under rule of law regimes. To call on them to make determinations as to unlawfulness is to privatize judicial decisions that should be made within courts of law.

Of course, governments or legislators can always build into Regulatory-C, as is true of the NetzDG Act in Germany (s. 4(5)), a legal protocol whereby governmental agencies such as the ministry of justice or public prosecutors must seek judicial approval—such as in administrative court—for imposing fines on Internet platforms and whereby Internet platforms have a judicial means of objecting to or challenging the levying of fines on them. But this is not the same as a prior court proceeding as to the unlawfulness of the content in question.

Moreover, this in turn raises an issue over due process if the Internet platform is denied a procedure for appealing against the ruling of an administrative court to reject its objections and uphold the fine (Article 19 2017: 18). These issues are depicted in Table 24 below.

Table 24. Regulatory-C: Legal responsibility to remove unlawful hate speech enforced with fines (right of appeal variants)

Tool			
Regulatory-C (*continued*) Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified period following notice (e.g. “manifestly” unlawful within 24 hours, otherwise unlawful within 7 days) with liability to fines for a pattern of failure to comply with that responsibility			
Tool variants		Strengths	Weaknesses
Right of appeal	Governmental authorities must seek judicial approval for imposing fines on Internet platforms and the latter have judicial means of objecting to the fines, but have no right of appeal under administrative law if their objections are rejected	- Allows for a judicial process but at the same time limits the process thus speeding up the levying of sanctions	- Falls short of highest standards of due process given the absence of a right of appeal
	Governmental authorities must seek judicial approval for imposing fines on Internet platforms and the latter have judicial means of objecting to the fines, moreover they also have a right of appeal under administrative law if their objections are rejected	- Satisfies a higher standard of due process	- Greater resource burden on administrative courts - Greater resource burden on governmental authorities - Potentially adds lengthy delays to sanctions against Internet platforms for a pattern of failure to remove unlawful content

A second potential problem with Regulatory-C set out in Table 23 is that the legal responsibility does not compel Internet platforms to proactively seek out unlawful hate speech content. Rather, it requires removal of content that has been reported or flagged to Internet platforms as clearly or potentially unlawful or illegal. This places the reporting burden on users and trusted flaggers, for example. It also potentially slows down the speed with which content is likely to be removed, if at all. The idea that Internet platforms are not obliged to be proactive in this way is a feature of current laws on the liability of Internet platforms that has been challenged by the European Court of Justice in *Eva Glawischnig-Piesczek v. Facebook Ireland Limited*.[133]

One way to address this would be for Internet laws to impose a legal responsibility on Internet platforms to proactively identify and remove unlawful hate speech content. The potential strengths and weaknesses of this variant are outlined in Table 25 below.

Table 25. Regulatory-C: Legal responsibility to remove unlawful hate speech enforced with fines (extent of legal responsibility variants)

Tool			
Regulatory-C (*continued*) Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified period following notice (e.g. “manifestly” unlawful within 24 hours, otherwise unlawful within 7 days) with liability to fines for a pattern of failure to comply with that responsibility			
Tool variants		Strengths	Weaknesses
Extent of legal responsibility	Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified period following receipt of reports or flags	- Places limits on the extent of legal responsibility and thereby limits the size of resource burden on Internet platforms	- Does not compel Internet platforms to proactively seek out unlawful hate speech content - Places the resource burden of flagging unlawful hate speech content on users and trusted flaggers - Makes removal contingent on reports and flags and as a result potentially slows down the speed and rate at which content is removed, if at all
	Impose a legal responsibility on Internet platforms to proactively identify and remove unlawful hate speech content	- Potentially places a significant technological and resource burden on Internet platforms and may undercut the commercial viability of some Internet platforms	- Any responsibility to remove “identical” or “equivalent” hate speech content potentially creates issues where such content is present in another country where it is not unlawful - A responsibility to proactively remove content could incentivise over-removal of lawful hate speech content

The decision in *Eva Glawischnig-Piesczek v. Facebook Ireland Limited* affirms a legal responsibility on Internet platforms to be proactive in removing “identical” wording and also “equivalent” content (i.e. language translations). For all intents and purposes this necessitates Internet platforms using text extraction and machine learning tools or algorithms. Given the amount of content to be searched through, it is not a task that could be done by humans alone. Then again, when it comes to language translations, nuances like idioms and slang, for example, mean that Internet platforms will also need to hire language specialists either to make final decisions as to semantic equivalence or to create “training data” for programmers.[134]

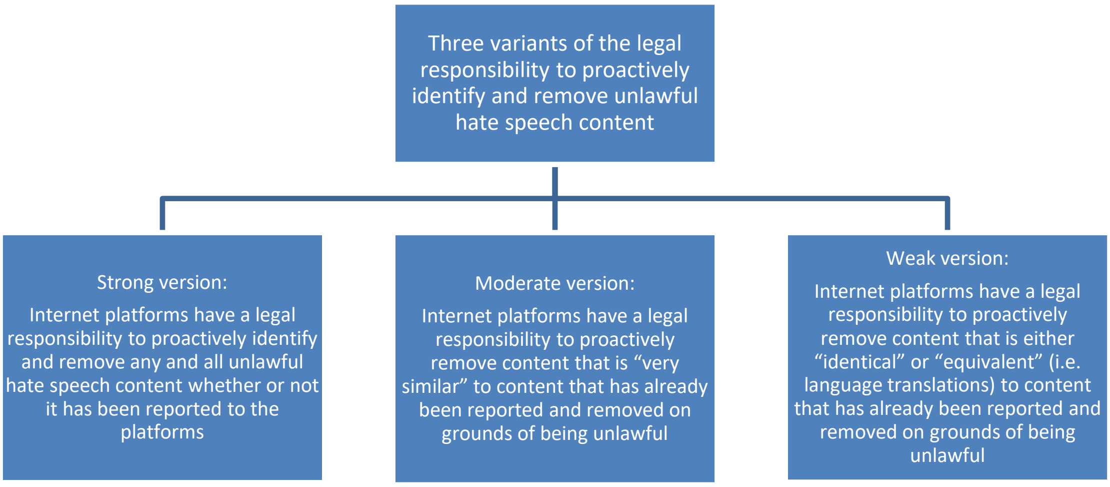
In a similar vein, Art. 2 of the Avia Bill in France places a responsibility on Internet platforms to “implement the appropriate means to prevent the redistribution of content” that has already been taken down on grounds of being “clearly” illegal. This provision speaks directly to the problem of hate speakers responding to the removal of hate speech content they have previously posted on Internet platforms by simply reposting the identical or equivalent content.

[133] C-18/18.

[134] 1st consultative meeting, London, 17-18 October, 2019.

Nevertheless, it is important to recognise that the extent of the legal responsibility on Internet platforms to proactively remove unlawful hate speech content could be specified more or less strongly, as depicted in Diagram 1.

Diagram 1. The legal responsibility to proactively identify and remove unlawful hate speech content


Three variants of the legal responsibility to proactively identify and remove unlawful hate speech content
Strong version:
Internet platforms have a legal responsibility to proactively identify and remove any and all unlawful hate speech content whether or not it has been reported to the platforms
Moderate version:
Internet platforms have a legal responsibility to proactively remove content that is "very similar" to content that has already been reported and removed on grounds of being unlawful
Weak version:
Internet platforms have a legal responsibility to proactively remove content that is either "identical" or "equivalent" (i.e. language translations) to content that has already been reported and removed on grounds of being unlawful

As shall be discussed in section VII below, taking a victim-sensitive approach to Internet governance for online hate speech could point in the direction of expecting Internet platforms to be proactive in the strong or moderate sense, and not merely in the weak sense. This would constitute a "gold standard" of victim-sensitivity, albeit it may be out of reach for the vast majority of Internet platforms for practical, technical and financial reasons.

Nevertheless, it is recommended that national governments and intergovernmental organisations should place a legal responsibility on Internet platforms to be proactive in identifying unlawful or illegal hate speech content, but that the precise extent of that responsibility, the methods of enforcement and the use of any exceptions, exemptions or leniency programs should also reflect country context [see sections I.A(ii), I.D(vi), I.D(vii)].

A third potential problem with Regulatory-C set out in Table 23 above is that the choice of threshold at which a pattern of failure to remove illegal content attracts fines may seem arbitrary or could be perceived as evidence of regulatory capture by industry interests. For example, suppose an Internet platform demonstrates a pattern of limited removal of illegal content over a prolonged period of time, say 12 months. During the 12 months it has removed only 28 percent of illegal content within a specified time frame (e.g. 24 hours). The relevant governmental agency might deem this pattern of failure to remove illegal content sufficiently serious to warrant imposing a fine. But suppose another platform has removed 50 percent of illegal content over the same 12 month period. Is that pattern of failure also sufficiently serious to warrant a fine? What if the removal rate is 60 percent, 70 percent, or 80 percent? At what level would the pattern of failure no longer be serious enough to warrant a fine? The public perception of such thresholds could be that they are arbitrary. Moreover, if the threshold selected seems to favour large, mainstream Internet platforms over small, artisanal platforms, for example, then the public perception might also be that the decision to impose fines is subject to political lobbying or regulatory capture by industry interests.

A fourth potential problem with Regulatory-C set out in Table 23 above is that it could create a financial incentive for Internet platforms to act "on the safe side" from a legal perspective when it comes to content removal. This in turn might lead to an unwelcome unintended consequence: a tendency among Internet platforms to over-remove hate speech content, specifically, a tendency to remove lawful as well as unlawful hate speech content. As Article 19 puts it in relation to NetzDG in Germany, "the Act […] creates a serious danger of further incentivising the over-removal of expression that should be considered lawful under international human rights law" (Article 19 2017: 12). In other words, "[t]he threat of sanctions incentivises platforms to be over-cautious and err towards removal or blocking (or remove functions for third party comments entirely) to avoid liability" (19).

Of course, in order to substantiate the current objection those making it have to provide concrete examples of the sorts of lawful hate speech content that self-evidently ought not to be removed. At this point it might be difficult to find examples that everyone would agree on. Nevertheless, this is not impossible. Most people would deem it problematic, for example, if Internet platforms, due to perverse incentives created by fines for the under-removal of illegal content, ended up removing content created by community leaders with the aim of highlighting the online hate speech that members of minority groups have suffered.[135]

More generally some people may deem the removal of *any* lawful hate speech content by Internet platforms a form of illegitimate "censorship".[136] Some academics take the opposite view that it is "hard" to call content moderation undertaken by Internet platforms "censorship in the strict sense" (Gillespie 2018: 176). But whether or not one calls it "censorship", what is clear is that if the aforementioned incentive is real it raises issues over Internet platforms' compliance with the Ruggie Principles.[137] These international human rights principles include the core principle that corporate enterprises, including Internet platforms, also have a responsibility to respect human rights, not least of which being the human right to freedom of expression.

Of course, robust empirical research is required based on data of actual practices among Internet platforms operating in Germany over time, to determine whether or not in fact the "serious danger" is materialising and Internet platforms are actually removing a higher proportion of lawful hate speech content in Germany than in countries without the fines.

This study suggests that empirical research is needed to test the hypothesis that in countries where governmental authorities (with judicial consent) have powers to levy administrative fines on Internet platforms for patterns of failure to remove online hate speech, Internet platforms in those countries have a greater tendency to over remove lawful hate speech content as compared to how those Internet platforms operate in countries where such fines are not levied.

Of course, any such empirical research faces several significant challenges. First, gaining access to accurate, relevant, and reliable evidence and data from Internet platforms about not just the amounts of content they remove but also the exact nature and substance of the content is not easy. Comparative data would also be needed to test the aforementioned hypothesis.

The second challenge is how to control for other potential variables (dependent and independent variables). For example, the nature and extent of each country's extant hate speech laws (e.g.

[135] Interview with Unia, 1 October, 2019.

[136] Interview with Article 19, 20 September, 2019.

[137] Interview with Article 19, 19 June, 2019.

the breadth of the hate speech laws) might also impact any tendency on the part of Internet platforms to over-remove lawful or legal content quite apart from any regulatory interventions for online content.

The third, related challenge is how to interpret the evidence that is available in instances where actual Internet regulations are complex and multifaceted, such as when the Internet regulations impose various procedural responsibilities on Internet platforms, and do far more than simply allow the imposition of administrative fines for patterns of failure to remove unlawful hate speech content.

Consider the NetzDG Act which came into force in January 2018. What evidence would confirm the hypothesis that the introduction of NetzDG incentivised the over-removal of lawful content? What evidence would confirm the more specific hypothesis that NetzDG's imposition of a responsibility to remove unlawful hate speech content in particular created an incentive for over-removal of lawful content as opposed to other, more procedural aspects of NetzDG? For instance, if the hypothesis is correct then would it predict that Internet platforms would have reported a significant increase in their removal of allegedly unlawful hate speech content post January 2018? Perhaps. But the available evidence is far from clear-cut. Facebook, for example, as part of its own transparency policy, publishes on its website statistics, broken down by countries around the world, on the number of instances per half year when it has "limited access to content based on local law".[138] (Facebook mentions removing content in response to reports about content "alleged to be illegal" received from "governments and courts, as well from non-government entities such as members of the Facebook community and NGOs".[139]) The number of instances when Facebook removed allegedly illegal content in Germany are as follows: 2019-H1: 937, 2018-H2: 1148, 2018-H1: 1764, 2017-H2: 1893, 2017-H1: 1297.[140] Nevertheless, these figures neither confirm nor disprove the aforementioned hypothesis, and for several important reasons. First, the figures include all allegedly unlawful content, not just allegedly unlawful hate speech. Second, the change in removal numbers between the second half of 2017 and the first half of 2018, down from 1893 to 1764, is not significant, albeit the numbers do continue to decline into the first half of 2019. Third, the numbers relate to removal of allegedly unlawful content, but because the lawfulness of the content has not actually been tested in a court of law, it cannot be safely assumed that these numbers indicate or signpost the over-removal of lawful content (any more than these numbers can be safely assumed to equate exactly to the proper removal of only unlawful content from Facebook). Fourth, any sustained decrease in removal numbers post-January 2018 cannot reasonably be interpreted as disproving the aforementioned hypothesis; that is to say, they cannot reasonably be interpreted as demonstrating that NetzDG did not incentivise the over-removal of content. The decrease could be explained just as easily by an increase in Facebook removing content based on its own community standard on hate speech, with the effect that Facebook removed content based on its own community standard before that content needed to be removed based on Facebook receiving reports that the content allegedly violated local laws. And it is possible that an increase in removal based on Facebook's own community standard on hate speech did itself have something to do, directly or indirectly, with NetzDG incentivising the over-removal of content (just as it is possible that an increase had nothing to do with NetzDG).

[138] Statistics published by Facebook Transparency. Available at: https://transparency.facebook.com/content-restrictions [last accessed 17 April, 2020].

[139] Ibid.

[140] Ibid.

Notwithstanding these points about evidential challenges, this study recommends that the credible risk that the regulatory imposition of a responsibility on Internet platforms to remove unlawful hate speech backed up with a fines regime for patterns of failure in this responsibility could create an incentive for platforms to over-remove lawful content is a risk that needs to be taken seriously and managed as a matter of precautionism. Precautionism in this sense is about devising and adopting policies to mitigate the credible risk, even when the exact nature of the risk is unknown due to the absence of complete or perfect data.[141] This study will consider ways of mitigating this serious danger in section IV.D below.

(i) Exceptions from legal responsibilities in the case of journalistic content

One notable variant of Regulatory-C addresses the potential weakness identified in Table 23 above that it might potentially give insufficient protection for journalistic content, which is typically deemed to possess high free speech value for reasons of truth discovery, democratic legitimacy, speaking truth to power, dissent, uncovering abuse of power, social injustice and oppression, and so on. Of course, in some local contexts hate speech laws might already give exceptions to journalistic content, meaning that certain forms of journalistic content could not be prosecuted under local hate speech laws. However, that might not be the case in all countries. And so if Internet platforms must operate under a legal responsibility to remove illegal hate speech content within specified time frame under threat of regulatory sanctions, then there could be danger that journalistic content gets swept up in the rush to removal. This might be especially problematic in circumstances where an online journalist is mentioning, as opposed to using, hate speech in the course of uncovering prejudice or reporting the scale of the problem of hate speech in the country. Potential strengths and weaknesses of a variant of Regulatory-C that addresses this particular issue are set out in Table 26 below.

Table 26. Regulatory-C: Legal responsibility to remove unlawful hate speech enforced with fines (exceptions for journalistic content)

Tool		
Regulatory-C (*continued*) Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified period following notice (e.g. “manifestly” unlawful within 24 hours, otherwise unlawful within 7 days) with liability to fines for a pattern of failure to comply with that responsibility		
Tool variants	Strengths	Weaknesses
Exceptions from legal responsibility to remove hate speech in the case of journalistic content	- Protects journalistic content that often has high free speech value - Clarifies the legal responsibilities (or lack thereof) of Internet platforms that only carry journalistic content - Shows sensitivity to the fact that journalists already operate under editorial codes of ethics	- Creates ambiguity as to liability of Internet platforms with websites, services and platforms that carry both ordinary and journalistic content - Gives an incentive to Internet platforms with websites, services and platforms that carry journalistic content to remove “below the line” public comment sections - Potentially redundant depending on how unlawful hate speech content is defined in some local contexts - Potentially redundant if legal compliance teams have a clear understanding of the scope of hate speech laws - Could potentially enable the spread of hate speech content by unscrupulous or unethical journalists - Could potentially enable the spread of hate speech content by citizen journalists who are not subject to the same editorial codes of ethics as professional journalists

[141] For more on precautionism as a general reason not to regulate hate speech, see Schauer (2009). For more on precautionism as an approach to justifying the regulation of hate speech, see Brown (2015, 2017d).

(ii) Exceptions from legal responsibility for Internet platforms that refer grey area cases to competent independent institutions and abide by the decisions

Another notable variant of Regulatory-C involves granting Internet platforms an exception to the legal responsibility to remove potentially unlawful hate speech content within specified time frames if instead, within those time frames, the Internet platforms refer the cases to a competent independent institution that will form a judgment about the legality or illegality of the content, and if the Internet platforms also agree to abide by those decisions. As a real world illustration of this variant, under s. 3(2)3.b) of the NetzDG Act, the general requirement for Internet platforms to remove potentially unlawful content within 7 days of receiving a report or complaint about the content does not apply *inter alia* in circumstances where the platform refers the case to a competent independent institution ("institution of regulated self-governance") within 7 days of receipt, and agrees to accept the decision of that institution. Potential strengths and weaknesses of this variant are set out in Table 27 below.

Table 27. Regulatory-C: Legal responsibility to remove unlawful hate speech enforced with fines (exceptions for referrals to competent independent institutions)

<table>
<tr><td colspan="3">Tool</td></tr>
<tr><td colspan="3">Regulatory-C (<i>continued</i>) Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified period following notice (e.g. "manifestly" unlawful within 24 hours, otherwise unlawful within 7 days) with liability to fines for a pattern of failure to comply with that responsibility</td></tr>
<tr><td>Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td>Exceptions from legal responsibility for Internet platforms that refer grey area cases to competent independent institutions and abide by the decisions</td><td>- Potentially mitigates incentives for over-removal of content (a "remove first" policy)
- Ensures that legal opinions formed by Internet platforms' legal teams are tested by independent competent institutions ("second opinions")
- Guarantees that decisions are based on expertise about local hate speech laws
- May lend added credibility to removal decisions, especially in grey area cases</td><td>- Still does not reach the same high standards of due process as a judicial decision
- Creates an incentive for Internet platforms to flood the competent independent institution with cases
- Potentially redundant if Internet platforms already typically instruct the services of external legal counsel in local contexts</td></tr>
</table>

Among the various potential drawbacks with this variant of Regulatory-C is that it could create an incentive for Internet platforms to send a large number of cases to the competent independent institution ("institution of regulated self-governance") merely so as to avoid failing to comply with the 7 days requirement, and to avoid paying fines.[142] Arguably this would constitute the platform taking advantage of or exploiting the exception in ways that undermine the regulatory purpose being pursued by governmental authorities.

To mitigate this incentive, this study recommends that competent independent institutions should be granted the power to select the cases they will hear, so as to prevent Internet platforms from inundating or flooding these institutions with cases simply to qualify for exceptions and to avoid fines. The Internet platforms and the competent independent institutions should work closely to set standards for when cases really are difficult cases that genuinely merit referral to the competent independent institution.

[142] Interview with Laëtitia Avia, 28 October, 2019.

(iii) Exemptions from liability for Internet platforms granted "responsible platform" status

One response to the above weaknesses might be to say that Internet laws giving governmental authorities to impose fines on Internet platforms for a pattern of failure to remove unlawful hate speech content should be repealed or else never introduced in the first place. However, this might drastically weaken the relevant Internet laws and allow irresponsible Internet platforms to continue to fail to remove unlawful hate speech content with impunity.

An alternative way of mitigating the problems—and one that involves moving to a process-based approach to regulation [see section I.C]—could be to build a "responsible platform" clause into any Internet law that imposed a legal responsibility on Internet platforms to remove unlawful hate speech content enforced through a system of fines. To explain, the responsible platform clause would make Internet platforms *ex ante* exempt from such fines for the period of the exemption, if they can demonstrate a high degree of responsible conduct. Potential strengths and weaknesses of this variant appear in Table 28 below.

Table 28. Regulatory-C: Legal responsibility to remove unlawful hate speech enforced with fines (exemptions for Internet platforms granted "responsible platform" status)

Tool		
Regulatory-C (*continued*) Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified period following notice (e.g. "manifestly" unlawful within 24 hours, otherwise unlawful within 7 days) with liability to fines for a pattern of failure to comply with that responsibility		
Tool variants	Strengths	Weaknesses
Exemptions from liability for Internet platforms granted "responsible platform" status	- A shift from outcome-based regulation to process-based regulation - Mitigates incentives for over-removal of content - Removes the adversarial relationship created by the imposition of fines - Potentially builds trust between governmental agencies and Internet platforms - Reduces resource burden on governmental agencies in continuously investigating whether in fact Internet platforms are removing unlawful hate speech content as required - Reduction of resource burden on governmental agencies especially important in times of economic austerity and limited budgets across the justice system as a whole - Internet platforms are not compelled to apply for "responsible platform" status - Promotes definitional harmonisation	- Resource burden on government authorities in assessing applications for responsible platform status - Incentivises Internet platforms to game the application system (i.e. improving legal compliance but only during the period of assessment of their applications) - Appearance of treating some Internet platforms as "second-class" businesses - Removes the possibility of imposing fines as a last resort during the period of exemption

On a practical level, Internet platforms could apply for this *ex ante* exemption status, the "responsible platform" status, periodically, say, every two, three or five years, and during the period of exemption would not be liable for the relevant fines (i.e. fines for a pattern of failure to remove unlawful hate speech content). For some platforms this may be preferable to a situation in which they find themselves in administrative court on a regular basis launching objections to fines that governmental authorities are seeking to have imposed on them.

Internet platforms, like all commercial enterprises, have limited financial resources, and so need to decide how to spend money on governance in ways that best reflect not merely their corporate mission but also facts about value for money or efficiency. Some platforms may take the view that it is more beneficial that the lion's share of resources should go into supporting their legal compliance teams and moderation teams make the right decisions about potential hate speech content the first time around rather fighting legal cases in administrative courts (i.e. launching objections to fines) and paying fines when legal cases are lost.

Another potential benefit of the responsible platform clause being proposed in this study is that it is voluntary in the sense that Internet platforms are under no legal obligation to apply for the responsible platform status.

Granting responsible platform status would also have potential cost savings for governmental authorities (e.g. Internet regulators, ministry of justice), who could focus on the assessment of periodic applications by Internet platforms for the responsible platform exemption status, rather than spending money on (a) continuous investigations as to whether Internet platforms are in fact meeting their legal responsibilities to remove unlawful hate speech, and (b) imposing fines on Internet platforms and defending these fines in administrative court on a regular basis.

What standards should the “responsible platform” status seek to capture? From the perspective of governmental agencies the acid test is “governance alignment”. The basic idea is that there is less urgency for governmental authorities to fine Internet platforms for a pattern of failure to remove hate speech content if the platforms are pursuing moderation and oversight of moderation in ways, procedurally speaking, that are reasonably closely aligned to the government’s own regulatory procedures and goals for tackling online hate speech.

Given this acid test, how should the “responsible platform” standard be operationalized? In abstract or general terms the standard might say that an Internet platform should be entitled to an exemption from fines provided that the Internet platform acts in good faith or makes a good faith effort to meet its legal responsibilities to remove unlawful hate speech content. But what, more precisely, does a good faith effort look like? What can authorities reasonably expect?

In more concrete of practical terms, the “responsible platform” standard could be translated using four main operational tests or qualifying criteria:

(1) Transparency of Moderation and Legal Compliance
(2) Proper Resourcing of Moderation and Legal Compliance
(3) Robust Oversight of Moderation and Legal Compliance
(4) Definitional Harmonisation

Starting with (1) Transparency of Moderation and Legal Compliance, the requirement is for Internet platforms to produce regular public reports detailing, amongst other things, statistics on the quantities of notifications, reports or flags concerning hate speech content that have been handled or dealt with by their content moderation teams and legal compliance teams respectively, and also statistics on the decisions reached broken down according to decision to take down content and decisions to leave up content, for example.

In terms of (2) Proper Resourcing of Moderation and Legal Compliance, the requirement would be for Internet platforms to properly fund their content moderation teams and legal compliance teams, taking into account such factors as the size of the teams relative to the size of the platform, the level of expertise and competence of the teams, the quality of training and support provided to them, and striking a reasonable balance between full-time employees and contract labour.

In order to make the assessment of this qualifying criterion more feasible, a proxy measure could be used such as the amount of money spent on funding content moderation teams and legal compliance teams. Although crude, the proxy has the benefit of monitorability. The proxy should be understood not as an absolute amount of money but as an appropriate amount relative to the nature and size of the Internet platform. This might be operationalised as a requirement to spend a minimum percentage of revenue on moderation and legal compliance.

Put simply, rather than seeking to fine Internet platforms up to 4 percent of revenues for a failure to achieve some arbitrarily fixed target rate for the removal of unlawful hate speech, governmental authorities could instead invite platforms to spend a minimum percentage of their revenues on funding the work of their content moderation teams and legal compliance teams to achieve the responsible platform status which gives them an exemption from fines. If Internet platforms elect to put the money that they might otherwise have spent on paying or challenging regulatory fines into the work of their content moderation teams and legal compliance teams, then this is potentially a win-win for both sides. Governmental authorities promote responsible conduct on the part of Internet platforms and for their part platforms do not waste money on paying or challenging fines because the money goes directly into supporting the work of their content moderation teams and legal compliance teams. As with fines themselves, potentially the minimum percentage of revenues that Internet platforms are expected to invest into the work of their content moderation teams and legal compliance teams could be set as a sliding-scale depending on the nature and size of the platform.[143]

At this stage, it might be objected that making the minimum investment of revenues in supporting the work of moderation teams and legal compliance teams a qualifying criterion for the responsible platform status is arbitrary. For example, why not instead make Internet platforms donating a minimum percentage of their revenues to NGOs or charities that assist victims of online hate speech a qualifying criterion for the responsible platform status?[144] An obvious answer is that this smacks of allowing Internet platforms to leave up hate speech content in return for "paying off" the victims—a kind of blood money. Surely it is better to properly fund teams that remove content in the first place than to use that money to pay organisations to deal with the "fallout" of failures to remove content.

Turning to (3) Robust Oversight of Moderation and Legal Compliance, this requires Internet platforms to achieve an adequate degree of robustness of oversight. Take the case of Internet platforms that have an independent supervisory council, steering committee or oversight board. Robustness of oversight might be a function of process factors, such as, for example, the amount and quality of external input that the Internet platform's independent supervisory council, steering committee or oversight board has into the moderation and legal compliance practices of the Internet platform. And also a function of outcome factors, such as the extent of actual impact that the independent supervisory council, steering committee or oversight board has over the moderation and legal compliance practices of the Internet platform.

As a rule of thumb, the higher the degree of robust oversight an Internet platform has, the less restrictive the regulation should be, meaning greater exemptions for the Internet platform. This rule of thumb is depicted in Diagram 2, which shows optimal (green), partially optimal (orange) and sub-optimal (red) combinations of oversight and regulation.

[143] 1st consultative meeting, London, 17-18 October, 2019.
[144] Ibid.

Diagram 2. More or less optimal combinations of oversight and moderation

	No oversight	Some oversight	Significant oversight	Robust oversight
Exemption from regulation	🟥	🟥	🟧	🟩
Permissive regulation	🟥	🟧	🟩	🟧
Moderately restrictive regulation	🟧	🟩	🟧	🟥
Highly restrictive Regulation	🟩	🟧	🟥	🟥

However, merely achieving transparency, properly resourced content moderation teams and legal compliance teams, and robust oversight of moderation and legal compliance should not be sufficient for an Internet platform to be granted the responsible platform status. It might well be that an Internet platform is being transparent, properly funding its content moderation teams and subjecting its moderation practices to robust oversight from an independent supervisory council, steering committee or oversight board, but this alone would not guarantee that the Internet platform is getting its moderation right as far as governmental agencies are concerned. This is due to the potential for definitional divergence, that is, discordant or mismatched definitions of hate speech between the Internet platform and local hate speech laws.

If the Internet platform undertakes content moderation based on a “community standard” or “content policy” that defines impermissible hate speech without any formal reference to unlawfulness and without substantive overlap with legal definitions, then not even properly resourced moderation and robust oversight of moderation is likely to satisfy governmental agencies that adequate steps are being taken to tackle unlawful hate speech content.

Of course, the Internet platform might insist that within its management structure there is a sort of division of labour between its content moderation teams who focus on applying its community standards on hate speech, on the one hand, and its legal compliance teams who concentrate on removing illegal content including illegal hate speech, on the other hand.

However, as discussed in section I.F, there remains a significant risk that on a day-to-day level content moderation teams will end up dealing with a substantial amount of content that also happens to raise issues of legality. If there is a good chance that the content moderation teams do not refer, upscale or escalate every single case of potentially unlawful hate speech content to the legal compliance teams, then the content moderation teams will be, in effect, making decisions on cases that raise legal issues.

And so if governmental authorities are serious about getting Internet companies to tackle unlawful hate speech content, then the former will have an interest in ensuring not only that Internet platforms' legal compliance teams are working effectively but also that their content moderation teams are also applying community standards or content policies on hate speech that do not diverge too far from local legal definitions of hate speech. Where the definitions of hate speech being used by the Internet platforms' content moderation teams are divergent from legal definitions, in other words, it is not clear that governance alignment has been achieved. In that scenario regulation is not redundant and exemptions from fines should not be granted.

Therefore, achieving the responsible platform exemption status should also require (4) Definitional Harmonisation, namely, that the definition of hate speech used by the Internet platform is in reasonable harmonisation with local hate speech laws, meaning that the two show a sufficient degree of convergence. This could be achieved in two ways. First, through formal definitional harmonisation: that is, the Internet platform employs a definition of hate speech that makes reference to "unlawful hate speech as defined in domestic law", for instance. Or, second, by means of substantive definitional harmonisation: the Internet platform employs a definition or characterisation of hate speech that is broadly similar to the ways governmental agencies and local hate speech laws define or characterise hate speech.

In countries without well developed statutory or common law definitions of hate speech and without widely recognised hate speech laws, it will be difficult for governmental authorities to fairly seek to impose fines on Internet platforms for an alleged pattern of failure to remove unlawful hate speech content. In these countries primary legislation on hate speech may be a necessary preliminary step prior to enacting the current version of Regulatory-C or something like it.[145]

To recap, the current proposal, original to this study, is that Regulatory-C should incorporate a responsible platform exemption status that Internet platforms may voluntarily apply for. Moreover, the exemption status should depend on four operational tests or qualifying criteria: (1) transparency in moderation and legal compliance, (2) spending a minimum percentage of revenues on funding the work of its content moderation teams and legal compliance teams, (3) ensuring robust oversight, and (4) achieving definitional harmonisation.

This proposal has several important implications. First, it shifts the focus of governance of online hate speech partly away from outcome-based governance to process-based governance.

Second, it does not do away with fines entirely but instead creates a kind of conditional liability for fines. Internet platforms are only liable to fines for a pattern of failure to remove unlawful hate speech content if they choose not to apply for, or fail to achieve, responsible platform exemption status, and then also demonstrate a pattern of failure to remove content. Importantly, the responsible platform exemption status is a voluntary scheme intended for use in countries where Regulatory-C or something like it is in operation. Internet platforms are free to apply for the exemption status or not. It is not the case that they will be automatically fined if they do not apply for the status or else have their application rejected. Rather, it is that if they do not apply for, or

[145] 1st consultative meeting, London, 17-18 October, 2019.

are not granted, the status, they will not gain an exemption from the default fines regime and could potentially be subject to fines.

Of course, an Internet platform might take the view that it is very unlikely to be fined and so does not wish or need to apply for the responsible platform exemption status. In other words, an Internet platform might decide that it does not need the certainty that the exemption status confers. So it will not seek to demonstrate it meets the four operational tests or qualifying criteria. Or instead an Internet platform could calculate that it would be infeasible or unbeneficial to meet the tests. For example, an Internet platform might judge that if it can spend less than the minimum percentage of revenues on funding the work of its content moderation teams and legal compliance teams, then it would be suboptimal to spend more than is strictly needed.[146] However, other Internet platforms may take the view that the exemption status is worth applying for perhaps because they fear they will be fined, or because they already meet the qualifying criteria or because meeting the qualifying criteria would not be too onerous. There might also be a reputational benefit to gaining the exemption status.

Third, in order to obtain the responsible platform exemption status Internet platforms may have to make significant changes to their own definitions of hate speech in order to pass test (4) Definitional Harmonisation. This could necessitate a move from global to local definitions of hate speech. Because of this, Internet platforms would also have an incentive not to attempt to build a governance firewall between their moderation and oversight procedures, on the one hand, and their legal compliance procedures, on the other hand [see section I.F].

Finally, for some Internet platforms the responsible platform proposal substitutes continuous monitoring of Internet platforms' compliance and imposition of fines with periodic assessments of Internet platforms' applications for the responsible platform exemption status. Of course, a continuous monitoring process would still be needed in the case of Internet platforms that do not apply for the exemption states or whose applications are rejected.

[146] 1st consultative meeting, London, 17-18 October, 2019.

(iv) Leniency programmes that give Internet platforms reductions in fines if they fully cooperate with governmental authorities

Yet another variant of Regulatory-C involves leniency programmes that grant Internet platforms reductions in fines if they cooperate with governmental authorities during investigations into a suspected pattern of failure to remove unlawful hate speech content within specified time frames. Unlike the *ex ante* exemptions discussed in section IV.C(iii) above, the leniency programmes would be assessed and administered *ex post*, that is, after the fact of any failures to comply with legal responsibilities to remove illegal hate speech content.

The leniency programmes might grant reductions in fines, say, of 30 percent, 50 percent or even higher, to Internet platforms that cooperate with governmental authorities, such as by coming forward with information about the true extent of illegal hate speech content on their platforms (full disclosure). Potential strengths and weaknesses of this variant appear in Table 29 below.

Table 29. Regulatory-C: Legal responsibility to remove unlawful hate speech enforced with fines (leniency programmes)

Tool			
Regulatory-C (*continued*) Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified period following notice (e.g. “manifestly” unlawful within 24 hours, otherwise unlawful within 7 days) with liability to fines for failure to comply with that responsibility			
Tool variants		Strengths	Weaknesses
Leniency programmes	No leniency programmes	- More congruent with systems in which Internet platforms are fined precisely for failures to provide full information	- Lost opportunity to incentivise Internet platforms to cooperate fully with investigations into their failures to remove unlawful hate speech content
	Leniency programmes that grant reductions in fines to companies that fully cooperate with governmental authorities	- Potentially provides governmental authorities with greater access to information about the conduct of Internet platforms in circumstances where it is technically difficult or expensive for them to obtain accurate information by themselves - Incentivises Internet platforms to “break ranks” with other platforms by being more transparent than other platforms about their failures to remove unlawful hate speech content - May lead to greater enforcement of responsibilities and fines being levied on more Internet platforms overall	- Incongruent with systems in which Internet platforms are fined precisely for failures to provide full information - Danger that reductions in fines for admission of failure might reduce the deterrence effect of the fines themselves, leading to recidivism

One potential weakness with leniency programmes is that in some governance systems (e.g. under the NetzDG Act in Germany) Internet laws also place a legal responsibility on Internet platforms to provide reports on the extent of hate speech content on their platforms or services and on the removal of that content, i.e. transparency requirements, and, importantly, grant governmental authorities powers to seek to levy fines on Internet platforms for failures of transparency. Leniency programmes would in effect grant Internet platforms reductions in fines for a pattern of failure to remove unlawful hate speech content in exchange for providing governmental authorities the very reports or information that they also have a legal responsibility to provide. This makes for incoherence in governance.

Nevertheless, leniency programmes may be suitable where a fines regime exclusively targets a pattern of failure to remove unlawful content, and in circumstances where it is technically difficult or expensive for governmental authorities to obtain accurate information themselves.

D. Legal responsibility not to over-remove lawful hate speech enforced with fines

Regulatory-D involves using legislation to impose a legal responsibility on Internet platforms not to undertake over-removal of lawful hate speech content, and to grant powers to governmental authorities to impose fines, or apply to courts for permission to impose fines, on Internet platforms for a pattern of failure to comply with this responsibility. Regulatory-D is one solution, albeit a radical solution, aimed at reducing or mitigating the tendency to risk-aversion potentially caused by the imposition of fines for a pattern of failure to remove of unlawful hate speech. Potential strengths and weaknesses of Regulatory-D are set out in Table 30 below.

Table 30. Regulatory-D: Legal responsibility not to over-remove lawful hate speech enforced with fines

Tool	Strengths	Weaknesses
Regulatory-D: Impose a legal responsibility on Internet platforms not to over-remove lawful hate speech content with liability to fines for a pattern of failure to comply with that responsibility	- Mitigates the incentive to take a "safety first" approach to removing hate speech - Discourages Internet platforms from what some people deem "censorship" - Protects the human right to freedom of expression	- Potentially undermines commercial freedom (the freedom to conduct a business) - Puts even more power in the hands of governmental authorities to control online content - Potentially a disproportionate intervention to the problem being addressed - Both redundant and potentially not the least restrictive governance tool if other corporate, regulatory, judicial, or reputational protections of free speech are available - Danger that Internet platforms would be fined no matter what they did thus giving the appearance of a "stealth tax" on platforms - Risk that in practice most fines would be levied for under-removal of unlawful hate speech content
Collaboration potential		
Medium (Governmental agencies; legislature; Internet platforms)		

Whilst this study is unaware of any governmental authorities or courts that have already imposed fines on Internet platforms for over-removal of lawful hate speech content, the idea is hinted at in Art. 4 of the Avia Bill in France. The Bill aims to give the regulator ("Superior Audiovisuel Council") the power to impose fines on Internet platforms of up to 4 percent of turnover for serious and repeated failures in tackling illegal hate speech, but, importantly, states that the regulator may base its decisions on its "appreciation" of "the insufficiency *or excessive* nature of the operator's withdrawal" of content (emphasis added). The intention is that the regulator will consider not specific content removal decisions but more general "patterns" or "tendencies" in those decisions in terms of insufficient or excessive removal.[147]

Philosophically, the case for imposing fines for over-removal of lawful hate speech content might be put like this. Perhaps in an ideal world there should be no fines for the under-removal of unlawful hate speech content. But if in the real world fines are levied for under-removal and this has unwelcome consequences, and if the fines are not going to be repealed any time soon due to political realities, then extending the fines could become a second-best outcome. If the first choice of repealing fines for under-removal is simply not going to happen and Internet platforms behave as they do, then it might become necessary as a remedial step to add an extra layer of fines to counteract the tendency to over-removal. From this perspective the extra layer of fines could make sense as part of non-ideal theory or the art of the second-best.

[147] Interview with Laëtitia Avia, 28 October, 2019.

Furthermore, the particular line in the sand that Regulatory-D relies on, namely, fines for a pattern of over-removal of lawful hate speech content, seems no more or less fuzzy than the line relied upon by Regulatory-C, namely, fines for patterns of failure to remove unlawful hate speech content. If it is reasonable—which is, of course, debatable—to expect Internet platforms to make assessments of unlawfulness of hate speech content, then surely it is no more or less reasonable to expect them to make assessments of lawfulness of hate speech content.

Moreover, the possible extension of fines has the process-aesthetic merit of being "a more balanced approach".[148]

However, as set out in Table 30 above, there are several potential problems with Regulatory-D. For one thing, it restricts the freedom of Internet platforms to go further than local hate speech laws and to remove lawful hate speech content in line with their own community standards or content policies, and based on their specific corporate values and business models. As such it potentially raises commercial freedom issues, namely, the freedom to conduct a business.

Similarly, the public might think that an Internet platform should not be subject to fines or other administrative sanctions for over-removal of lawful hate speech content if they acted in good faith to remove or restrict access to hate speech content in order to protect the human right not to be targeted by hate speech or associated human rights (a good Samaritan clause).

For another thing, the proposal to impose fines for over-removal of lawful hate speech content puts even more power in the hands of governmental agencies to control content. Even if it achieves the aim of discouraging Internet platforms from engaging in what some people deem "censorship", it comes at the price of installing governmental agencies in the role of "content police" who can fine Internet platforms both for what they do not remove and for what they do remove. Some people believe that governmental agencies should simply not be in the business of curating the Internet, especially when it comes to grey area cases of hate speech.[149]

A related worry is that in the real world fines are more likely to be imposed for under-removal of unlawful hate speech content than for over-removal of lawful hate speech content.[150] This might not be by intention, but may be a de facto outcome given current sensibilities and pressures on governmental agencies from the media and the general public (e.g. tech backlash).

Moreover, the levying of fines for both under-removal and over-removal of hate speech content means that Internet platforms may find it difficult to tread the line perfectly and avoid fines. They might end up simply being fined whatever they did.[151] As a result the fines might have the appearance, deserved or not, of a "stealth tax" on Internet platforms.

Furthermore, the proposed extension of fines is unlikely to be the least restrictive governance tool available to mitigate the tendency to over-removal of lawful hate speech potentially caused by Regulatory-C. Diagram 3 below sets out a range of other governance tools that would be less restrictive but could potentially achieve the same end of protecting free speech.

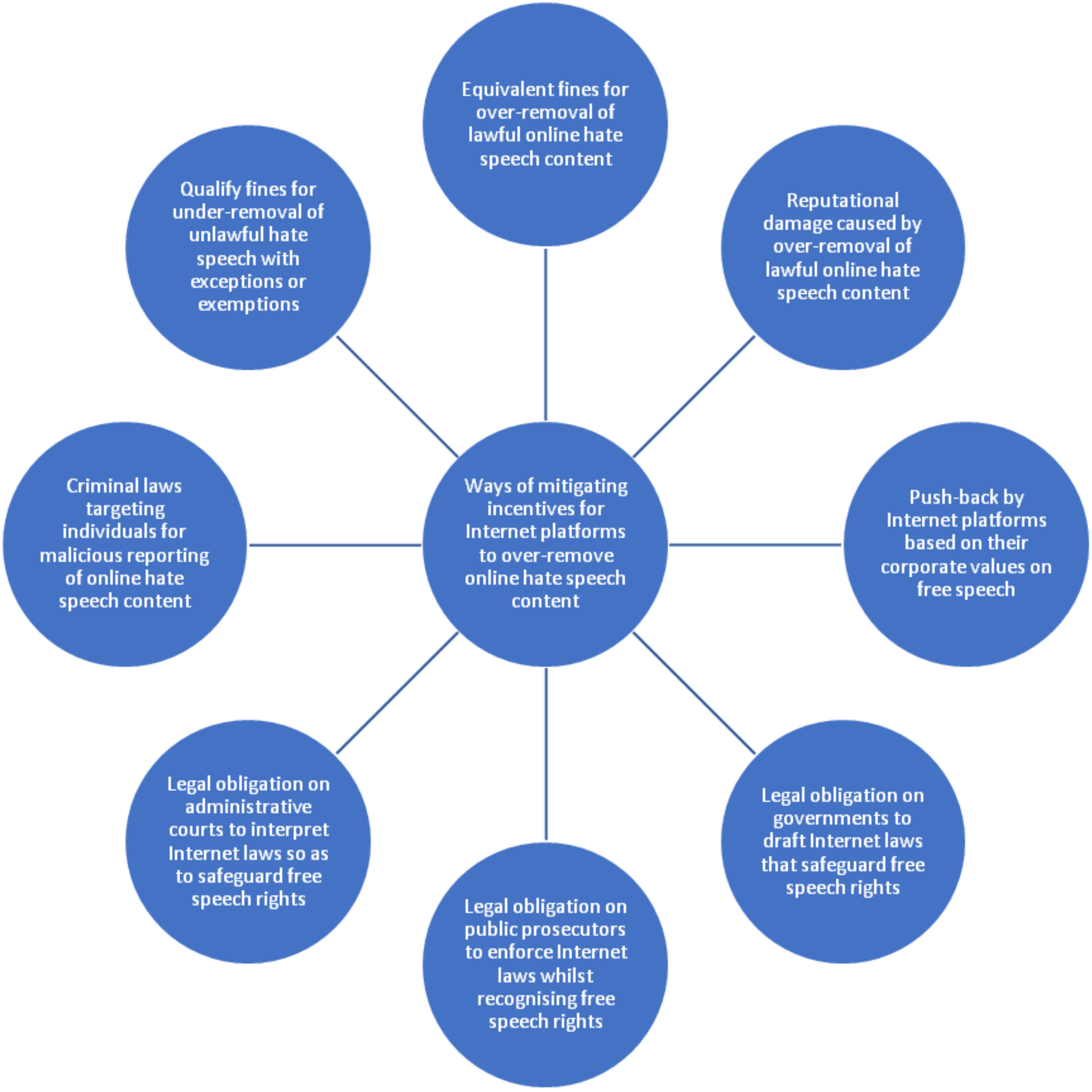
[148] Interview with representatives of civil society organisations and research centres, 29 July, 2019.

[149] Ibid.

[150] Ibid.

[151] Ibid.

Diagram 3. Ways of mitigating the tendency to over-removal of lawful online hate speech content


Equivalent fines for over-removal of lawful online hate speech content
Reputational damage caused by over-removal of lawful online hate speech content
Push-back by Internet platforms based on their corporate values on free speech
Legal obligation on governments to draft Internet laws that safeguard free speech rights
Legal obligation on public prosecutors to enforce Internet laws whilst recognising free speech rights
Legal obligation on administrative courts to interpret Internet laws so as to safeguard free speech rights
Criminal laws targeting individuals for malicious reporting of online hate speech content
Qualify fines for under-removal of unlawful hate speech with exceptions or exemptions
Ways of mitigating incentives for Internet platforms to over-remove online hate speech content

Notable among these alternative tools are those already discussed in section IV.C, which involve qualifying fines for under-removal of unlawful hate speech with exceptions or exemptions. Specifically, allowing exceptions for Internet platforms that refer difficult cases to competent independent institutions and abide by the decisions [see section IV.C(ii)]; and granting exemptions from regulatory fines for Internet platforms granted “responsible platform” status [see section IV.C(iii)]

No doubt these other governance tools also have their own weaknesses. For example, some people might question the potential effectiveness of reputational damage for over-removal of lawful hate speech content since the current focus of media attention and the trend of public disquiet is against under-removal of unlawful hate speech content.[152]

So there is always a trade-off between effectiveness and restrictiveness. But the important point here is that (some of) these alternative tools [see Diagram 3] are arguably less restrictive but potentially as effective when used in combination or acting together.

[152] Ibid.

E. Statutory duty of care and/or code of practice

Regulatory-E involves an Internet regulator establishing and enforcing a statutory duty of care and/or code of practice for Internet platforms with specific guidance on how they should handle flags or reports of hate speech content on their platforms, services, websites and products. The potential strengths and weaknesses of Regulatory-E are set out in Table 31 below.

Table 31. Regulatory-E: Statutory duty of care and/or code of practice

<table>
<tr><td colspan="2">Tool</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td colspan="2">Regulatory-E: Internet regulator establishes and enforces a statutory duty of care and/or code of practice with guidance on how Internet platforms should handle hate speech content</td><td>- Potentially reduces the burden on the courts and wider criminal justice system in dealing with huge volumes of unlawful hate speech content online
- Potentially strong policy influence or impact on Internet platforms
- Promotes collaboration rather than adversarial relationship between government authorities and Internet platforms
- Congruence with media regulation in general</td><td>- Possible divergent definitions of hate speech between Internet regulator and Internet platforms
- Devolution of quasi-legislative powers to the Internet regulator without same levels of debate, representation and transparency as the legislature
- Application of the concept of "duty of care" to Internet platforms potentially incongruent with usage of that concept in other areas of law (e.g. lack of focus on physical harm)</td></tr>
<tr><td colspan="4">Collaboration potential</td></tr>
<tr><td colspan="4">Medium to high (Internet regulator; Internet platforms; stakeholders)</td></tr>
<tr><td colspan="2">Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td rowspan="2">Who decides the content of the duty of care and/or code of practice?</td><td>Internet regulator has final say over the content of the duty of care and/or code of practice but consults with Internet platforms, civil liberties organisations, minority rights organisations, equality boards, NGOs and other stakeholders</td><td>- Potentially increases the credibility of the Internet regulator
- Independence of the Internet regulator</td><td>- Potentially weak policy influence or impact
- Risk of lack of transparency in deliberations about the content of the duty of care and/or code of practice
- Risk of powerful organisations exerting more influence than others
- Risk of reputational damage to the Internet regulator if organisations make public complaints of being ignored or side lined
- Potential appearance of a PR exercise</td></tr>
<tr><td>A steering committee composed of the Internet regulator, Internet platforms, civil liberties organisations, minority rights organisations, equality boards, NGOs and other stakeholders has final say over the content of the duty of care and/or code of practice</td><td>- Democratisation of Internet regulation
- Potentially more transparent in its deliberations
- Dilution of the control of the Internet regulator
- Ambiguity over the secondary authority of the Internet regulator</td><td>- Challenge of determining the composition of the steering committee
- Risk of regulatory capture, i.e. control of the steering committee by more powerful organisations
- Risk of instability or collapse of the steering committee due to failure to agree on the content of the duty of care and/or code of practice</td></tr>
</table>

An illustration of Regulatory-E can partly be seen in some of the instruments outlined in the UK government's recent Online Harms White Paper (s. 7.19). One general concern raised by lawyers has been whether the very idea of a "duty of care" is fitting in areas of regulation not having to do with physical harm (see Hurst 2019). Of course, one response would be to defend the need for parity of concern between physical and emotional harm, the latter being especially relevant to online hate speech (see Citron 2014; Brown 2015; Gelber and McNamara 2016).

Notable variants of Regulatory-E include whether its working methods allow for it to operate as a “complaints body”, potentially including both complaints against individual content removal decisions by the Internet platform as well as submissions or reports about patterns of conduct. These variants are set out, along with potential strengths and weaknesses, in Table 32 below.

Table 32. Regulatory-E: Statutory duty of care and/or code of practice (working methods variants)

Tool			
Regulatory-E (*continued*) Internet regulator establishes and enforces a statutory duty of care and/or code of practice with guidance on how Internet platforms should handle hate speech content			
Tool variants		Strengths	Weaknesses
Working methods	Internet regulator identifies and investigates cases of non-compliance acting on its own initiative and working by itself	- Focus on cases the Internet regulator deems important	- Subject to silo thinking and limitations of knowledge and expertise within the Internet regulator’s office
	Internet regulator takes on the role of a “complaints body”, hearing complaints against Internet platforms made by individual users that have exhausted the Internet platform’s internal appeals process and oversight board (if applicable).	- Potentially gives victims of hate speech a powerful remedy - Provides creators of content a right of appeal against removal decisions to an external body - Internet regulator as a “one-stop-shop” complaints service, meaning that (i) users could launch group or “class action” complaints, and (ii) users could launch simultaneous complaints against multiple Internet platforms	- Potential massive increase in the workload of the Internet regulator - Risk of unrealistic, frivolous or malicious complaints - Potentially redundant if the Internet platform has an internal appeals process and refers cases to an oversight board (if applicable)
	Internet regulator acts as a “complaints body” hearing complaints, reports or submission about non-compliance made against Internet platforms by “recognised organisations” such as equality bodies, civil liberties organisations, minority rights organisations, equality boards, NGOs or other stakeholders	- Allows independent knowledge and expertise to influence case selection - Mitigates risk of unrealistic, frivolous or malicious complaints	- Challenges in determining the list of “recognised organisations” - Restricting complainants to “recognised organisations” limits opportunities for individual users to seek redress themselves - Cuts off one potentially important avenue for alerting the regulator to systematic breaches of a duty of care around online harms based on the particular experiences and needs of victims

The idea of an Internet regulator acting as a “complaints body” is briefly touched upon in the above-mentioned Online Harms White Paper (ss. 3.27-3.28). That said, it limits the scope of this “super-complaints” function to “recognised bodies”. “We do not envisage a role for the regulator itself in determining disputes between individuals and companies” (s. 3.30). Potential lost opportunities in this exclusion are outlined in Table 32 above. For one thing, this exclusion raises issues of victim-sensitivity at the regulatory level, discussed in section VII.C(iii) below.

In terms of the Internet regulator acting as a complaints body that hears complaints, reports or submissions about non-compliance made against Internet platforms by “recognised organisations”, the question arises as to which body would grant the “recognised organisations” status. Potentially this could be done by (i) the Internet regulator, (ii) administrative court, or (iii) a steering committee composed of the Internet regulator, Internet platforms, civil liberties organisations, minority rights organisations, equality boards, NGOs and other stakeholders.

Other important variants of Regulatory-E concern different methods enforcement. The potential strengths and weaknesses of these methods are outlined in Table 33 below.

Table 33. Regulatory-E: Statutory duty of care and/or code of practice (enforcement variants)

Tool			
Regulatory-E (*continued*) Internet regulator establishes and enforces a statutory duty of care and/or code of practice with guidance on how Internet platforms should handle hate speech content			
Tool variants		Strengths	Weaknesses
Enforcement of the duty of care and/or code of practice	Fines	- Potentially medium to strong policy influence or impact - Less restrictive than banning, blocking or otherwise disrupting Internet platforms	- Risk that an Internet platform could treat this simply as a "business tax" rather than a reason to change - Devolves quasi-judicial powers to Internet regulators that potentially lack the same high standards of due process - Potentially reduces breathing space for Internet platforms to gradually reform and change their business models and user usage patterns over time, and to flourish financially and in so doing continue to support the digital economy - Excessive fines might be deemed disproportionate
	Naming and shaming Internet platforms for non-compliance	- Less restrictive - Makes use of reputational damage as an incentive - Potentially limits the credibility of the Internet regulator	- Potentially weak policy influence - Reliant on unpredictable and inconsistent reputational damage given contingencies of media coverage and public mood
	Banning, blocking or otherwise disrupting Internet platforms	- Potentially strong policy influence or impact	- Highly restrictive - Potentially disproportionate intervention - Incentivizes Internet platforms to adopt an extreme precautionary approach and potentially diminishes innovation and creativity in the sector

F. Special public prosecutor

Regulatory-F involves a special public prosecutor notifying Internet platforms about particular bits of content that the special public prosecutor deems to be unlawful or illegal hate speech content. Potential strengths and weakness are set out in Table 34 below.

Table 34. Regulatory-F: Special public prosecutor

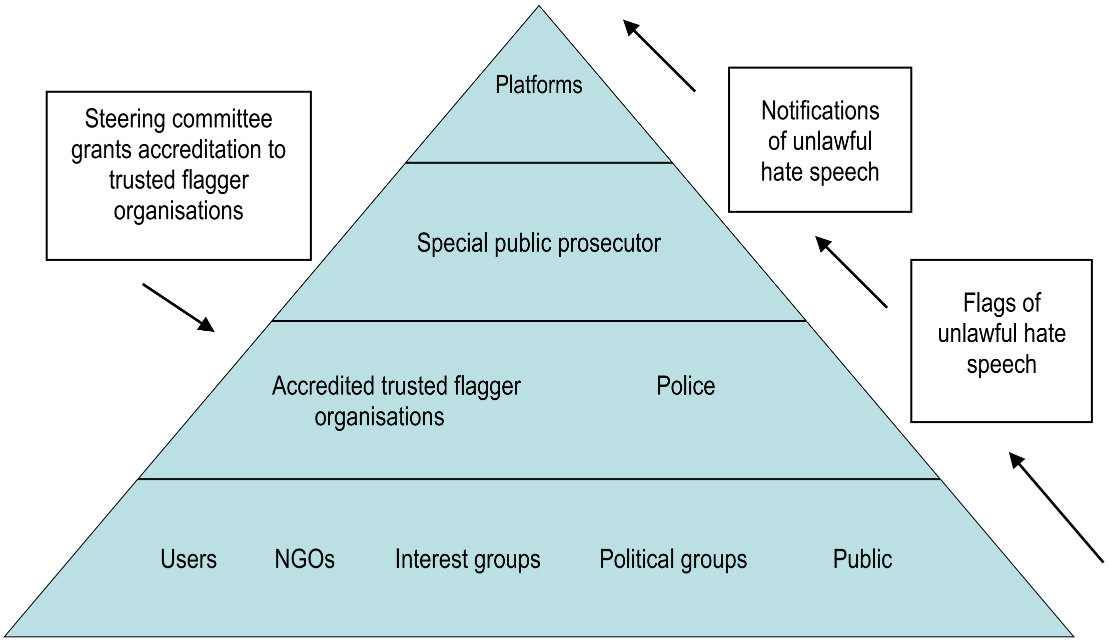
<table>
<tr><td colspan="2">Tool</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td colspan="2">Regulatory-F: Special public prosecutor notifies Internet platforms about particular bits of content that the public prosecutor deems to be unlawful or illegal hate speech content</td><td>- Internet platforms retain ultimate control over content removal decisions
- Filters the flags progressing through to Internet platforms thus potentially increasing quality
- Potentially lends greater credibility to notifications given the expertise and powers of public prosecutors
- Adds another filter stage at which human rights are considered including free speech
- Reduces resource pressures on Internet platforms
- Builds expertise, capacity and engagement with the aim of tackling online hate speech among public prosecutors
- Promotes definitional harmonisation across internet platforms, trusted flaggers and public prosecutors</td><td>- Outsourcing of quasi-judicial powers to special public prosecutor that may lack highest standards of due process
- Potentially leads Internet platforms to deprioritize reports from users
- Potentially creates confusion and ambiguity as to whether the notifications are administrative or quasi-judicial in character</td></tr>
<tr><td colspan="4">Collaboration potential</td></tr>
<tr><td colspan="4">Medium to high (Special public prosecutor; the police; Internet platforms; recognised trusted flagger organisations)</td></tr>
<tr><td colspan="2">Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td rowspan="2">Genesis of notifications made by the special public prosecutor to the Internet platform</td><td>Special public prosecutor makes notifications based on its own assessments but working with a body of cases flagged to it by (i) the police</td><td>- Draws on a body of cases that emerge from the work of well-trained and professional police staff</td><td>- Significant investigatory burden on the police
- Referrals reflect the limitations in investigatory powers of the police and the knowledge and expertise of the special public prosecutor</td></tr>
<tr><td>Special public prosecutor makes notifications based on its own assessments but working with a body of cases flagged to it by both (i) the police and (ii) accredited trusted flagger organisations</td><td>- Draws on a body of cases that emerge from both the police and trusted flagger organisations thus widening the field of expertise
- Still restricts the number of flags coming through to the special public prosecutor thus reducing resource burden</td><td>- Burden of identifying organisations that have the capacity, expertise and independence to act as accredited trusted flaggers
- Risk of missing out on insights from non-accredited organisations</td></tr>
</table>

The variant of Regulatory-F which involves the special public prosecutor receiving flags from both the police and accredited trusted flagger organisations, and in turn, after the special public prosecutor making its own assessments, making notifications to Internet platforms can be illustrated by a forthcoming agreement in a member state of the European Union (anonymous). Governmental authorities in this country are establishing a working procedure for tackling hate speech between accredited trusted flagger organisations ("reliable flaggers"), a special public prosecutor for digital crimes ("point of contact") and Internet platforms.

This variant of Regulatory-F creates a kind of flagging and notification pyramid in which cases progress upwards through a series of filters from users and other stakeholders through to accredited trusted flagger organisations and the police, and from accredited trusted flagger organisations and the police through to the special public prosecutor, after which the special public prosecutor can send notifications to the Internet platforms themselves. This is depicted in

Diagram 4 below. The fact that the pyramid is wider at the bottom depicts the fact that there will be a greater number of flags at the bottom than notifications at the top.

Diagram 4. Model flagging and notification pyramid


Steering committee grants accreditation to trusted flagger organisations
Platforms
Notifications of unlawful hate speech
Special public prosecutor
Flags of unlawful hate speech
Accredited trusted flagger organisations
Police
Users
NGOs
Interest groups
Political groups
Public

Because the special public prosecutor only accepts hate speech flags from accredited trusted flagger organisations and from the police, cases have already passed through at least one filter stage before they even get to the special public prosecutor. The prosecutor itself represents another filter stage. So by the time the special public prosecutor's notifications reach Internet platforms the relevant cases have passed through at least two filter stages.

Under the proposed Spanish system, notifications from the special public prosecutor ("Point of contact") take the form of administrative notifications or requests following this broad form: "The special public prosecutor hereby notifies the Internet platform that it deems certain content which has been posted or shared on the platform to be unlawful or illegal, and requests that the Internet platform remove it." In addition, the special public prosecutor also has the option of passing the details of the case onto a regular public prosecutor, who can launch a full investigation. In such cases the regular public prosecutor could notify the platform that a full investigation has been launched and can also request that the platform remove the content at that stage. If the regular public prosecutor obtains a judicial order from the court, then a "notice and take down" request can be send to the platform ordering the removal of the content.[153]

A potential benefit of Regulatory-F is that, as a rule of thumb, there is potentially greater protection of the human right to freedom of expression the more filter stages a case must pass through before reaching the Internet platform. At each filter stage there is a chance that the content will be deemed too important on free speech grounds to be flagged or notified to the Internet platform. In other words, each filter stage represents yet another opportunity to *un*flag or to *not* notify thus stopping the case from ascending upwards.

[153] 1st consultative meeting, London, 17-18 October, 2019.

Another potential benefit of Regulatory-F is the reduction of resource burden on organisations higher up the pyramid. From the perspective of Internet platforms, for example, the more steps or filter stages that hate speech flags or notifications must pass through before reaching them, the fewer the number of cases that make it all the way. This potentially reduces the resource burden on Internet platforms. Likewise, the resource burden is reduced on the special public prosecutor if it only accepts flags from accredited trusted flagger organisations and the police.

At first glance, it might seem unfair if the practical burden of flagging hate speech content is, in effect, being pushed down the pyramid to stakeholders at the bottom. However, two things mitigate this. One is that although there are a greater number of flags at the bottom there are also a greater number of stakeholders doing the flagging. Another is that the lower down the pyramid the lower the expectation is concerning standards of assessment and due process. Since the special public prosecutor will be expected to achieve higher standards of assessment and due process, it seems fitting that the number of cases it handles are lower.

An important feature of Regulatory-F is that arguably it shifts some of the responsibility for assessing potentially unlawful or illegal hate speech content away from the Internet platform and places it with the special public prosecutor. This feature could be seen as a strength or weakness depending on one's perspective and priorities. On the one hand, if the public believes that decisions to take down a given bit of content on the grounds of it being putatively unlawful hate speech should always be the result of the application of high standards of investigation, assessment and due process, then the public might believe that the involvement of the special public prosecutor is a step in the right direction. The public might prefer this to be done by a court, but if this is not feasible then having a special public prosecutor involved is a second-best option, better than it being left entirely to the Internet platform.

On the other hand, if the public believes that it is only fitting that Internet platforms take the lion's share of the responsibility for removing unlawful hate speech content—perhaps because their services and platforms help to not merely facilitate but also to drive the production and accelerate the distribution of this content and/or because of any financial gains they are making from the content—then the public may think it a retrograde step to relieve the Internet platform of some of that responsibility. "Why should the taxpayer have to fund a special public prosecutor to do some of the heavy lifting for Internet platforms when Internet platforms should be responsible for the content that appears on their platforms?", the public might ask.

Regulatory-F also faces some technical challenges. One is the genesis of cases reported to the special public prosecutor, that is, whether the reports or flags should come from only the police or also from organisations with trusted flagger status or accreditation. The strengths and weaknesses of these variants are set out in Table 34 above.

The variant which includes accredited trusted flagger organisations bringing cases to the attention of the special public prosecutor raises some technical question. First, on what basis should trusted flagger accreditation be granted? Standards of assessment would need to be agreed, such as relating to the competence, independence, capacity, and so on of the trusted flagger organisation. It could be that accreditation is granted for fixed periods only, and trusted flaggers would be assessed based on performance-indicators like accuracy.

Second, which body should grant trusted flagger accreditation? In principle the granting of "trusted flagger" accreditation could be done by (i) the special public prosecutor, (ii) administrative courts, (iii) a steering committee composed of the special public prosecutor, Internet platforms, civil liberties organisations, minority rights organisations, equality boards, NGOs and other stakeholders, or (iv) a committee of elected representatives in the legislature.

No doubt a case could be made for each of these bodies, but arguably given the focus on accredited organisations flagging unlawful or illegal hate speech content to the special public prosecutor, it might be fitting for (ii) administrative courts to perform this function given their expertise in hate speech law, independence and higher standards of due process. Then again, in order to achieve policy influence or impact the special public prosecutor will need to win the confidence and trust of Internet platforms. This might be easier to achieve if trusted flagger organisations are accredited by (iii) a steering committee that includes Internet platforms.

Another potential weakness of Regulatory-F is that it could, if not properly developed and not accurately understood, create confusion and uncertainty among Internet platforms and the public about whether the special public prosecutor's notifications to the Internet platform have the status of administrative notifications (non-binding), quasi-judicial notifications (semi-binding) or judicial notifications (binding). It might be in the interests of the special public prosecutor to maintain a creative ambiguity in which Internet platforms fear that the notifications have a judicial and binding character when they are actually only administrative and non-binding. Therefore, Internet platforms should seek out advice and clarification on the legal standing of the special public prosecutor's notifications.

Moreover, whatever the legal standing of the special public prosecutor's notifications, this study recommends that Internet platforms should be willing, if necessary, to reject or challenge these notifications if the Internet platforms' own legal team thinks they are unsound. For example, an Internet platform might deem that there are strong free speech reasons to reject or challenge a notification requesting content to be removed, such as if it is a grey area case.

G. Bespoke criminal offences

Regulatory-G involves creating bespoke criminal offences relating to online hate speech punishable by fines and/or custodial sentences. The potential strengths and weaknesses of Regulatory-G are set out in Table 35 below.

Table 35. Regulatory-G: Bespoke criminal offences

Tool	Strengths	Weaknesses
Regulatory-G: Create bespoke criminal offences relating to online hate speech punishable by fines and/or custodial sentences	- Provides the police and public prosecutors with the tools to target a range of agents - Potentially fills a void in countries where governance tools targeting Internet platforms are ineffective or absent for some reason	- Places significant resource burden on the criminal justice system in identifying and prosecuting perpetrators - A relatively restrictive form of governance - Burdens and obstacles of introducing new legislation
Collaboration potential		
Low (Government; the legislature)		

Tool variants		Strengths	Weaknesses
Scope of offences	Offences relating to authors or creators who post or share unlawful hate speech	- Tackles the root source of unlawful online hate speech - Parity between offline and online offences	- Challenge of obtaining information about individual users from Internet platforms - Could send a symbolic message that only the authors of online hate speech are responsible
	Offences relating to conduct of senior managers of Internet platforms or platforms as corporate entities, including (i) systemic failures to remove unlawful hate speech content, (ii) failure to disclose identity of users who post unlawful content, and (iii) failure to ensure that unlawful content is not only removed but also archived and stored securely for use in future prosecutions	- Strong deterrence against senior managers and corporate entities failing to act responsibly - Parity with other forms of senior manager and corporate liability	- Incentivises risk-aversion and a “safety first” approach to content removal - Potentially sets unwanted precedent for liability of senior managers and corporate entities in the tech sector - Potential lack of predictability of liability - Potentially disproportionate intervention
	Offences relating to the conduct of individuals who maliciously report or flag content as being manifestly unlawful (i.e. reporting content as being manifestly unlawful whilst knowing that it is not or failing to take due care to check, and for the purposes of bringing another person into legal jeopardy)	- Provides a deterrence against malicious reporting or flagging - Promotes free speech	- Potentially redundant if system of granting “trusted flagger” status is operative and effective - Potentially places too high a burden on reporters or flaggers of hate speech content to determine whether or not it is unlawful - Potentially disincentivises fair reporting thus exacerbating the existing problem of under-reporting

As set out in Table 35, one variant of Regulatory-G focuses on the root source of unlawful online hate speech content, namely, its authors or creators. According to some civil society organisations, including Article 19, for instance, this is a more appropriate and fitting regulatory response to the problem than putting pressure on Internet platforms to “censor” hate speech (Article 19 2016: 16). It also addresses the problem of creators of unlawful hate speech “shopping around” to find permissive Internet platforms (ibid.).

Nevertheless, another noteworthy variant of Regulatory-G is the creation of offences relating to the conduct of senior managers of Internet platforms or platforms as corporate entities. The function of such offences might be to provide a stronger deterrence against Internet platforms failing to remove content. Potentially Internet platforms could treat regulatory fines simply as a "business tax" rather than a reason to change. Criminalising the conduct of senior managers might prove a stronger deterrent. It could also reflect the central role that some senior managers play in determining the Internet platform's policies and practices on content removal.

Another variant of Regulatory-G involves the creation of new criminal offences of "maliciously" reporting or flagging content as being manifestly unlawful. This relates to the conduct of persons who report content as being manifestly unlawful whilst knowing that it is not or failing to take due care to check, and for the purposes of bringing another person into legal jeopardy. An illustration of this variant can be found in the Avia Bill in France [English translation]:

> Article 6-2 of the aforementioned Law n°2004-575 of 21 June 2004, as it appears from Articles 1, 1a and 1b of this Law, is supplemented by a III which reads as follows:
>
> > « III. - The fact, for any person, to present to the operators mentioned in the first paragraph of I of this article content or activity as being illegal under the same I for the purpose of withdrawing or terminating dissemination, while this person knows this information inaccurate, is punishable by one year imprisonment and 15 000 euros fine. »

At first glance, some people might view the proposed punishment for this new offence disproportionate to the problem it seeks to address. But disproportionate or not, arguably the offence is rationally related to a legitimate interest on the part of governmental agencies, Internet platforms and users alike. The interest is deterring persons from maliciously reporting content as manifestly illegal hate speech in circumstances where this could waste police time, where there are serious legal consequences for users posting illegal hate speech and where this is legal jeopardy for senior managers of Internet platforms and for platforms as corporate entities for failing to put in place satisfactory procedures for removing illegal hate speech.

Whether this new offence is the least restrictive means available of addressing the problem is debatable. Indeed, some people might argue that the best solution to the problem of malicious flagging or reporting of illegal hate speech is not additional criminal offences but decriminalising both hate speech and the failure to remove it. But this "solution" is equally debatable.

Nevertheless, it would be wrong to underestimate the extent of the problem. Tarleton Gillespie offers this illuminating account of the current state of play:

> A flag, in its purest form, is an objection. [...] [But] platform operators cannot glean much about the particular nature of the user's objection from just a single flag. One might imagine a flag as meaning "I have judged this to have violated the posted community guidelines"—but platform moderators know that this would be naïve. [...] A complaint could be fuelled by the deepest sense of moral outrage, or the flimsiest urge of puerile chicanery, and it is nearly impossible to tell the difference. (Gillespie 2018: 91)
>
> Flags can be a playful prank between friends, part of a skirmish between professional competitors or rival YouTubers, or retribution for a social offense that happened elsewhere. Flagging may even help generate interest in and publicity around racy content [...] Flagging systems can also be gamed, weaponized to accomplish social and

political ends. There is evidence that strategic flagging has occurred, and suspicion that it has occurred widely. Users will flag things that offend them politically, or that they disagree with; whether a particular site guideline has been violated can be irrelevant. The hope is that enough flags might persuade platform moderators to remove it. Even if a platform is diligent about vindicating content that’s flagged inappropriately, some content may still be removed incorrectly, some accounts may be suspended. (92)

H. Reform of sentencing guidelines

Regulatory-H involves reform of sentencing guidelines for hate speech offences and hate crimes committed online. Potential strengths and weaknesses are listed in Table 36 below.

Table 36. Regulatory-H: Reform of sentencing guidelines

<table>
<tr><td colspan="2">Tool</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td colspan="2">Regulatory-H: Reform of sentencing guidelines for hate speech offences and hate crimes committed online</td><td>- A way of achieving policy goals without burden and obstacles of introducing new legislation</td><td>- Usefulness depends on the usefulness of the existing offences</td></tr>
<tr><td colspan="4">Collaboration potential</td></tr>
<tr><td colspan="4">Low (Government; sentencing council; the courts)</td></tr>
<tr><td colspan="2">Tool variants</td><td>Strengths</td><td>Weaknesses</td></tr>
<tr><td>Strength of sentences</td><td>Lesser sentences (e.g. suspended sentences and non-custodial sentences)</td><td>- Reduces resource burden on some parts of the criminal justice system (e.g. prisons)
- A way of reflecting the importance of promoting free speech online</td><td>- Symbolic of a failure to take the problem of online hate speech sufficiently seriously
- Does not reduce pressure on other parts of the criminal justice system (e.g. police, probation service)</td></tr>
<tr><td></td><td>Greater sentences (e.g. Internet as an aggravating factor in sentencing)</td><td>- Potential deterrent effect
- Redress for targets of hate speech through retributive justice
- Sends a message that governmental authorities take offences committed online seriously</td><td>- Potentially increases resource burden on some parts of the criminal justice system (e.g. prisons)
- Might be deemed a disproportionate response to the conduct
- Deterrent effect highly dependent on the police and public prosecutors being able to bring successful cases in the first place
- Unlikely to deter hard-core offenders</td></tr>
</table>

Interestingly, Spain’s Attorney General, Dona Maria Jose Segarra Crespo, has proposed lesser sentences for persons convicted of hate speech offences under Art. 510 of the Spanish criminal codes if the offences were committed on social media. Her reasoning picks up several of the points outlined in Table 36 above, including proportionality, promoting free speech online and the burden of enforcing long prison sentences. She writes [English translation]:

> A modification of the severe punitive regime provided for in article 510 of the Criminal Code is proposed, whose imprisonment reaches 4 years in prison, and which must be applied in its upper section (a minimum of 2 years and 6 months of prison) when the facts are carried out through a means of social communication, through the Internet or through the use of information technologies (art. 510.3 CP), so that the contents are accessible to a large number of people.
>
> The reform proposal is intended to introduce greater respect for the principle of proportionality of penalties in those cases of public dissemination of messages or content that, although objectively publicly promote, promote or incite, directly or indirectly, the hatred, hostility, discrimination or violence, however, due to its context, content, lack of repetition or personal characteristics or circumstances of the author, have less entity and should not have such a high reproach […].
>
> Experience shows that many of these cases are committed by persons not belonging to criminal groups or organizations, and that they insert deeply offensive or humiliating comments on social networks for certain groups of people for racist, xenophobic,

religious, homophobic or other reasons, and that they have acted impulsively and thoughtlessly. A good part of the authors of these facts, when their identification is achieved, are willing to acknowledge the facts, even in the guard service itself, but compliance cannot be achieved, first because the penalty in the abstract exceeds the threshold of 3 years in prison provided for in 801.1 2nd of the Criminal Procedure Act, and secondly to the high penalties provided for these cases since they entail the inevitable imprisonment. (Segarra Crespo 2018: 970)

I. Special police unit

Regulatory-I involves the creation of a special online hate speech and hate crime police unit (i.e. central police bureau or national hub for combating online hate speech and hate crime). Potential strengths and weaknesses are set out in Table 37 below.

Table 37. Regulatory-I: Special police unit

Tool	Strengths	Weaknesses
Regulatory-I: Special online hate speech and hate crime police unit (i.e. central police bureau or national hub for combating online hate speech and hate crime)	- Does not require new legislation or changes to sentencing guidelines - Builds law enforcement capacity nationally - Increases the public profile of online hate speech and hate crime more generally - Symbolic of governmental authorities taking the problem seriously	- In order to be effective may require significant resources - If the unit is under-resourced and remains ineffectual, this may have negative symbolism about the lack of importance attached to the problem
Collaboration potential		
Medium to high (Government; police; public prosecutors; Internet platforms)		

Tool variants		Strengths	Weaknesses
Powers and capabilities	Coordinating law enforcement of online hate speech and hate crime offences across subnational or regional police forces	- Potential to improve consistency of relevant law enforcement across subnational police forces - Potential to address crimes committed across multiple subnational regions	- Potentially redundant if subnational or regional police forces can do the same
	National training centre for police officers to learn about best practice in enforcement of online hate speech and hate crime offences	- Potential to improve law enforcement	- Potentially redundant if subnational or regional police forces can do the same
	Special investigatory powers including development of text extraction and machine learning tools or algorithms	- Potential to improve law enforcement	- Resource, skills and technological challenges in development of text extraction and machine learning tools or algorithms
	Special investigatory powers including seeking court orders to require Internet platforms to supply information	- Potential to improve law enforcement	- Potential conflict with users' privacy rights - Potential conflict with Internet platforms' responsibility to protect users' private data
	Send notifications to Internet platforms about unlawful hate speech content	- A means of redress for victims of online hate speech	- Outsourcing of quasi-judicial decisions to police unit that may lack high standards of due process - Ambiguity as to whether the notifications are administrative or quasi-judicial in character
	Make referrals of suspected unlawful hate speech to public prosecutors	- Appropriate and familiar role of police units	- Potentially redundant if subnational or regional police forces are capable of the same
	Powers to seek court orders requiring Internet platforms to remove content	- Potential strong policy influence or impact	- Potentially redundant if subnational or regional police forces are capable of the same

V. ON THE BENEFITS AND CHALLENGES OF COLLABORATIVE APPROACHES TO THE GOVERNENCE OF ONLINE HATE SPEECH

At first glance, it might be tempting to understand how Internet platforms would see the benefits of collaboration in tackling online hate speech as a matter of the rational pursuit of corporate self-interest. For some Internet platforms either the type of users they attract or simply their vast scale means that they are dealing with potentially tremendous volumes of flags or reports of hate speech content. As such, prompt as well as accurate moderation is no mean feat. If collaboration with trusted flaggers helps to filter reports, and if this in turn makes the job of moderation feasible within the resource and business model constraints of the Internet platforms, which are commercial enterprises not public bodies after all, then so much the better.

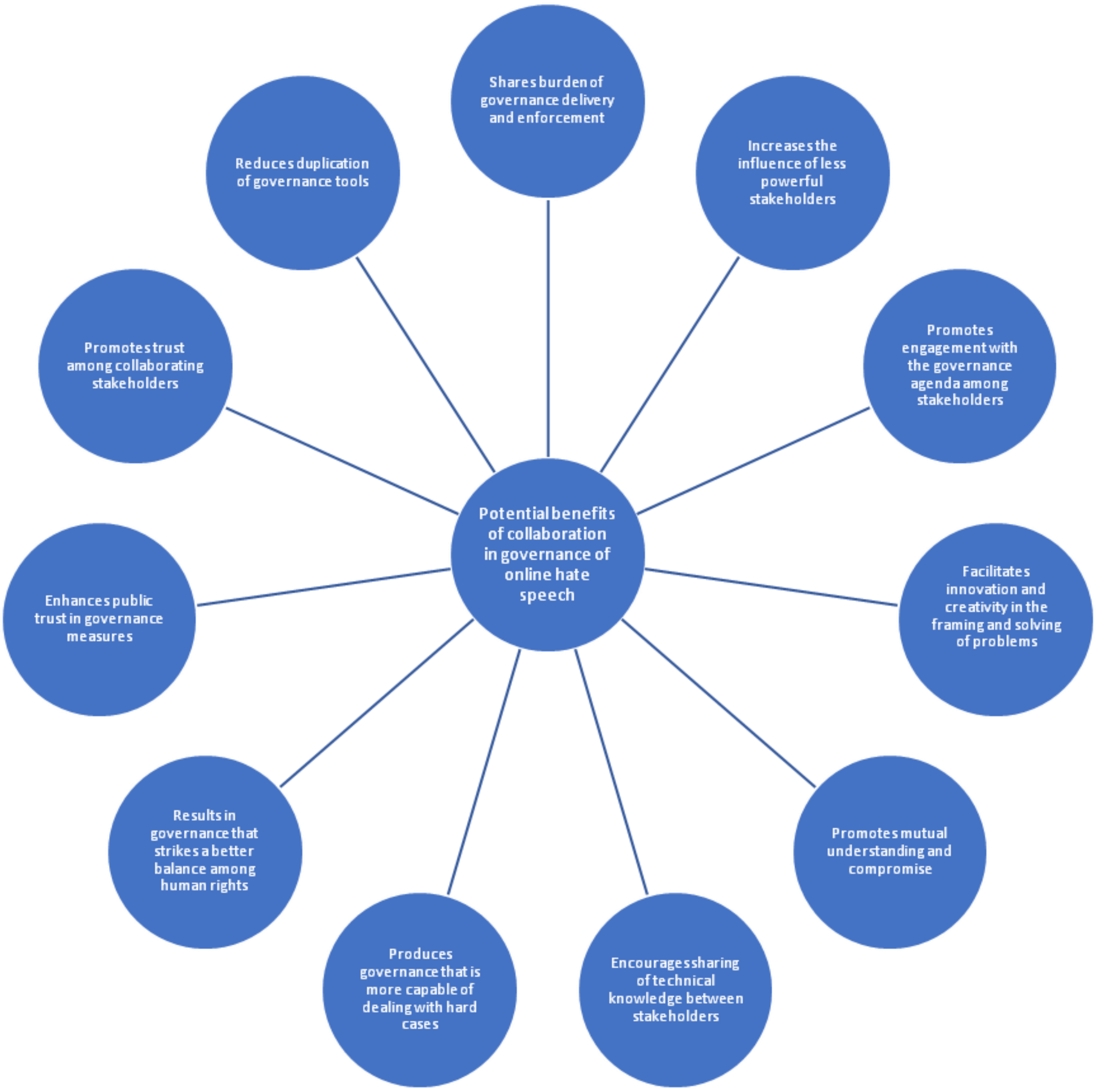
Likewise, it may be hard to resist the thought that where national governments and even intergovernmental organisations indicate a willingness to "fill the vacuum" of governance by imposing legal responsibilities on Internet platforms to remove illegal hate speech content enforced by a system of administrative fines, for example, then it stands to reason that this could act as a powerful catalyst for Internet platforms to become more open to "working with" national governments and intergovernmental organisations.

But are there other sorts of benefits of cooperation and collaboration in the governance of online hate speech that go beyond the aforementioned modus vivendi?

A. Potential benefits of collaboration in the governance of online hate speech

A broad range of potential benefits of cooperation and collaboration in the governance of online hate speech are depicted in Diagram 5 below.

Diagram 5. Potential benefits of collaboration in the governance of online hate speech


Potential benefits of collaboration in governance of online hate speech
Shares burden of governance delivery and enforcement
Increases the influence of less powerful stakeholders
Promotes engagement with the governance agenda among stakeholders
Facilitates innovation and creativity in the framing and solving of problems
Promotes mutual understanding and compromise
Encourages sharing of technical knowledge between stakeholders
Produces governance that is more capable of dealing with hard cases
Results in governance that strikes a better balance among human rights
Enhances public trust in governance measures
Promotes trust among collaborating stakeholders
Reduces duplication of governance tools

For all of the reasons, or potential benefits, listed in Diagram 5, inter-platform collaboration, governmental agency-platform collaboration, and governmental agency-platform-civil society collaboration could be a good thing. For example, in the case of “Promotes engagement with the governance agenda among stakeholders”, some representatives of major Internet platforms who participated in the study reported that the collaborative nature of the European Commission’s Code of Conduct on Countering Illegal Hate Speech Online helped them to convince senior managers to engage with the governance agenda.[154] To give another example, some representatives of trusted flagger organisations and monitoring bodies suggested that collaborating in the implementation and monitoring of the European Commission's Code of Conduct gave them greater influence over the governance of online hate speech than they would have otherwise enjoyed because they are less powerful players.[155]

That being said, these remain potential benefits. More research is needed to see whether or not these benefits are realised equally for different kinds of collaborative governance tools for online hate speech, under what circumstances and in what contexts they can be realised, and which forms of collaboration tend to realise the greatest benefits. For example, it may be that some of these benefits are more difficult to realise in the absence of transparency.[156] Similarly, it could be that some of these benefits are significantly less realisable if there is a significant imbalance of power among the stakeholders involved in collaboration from the start.[157]

[154] 1st consultative meeting, London, 17-18 October, 2019.

[155] Ibid.

[156] Ibid.

[157] Ibid.

B. Potential disbenefits of collaboration in the governance of online hate speech

Another key question that needs to be addressed is this: Is greater collaboration always a good thing when it comes to the governance of online hate speech? Are there any challenges to collaboration? One potential disbenefit of collaboration could be a reduction in the variety of governance models such as if more powerful stakeholders use their position to persuade or even force others to accept one governance model. If pluralism in the governance of online hate speech is replaced with a single model across Europe, for example, then there is a danger that local context could be ignored and group think sets in. To mitigate this tendency it is important for any pan-European governance model to include some decentralization.

A second potential disbenefit occurs if there is a significant imbalance of power among the collaborating partners. If weaker partners lack influence (or "voice") within the collaborative structure and also lack a viable means of exit from the relevant governance arrangements, this may cause them to suffer a net reduction in their control compared to non-collaboration.

A third disbenefit comes to the fore in circumstances where collaborative arrangements collapse for some reason, such as due to creative disagreements or to a break down in good will or trust between stakeholders. This collapse could result in a significant reduction in public trust in organisations involved in the governance of online hate speech across the board.

A fourth potential disbenefit relates to challenges and pitfalls in managing relationships among the participating stakeholders. These challenges are outlined in the remaining subsections.

(i) Challenges in collaboration between Internet platforms and trusted flaggers and monitoring bodies

In its Communication on Tackling Illegal Content Online of 28 September 2017 the European Commission put significant emphasis on the useful role that trusted flaggers can play in the governance of online hate speech.

> The removal of illegal content online happens more quickly and reliably where online platforms put in place mechanisms to facilitate a privileged channel for those notice providers which offer particular expertise in notifying the presence of potentially illegal content on their website.[158]

Yet both Internet platforms and trusted flaggers face several challenges in managing their relationship. One challenge arises from potential "creative differences" about what constitutes impermissible or removable hate speech given fundamental differences in mission. Because Internet platforms are commercial enterprises not public bodies, each platform will also have its own sense of the unique offering it wants to give its users and potential users, including a unique "contract" it seeks to make with its users about how content moderation will be done and according to which particular community standards or content policies. This offering might include a definition of hate speech that is less expensive or more expensive than that which might otherwise be favoured by the trusted flagger organisations it works with.

[158] Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online.

By way of illustration, if a trusted flagger organisation also happens to be a minority rights organisation or some other form of non-governmental organisation (NGO) that focuses on protecting the rights of a particular group of people in society, then its own preferred definition of hate speech might be tailored to capturing forms of hate speech most often directed at the group it seeks to protect. If the trusted flagger also happens to be an equality board and its mission is to promote the elimination of discrimination and segregation, then its own preferred definition of hate speech is likely to be oriented towards these broad policy goals. And if a trusted flagger is a civil liberties organisation that seeks to promote and secure human rights with a particular emphasis on the human right to freedom of expression, then its own preferred definition of hate speech might be narrowly tailored to achieving maximum security for that right. However, when an organisation becomes a trusted flagger for an Internet platform, it will typically be expected to work with the platform's own community standard or content code on hate speech (and perhaps also local legal definitions of hate speech). This in itself may potentially lead to divergence in the way trusted flaggers interpret the Internet platforms community standard on hate speech and the way the platform interprets it.

Another potential source of tension in the relationship comes from the fact that the Internet platform is expected to simultaneously support the trusted flagger organisations it works with and to respect their independence. This tension is implicit in the following passage from the European Commission's Code of Conduct:

> The IT Companies to encourage the provision of notices and flagging of content that promotes incitement to violence and hateful conduct at scale by experts, particularly via partnerships with CSOs, by providing clear information on individual company Rules and Community Guidelines and rules on the reporting and notification processes. The IT Companies to endeavour to strengthen partnerships with CSOs by widening the geographical spread of such partnerships and, where appropriate, to provide support and training to enable CSO partners to fulfil the role of a "trusted reporter" or equivalent, with due respect to the need of maintaining their independence and credibility.[159]

In practice it is typically Internet platforms that approach organisations about becoming trusted flaggers, and it is in the gift of Internet platforms to offer or withhold that status. In one sense this is a relationship in which nearly all the power lies with the Internet platform. Moreover, even once an organisation has acquired trusted flagger status, it is still at the discretion of the Internet platform to determine the extent of the access given to the trusted flagger, such as in terms of how and where its flags enter the Internet platform's moderation process chain. Flags from trusted flaggers can go into the general flagging system, into the group email addresses of special content moderation teams or to more senior members of staff ("points of contact"). Sometimes a trusted flagger organisation will need to lobby to get better quality access.[160]

Then there are questions about transparency and impartiality over the choice of trusted flagger organisations by Internet platforms. Do they choose based on commercial self-interest or the best interests of their users (assuming the two things can be detached)? And given that trusted flaggers enter into informal arrangements with Internet platforms over which the former typically have limited say, questions might be asked as to the independence and neutrality of trusted flaggers over the long term. Might their flagging decisions start to be influenced, consciously or unconsciously, by the responses or lack thereof they get from Internet platforms?

[159] European Commission's Code of Conduct on Countering Illegal Hate Speech Online, 2-3.
[160] Interview with a trusted flagger organisation, 1 October, 2019.

Another set of concerns has to do with balance and the appearance of balance in the sorts of trusted flagger organisations that Internet platforms work with. Potentially Internet platforms could be cautious about granting trusted flagger status to NGOs working specifically on behalf of particular communities or groups. Platforms might be concerned that such NGOs would err on the side of, or might be perceived as erring on the side of, over-flagging when it comes to content directed at particular communities or groups. If the corporate values and business mission of the Internet platform includes promoting free speech, balance and neutrality, for example, the platform might feel the need to either refrain from using any NGOs or instead seek to use as many NGOs as possible covering the full range of different vulnerable communities and groups in society. But achieving this sort of balance, and the appearance of balance, against a backdrop of the politics of identity may not be straightforward.

Moreover, if Internet platforms made more extensive use of NGOs as trusted flaggers, whilst at the same time recognising that certain NGOs might be more likely to flag suspected hate speech content that targets a particular vulnerable community or minority group, would there be a need for Internet platforms to seek or invite trusted flagger "second opinions" from other civil society organisations? For example, should Internet platforms seek second opinions from organisations that place an emphasis on promoting and securing human rights, including civil liberties organisations that highlight and promote the special importance of the human right to freedom of expression? On this model, when the Internet platform receives a flag from an NGO the platform would send the same content and other relevant contextual information to a civil liberties organisation, or trusted *un*flagger, for it to pass comment. In this way the Internet platform is receiving assessments from two different organisations looking at the same content with different goals, perspectives and prioritisations of values.

Then again, arguably the "second opinion" system would be redundant if the experience and expertise of the trusted flaggers is high. For example, if the internal policies and procedures of trusted flagger organisations protect free speech, then no second opinion would be required. One thing that trusted flaggers, such as national equality boards, currently do is consider cases in the light of international human rights standards including the right to freedom of expression. Moreover, decisions to flag content for removal are often collective and majority-based decisions: at least three members of staff look at the case and reach a majority opinion.[161]

Yet more challenges in maintaining good relations between Internet platforms and trusted flagger organisations emerge when the trusted flagger also operates as a monitoring body. Consider the European Commission's Code of Conduct and its allied monitoring system (or "common methodology") agreed on 5 October 2016 by several Internet platforms and the European Commission's sub-group on combating illegal online hate speech.[162] A key part of the monitoring system is the process by which the European Commission appoints and manages monitoring bodies to undertake periodic monitoring cycles of the success or diligence of Internet platforms with regards to fulfilling the demands of the Code. Typically these monitoring bodies are also existing trusted flagger organisations for the platforms.[163]

This dual role complicates the task of relationship management between Internet platforms and trusted flaggers in several ways. For one thing, there may be instances in which the sort of

[161] Interview with Unia, 1 October, 2019. Interview with Gitanos, 3 October 2019.

[162] Minutes of meeting of the European Commission's sub-group on countering illegal hate speech online concerning the monitoring process and methodology, 5 October, 2019.

[163] Interview with trusted flagger, 1 October, 2019. Interview with trusted flagger, 3 October, 2019. 1st consultative meeting, London, 17-18 October, 2019.

advice, training and material support that Internet platforms provide to trusted flagger organisations may cut against the independence, or the appearance of independence, of the work of the trusted flagger as a monitoring body. For example, some trusted flaggers report that during meetings and training sessions provided by Internet platforms, the latter will typically offer "advertising grants" to trusted flaggers that enable them, for example, to run their own campaigns on the platforms free of charge as a sort of "goodie bag" for participating in the meeting or training session.[164] These advertising grants may be extremely valuable to NGOs both because they depend on access to Internet platforms to advertise their campaigns and because they tend to have limited financial resources. This sort of close relationship may or may not hinder the independence of organisations working as trusted flaggers but it arguably does reduce the independence, or reduce the appearance of independence, of these organisations as monitoring bodies. Thus, it is recommended that monitoring bodies must publicly declare any "in kind" benefits received from Internet platforms.

For another thing, when Internet platforms already have a very close working relationship with trusted flaggers, this could make it easier for the platforms to gain information, directly or indirectly, about when the monitoring period is underway. For example, some platforms may have a sense from the conduct of a trusted flagger that it is not "business as usual": that the nature, frequency or extent of flags indicates that a monitoring cycle is underway or nearing its conclusion.[165] For more discussion of these and related issues, see section IV.B above. Thus, this study also recommends that reforms are made to the monitoring system to ensure that Internet platforms are not made aware—deliberately or structurally—of the period of monitoring, such as by extending the monitoring period to 12 months of the year.

(ii) Challenges in collaboration between Internet platforms and independent supervisory councils, steering committees or oversight boards

Internet platforms that establish relationships with independent supervisory councils, steering committees or oversight boards face several challenges in managing this relationship. One challenge for the Internet platform is to keep members of the council, committee or board regularly and fully informed as to developments happening at the platform. On the one hand, because the Internet platform needs to protect the privacy of its users and its users' personal data and needs to protect its own commercially sensitive information including its trade secrets and intellectual property—an Internet platform will have fiduciary responsibilities to its shareholders not to act in ways that clearly damage the company—it might be tempted to err on the side of caution and provide limited information to members of the council, committee or board. On the other hand, if members of the council, committee or board are not regularly updated and do not receive sufficient information, they will inevitably have less capacity to provide well-informed and up-to-date general recommendations [see section III.C]. Moreover, if they feel that they are being kept "out of the loop", either consciously or through poor management, there is a risk that members could quit or else "go public" with their misgivings. This might inflict reputational damage on the Internet platform (Matsakis 2019).

Internet platforms that refer particular cases (moderation decisions) to independent supervisory councils, steering committees or oversight boards also face unique challenges. The Internet platform will wish to refer cases based on matters of general importance to it, such as referring cases that are representative of broader categories of grey area or difficult cases it is struggling

[164] Interview with trusted flagger and monitoring body, 21 October, 2019.

[165] 1st consultative meeting, London, 17-18 October, 2019.

with. Then again, individual users will want their cases referred simply because they have a personal stake in the matter. If the Internet platform allows users to also refer cases, albeit only after any internal appeals process has been exhausted, then this has the merits of giving users a right of referral. Then again, since not all referred cases are actually looked at by the council, committee or board—after all, it will have limited capacity and its own case selection criteria—this means in effect limited or contingent access to a right of appeal for individual users. Given the fact that the council, committee or board will hear relatively few cases, it is especially important that the findings it does reach on given cases exert maximum policy influence over the Internet platform. If not, then the rejection of a user's referral will not be "compensated" by the fact that similar cases have been heard and did achieve policy influence or impact. The surest way to guarantee policy influence is for the decisions of the council, committee or board to be binding on the Internet platform.

Furthermore, if the decisions of the council, committee or board are non-binding, then it makes their policy influence or impact uncertain at best. Consistent failure to achieve policy influence or impact could make the council, committee or board seem at best redundant and at worst a cynical ploy or PR exercise designed to give the appearance of oversight. If the decisions are binding, however, then potentially it could make it harder for the Internet platform to pursue its mission, honour its corporate values, or execute its business model. It would be sacrificing or handing over some of its "creative control" to a third party that might not fully share its vision. There is an unavoidable trade-off to be made here.

Moreover, if the decisions are binding this could potentially change fundamentally the nature of the relationship between the Internet platform and its users. To expand on this point, users might not expect that third parties would be involved in making ultimate content decisions. There may be an expectation or even a bond of trust between the user and the Internet platform that it will not outsource content decision-making to a third party. In the words of one representative from a major online platform:

> When you sign up to use [an Internet platform] you are agreeing […] a terms of service contract with the [platform], so it's ultimately up to the user and the [platform] what each other does about the content on the site. […] A third party being involved from an advisory point of view is interesting and valuable but it's not part of [existing] terms of service, and nor would users expect to, and nor would users be even aware of the fact a third party is making decisions on their content outside of the terms of service they agreed to with the platform. […] Ultimately a platform […] would have an obligation to the users, arguably a legal obligation, to make the final call [on content].[166]

Then again, the Internet platform might be able to change its terms of service in a very clear way, to make it obvious that it will refer cases to an independent advisory council, steering committee or oversight board, and that it will treat the decisions as binding. This would undoubtedly make the relationship between the Internet platform and its users more complicated, more mediated, but it would not necessarily amount to a breach of trust, provided that users were made fully aware of the situation beforehand.

As discussed in section IV.D above, Facebook has responded to these challenges and trade-offs by adopting the following nuanced approach to oversight. First, Facebook maintains that it has a general responsibility to "own" its policies on content moderation based on the principle

[166] Interview with a major Internet platform, 12 July, 2019.

that because Facebook made the platform then it should be responsible for key policies concerning what users can and cannot post or share on the platform; and this general responsibility on the part of Facebook is also something that its users would fully expect when signing up to the platform. But second, Facebook makes an important qualification to its general responsibility based on the principle that whilst it should "own" its policies it is not required, and its users would not expect (and may not want), the platform to make literally "all the calls" concerning how the policies are applied to given cases.[167]

(iii) Challenges in collaboration between Internet platforms and the police and public prosecutors

Relationships between Internet platforms, on the one side, and the police and public prosecutors, on the other side, face several challenges. One main challenge lies in navigating the different sorts of notifications that the police and/or public prosecutors can give to Internet platforms about suspected unlawful or illegal hate speech content.

On the one hand, notifications made by the police and/or public prosecutors to Internet platforms could be non-binding or advisory only (administrative notifications). One potential benefit is that the police and/or public prosecutors are able to advise Internet platforms swiftly and without undertaking full investigations and seeking court orders. This may create opportunities for expedited interventions. These sorts of notifications also leave the Internet platforms (legal compliance teams, for example) with control over the final decision about whether to take down content based on the notifications but also their own assessments. Then again, because administrative notifications do not reflect full investigations and legal hearings, they are not based on the highest standards of due process. Moreover, they might give the impression of the police and/or public prosecutors exerting undue influence over Internet platforms.[168]

On the other hand, notifications made by police or public prosecutors to Internet platforms could be legally binding insofar as they are based on court rulings obtained by the police or public prosecutors (judicial notifications).[169] This may ensure higher levels of due process, and remove any uncertainty or ambiguity about whether Internet platforms have an obligation to remove the content upon notification because these would be judicial "notice and take down" orders. Then again, this could be a slower process and may impede swift interventions. It also removes control from Internet platforms, which, depending on one's perspective, may not be a good thing.[170] These and other strengths and weaknesses of a system of notifications sent by a special public prosecutor are outlined in section IV.F above. A system of notifications (and other sorts of interventions) by a special police unit are discussed in section IV.I.

[167] Comments by Facebook, 1st consultative meeting, London, 17-18 October, 2019.
[168] 1st consultative meeting, London, 17-18 October, 2019.
[169] Ibid.
[170] Ibid.

VI. ON THE MOTIVATIONS, GOALS, VALUES AND EXPECTATIONS OF GOVERNMENTS, INTERNET PLATFORMS, CIVIL SOCIETY ORGANISATIONS AND THE GENERAL PUBLIC CONCERNING THE GOVERNANCE OF ONLINE HATE SPEECH

This section of the study seeks to identify and clarify the motivations, goals, values and expectations of governmental agencies, Internet platforms, civil society organisations and the general public concerning the governance of online hate speech.

A. Governmental agencies

This subsection focuses on governmental agencies, including government departments or ministries, the police, public prosecutors, and Internet regulators. Reviews of existing studies and reports, along with interviews, questionnaires and consultative meetings conducted as part of this study suggest that when it comes to designing, managing and implementing regulatory governance tools for illegal hate speech online, governmental agencies, on average, rate as important the following motivations, goals, values and expectations, in no particular order:

- fulfilling any responsibilities relating to the governance of illegal hate speech online that follow from the particular statutory role or devolved duties of the governmental agency;
- a sense of responsibility to protect not only users of Internet platforms against exposure to illegal hate speech but also to protect society as a whole against the negative effects, direct and indirect, of such content;
- a commitment to the principle that what is illegal offline must also be illegal online;
- a sense of responsibility to promote and protect free speech under a human rights framework;
- the aim of building or improving public trust in, and satisfaction with, governance tools for tackling illegal hate speech online;
- a commitment to ensuring that redress mechanisms made available to victims of online hate speech are user-friendly and accessible;
- the view that governmental agencies should seek input from, and collaboration with, Internet platforms, civil society organisations, other stakeholders and the public in developing governance tools for tackling illegal hate speech online;
- a sense that governmental agencies cannot, and should not, tackle the problem of illegal hate speech online by themselves and that Internet platforms must take their equitable share of responsibility;
- to spend tax payers money responsibly in the governance of illegal hate speech online;
- to obtain full and accurate information from Internet platforms about individuals or groups posting illegal hate speech online;
- to obtain full and accurate information from Internet platforms about total amounts of illegal hate speech being posted or shared online;
- responding to policy goals set down by the ruling government, such as in response to manifesto commitments, to ideological motivations or to more sudden changes in the political climate or simple political expediencies (“serving political masters”);

- seeking to reflect how Internet platforms, civil society organisations and the public at large think and feel about the proper governance of illegal hate speech online and any evolving social norms around this issue ("reflecting the public mood");
- aiming to collaborate with governmental agencies in other countries and with intergovernmental organisations, such as in terms of sharing information and technological expertise, establishing best practice, learning lessons, dealing with jurisdictional issues, and so on;
- working to ensure that the staff within governmental agencies are well trained, properly supported and protected, especially staff who handle hate speech content on a regular basis.

Of course, different governmental agencies will differ in the specific roles they play in the story of governance of online hate speech and, therefore, are likely to differ in how they rate the importance of the aforementioned motivations, goals, values and expectations. Nevertheless, it is interesting to note that all of the governmental agencies that provided questionnaire responses to this study rated "A sense of responsibility to protect not only users of Internet platforms against exposure to illegal hate speech but also to protect society as a whole against the negative effects, direct and indirect, of such content" the most important factor when it comes to designing, managing and implementing regulatory governance tools for illegal hate speech online.

Two specific questionnaire responses are also worth highlighting for other reasons. One questionnaire response rated as highly important the fact that governmental agencies should seek input from, and collaboration with, Internet platforms, civil society organisations, other stakeholders and the public in developing governance tools for illegal hate speech online.

> To be a tool for cooperation and coordination between state authorities responsible for enforcing legislation against online hate crimes, and also those authorities that fight illegal hate speech online in areas other than criminal, as well such as coordination with CSO and internet platforms. To be a platform for collaboration, facilitation of dialogue, exchange of challenges and formulation of solutions among the actors involved in the fight against online speech. To create a common and homogeneous working procedure to ask for the removal hate speech online by the different actors of the agreement.[171]

A second questionnaire response touched on the role played by the media in framing public debate about online hate speech and influencing public opinion on how to deal with it, which in turn can impact how governmental agencies undertake the governance of online hate speech.

> It is important to understand the role of both levels: The political level has the task of giving directives, through its laws, and ensuring that the population is protected, while respecting fundamental rights; The police services must protect the population, on the basis of existing laws and must report offences to the competent judicial authorities. The police evaluation is based on the number of facts observed and reported. The political assessment uses these statistics to try to respond to society's expectations, but also to, if necessary, amend a law. Therefore, an important factor for the political level is the Society's opinion, which is regularly given in case of shocking or serious events, which will sometimes have repercussions on the police strategy. An important factor is therefore the press, which is the means of communication to the public. It is therefore important, in my opinion, that the press be included in the fight process. Indeed,

[171] Governmental agency questionnaire response 2, 16 November, 2019 [Anonymised].

depending on how an article is treated (not to mention censorship, but with a protectionist view), the public reaction will be different, and political and police actions could be better perceived.[172]

[172] Governmental agency questionnaire response 1, 12 November, 2019 [Anonymised].

B. Internet platforms

This subsection focuses on Internet platforms. Reviews of existing studies and reports, along with interviews, questionnaires and consultative meetings conducted as part of this study suggest that when it comes to designing, managing and implementing governance tools for online hate speech, Internet platforms, on average, rate as important the following motivations, goals, values and expectations, in no particular order:

- a sense of responsibility to protect users from (illegal and legal) hate speech;
- a sense of responsibility to promote and protect free speech under a human rights framework;
- the practical aim simply to make it clearer to users where the platform draws the line between permissible and impermissible content;
- to not only remove hate speech content but also where appropriate to use content management tools to reduce assess to hate speech content;
- a commitment to ensuring that redress mechanisms made available to victims of online hate speech are user-friendly and accessible;
- the view that Internet platforms should seek input from other stakeholders and the public about their community standards or content policies on hate speech;
- a sense that Internet platforms need not, and should not, seek to make every single decision on particular bits of content as being or not being hate speech;
- to spend whatever is the right or responsible amount of money on the governance of online hate speech, within budgetary constraints;
- to achieve legal compliance both as an end in itself and so as to avoid adverse legal judgments, fines and negative publicity;
- to achieve legal compliance but only where compatible with the platform's own corporate values;
- being sensitive to facts about the Internet platform's own organisational capacity, human resources, and technologies available to deal with potentially large volumes of reports, notifications or flags of hate speech content;
- being sensitive to facts about the technological challenges and limits of accurately identifying hate speech content using automated text extraction and machine learning tools or algorithms;
- seeking to reflect how governments, civil society organisations and the public at large think and feel about the proper governance of online hate speech and any evolving social norms around this issue ("reading the room");
- aiming to match users' expectations about proper governance of online hate speech either as an end in itself or to optimise the user experience;
- striving to undertake governance of online hate speech better than other platforms that are competitors for users' time and for advertising revenues;
- working collaboratively with other Internet platforms and with national governments, intergovernmental organisations and other stakeholders to achieve shared governance aims for tackling online hate speech;
- working to ensure that the Internet platform's staff are well trained, properly supported and protected, especially staff who handle hate speech content on a regular basis.

Several of the above motivations, goals, values and expectations were reflected in questionnaire responses. For example, one questionnaire response emphasised that the Internet platform does much more than remove hate speech content, it also takes steps to manage content by reducing access to it.

> We value keeping content accessible and want to be careful to only remove content from our services when it crosses the line. We complement removals with other approaches […]. In 2017, we introduced a tougher stance towards videos with supremacist content, including limiting recommendations and features like comments and the ability to share the video. This step dramatically reduced views to these videos (on average 80 percent).[173]

This questionnaire response also highlighted the challenges of using text extraction and machine learning tools or algorithms to identify hate speech content, underscoring that governance of such content is constrained by limits in technological advancement.

> We're investing in machine learning technology, but detection of hate speech remains a significant challenge for technology. We deploy machine learning to better detect potentially hateful content to send for human review, applying lessons from our enforcement against other types of content, like violent extremism. Using a combination of smart detection technology and human reviewers, our teams routinely remove hundreds of millions of comments each quarter. […]
>
> However, it is important to recognize that the technology is still early stages and certain ML challenges that apply to hateful content—like understanding context, history, spoken cues, or replicating human judgement—are still a long way off. This is particularly true for hate speech, where the slang and slurs deployed are constantly changing, where dog whistles are deployed, and where commentary can fall in the grey zone.
>
> Machine learning/AI experts talk of a classifier's "precision" and "recall," to examine questions like "Of the content retrieved, how much was relevant?" or "What did the classifier miss?" In hate speech, the classifiers can retrieve a lot of irrelevant material, and miss other relevant content. As one example, a classifier that searches [content] for the slur "cracker" will also pull up a lot of perfectly valid content on snacks or the ITV show of the same name.[174]

Finally, this questionnaire response also emphasised that the Internet platform takes seriously the need to both engage in consultation with third parties but also to regularly update its policies to reflect new phenomena in the area of online hate speech.

> We review our policies on an on going basis to make sure we are drawing the line in the right place: In 2018 alone, we made more than 30 policy updates. One of the most complex and constantly evolving areas we deal with is hate speech. We've been taking a close look at our approach towards hateful content in consultation with dozens of experts in subjects like violent extremism, supremacism, civil rights, and free speech.[175]

[173] Internet platform questionnaire response 1, 19 November, 2019 [Anonymised].

[174] Ibid.

[175] Ibid.

C. Trusted flaggers and monitoring bodies

This subsection focuses on trusted flaggers, that is, organisations that flag hate speech content to Internet platforms, and monitoring bodies, such as organisations that work with the European Commission to help monitor the success or progress of Internet platforms in meeting their obligations under the Code of Practice. Reviews of existing studies and reports, along with interviews, questionnaires and consultative meetings conducted as part of this study suggest that when it comes to designing, managing and implementing governance tools for online hate speech, trusted flaggers and monitoring bodies, on average, rate as important the following motivations, goals, values and expectations, in no particular order:

- fulfilling any responsibilities relating to the governance of illegal hate speech that follow from any statutory role or devolved duties of the organisation, if it has such a role or duties (e.g. equality board, monitoring body);
- giving more visibility to the problem of online hate speech targeted at any particular groups or communities the organisation was set up to advocate on behalf of;
- a sense of responsibility to protect users of Internet platforms from (illegal and legal) hate speech;
- a sense of responsibility to protect the wider society from the negative effects, direct and indirect, of online hate speech;
- a sense of responsibility to promote and protect free speech under a human rights framework;
- the goal of providing redress to victims of hate speech, including through reporting and appeals mechanisms provided by Internet platforms but also by building cases to go forward into administrative, civil and criminal proceedings (legal remedies);
- the objective of gathering information or data on online hate speech so as to improve governance of online hate speech (as well as to improve education and counter-narrative schemes);
- the aim of encouraging Internet platforms not merely to remove or reduce access to hate speech content but also to inform users in clear and visible ways about community standards or content policies on hate speech;
- working to influence policy making and to improve the design and implementation of governance tools operated by governmental agencies and intergovernmental organisations;
- a commitment to ensuring that redress mechanisms made available to victims of online hate speech are user-friendly and accessible;
- the aim of building or improving public trust in, and satisfaction with, governance tools for online hate speech;
- the view that civil society organisations and bodies participating in governance tools for online hate speech should also seek input from other stakeholders including civil liberties organisations and the public in carrying out the work they do;
- a sense that civil society organisations and bodies cannot, and should not, tackle the problem of online hate speech by themselves and that both Internet platforms and governmental agencies (e.g. government ministries, police, public prosecutors, regulators) must take their equitable share of responsibility;
- the aim of maintaining a degree of independence from both Internet platforms and governmental agencies, including a degree of financial and management independence, so as to ensure the credibility and integrity of the work of civil society organisations participating in the governance of online hate speech;

- the need to build trust and good working relationships with Internet platforms and with governmental agencies, so as to be able to participate properly and effectively in the governance of online hate speech;
- the need to build trust and good working relationships with other civil society organisations and NGOs including organisations that are and organisations that are not trusted flaggers or monitoring bodies;
- the need to build trust and win the confidence of victims of online hate speech so as to be able to participate properly and effectively in governance of online hate speech;
- the expectation that Internet platforms will take seriously any reports or flags about hate speech content that the organisation sends to them, and would treat them as credible and deserving urgent action;
- working to ensure that Internet platforms also take seriously any reports or flags about hate speech content made by ordinary users;
- undertaking governance in ways that also reflect how Internet platforms, governmental agencies, other civil society organisations and the public at large think and feel about the proper governance of online hate speech and any evolving social norms around this issue ("reflecting the public mood");
- working to ensure that the organisation's staff are well trained, properly supported and protected, especially staff who handle hate speech content on a regular basis.

Several of these goals and expectations among trusted flaggers and monitoring bodies are articulated in the following questionnaire response.

> Our priority is having online spaces that are inclusive and where the extent to which users are exposed to hate speech is limited.
> Our core needs are:
>
> - an easy-to-use interface for reporting hate speech
> - that triggers a transparent and fast decision-making process
> - the possibility to appeal and the appeal should be judged by humans who understand local context and subtext
> - user who submitted the report should be notified of the action taken by the internet intermediary and the reasoning behind its decision.
>
> CSO-s should be given opportunity to scrutinize practices of intermediaries and contribute to regular evaluation of their operation.[176]

Other goals and expectations figure in the following questionnaire responses.

> Goals:
> - get as much hate content removed as quickly as possible
> - updating the methodology of monitoring exercises
> - improve the platforms' responsiveness to reports by "ordinary" users, i.e. close the gap between them and trusted flaggers
> - get involved in policy making, the design of governance tools and management structures
> - play a key role in training moderators and other key staff[177]

[176] Trusted flaggers and monitoring bodies questionnaire response 1, 22 November, 2019 [Anonymised].

[177] Trusted flaggers and monitoring bodies questionnaire response 2, 4 December, 2019 [Anonymised].

More involvement of IT platforms to inform users in a clear and visible way about community rules, hate speech and respect.[178]

[178] Trusted flaggers and monitoring bodies questionnaire response 3, 4 December, 2019 [Anonymised].

D. The general public

This subsection focuses on the general public. In order to fill a gap in existing public opinion surveys on attitudes to governance of online hate speech, this study commissioned YouGov to undertake an opinion survey in several countries. However, one technical challenge was to ensure that respondents had a good understanding of what was being asked of them. Asking the public abstract questions about what values or goals should inform governance of online hate speech may not have elicited useful or illuminating results. It would have been difficult to know what the relevant values or goals actually meant to people taking the poll. Therefore, this study commissioned a more focused poll of public attitudes towards some of the governance tools discussed in section IV. The poll was done in the UK, France and Germany.

(i) UK public opinion poll results (YouGov, sample of 1,633 UK adults, November 2019)

> When it comes to government agencies (e.g. ministry of justice, police, public prosecutors, regulators etc.) tackling illegal hate speech that is posted or shared on Internet platforms, how important, if at all, would you rate each of the following measures?

- Identifying and prosecuting individuals who post or share illegal hate speech on Internet platforms

<1> Very important	49 percent
<2> Fairly important	30 percent
<3> Not very important	7 percent
<4> Not important at all	3 percent
<5> Don't know	12 percent

- Requiring Internet platforms to remove illegal hate speech within 24 hours of it being reported to them and imposing fines on Internet platforms that fail to comply with that requirement

<1> Very important	58 percent
<2> Fairly important	22 percent
<3> Not very important	6 percent
<4> Not important at all	2 percent
<5> Don't know	12 percent

- Offering exemptions from the aforementioned fines in the case of Internet platforms that can show they are devoting reasonable resources to removing illegal hate speech

<1> Very important	11 percent
<2> Fairly important	37 percent
<3> Not very important	17 percent
<4> Not important at all	9 percent
<5> Don't know	26 percent

- Offering reductions in the aforementioned fines in the case of Internet platforms that provide full disclosure about the amounts of illegal hate speech content on their platforms

<1> Very important	12 percent
<2> Fairly important	32 percent
<3> Not very important	19 percent
<4> Not important at all	10 percent
<5> Don't know	27 percent

(ii) France public opinion poll results (YouGov, sample of 1008 French adults, November 2019)

When it comes to government agencies (e.g. ministry of justice, police, public prosecutors, regulators etc.) tackling illegal hate speech that is posted or shared on Internet platforms, how important, if at all, would you rate each of the following measures?

- Identifying and prosecuting individuals who post or share illegal hate speech on Internet platforms

<1> Very important	48 percent
<2> Fairly important	23 percent
<3> Not very important	7 percent
<4> Not important at all	4 percent
<5> Don't know	17 percent

- Requiring Internet platforms to remove illegal hate speech within 24 hours of it being reported to them and imposing fines on Internet platforms that fail to comply with that requirement

<1> Very important	46 percent
<2> Fairly important	24 percent
<3> Not very important	7 percent
<4> Not important at all	5 percent
<5> Don't know	19 percent

- Offering exemptions from the aforementioned fines in the case of Internet platforms that can show they are devoting reasonable resources to removing illegal hate speech

<1> Very important	14 percent
<2> Fairly important	30 percent
<3> Not very important	17 percent
<4> Not important at all	9 percent
<5> Don't know	31 percent

- Offering reductions in the aforementioned fines in the case of Internet platforms that provide full disclosure about the amounts of illegal hate speech content on their platforms

<1> Very important	15 percent
<2> Fairly important	30 percent
<3> Not very important	16 percent
<4> Not important at all	19 percent
<5> Don't know	31 percent

(iii) Germany public opinion poll results (YouGov, sample of 2,055 German adults, November 2019)

When it comes to government agencies (e.g. ministry of justice, police, public prosecutors, regulators etc.) tackling illegal hate speech that is posted or shared on Internet platforms, how important, if at all, would you rate each of the following measures?

- Identifying and prosecuting individuals who post or share illegal hate speech on Internet platforms

<1> Very important	58 percent
<2> Fairly important	22 percent
<3> Not very important	8 percent
<4> Not important at all	4 percent
<5> Don’t know	9 percent

- Requiring Internet platforms to remove illegal hate speech within 24 hours of it being reported to them and imposing fines on Internet platforms that fail to comply with that requirement

<1> Very important	58 percent
<2> Fairly important	21 percent
<3> Not very important	8 percent
<4> Not important at all	5 percent
<5> Don’t know	9 percent

- Offering exemptions from the aforementioned fines in the case of Internet platforms that can show they are devoting reasonable resources to removing illegal hate speech

<1> Very important	26 percent
<2> Fairly important	34 percent
<3> Not very important	16 percent
<4> Not important at all	8 percent
<5> Don’t know	15 percent

- Offering reductions in the aforementioned fines in the case of Internet platforms that provide full disclosure about the amounts of illegal hate speech content on their platforms

<1> Very important	22 percent
<2> Fairly important	31 percent
<3> Not very important	18 percent
<4> Not important at all	13 percent
<5> Don’t know	17 percent

(iv) Analysis

The public opinion survey method has the advantage of getting access to the opinions of reasonably large and representative samples of the general public in the relevant countries. But the well-known trade-off is that the questions tend to be fewer, more focused and often closed, in comparison to the semi-structured interviews and questionnaires used by this study for governmental agencies, Internet platforms and civil society organisations.

Based on these limitations, the study commissioned YouGov to ask members of the public about two specific regulatory governance tools, namely, identifying and prosecuting individuals who post or share illegal hate speech on Internet platforms (Regulatory-G, Regulatory-H and Regulatory-I) and requiring Internet platforms to remove illegal hate speech within 24 hours of it being reported to them and imposing fines on Internet platforms that fail to comply with that requirement (Regulatory-C). In the case of regulatory fines, the study also commissioned YouGov to ask members of the public about how they rated the importance of giving exemptions from fines to Internet platforms if they devoted reasonable resources to removing illegal hate speech, and of giving reductions in fines to Internet platforms in return for them providing full disclosure about the amounts of illegal hate speech content on their platforms. Because of the focus on these specific regulatory governance tools, the study opted to target countries in which either these governance tools were already in place or where these governance tools were being actively considered or developed by legislators and policymakers (e.g., the UK, France, Germany). The finite financial resources of the study did not permit commissioning similar polls in every member state of the Council of Europe, for instance.

Overall the survey results suggest that the general public in the UK, France and Germany agree that it is as important for regulators to impose legal responsibilities on Internet platforms to remove illegal hate speech content as it is for authorities to prosecute the creators of such content. This suggests that the public do not accept the notion that Internet platforms should be immune from legal responsibilities and regulatory sanctions, and that Internet platforms also need to be held accountable for the hate speech that is posted or shared on their platforms.

In addition, however, the survey results also suggest that the public are willing to cut some slack to Internet platforms, such as by giving them exemptions from regulatory fines if they devote reasonable resources to removing illegal hate speech or by giving them reductions in fines in return for them providing full disclosure about the amounts of illegal hate speech on their platforms. This suggests a balanced, non-punitive or non-absolutist view on how to regulate Internet platforms. The ethos seems to be that if necessary Internet platforms should be fined for a pattern of failure to remove illegal hate speech but that leeway should be given to Internet platforms that fail but which are nonetheless making a good faith effort.

What is unclear is what lies behind this balanced view. Clearly this is one of the drawbacks of public opinion surveys as opposed to semi-structured interviews or focus groups, for instance. Could it be that the public are aware of the practical and technical challenges of regulating Internet platforms and take a pragmatic view that if Internet platforms “play ball” then they should be cut some slack simply because this creates trust and might achieve greater compliance in the long run? Or instead could it be a more principled position that what really matters when it comes to regulating how Internet platforms tackle illegal hate speech content is less the brute outcomes and more that platforms make a good faith effort? Maybe this reflects deeper views about fair play or reasonableness, namely, that it is wrong to punish agents for failures to achieve certain desired outcomes so long as they are trying to do the right thing?

In terms of differences in the results across countries, it is interesting that German respondents placed a bit more importance on both prosecuting creators of illegal hate speech content (very important, 58 percent) and on requiring Internet platforms to remove illegal hate speech content (very important, 58 percent) than respondents in France (48 and 46 percent respectively). Could this have something to do with especially strong perceptions of the historical legacy of, and connections between, anti-Semitic hate speech and the Holocaust in Germany (see also Brown and Sinclair 2019: 81-85)? Or to do with the constitutional culture in Germany where protection of dignity is a driving force within the legal system? Or instead has it to do with the simple fact the NetzDG Act is already in play within Germany and so the public are more familiar with this regulatory intervention, and so less suspicious and apprehensive of it?

The UK results were interesting because, unlike in Germany and France, the respondents placed a bit less importance on prosecuting creators of illegal hate speech content (very important, 49 percent) than on requiring Internet platforms to remove illegal hate speech (very important, 58 percent). This suggests that the British public think it especially important to impose legal responsibilities and regulatory sanctions on Internet platforms in relation to the removal of illegal hate speech. Once again, it is unclear whether this reflects a pragmatic view that Internet platforms are simply in a better position to crack down on this content than, say, the police and public prosecutors, or instead a principled view that somehow Internet platforms are especially to blame because they provide a vehicle for this content in the first place.

It is perhaps also worth mentioning that existing surveys on public attitudes to hate speech regulations suggest solid support in the UK for such regulations in general. For example, a 2017 YouGov UK poll on public attitudes to the stirring up hatred offences found that 63 percent of respondents thought it should be against the law to use threatening words or behaviour with intent stir up hatred against other people on grounds of their race, religion and sexual orientation, while 65 percent supported such laws in the case of a newly proposed characteristic, mental or physical disabilities (Brown 2017f). The same survey was repeated in 2019 and support had increased to 66 and 71 percent respectively, potentially reflecting heightened visibility of, and sensitivity to, hate speech across the board during and after the 2016 EU referendum (Brown 2019b).

VII. A VICTIM-SENSITIVE APPROACH TO GIVING REDRESS TO TARGETS OF ONLINE HATE SPEECH

One of the first academic articles to use the term “hate speech” was an article by the renowned critical race theorist Mari Matsuda titled ‘Public Response to Racist Speech: Considering the Victim’s Story’ (Matsuda 1989b). In the article she sought to put the experiences and needs of victims of hate speech front and centre in a wider constitutional debate within America about whether criminal and civil remedies for victims of hate speech can be justified given the First Amendment. Keeping in mind this intellectual tradition it seems natural to similarly ask whether some governance tools for online hate speech are more “victim-sensitive” than others and what place a victim-sensitive approach should have within a wider human rights framework.

Now it might seem axiomatic that governance tools for online hate speech are by nature “victim-centred”. However, the reality is rather more complicated. A governance tool might be fixated on “dealing with” or “confronting” the conduct of creators or authors of online hate speech and on establishing the legal responsibilities of the Internet platforms on which this content is posted or shared, but at the same time properly understanding the experiences and needs of individuals and groups who are targeted by online hate speech might become an afterthought. As outlined in section I.C above, arguably one of the core functions of the governance of online hate speech is to provide means of redress (e.g. reporting content, appealing non-removal decisions, legal remedies of different kinds) for individuals or groups who are targeted or adversely affected by online hate speech. This is a good starting point for thinking about what a victim-sensitive approach might look like but it is certainly not the end of the story. After all, there are many ways giving redress to victims of online hate speech. So the more difficult question is about which ways are more victim-sensitive than others.

A. The need for a victim-sensitive approach

It is perhaps indicative that among the key texts at the European level on tackling online hate speech discussed in section I.B above,[179] while all of them speak at length about the need for Internet platforms to remove unlawful hate speech content because of the harm it can do to those targeted, none of them contain any explicit reference to “victim-centred approaches” to tackling online hate speech. It is not enough to recognise the fact that online hate speech represents “a serious threat to […] the dignity of victims”.[180] There must be active steps to embed within governance regimes ways of properly understanding and meeting the particular experiences and needs of victims of online hate speech. Indeed, one document on hate speech that does show greater awareness of the victim-sensitivity agenda is ECRI’s General Policy Recommendation No. 15. For example, its Preamble contains the following paragraph:

[179] Specifically, Directive 2000/31/EC of the European Parliament and of the Council on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market (“Directive on Electronic Commerce”) of 8 June 2000, ECRI’s General Policy Recommendation No. 15 on Combating Hate Speech of 8 December, the European Commission’s Code of Conduct on Countering Illegal Hate Speech Online of May 2016, the European Commission’s Communication on Tackling Illegal Content Online of 28 September 2017, Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Role and Responsibilities of Internet Intermediaries of March 2018, and the Revised EU Audiovisual Media Services (“AVMS”) Directive of 14 November 2018.

[180] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online.

> Recognising that the use of hate speech appears to be increasing, especially through electronic forms of communication which magnify its impact, but that its exact extent remains unclear because of the lack of systematic reporting and collection of data on its occurrence and that this needs to be remedied, *particularly through the provision of appropriate support for those targeted or affected by it*; (emphasis added)[181]

Focusing on the victims of online hate speech is partly about being able to accurately identify the people who are subject to online hate speech and about recognising the real harms wrought on them. But it is also, and this point is crucial, about reflecting on how the design, operation or implementation of given governance tools itself impacts victims and about seeing whether these tools can rise to challenges of their lived experiences and can meet their needs.

A wider point worth emphasising here is that within the general category of targeted or adversely affected individuals or groups (or "victims") there might exist individuals or groups who are particularly vulnerable, either because they are subjected to greater amounts of, or more severe forms of, online hate speech or because for some reason they face greater obstacles in using redress mechanisms. To give one example, some victims might find it especially difficult to report content to Internet platforms if they face learning difficulties in using complex online reporting systems.[182]

By analogy, under a victim-centred approach to policing, criminal justice and human rights, the victim's wishes, safety and well-being is given special priority in all matters of both policy and procedure that concern the victim. This means a variety of different things in different contexts, but in the area of criminal justice and public prosecutions for rape, for example, it can mean things like keeping the accuser informed at key points in the criminal law process, avoiding retraumatisation of the accuser during the police investigation and trial, and protecting the accuser from further intimidation or distress at the hands of the person who is accused.

Interestingly, the European Commission has embraced the need for a victim-centred approach in the area of human trafficking. Whilst not strictly analogous, this shows that victim-centred approaches in the area of human rights are not unprecedented. In its report on combating human trafficking, for example, the European Commission writes:

> A victim-centred approach is at the heart of the EU anti-trafficking legislation and policy. This means establishing appropriate mechanisms for the early identification of victims and provision of assistance and support, in cooperation with the relevant support organisations. (European Commission 2016a: 11)

Because people who have been trafficked are often heavily controlled and exist "under the radar", part of the victim-centred approach here is about proactively identifying them and then giving them the right kind of support that is sensitive to their particular needs.

To give another example, whilst the European Commission's Strategic Engagement for Gender Equality 2016-2019 document does not explicitly refer to a "victim-centred" or "victim-sensitive" approach to tackling gender violence, many of its recommendations clearly speak to this sort of agenda. For example, one of its five priority areas is "combating gender-based violence and *protecting and supporting victims*" (European Commission 2016c: 9, emphasis added).

[181] CRI(2016)15, Strasbourg, 8 December 2015. Available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 [last accessed 7 October, 2019].

[182] 2nd consultative meeting, Berlin, 26 November, 2019.

Perhaps unsurprisingly, some academics have advocated a victim-centred approach to combating hate crime. Kees et al. (2016), for example, have put together a set of principles and practical guidelines designed for those supporting victims of hate crime based on the fundamental requirement "to establish a visible and explicitly victim-centred approach to the provision of effective support services for those who suffer hate violence in Europe" (Kees et al. 2016: 7). As best practice they argue that support services should recognise that victims of hate crime may need urgent support; need to be believed; require time to articulate and communicate their needs and may need an advocate to help them do this; should have their needs recognised, acknowledged, and addressed; should receive assistance with overcoming the consequences of what has happened to them; and assistance that is specific to their identity (Kees et al. 2016: 23-26).

All of this raises the following set of questions. Is there a case for adopting a victim-centred approach, or something like it, to the governance of online hate speech? Is the notion of a "victim" (more) problematic when applied to online hate speech than to hate crime? What would a victim-centred approach even look like? Among the more than 20 different model types of governance tools discussed in sections II, III and IV above, which, if any, can be plausibly considered more or less victim-sensitive and why? The remainder of this section is devoted to answering these questions; or at least, making a start at answering these questions.

However, before beginning it is important when seeking to motivate the need for a victim-sensitive approach to the governance of online hate speech to devote attention to some personal experiences or narratives ("your stories") from victims gathered as part of this study. The following statement, for example, focuses on potential barriers to victims pursuing cases through the legal system. The story underscores the importance of not merely reforming those regulatory governance tools that involve the police and public prosecutors, so as to make them more victim-sensitive, but also of ensuring that victims also have access to other, simpler ways of seeking redress, either as a first line of defence or as a backup, such as via content reporting mechanism and appeals processes provided by Internet platforms themselves.

> I was targeted online and reported various content, comments and photos of me to [a major Internet platform in Cyprus]. I reported it to Aequitas, which is a trusted flagger of [the Internet platform]. [I felt] [r]eassured. Less anxious. Feeling more safe. […] I had some contact with the police but I did not push the issue further after my initial contact. […] I did not press forward with my complaints so I received no redress. The only redress I received was deletion of some of the posts and photos by the online platform. To the extent that it was quick and effective deletion of posts I am satisfied (from technical perspective). To the extent of feeling that 'justice was done' probably not at all. […] Given the circumstances and climate of that time (incident happened in summer of 2017), I would say that everyone handled the issue very well, because the climate of Cyprus at the time was that it made more sense to hush the issue up rather than magnify it and make it a cause celebre. This was the advice given by everyone who I spoke to. Advice was of the form: for this to be followed up legally then the case needs to go up to the attorney general and the police is not really well equipped and it will be a long arduous process etc. That was probably good advice at the time. (And this highlights the need for institutions to be more efficient or to have better way of quickly dealing with such cases.) However, I think that something has significantly changed in Cyprus in the last 2 years. […] [Recent] cases have now formed a 'social precedent': if my case had happened today, I would be much more confident to openly say what happened and

combat the voices abusing me, and campaign for proper redress. Basically the point is that my case could also have generated dialogue, made an impact, changed views etc, so with hindsight, it may have been good advice at that time, before Cypriot society because more progressive, but in long term I think it was bad advice to 'just do nothing and it will blow over' and I would hope that in future NGOs, civil society and victims are empowered more to speak up about their stories of abuse. This can happen only when NGOs, police, and institutions are ready to back victims up. So really, my case was a missed opportunity to really change perceptions in Cyprus. It's the responsibility of society, institutions, police and particularly NGOs to keep pushing for this social precedent to be solidified.[183]

[183] Personal narrative statement 1, 8 November, 2019 [Anonymised].

B. Some important qualifications about a victim-sensitive approach

It is important to add the qualification that a “victim-centred” or “victim-sensitive” approach to tackling online hate speech is not itself a governance tool. Rather, it is an overarching framework for selecting, planning, shaping, reforming and implementing governance tools. A “victim-sensitive” approach could, and should, also be used a means of assessing the success or progress of governance tools; that is, one of the many measures or indicators that might be utilised in the evaluation of governance tools to be discussed in section VIII below.

It is also fair to point out that victim-centred approaches may not be uncontroversial. For one thing, governance tools that incentivise over-removal of hate speech content could be plausibly said to create, perhaps as an unintended consequence, “victimised content”.[184] In other words, the public may think it important for governmental agencies and Internet platforms to take seriously not only the needs and experiences of persons that have been targeted or adversely affected by illegal hate speech posted or shared online but also the issue of content that has been removed even though it is not unlawful and even though it might have free speech value. Insofar as it makes sense to speak of content that has been removed due to bogus, illegitimate or vague content policies or simply due to flawed or mistaken moderation decisions as victimised content, for example, then arguably such removal involves victims.

Indeed, at the same time there may be scenarios in which a person is doubly victimised, not only by being targeted by online hate speech but also by having their own content removed due to malicious reporting and flawed moderation decisions, for example.[185]

Therefore, whatever the term “victim-centred” means it should not exclude the idea that not only targets of online hate speech but also victims of over-removal of content deserve to be protected.

Furthermore, some civil liberties organisations, whilst accepting the importance of protecting persons against clearly or manifestly illegal hate speech content posted or shared on Internet platforms, might find the very phrase “victim-centred approach” objectionable if the phrase somehow implied that the experiences and needs of targets of online hate speech are automatically given priority or preferential treatment over authors or creators of online content even in circumstances where the content is actually lawful or legal content but just so happens to fall foul of Internet platforms’ potentially egregious or over-zealous content moderation.

For these reasons it might be better to speak in terms of a “victim-sensitive approach”: this means simply that the wishes, safety and well-being of persons that are targeted or adversely affected by hate speech posted or shared online are given their proper importance and due consideration in all matters of moderation, oversight and regulation that directly concern them. This is perfectly compatible with simultaneously holding that both speakers’ interests and Internet platforms’ interests are also given their proper importance and due consideration in matters of Internet governance that directly concern them. For more on the interests, expectations, goals and values of different stakeholders, see section VI above.

Another objection that could potentially be levelled even at the notion of a “victim-sensitive approach”, is the claim that when it comes to online hate speech this is “victimless” in the strict sense. In his seminal text on the criminal law Herbert Packer described victimless crimes as “offences that do not result in anyone’s feeling that he has been injured so as to impel him to

[184] Interview with representatives of civil society organisations and research centres, 29 July, 2019.
[185] Ibid.

bring the offense to the attention of the authorities" (Packer 1968: 151). Such crimes typically occur in circumstances where another person has consented (e.g. adultery) or else where the effects of the offences are so diffuse and indirect that no particular individual feels (or could reasonably feel) that he or she has been injured or victimised (e.g. certain public order offences). Arguably some hate speech-based public order offences might be put into this category. Take the offence of stirring up religious hatred in England and Wales.[186] The offence requires proof of intent to stir up hatred but not proof that that any particular individuals or groups actually had hatred stirred up against them and not proof that any particular individuals or groups suffered specific harms such as assault or damage to property as a result.

However, the current objection over-generalises from the case of incitement to hatred to all kinds of online hate speech. In order to see this point, consider the fact that in 1985 the General Assembly of the United Nations adopted the following definition of "victim" in its Declaration of Basic Principles of Justice for Victims of Crime and Abuse of Power: "'Victims' means persons who, individually or collectively, have suffered harm, including physical or mental injury, emotional suffering, economic loss or substantial impairment of their fundamental rights, through acts or omissions that are in violation of criminal laws operative within Member States, including those laws proscribing criminal abuse of power."[187] There is good reason to think that incitement to hatred along with many other forms of hate speech—including abusive slurs, negative stereotypes, derogatory remarks, dehumanising images, mocking imitations, claims about lower fundamental moral worth or dignity, false rumours, group libels, threatening words or behaviour, denials or glorifications of atrocities, incitement to hatred, discrimination or violence based on protected characteristics [see section I.E]—does create victims in the aforementioned sense, not only collectively but also individually, whether offline or online.

In fact, there is a growing body of evidence, not only from legal cases,[188] from anecdotal reports and testimonies, from the extensive country monitoring reports[189] and the General Policy Recommendations of ECRI,[190] but also from social scientific studies undertaken by academics, some of which are based on semi-structured interviews and/or questionnaires, that points to many kinds of harmful consequences suffered by victims of hate speech, offline and online. The measures of these harmful consequences include but are not limited to mental and emotional health and well-being, standing in society, safety and security, discrimination, oppression, subordination, and even access to public discourse and democratic decision-making (see Matsuda 1989b; Lawrence 1990, 1992; Nielsen 2002; Tsesis 2002, 2017; Delgado and Stefancic 2004; Parekh 2005–6, 2012; Langton 2012; Langton et al. 2012; Brown 2015: chs. 3 and 7, 2017d; Gelber and McNamara 2016; Gelber 2017, 2019).

Furthermore, victim identifiability may be even greater in the online context compared to the offline world. There are, of course, many forms of hate speech and some governance tools cover a range of forms of online hate speech including lawful hate speech, whereas others concentrate on unlawful hate speech. But if we just think for a moment about the use of abusive slurs,

[186] Part 3A of the Public Order Act 1986.

[187] General Assembly resolution 40/34 of 29 November 1985.

[188] Consider the ECtHR cases Jersild v. Denmark, Aksu v. Turkey, and Vejdeland and Others v. Sweden. Or consider the UK cases R. v. El-Faisal No. T20027343 (Central Criminal Court, 7 March, 2003), R. v. Ali, Javed, and Ahmed No. T20110109 (Derby Crown Court, 10 February, 2012), and R. v. Sheppard and Whittle II [2010] EWCA Crim 65. In all of these cases the judges made reference to the harmful effects of illegal hate speech, not only to social cohesion but also to people's sense of security and to their feeling equal members of society.

[189] Available at: https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/country-monitoring [last accessed 6 December, 2019].

[190] Consider, for example, ECRI, General Policy Recommendation No. 15.

derogatory remarks, and false rumours that might be lawful, it is clear that vast amounts of such content targets specific individuals. In addition, Internet websites, services and platforms often facilitate "piling on", whereby large numbers of other persons can join in the targeting of a specific individual by adding their own hate messages (Saccardo 2016).

In terms of unlawful online hate speech and hate crimes, we know that a great deal of online misogyny takes the form of harassment of identified women (see Citron 2014). More generally, consider hate crimes that are often committed online and that have a strong communicative dimension meaning they are typically performed with or through the use of hate speech.[191] In England and Wales examples include public order offences like fear or provocation of violence[192] or intentional harassment, alarm or distress,[193] offences of harassment,[194] and offences of malicious communications,[195] when committed online and aggravated by hostility (motivated by hostility or demonstrates hostility) towards the victim's possession or perceived possession of a protected characteristic. All of these offences make reference to actions being done to "a person" or "another person", and therefore have identifiable victims.

Nevertheless, the objector might insist that talk of "victims" could have other connotations that are potentially counter-productive. For example, talk of "victims" might inadvertently convey an idea of people being "different", "weak", "powerless" or "disliked", and that such connotations mirror and reinforce negative stereotypes that are characteristic of hate speech itself.

This and similar dangers of victim-centred approaches to hate speech laws are often-cited in the academic literature on the topic. Identifying individuals as members of "victim groups" allegedly risks simplifying, essentialising and fixing their identities in undesirable ways (Butler 1997); risks creating even greater antagonism against the target groups because they are given victim status (Neller 2018); risks incentivising a sort of victimhood competition, that is, "an unseemly battle of more-victim-than thou" (Heinze 2006: 568).

As far as hard evidence is concerned, however, these risks are often overblown (Brown and Sinclair 2019: 275-282). The facts suggest that in practice laws which seek to protect persons who are targeted by hate speech are no more or less likely to have the above-mentioned consequences than other sorts of public policies. Moreover, even where hate speech laws could have some of these consequences, this is offset by the power of well-designed, properly drafted and sensitively enforced hate speech laws to, at the same time, also respect people's multiple identities, reduce inter-identity antagonism and discourage competition.

A defining feature of hate speech is its tendency to misrecognise its targets: to ascribe an identity to people often against their wishes; to tarnish that identity and bring it into contempt; and to oversimplify the identity of the individual (see Parekh 2005-6; Brown 2015: 166-174). Whatever else it does, a governance tool that by design or effect gives redress to targets of online hate speech must not perpetuate the misrecognition associated with hate speech itself.

[191] These are sometimes referred to as "expression-oriented" hate crimes (Brown 2015: 35-38).

[192] s. 4A of the Public Order Act 1986.

[193] s. 4A of the Public Order Act 1986.

[194] ss. 1 and 2 of the Protection from Harassment Act 1997.

[195] s. 1 of the Malicious Communications Act 1988 and s. 127 of the Communications Act 2003.

C. Guidelines for achieving a victim-sensitive approach

Turning next to questions of practical application, what would a victim-sensitive approach to the governance of online hate speech actually look like? In general or abstract terms it means focusing on the experiences and needs of those who have been subjected to online hate speech and treating them with compassion, respect, courtesy and above all else as individuals who have their own identities, who are capable of suffering but who are also able to give or withhold consent. But what do these ideas mean at a more concrete level?

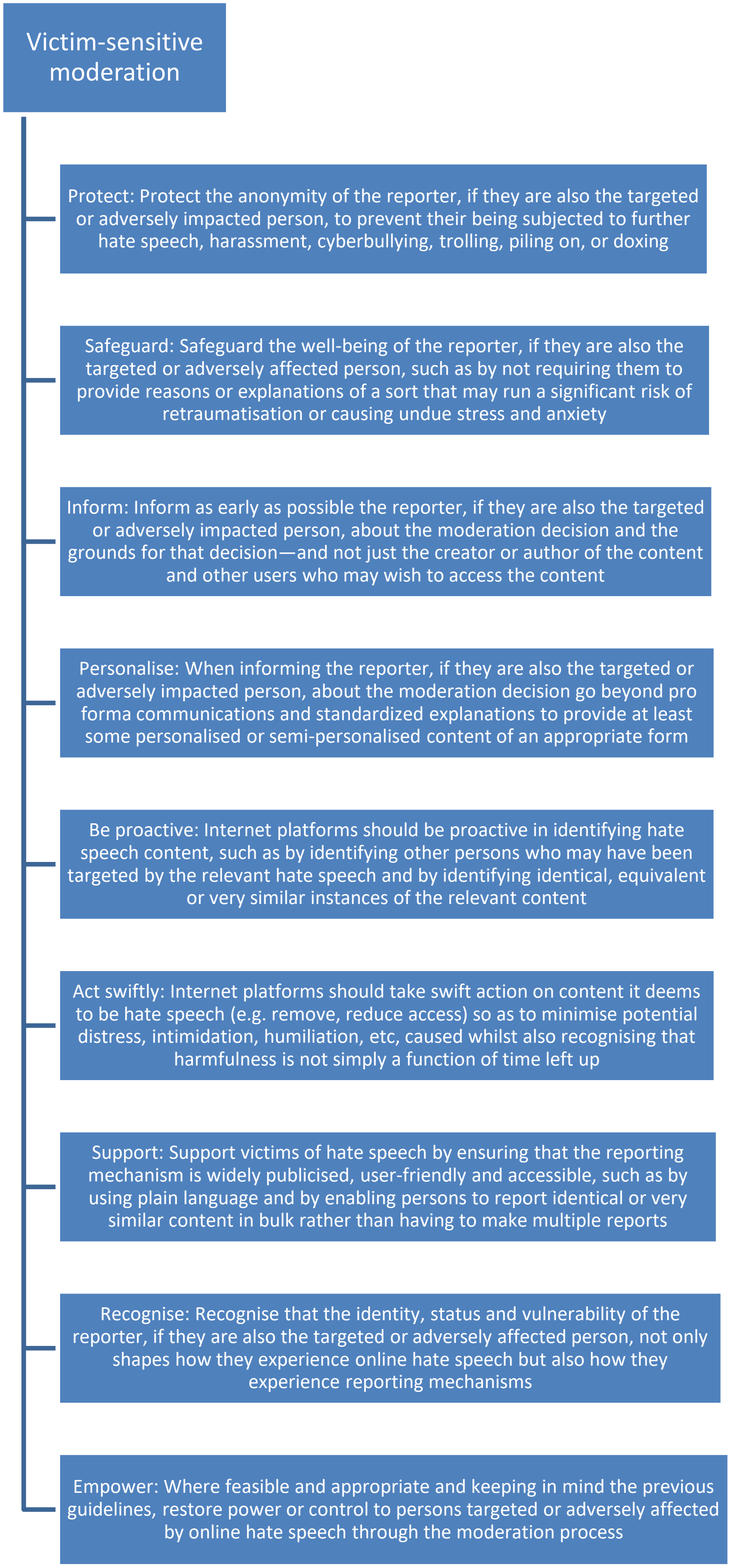
It goes without saying that a victim-sensitive approach to the governance of online hate speech should be applicable at each of the three levels of governance outlined in section I.C above: the moderation level, the oversight level and the regulatory level. In what follows, therefore, guidelines are proposed for what a victim-sensitive approach could mean at each of the three levels of governance. They are illustrative in nature and are not intended to be exhaustive or exclusive. They represent the beginning not the end of a conversation.

Importantly, some of the guidelines relate to features of existing governance tools for online hate speech already found in some countries. They are in that sense a matter of recognising the good in what is already being done, and categorising it as victim-sensitivity. In other instances the guidelines set down ideals or goals that are not currently being followed and should be. Some require just a little more effort to comply with, some prescribe more radical reform. Some guidelines flag up potential dangers by pointing to weaknesses in governance tools not as they are now but as they could be in the future if they are taken to an extreme.

(i) Victim-sensitivity at the moderation level

Starting at the moderation level, Diagram 6 below sets out some possible guidelines that Internet platforms should seek to follow as part of the process of moderation of online hate speech content.

Diagram 6. Guidelines for victim-sensitivity at the moderation level



Victim-sensitive moderation

Protect: Protect the anonymity of the reporter, if they are also the targeted or adversely impacted person, to prevent their being subjected to further hate speech, harassment, cyberbullying, trolling, piling on, or doxing

Safeguard: Safeguard the well-being of the reporter, if they are also the targeted or adversely affected person, such as by not requiring them to provide reasons or explanations of a sort that may run a significant risk of retraumatisation or causing undue stress and anxiety

Inform: Inform as early as possible the reporter, if they are also the targeted or adversely impacted person, about the moderation decision and the grounds for that decision—and not just the creator or author of the content and other users who may wish to access the content

Personalise: When informing the reporter, if they are also the targeted or adversely impacted person, about the moderation decision go beyond pro forma communications and standardized explanations to provide at least some personalised or semi-personalised content of an appropriate form

Be proactive: Internet platforms should be proactive in identifying hate speech content, such as by identifying other persons who may have been targeted by the relevant hate speech and by identifying identical, equivalent or very similar instances of the relevant content

Act swiftly: Internet platforms should take swift action on content it deems to be hate speech (e.g. remove, reduce access) so as to minimise potential distress, intimidation, humiliation, etc, caused whilst also recognising that harmfulness is not simply a function of time left up

Support: Support victims of hate speech by ensuring that the reporting mechanism is widely publicised, user-friendly and accessible, such as by using plain language and by enabling persons to report identical or very similar content in bulk rather than having to make multiple reports

Recognise: Recognise that the identity, status and vulnerability of the reporter, if they are also the targeted or adversely affected person, not only shapes how they experience online hate speech but also how they experience reporting mechanisms

Empower: Where feasible and appropriate and keeping in mind the previous guidelines, restore power or control to persons targeted or adversely affected by online hate speech through the moderation process

The "Inform" guideline further clarifies (or goes beyond?) an important guideline set out in the Committee of Ministers to Member States of the Council of Europe Recommendation CM/Rec(2018)2 on the Roles and Responsibilities of Internet Intermediaries. It states:

> Content restrictions should provide for notice of such restriction being given to the content producer/issuer as early as possible […]. Information should also be made available to users seeking access to the content, in accordance with applicable data protection laws.[196]
>
> Notice should also be given to the user and other affected parties […].[197]

The "Inform" guideline makes clear that notification must be sent to the reporter of the content if that person is also the person targeted or adversely affected by the hate speech in question.

The "Personalise" guideline speaks to the need to treat persons as individuals and not mere statistics or faceless inputs into a moderation machine. When informing the reporter, if they are also the targeted or adversely impacted person, about the moderation decision, Internet platforms should go beyond pro forma communications and standardized explanations to provide messages that contain at least some personalised or semi-personalised content of a suitable form, where feasible and appropriate. By including some personalised or semi-personalised content within communications, this may help the Internet platform to create an end to end user experience, which is especially fitting for users that have had "bad experiences" of hateful content whilst using the platform through no fault of their own.

Personalised content might include: where appropriate recognition of the facts relating to the content that the victim has reported (e.g. acknowledging exactly what the content was such as by quoting the content); where appropriate affirmation of the validity of the victim's perception of the content as hate speech (e.g. using and repeating the descriptions used by the reporter such as "we as a company agree with you that the content was hate speech and that this sort of content is not allowed on the platform"); and where appropriate showing empathy for the feelings reported by the victim of hate speech (e.g. restating or recognising the emotional impact such as "we as a company understand how hurtful it was to see this content").

Semi-personalised content might include: where appropriate providing the victim with links to information on any administrative, civil or criminal remedies that might be available in the country or local area to persons targeted or adversely impacted by online hate speech content; where appropriate providing the victim with links to information on any victim-support organisations that might be able to assist persons targeted or adversely impacted by online hate speech content in the country or local area; and where appropriate providing the victim links to information on any psychological or emotional counselling services available in the country or local area in circumstances where the victim has made reference to ways in which exposure to the content has adversely affected them psychologically or emotionally.[198]

Of course, because of the large volume of reports, notifications and flags that Internet platforms may be dealing with at any given time, it would not be reasonable to expect them to always and in every case provide personalised or semi-personalised messages when notifying or informing

[196] Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, s. 1.3.7.

[197] Ibid., s. 2.3.3.

[198] Interview with Google, 6 August, 2019.

reporters of decisions about particular bits of content. The resource burden of such a requirement may be disproportionate to the policy goal (victim-sensitive moderation).

However, it does seem reasonable to expect Internet platforms to provide at least some personalised or semi-personalised content of a suitable form in some circumstances.[199] Therefore, the qualification "where appropriate" is important. It could mean that the duty to provide communications or messages that include at least some personalised or semi-personalised content would only apply where (1) the person reporting the content has self-identified as having been targeted and/or adversely affected by the hate speech content in question, and (2) the Internet platform has found in favour of the reporter of the content and is communicating its positive decision to remove the content because the platform deems the content to breach its own community standard or content code on hate speech.

To clarify, this requirement might be applicable where reporters self-identify as having been targeted by hate speech in the sense that they have been named in the hate speech, have had hate speech addressed to them through direct messaging, or have had hate speech directed at them in some other clear and obvious manner. It might also be applicable in the case of persons that self-identify as being members of groups or communities of people (e.g. Muslims, transsexuals) that have been targeted by hate speech naming the group or community. But the requirement would not apply in the case of persons who report hate speech content but who are witnesses or third parties but not targeted or negatively impacted themselves.

Most importantly, it should be up to the reporter whether or not to self-declare as having been personally targeted or adversely impacted by online hate speech. In the circumstances that they have not declared themselves to have been personally targeted or adversely impacted by the hate speech in question, then pro forma communications and standardized explanations about content removal decisions should be used. Platforms should not presume that persons reporting hate speech are "victims", not least because doing so may risk retraumatising persons that have for some reason made the decision not to self-declare as victims.

Likewise, although Internet platforms should be proactive in identifying and removing identical, equivalent or very similar hate speech content to that which has been reported—more on this guideline below—this does not mean that platforms should automatically bring this larger body of content to the attention of the reporter, if the reporter is also the targeted or adversely affected person. To do so may also risk causing further emotional distress to victims. A better approach would be to give the reporter an "opt-in" function whereby they can click on a button or icon to see content that is identical, equivalent or very similar to the hate speech they are in the process of reporting.[200] This provides both user-friendliness but also control.

The "Personalise" guideline echoes a general recommendation about the need to offer counselling services to targets of hate speech that has already been made by ECRI in its General Policy Recommendation No. 15 (para. 106), and a similar recommendation in the case of victims of hate crime (Kees et al. 2016).

Importantly, the "Personalise" guideline also potentially goes further than Art. 3 of the so-called Avia Bill in France, both in the kind and range of information that platforms should provide to "victims". For example, the Bill makes reference to informing users using generally accessible

[199] Interview with Unia, 1 October, 2019.

[200] Comments by Twitter, 1st consultative meeting, London, 17-18 October, 2019.

public information, whereas the “Personalise” guideline recommends informing victims via direct correspondence containing personalised or semi-personalised content.

It is also worth repeating that adopting a victim-sensitive approach to moderation does not mean giving absolute priority to the experiences and needs of targeted or adversely affected individuals and groups over all other users, or paying no regard whatsoever as to the experiences and needs of other users. Clearly no Internet platform would stay in business very long if people were made to feel like “second-class” users of services and platforms.

The “Be proactive” and “Support” guidelines converge around the idea of Internet platforms taking some of the burden off the reporter, if they are also the targeted or adversely affected person. For one thing, reporting mechanisms should use plain language and should be made available in multiple languages and formats so that, for example, poor literacy or language proficiency, learning difficulties or visual impairment, are not barriers to using reporting mechanisms. This is obviously vitally important in cases where, for instance, a person with learning difficulties has been targeted by online hate speech and wishes to report this.

In addition, a targeted individual should not be expected to fill out large numbers of online report forms for identical, equivalent or very similar hate speech content. Forms should enable bulk reporting, such as by allowing victims to highlight and report several bits of content at the same time.[201] (Indeed, some Internet platforms (e.g. YouTube) already provide their “trusted flaggers” with a tool that enables them to bulk flag.[202] This tool should be extended to users.)

It might be objected at this point that allowing bulk reporting could also be misused by persons wishing to maliciously report other users.[203] (Indeed, as mentioned in section VII.B above, malicious reporting might also be used by persons wishing to silence individuals or groups who are already victims of hate speech, thus creating a sort of double victimisation.)

This is always a risk, of course. But this risk is not a decisive reason to withhold bulk reporting functionality. For one thing, bulk reporting functionality could be used in conjunction with other measures designed to counteract malicious reporting (e.g. Internet platforms training staff to identify malicious reporting, Internet platforms banning malicious reporters, governments introducing criminal offences of malicious reporting).

Furthermore, the status quo situation in which users must report each bit of content separately already favours malicious reporters over genuine victims. This is because genuine victims of online hate speech may be unlikely to report all the content targeting them because doing so may be distressing or because they are seeking to “forget and move on”. Malicious reporters, by contrast, tend to be highly motivated and may find it easy to spend time reporting many bits of content separately. In that sense bulk reporting functionality levels the playing field.

Similarly, the “Be proactive” guideline enshrines the idea that individuals or groups who are targeted or adversely affected by online hate speech can, and should, reasonably expect that once they have reported hate speech content and it has been removed, then the Internet platform

[201] Note, a similar recommendation appears in the UK government (Department for Digital, Culture, Media & Sport) *Online Harms White Paper*, April 2019 in relation to online harassment and online abuse of public figures (paras. 7.24 and 7.37). Given the phenomenon of “piling on” in relation to online hate speech, for example, it is arguable that this recommendation is highly relevant for this other form of harmful speech.

[202] 1st consultative meeting, London, 17-18 October, 2019.

[203] Ibid.

should be proactive in identifying and removing “identical”, “equivalent” (e.g. language translations), or “very similar” content, including content that the reporter may have missed.

The fact that victims of online hate speech are likely to miss identical, equivalent or very similar content that has been posted elsewhere on an Internet platform is not only down to the huge volume of content but also due to the existence of “closed” or “private” groups available on many platforms. Nevertheless, content that contravenes the Internet platform’s community standard or content policy on hate speech in an area of public content also contravenes the same standard or policy when it appears in a closed or private group on the platform, and both can have adverse consequences for victims, both directly and indirectly.

Finally, the “Empower” guideline speaks to potentially one of the key benefits of redress mechanisms, namely, that redress enables the victim not merely to bring a grievance and seek resolution but also empowers the victim, putting them back in control. Whereas victims of online hate speech are typically made to feel out of control or powerless, a redress mechanism can potentially restore a degree of control, such as, for example, by enabling the victim to report hate speech content to the relevant Internet platform.

Of course, empowering victims of online hate speech cannot be a purely formal or superficial act. If Internet platforms simply gave victims the opportunity to report online hate speech but never actually removed content, even when it violated the platforms’ community standards or content codes, then the empowerment would be in name only.

What about when an Internet platform has a range of enforcement options at its disposal but in some cases chooses content management over content removal because of who the speaker is? Interestingly, Twitter has recently adopted the practice of using warning labels (a type of content management) for tweets that violate its content policy on hate speech when the tweets originate from the accounts of political figures (e.g. public officials, political representatives, candidates for political office).[204] The rationale given by Twitter is the public interest in protecting the free speech of political figures. However, it is far from obvious that such public interest rationales can, or should, trump the needs and experiences of victims. After all, the special position and power of political figures and the amplifying effect this can have on the harmfulness of hate speech is well document (Brown and Sinclair 2019: ch. 7). By prioritising the free speech of political figures, such Internet platforms risk downgrading the needs and experiences of victims of online hate speech especially in terms of empowerment.

More boldly, Internet platforms could go one step further by giving reporters of hate speech, if they are also the targeted or adversely affected individuals or groups, the power to decide what happens once content has been deemed to be hate speech by the Internet platform. The option range might include removal but also certain forms of content management such as reduced distribution, warning labels, preventing users sharing or adding comments, making the content ineligible for advertising, sponsorship, promotion or recommendation. Other options might include that the Internet platform posts a counter-narrative alongside the content, requires the user who created the post to take an online hatred awareness training course (education), or requires the user to read a victim impact statement (restorative justice). Giving the victim the power to select one or more of these other options, whether as alternatives to content removal or in addition to content removal, could potentially give them back a sense of control. Of course, this sort of radical empowerment is not uncontroversial. What if a victim decides that the content

[204] See Twitter Rules, About Public-Interest Exceptions on Twitter. Available at: https://help.twitter.com/en/rules-and-policies/public-interest [last accessed 15 April, 2020].

should not be removed, only for the content to hurt other victims? What if one victim opts for removal and another victim opts for non-removal of the same content? This sort of empowerment might also cut against other guidelines such as "Act swiftly".[205] Nevertheless, even if most victims would opt for removal, giving them some choice over what happens remains arguably a stronger form of empowerment than merely reporting.[206]

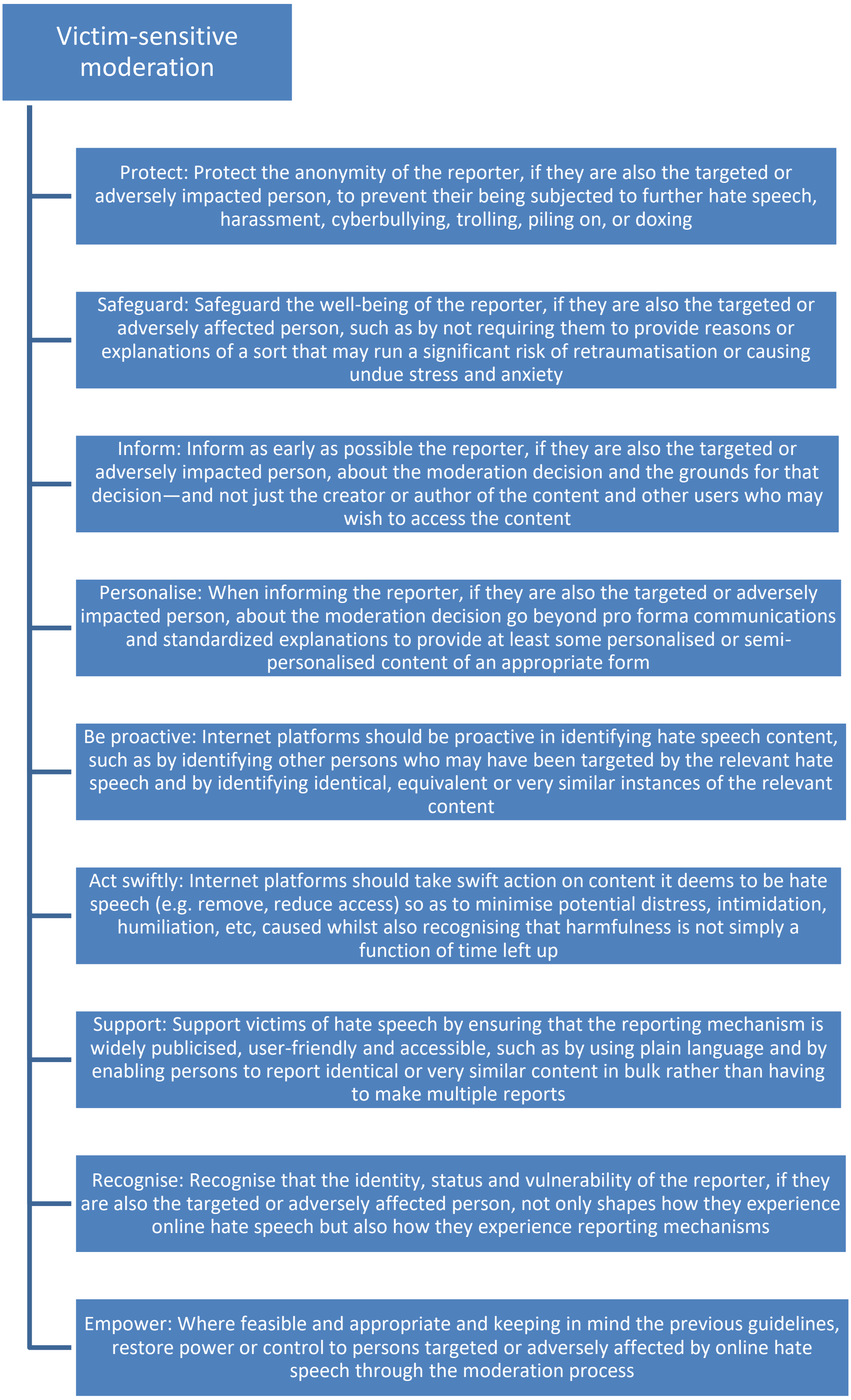
(ii) Victim-sensitivity at the oversight level

Turning to the oversight of moderation, Diagram 7 below sets out some possible guidelines that organisations responsible for the oversight of moderation relating to online hate speech (e.g., Internet platforms, independent supervisory councils, steering committees or oversight boards, fully independent dispute resolution services) should seek to follow.

[205] 2nd consultative meeting, Berlin, 26 November, 2019.

[206] Ibid.

Diagram 7. Guidelines for victim-sensitivity at the oversight level



Victim-sensitive moderation

Protect: Protect the anonymity of the reporter, if they are also the targeted or adversely impacted person, to prevent their being subjected to further hate speech, harassment, cyberbullying, trolling, piling on, or doxing

Safeguard: Safeguard the well-being of the reporter, if they are also the targeted or adversely affected person, such as by not requiring them to provide reasons or explanations of a sort that may run a significant risk of retraumatisation or causing undue stress and anxiety

Inform: Inform as early as possible the reporter, if they are also the targeted or adversely impacted person, about the moderation decision and the grounds for that decision—and not just the creator or author of the content and other users who may wish to access the content

Personalise: When informing the reporter, if they are also the targeted or adversely impacted person, about the moderation decision go beyond pro forma communications and standardized explanations to provide at least some personalised or semi-personalised content of an appropriate form

Be proactive: Internet platforms should be proactive in identifying hate speech content, such as by identifying other persons who may have been targeted by the relevant hate speech and by identifying identical, equivalent or very similar instances of the relevant content

Act swiftly: Internet platforms should take swift action on content it deems to be hate speech (e.g. remove, reduce access) so as to minimise potential distress, intimidation, humiliation, etc, caused whilst also recognising that harmfulness is not simply a function of time left up

Support: Support victims of hate speech by ensuring that the reporting mechanism is widely publicised, user-friendly and accessible, such as by using plain language and by enabling persons to report identical or very similar content in bulk rather than having to make multiple reports

Recognise: Recognise that the identity, status and vulnerability of the reporter, if they are also the targeted or adversely affected person, not only shapes how they experience online hate speech but also how they experience reporting mechanisms

Empower: Where feasible and appropriate and keeping in mind the previous guidelines, restore power or control to persons targeted or adversely affected by online hate speech through the moderation process

The "Improve" guideline can be followed in many different ways. One way is through appeals processes, either internal or external, that have the effect of making the reporter, if they are also the targeted or adversely impacted person, feel as though they have "had their day in court" or that in some sense justice has been done. Psychologically for some people this can be an important part of the process of healing or recovery.

Another way to put the "Improve" guideline into action is by giving victims access to some form of restorative justice, such as providing them with access to a dispute resolution procedure or mediation process, as per Oversight-E discussion in section III.E above. The goal would be for victims to have an opportunity to share their experience of what happened, to discuss how they were "harmed" by the hate speech content or hate crime, and to seek consensus through conciliation over what should happen next. This might also include an opportunity to speak with the author of the content in a safe, secure and respectful environment, online or offline. Whereas Internet platforms facilitate and arguably even encourage "forms of hate speech that are spontaneous in the sense of being instant responses, gut reactions, unconsidered judgments, off-the-cuff remarks, unfiltered commentary, and first thoughts" (Brown 2018: 304), dispute resolution procedures or mediation processes give the parties a chance to reflect more slowly and carefully about what they have said and how it impacted other people.

The "Inform" guideline at the oversight (and regulatory) level, adds an additional requirement to provide information on both the grounds and reasoning behind the oversight decision. So, for example, if a case is referred to an independent supervisory council, steering committee or oversight board (e.g. Facebook's Oversight Board), the oversight board should not merely explain which community standard or content policy was relevant to the bit of content at issue, as should happen when Internet platforms inform users of moderation decisions, but should also provide an account of the reasoning behind the oversight decision. This might include reference to relevantly similar or "precedential" oversight decisions, to expert interpretations of the language at issue, to wider social contexts or social values at play, and so on.

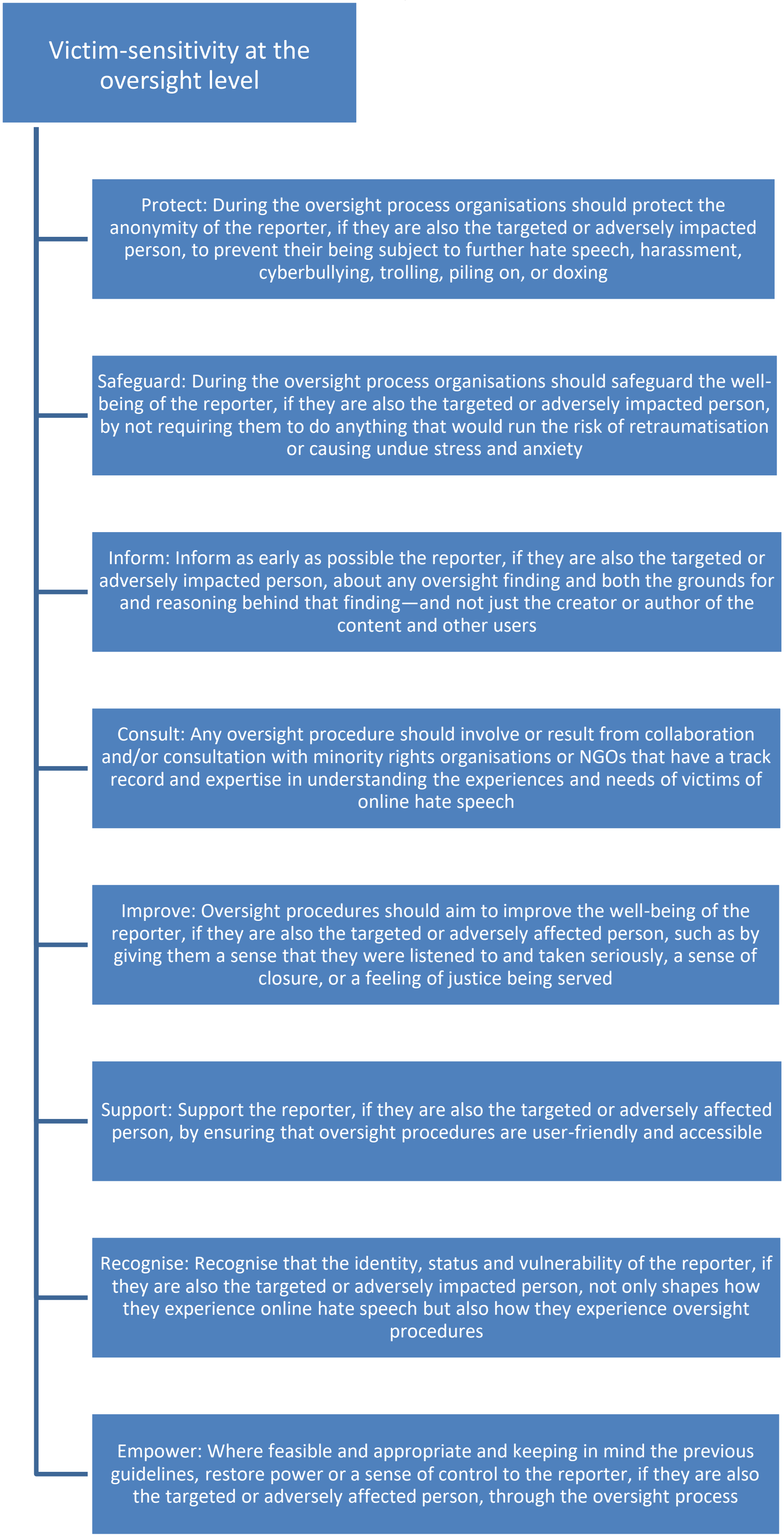
The "Recognise" guideline might be illustrated with the example of cases being referred to an independent supervisory council, steering committee or oversight board, as per Oversight-D discussed in section III.D above. Under a hybrid system of case selection, the independent supervisory council, steering committee or oversight board publishes a list of cases it is considering hearing and then invites from users, the public at large and also civil society organisations representing victims statements or "amicus briefs" setting out the merits of hearing certain cases. This could be a way of the oversight board coming to hear about the particular experiences and needs of victims of online hate speech.

Finally, the "Empower" guideline speaks to potentially one of the key benefits of redress mechanisms, namely, putting the victim back in control. Whereas victims of online hate speech are typically made to feel out of control or powerless, a redress mechanism can potentially restore a degree of control, such as, for example, by enabling the victim to appeal a moderation decision not to remove content or to refer a case to an oversight board.

(iii) Victim-sensitivity at the regulatory level

Turning finally to the regulatory level, Diagram 8 sets out some possible guidelines that stakeholders directly or indirectly involved in regulatory governance of online hate speech (e.g. Internet platforms, trusted flagger organisations, governmental agencies) should seek to follow.

Diagram 8. Guidelines for victim-sensitivity at the regulatory level



Victim-sensitivity at the oversight level

Protect: During the oversight process organisations should protect the anonymity of the reporter, if they are also the targeted or adversely impacted person, to prevent their being subject to further hate speech, harassment, cyberbullying, trolling, piling on, or doxing

Safeguard: During the oversight process organisations should safeguard the well-being of the reporter, if they are also the targeted or adversely impacted person, by not requiring them to do anything that would run the risk of retraumatisation or causing undue stress and anxiety

Inform: Inform as early as possible the reporter, if they are also the targeted or adversely impacted person, about any oversight finding and both the grounds for and reasoning behind that finding—and not just the creator or author of the content and other users

Consult: Any oversight procedure should involve or result from collaboration and/or consultation with minority rights organisations or NGOs that have a track record and expertise in understanding the experiences and needs of victims of online hate speech

Improve: Oversight procedures should aim to improve the well-being of the reporter, if they are also the targeted or adversely affected person, such as by giving them a sense that they were listened to and taken seriously, a sense of closure, or a feeling of justice being served

Support: Support the reporter, if they are also the targeted or adversely affected person, by ensuring that oversight procedures are user-friendly and accessible

Recognise: Recognise that the identity, status and vulnerability of the reporter, if they are also the targeted or adversely impacted person, not only shapes how they experience online hate speech but also how they experience oversight procedures

Empower: Where feasible and appropriate and keeping in mind the previous guidelines, restore power or a sense of control to the reporter, if they are also the targeted or adversely affected person, through the oversight process

The “Inform” guideline echoes an important guideline set out in s. 3(2)(5) of Germany’s NetzDG Act, namely, that “[t]he provider of a social network shall maintain an effective and transparent procedure for handling complaints about unlawful content”, and the procedure shall ensure that the provider of the social network “immediately notifies the person submitting the complaint and the user about any decision, while also providing them with reasons for its decision”.

The “Safeguard” guideline has various implications, including that the reporter of unlawful hate speech content, if they are also the targeted or adversely impacted person, should not be required to provide excessive personal information, complicated legalistic reasons for making the report or lengthy or otherwise burdensome victim impact statements if doing so would run a significant risk of retraumatisation them, that is, of causing undue stress and anxiety.

However, by way of clarification, this does *not* mean that the reporter or notifier cannot be asked to provide basic information. On the contrary. Take, for example, the following requirement present in the Avia Bill in France:

> The twenty-four hour period mentioned in the first paragraph of this article runs from the receipt by the operator of a notification including the following elements:
> « 1° If the notifier is a natural person: his surname, first name, e-mail address; if the notifier is a legal person: its corporate form, its corporate name, its e-mail address; if the notifier is an administrative authority: its name and e-mail address. These conditions are deemed to be satisfied if the notifier is a registered user of the online public communication service referred to in the same first paragraph, is connected at the time of notification and the operator has collected the necessary information, to his identification;
> « 2° The category to which the contentious content may be attached, the description of the content, the reasons for which it must be withdrawn, made inaccessible or dereferenced and, where applicable, the e-mail address(es) to which this content is made accessible.

This requirement has been commended by civil liberties organisations. In the words of Article 19, “the latest version of the Bill has restored a degree of procedural fairness by, among other things, requiring individuals or companies who notify content to set out the facts and reasons for notifying such content” (Article 19 2019). Insofar as providing information about one’s identity, the category to which the contentious content may be attached, the description of the content, and the reasons for which it must be withdrawn does not pose a significant risk of retraumatising the reporter or notifier, then safeguarding issues do not come into play.

Nevertheless, there may come a point at which, in the name of procedural fairness, placing additional and excessive demands on the reporter or notifier, if they are also the targeted or adversely impacted person, may pose a significant risk of damaging their emotional or psychological well-being, and under a victim-sensitive approach this should be avoided.

Another implication of the “Safeguard” guideline might be that governments should avoid legislation that creates unnecessary and detrimental friction for persons reporting content if they are also the targeted or adversely impacted persons, even if not intended to do so. For example, Regulatory-G discussed in section IV.G above includes as one possible variant the introduction of criminal offences relating to the conduct of individuals who maliciously report or flag content as being manifestly unlawful hate speech (i.e. reporting as manifestly unlawful whilst knowing it is not or failing to take due care to check). Whilst the legislative intent of this governance tool seems legitimate, aiming as it does to protect free speech by combating malicious reporting and flagging of hate speech, there are dangers. If the law is badly framed or overbroad (draconian), or imposes overly severe punishments or is misapplied by the courts (disproportionate), there is a risk that it could discourage honest reporting and flagging of hate speech. Genuine victims might refrain from flagging or reporting for fear of being prosecuted, thus potentially exacerbating an existing problem of under-reporting of online hate speech and blocking access to regulatory remedies.

In addition, the public might believe that Internet platforms should provide notices to *all* persons who report online hate speech about civil or criminal sanctions associated with malicious reporting, if such sanctions exist. This requirement is present in Art. 2.II. of the so-called Avia Bill in France, for example. In theory it might have the effect of not merely deterring the conduct in question but also providing balance in the treatment of reporters and notifiers ("fair warning"), and even a form of protection of people who maliciously report hate speech some of whom may be young, vulnerable or mentally ill.[207]

To illustrate the "Support" guideline, consider the variant of Regulatory-E outlined in section IV.E and Table 32 above, in which an Internet regulator takes on the role of a complaints body, hearing complaints, submissions or reports about non-compliance made against Internet platforms by individual users or by "recognised organisations", such as civil liberties organisations, minority rights organisations, equality boards, NGOs or other stakeholders. In the spirit of supporting victims of online hate speech, the Internet regulator could become a "one-stop-shop" complaints service, meaning that (1) victims could launch group or "class action" complaints, and (2) victims could launch simultaneous complaints against multiple Internet platforms, thereby reducing the burden of the regulatory process on individual victims.

In terms of the "Recognise" guideline, this might suggest that if the regulatory tool involves a regulator taking on the role of a complaints body, as with the variant of Regulatory-E set out in section IV.E and Table 32, then it would be better not exclude users, and therefore potential victims of online hate speech, from acting as complainants. To do so denies victims a means of redress ("their day in court" so to speak) and cuts off one potentially important avenue for alerting the regulator to systematic breaches of a duty of care around online harms based on the particular experiences and needs of victims of online hate speech.

Another implication of the "Recognise" guideline is more general but no less important. Understanding how victims might experience online hate speech in terms of the emotional or psychological toll it takes on them *qua* an attack on fundamental elements of their identity (e.g. stress, anxiety, fear, shame, humiliation, low self-respect, existential angst, sense of exclusion), points towards the need for regulatory tools to show parity of concern between physical and emotional harm (see Citron 2014; Brown 2015; Gelber and McNamara 2016).

Finally, the "Empower" guideline speaks to potentially one of the key benefits of redress mechanisms, namely, putting the victim back in control. Whereas victims of online hate speech are typically made to feel out of control or powerless, a redress mechanism can potentially restore a degree of control, such as, for example, by enabling the victim to report content to the police, public prosecutor or regulator and to play an active role in the case as it progresses through the

[207] Note, this practice may be a corollary of the practice of informing users whose content has been removed as unlawful hate speech about the potential legal consequences of posting or sharing unlawful hate speech online.

relevant legal or administrative processes including by giving evidence or testifying, where appropriate based on consent and with necessary legal and psychological support.

VIII. MEASURES OR INDICATORS AGAINST WHICH THE SUCCESS OR PROGRESS OF DIFFERENT INTERNET GOVERNANCE TOOLS FOR ONLINE HATE SPEECH CAN BE ASSESSED

The purpose of this study is not merely to map innovations in governance tools for online hate speech but also to provide ideas that can shape policy in this area across Europe. A key element of the second purpose is to set out possible measures or indicators against which the success or progress of different tools can be assessed going forward.

A. List of indicators

Below is a list of 30 measures or indicators that are worthy of consideration for use in assessing the degree of success or progress of the many different governance tools for online hate speech outlined above. The composition of the list is intended to reflect evidence, insights and arguments discussed above, and more generally information gathered from interviews and consultative meetings with participants in the study (including governmental agencies, Internet platforms, civil society organisations and users) as well as the existing and extensive body of reports, recommendations and studies on the governance of online hate speech.

- **Protection**: The extent to which the governance tool enhances the level of protection given to individuals or groups who are targeted or adversely affected by online hate speech, such as by reducing the chances of exposure to online hate speech or by reducing the harmfulness of any such exposure (e.g. degree of distress, traumatisation, intimidation, humiliation, threat, etc. caused).

- **Redress**: Whether or not the governance tool provides a means or mechanism specifically for individuals or groups who are targeted or adversely affected by online hate speech to report content, appeal decisions, assert grievances, lodge complaints, seek administrative, civil or criminal remedies, or in some other way claim or pursue resolution or rectification. (Note, the right of appeal against content removal for authors or creators of content is listed under the due process indicator below.)

- **Victim-sensitivity**: In addition to protection and redress, the extent to which the governance tool and its implementation responds to the experiences and needs of individuals or groups who are targeted or adversely affected by online hate speech.

- **Human rights**: The degree to which the governance tool promotes, secures and respects human rights based on local, regional and international human right standards, including but not limited to the human right to freedom of expression.[208]

- **Equity**: The extent to which the governance tool encapsulates or realises an equitable sharing of the practical burden and legal responsibility for tackling online hate speech

[208] See also Guidelines 1.1.2, 1.1.3, 1.1.4, 1.5.2, 2.1, 2.1.1, 2.1.2, 2.1.3, 2.1.4, 2.1.5, 2.2.2, 2.3.4, 2.3.5, and 2.4.6 of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Appendix: Guidelines for States on actions to be taken vis-à-vis internet intermediaries with due regard to their roles and responsibilities, 7 March 2018. Available at: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14 [last accessed 4 October, 2019].

content between governmental agencies, Internet platforms, civil society organisations, and users.[209]

- **Transparency**: The extent to which the relevant agency, organisation, body or company administering the governance tool provides adequate information to the public or to a suitable public body about its working practices and decisions.[210]

- **Independence**: The extent to which the relevant agency, organisation, body or company administering the governance tool is not subject to undue influence through "capture" by vested interests of various kinds (e.g. industry interests, elite interests).

- **Due process**: Insofar as the governance tool involves decisions on requests, claims, rights or liabilities of parties within proceedings or adjudications, the extent to which the tool encapsulates or achieves due process, including respecting the right of parties to submit evidence and argumentation, giving equal consideration to the evidence and argument offered by all parties, giving a right of appeal to parties, etc.[211]

- **Validity**: Insofar as the governance tool involves the weighing up of evidence and argumentation, the degree to which the relevant proceedings or adjudications are made on the basis of the best available (or reasonably sequestered) evidence and argumentation and in accordance with appropriate standards on evidence and argumentation, and are not based on discrimination, prejudice or animus.[212]

- **Accuracy and context**: Insofar as the governance tool involves a human or automated procedure, process or mechanism for applying a given definition, norm or law on hate speech (e.g. a community standard, a legal compliance rule, a hate speech law) to particular bits of content or to large bodies of content, the extent to which the procedure, process or mechanism has sound design and methodology and achieves reasonable levels of precision and recall. Accuracy is likely to depend in part on the ability of the human or automated mechanism to decipher semantics (and not just syntax), slang, the linguistic context of the speech, and the wider social and political context, etc.

[209] See also pp. 6-7 of Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online.

[210] See also p. 14 of Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online. And see Guidelines 1.1.4, 1.2, 1.2.3, 1.3.5, 1.5.2, 2.1.3, 2.2.2, 2.2.3, 2.2.4, 2.3.2, 2.4.6, 2.5.2, and 2.5.3 of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Appendix. And see the Santa Clara Principles on Transparency and Accountability in Content Moderation, May 2018. Available at: https://santaclaraprinciples.org/ [last accessed 14 November, 2019]. And see pp. 11-12 of Article 19's document, *Self-regulation and 'Hate Speech' on Social Media Platforms*, 2018. Available at: https://www.article19.org/wp-content/uploads/2018/03/Self-regulation-and-'hate-speech'-on-social-media-platforms_March2018.pdf [last accessed 16 October, 2019]. And see the European Commission's Code of Conduct on Countering Illegal Hate Speech Online of May 2016. Available at: http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf [last accessed 5 October 2019].

[211] See also p. 6 of Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online. And see para. 6 of Guidelines of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Preamble.

[212] See also Guideline 2.3.2 of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Appendix. And see the definition of "validity" on p. 149 of Facebook's document, Global Feedback and Input on the Facebook Oversight Board for Content Decisions, Appendix, 27 June, 2019. Available at: https://fbnewsroomus.files.wordpress.com/2019/06/oversight-board-consultation-report-appendix.pdf [last accessed 1 October, 2019].

- **Reliability**: Insofar as the governance tool involves assessing or monitoring the policies, practices or conduct of subjects of governance (e.g. Internet platforms), the extent to which it is capable of achieving a true and accurate picture, such as by minimising the capacity of the subjects of governance to game monitoring systems.

- **Enforceability**: Insofar as the governance tool involves laws, rules, guidelines, codes, policies, procedures, recommendations, adjudications or decisions, the extent to which these are capable of being enforced given industry, political, legal and economic constraints on enforcement.[213]

- **Practicality**: The degree to which the governance tool is feasible in a practical sense, including (1) whether it is equipped to meet the scale and complexity of the objects of governance given the levels of organisational, human and technological expertise and capacity needed to do the expected governance task, including but not limited to the extent to which it can deal with a high volume of content and the desired number of decisions, appeals or adjudications, and the extent to which it can handle both clear or manifest cases and grey area cases, (2) whether it is affordable given the budget or revenue of whichever agency, organisation, body or company is delivering or implementing the governance tool,[214] and (3) whether it represents good value for money in terms of the trade-off between amount spent and the actual levels of reach, performance and impact achieved (quality and quantity).

- **Impact and efficacy**: The extent to which the governance tool actually brings about the changes to policies, practices, conduct or decisions that it prescribes, and in doing so actually achieves its stated aim or purpose.

- **Legitimacy**: The extent to which the governance tool has legitimacy, that is, appropriate normative standing, including (1) whether the governance tool pursues or serves a legitimate aim or purpose,[215] (2) whether the agency, organisation, body or company that is delivering or implementing the governance tool has the right or authority to do so, and (3) insofar as the governance tool involves laws, rules, guidelines, codes, policies, procedures, recommendations, adjudications or decisions, the extent to which these are substantively legitimate in and of themselves.

- **Necessity**: Whether or not the governance tool is necessary, such as whether it is the least restrictive means available of effectively pursuing its stated aim or purpose.[216]

[213] See also Guideline 1.1.6 of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Appendix.

[214] See also the definition of "salience" on p. 149 of Facebook's document, *Global Feedback and Input on the Facebook Oversight Board for Content Decisions*, Appendix, 27 June, 2019. Available at: https://fbnewsroomus.files.wordpress.com/2019/06/oversight-board-consultation-report-appendix.pdf [last accessed 1 October, 2019].

[215] See also Guidelines 1.3.1, 1.4.1, 2.1.3, and 2.4.1 of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Appendix.

[216] See also Guidelines 1.1.1, 1.3.1, 1.3.8, 1.4.1, 2.3.1, and 2.3.2 of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Appendix.

- **Proportionality**: Whether or not the governance tool involves a level of restriction proportionate to its stated aim or purpose.[217]

- **Publicity and predictability**: Insofar as the governance tool involves laws, rules, guidelines, codes, policies, procedures, recommendations, adjudications or decisions, the extent to which parties subject to them are made aware of, and can understand, these things and can reasonably predict and therefore potentially avoid failures of compliance.

- **User-friendliness and accessibility**: Insofar as the governance tool involves mechanisms, processes or systems that are intended for use by ordinary people (e.g. reporting mechanisms for victims of hate speech, appeals processes, legal or administrative remedies), the degree to which the mechanisms, processes or systems are easy to understand and use, and, most importantly, are accessible and inclusive to all different types of users (e.g. use of plain language, information and mechanisms made available in multiple formats) so that literacy and disability are not barriers.[218]

- **Representativeness**: The extent to which any agencies, organisations, bodies or companies involved in delivering or implementing the governance tool are representative of society both demographically and in terms of prevalent social norms and values.[219]

- **Collaborativeness**: The extent to which the governance tool involves, promotes or facilitates fruitful cooperation and collaboration between different stakeholders involved in delivering or implementing the governance tool.[220]

- **Definitional harmonisation**: The extent to which the governance tool involves, promotes or facilitates desirable forms of coherence and convergence in definitions of hate speech both among and between Internet platforms, governmental agencies (lawmakers, government ministries, regulators, police, public prosecutors, courts), and civil society organisations, whether (1) at the national level or (2) at the international level.

- **Facts of pluralism and diversity within the sector**: The extent to which the governance tool is sensitive to and supportive of the facts of pluralism and diversity among different Internet platforms operating within the sector. A governance tool that had the effect, intended or otherwise, of creating an oligopoly of only a handful of Internet platforms or of flattening the different sorts of Internet platforms available would be undesirable according to this indicator.

- **Facts of technological advancement**: The extent to which the governance tool is responsive to the fact of technological advancement meaning it is almost inevitable that the objects of governance, or tech, will change in the future, sometimes both rapidly and radically. A governance tool that was not "tech neutral" but instead highly dependent on the existence of a particular bit of technology would run the risk of becoming obsolete if and

[217] See also Guidelines 1.1.1, 1.3.1, 1.3.6, 1.4.1, 2.1.5, and 2.4.1 of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Appendix.

[218] See also the Santa Clara Principles on Transparency and Accountability in Content Moderation.

[219] See also the definition of "broad representation" found on p. 11 of Article 19's document, *Self-regulation and 'Hate Speech' on Social Media Platforms*.

[220] See also p. 7 of Communication From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Tackling Illegal Content Online. And see the European Commission's Code of Conduct on Countering Illegal Hate Speech Online of May 2016.

when that bit of technology becomes obsolete, which would also have drawbacks in terms of the value for money and reach of the governance tool.

- **Democratization**: The degree to which the governance tool distributes, directly or indirectly, decision-making power over content across a variety of groups and society as a whole, rather than concentrating it in the hands of one organisation or elites.

- **Accountability**: The extent to which the governance tool involves some form of accountability to the public, whether directly through public consultation, for example, or indirectly through being accountable or answerable to other governmental agencies or civil society organisations that serve, represent or advocate on behalf of the public.[221]

- **Trust and confidence building**: The extent to which the governance tool commands and promotes trust and confidence, including (1) trust among the different agencies, organisations, bodies or companies that may be involved, or collaborate, in the implementation or delivery of the governance tool, (2) public trust and confidence in the governance tool itself, and (3) public trust and confidence in the different agencies, organisations, bodies or companies involved in the implementation or delivery of the governance tool.

- **Mental health and resilience**: Insofar as the governance tool involves human beings having to read, assess and examine substantial amounts of suspected hate speech content over a prolonged period of time (e.g. moderation teams, legal compliance teams, trusted flaggers, police, public prosecutors, oversight boards, dispute resolution services), the extent to which the relevant management and staff support structures are also geared up to promoting and protecting the mental health and resilience of these persons, such as through policies on work-life balance, training, counselling, peer support, sabbaticals, etc.

- **Avoidance of structural bias**: Whether or not the governance tool avoids being structurally biased against any particular group or community, such as by not disproportionately disadvantaging in an unfair or arbitrary way a particular group or community simply due to the way the tool is designed or implemented.

- **Compatibility**: The extent to which the governance tool is compatible with other regulatory aims and principles, such as respect for the privacy and personal data of users,[222] and recognition of the rights of Internet platforms to protect trade secrets and intellectual property.

[221] See also the Santa Clara Principles on Transparency and Accountability in Content Moderation.

[222] See also Guidelines 1.4.1, 1.4.3, and 1.4.4 of Recommendation CM/Rec(2018)2 of the Committee of Ministers to Member States on the Roles and Responsibilities of Internet Intermediaries, Appendix.

B. Some important qualifications about the indicators

It is important to recognise, however, that any such list cannot be taken at face value but must be explained and defended, which is the purpose of this section. Several considerations need to be borne in mind. First, the list contains many different sorts of indicators. For example, some are more oriented towards democratic values broadly construed, others are oriented towards issues of practicality, and yet others reflect legalistic doctrines and principles.[223] Furthermore, whilst some are more quantitative, others are more qualitative in nature.[224] Some seem to involve binary evaluation, whereas others involve sliding-scale assessments.

A second important consideration is that the list contains mostly highly abstract or general indicators that can be interpreted and fleshed out in different ways. Indeed, many of the indicators include value terms or normative concepts which are by nature subject to competing interpretations, and which are themselves rooted in value judgments (see Dworkin 2011).

Consider, for example, equitable sharing of the practical burden and legal responsibility for tackling online hate speech between governmental agencies, Internet platforms, civil society organisations, and users. The term "equitable" could be interpreted to mean an appropriate share relative to degree of causal responsibility for bringing about or facilitating online hate speech in the first place, or else an appropriate share relative to degree of capacity to tackle online hate speech, and so on. Likewise, the idea of sharing could mean each stakeholder undertakes a proportion of each element of governance, or instead it could mean a clear division of labour in which each stakeholder does a different element of governance.

To take a different example, the legitimacy of a governance tool partly depends on whether it serves a legitimate purpose, but the term "serves" can be understood in several different ways. It could mean "necessary for the achievement of the purpose" (strict), "substantially related to the legitimate purpose" (intermediate), or else "rationally related to the legitimate purpose" (weak). And, of course, the concept of a legitimate purpose can be variously interpreted.

A third important consideration has to do with precision. It is that the meaning and implications of each indicator in practice will differ depending on which type or subtype of governance tool is being assessed and depending on which level of governance is in play. This study has outlined three main categories of governance tools (moderation, oversight, regulatory), more than 20 different model types of governance tools split across the three main categories, numerous subtypes or variants of these main model types, and then 30 separate indicators or measures. To provide precise, bespoke definitions of every indicator for each of the possible governance tools, therefore, would require more than 1,000 extra definitions.[225] This study leaves open the further definition of indicators for organisations involved in monitoring, for example.

[223] 2nd consultative meeting, Berlin, 26 November, 2019.

[224] Ibid.

[225] Rather than attempt to furnish all of these definitions, it suffices to provide two illustrations. The first is democratization. At the moderation level, it is possible that democratization could mean the degree to which the governance tool distributes the task of moderation across a wide body of people (e.g. volunteer users) as opposed to concentrating it in the hands of a small body of professional moderators, for example. At the oversight level, an independent supervisory council, steering committee or oversight board might be said to achieve democratization due to (1) its making decisions about particular bits of content that Internet platforms agree to abide by, (2) its decision-making reflecting not merely the Internet platform's content policies and corporate values but also the norms of wider society including, for example, international human rights standards, and/or (3) its membership being subject to nomination by the public rather than by, say, Internet platforms. And, at the regulatory level, an Internet regulator, for example, might be said to achieve democratization insofar as (1) its functions derive from an agreement

A fourth important consideration is salience. It is that the relevance and importance of each indicator is likely to be more or less sensitive to which type or subtype of governance tool is being assessed and which level of governance is in play (moderation, oversight or regulatory). Some indicators, such as due process, for example, may be marginally more important at the oversight and regulatory levels than at the moderation level. Other indicators, such as accuracy, may be more important at the moderation level. Yet other indicators, such as legitimacy and transparency, will have roughly the same salience at each of the levels.

A fifth important consideration is that many of the indicators are highly interrelated meaning that for a governance tool to score well on some indicators would probably require it to score well on others or else would lead it to score well on others. Likewise, many of the indicators are overlapping meaning that two or more indicators have similar implications such that they would be likely to score the same governance tools in very similar ways. One obvious example is legitimacy which is likely to depend on indicators like independence, transparency, due process, valid adjudication, etc. Likewise, victim-sensitivity encompasses although is not exhausted by other indicators like protection, redress and user-friendliness.

A sixth consideration is compossibility, that is, whether it is possible in practice for a governance tool to achieve excellent or perfect scores on each of the measures or indicators at the same time. If not, then it means that two or more different indicators are not possible or compatible in conjunction with each other. For example, it seems likely that definitional harmonisation cuts against sensitivity to the facts of diversity and pluralism among Internet intermediaries.[226] This in turn means that assessing the success or progress of different governance tools is likely to be an all things considered calculation involving trade-offs between relative degrees of success or progress on different indicators at the same time.

A final and related point is that it may not always be possible to make straightforward comparisons of success or progress between different governance tools. For example, if indicators mean importantly different things at different governance levels (moderation, oversight, regulatory), then comparing the relative success or progress of different governance tools might be easier within the same governance level than across different governance levels. In other words, it might be more straightforward to judge which is the most successful governance tool among all the tools which operate at the moderation, for example, than it is to judge the relative success of a moderation tool which operates at the moderation level against another governance tool that operates at the oversight level or regulatory level.

among elected parliamentarians, and/or (2) its substantive policies and procedures are the result of a thorough process of public consultation. The second illustration is due process. At the moderation level, due process could mean that content removal decisions, for example, involve due consideration of any reasons for removal or non-removal provided by reporters and creators of content equally. At the oversight level, due process could mean that decisions about whether to uphold or overturn original content moderation decisions (e.g. internal appeals, oversight board decisions) are not taken without giving both sides the opportunity to present evidence and argumentation and that equal consideration is given to the evidence and argumentation of both sides. At the regulatory level, due process could mean that judicial decisions, for example, about imposing or not imposing fines on Internet platforms for a pattern of failure to remove illegal hate speech content are not taken without giving the platforms the opportunity to present evidence and argumentation in their defence and that any such decisions are subject to a right of appeal by the platforms.

[226] 1st consultative meeting, London, 17-18 October, 2019.

C. Selecting the right indicators

How should we judge these different measures or indicators? What makes an indicator worthy of inclusion within a list of indicators that are to be used by monitoring bodies, for example, to assess the success or progress of actual governance tools for online hate speech? Once again there may be several different factors to consider. However, below is a list of six factors that this study proposes and deems important based on the evidence.

- **Alignment with the core purposes of governance tools for online hate speech**: One factor is the degree to which a given indicator speaks to or captures one or more of the core purposes of governance tools for online hate speech outlined in section I.C above, or at least is sensitive to, and does not stand in strong conflict with, these core purposes. If the indicator or measure radically misunderstands or misconstrues the point of the governance tool, for example, it is likely to be a poor measure of its success or progress.

- **Meaningfulness and distinguishability**: A second factor is the extent to which a given indicator is or can be made meaningful and coherent in itself and also clearly distinguishable from other indicators. In terms of monitoring, for example, the acid test is whether the indicator can be properly understood by monitoring bodies that are tasked with applying the indicators, by organisations or Internet platforms who are subject to monitoring, and by the public at large, including understanding what the indicator means and how it is distinguished from other indicators.

- **Salience**: A third factor is the extent to which a given indicator is relevant to, and important for, the particular governance tools being assessed or evaluated. As made clear above, not all indicators will be equally salient to all governance tools.

- **Operationalizability**: A fourth factor is whether a given indicator can be operationalized, such as, for example, whether it can function as the sort of indicator, or can be translated into the sort of indicator, that monitoring bodies can use in the assessment of actual governance tools, given various technical, practical, legal and financial constraints faced by monitoring bodies to do with information gathering, data collection, accuracy and reliability of evidence, etc.

- **Proxies**: A fifth factor is the extent to which a given indicator can act as a proxy for other indicators in circumstances where it is not feasible to assess governance tools against all 30 indicators. This might be the case if, for example, one indicator is a node point or highly interrelated with several other indicators, or if the success or failure of a governance tool according to one indicator is a very good predictor of success or failure of a governance tool based on other indicators, such as if it reflects a deeper instrumental relationship between success according to one indicator and success according to others.

- **Support**: A final factor is the degree of support a given indicator commands among governmental agencies, Internet platforms, civil society organisations and users in any given country context and across different country contexts. In other words, could the indicator be the subject of deep and broad support from a wide variety of stakeholders? If not, that is, if the indicator is highly controversial and its value is strongly contested by numerous different kinds of stakeholders and on many different grounds, then monitoring bodies may face principled and practical resistance when attempting to apply the indicator to governance tools, and governmental agencies and the public at large might take less seriously any findings, reports or studies that utilised the indicator.

Transparency might be a good illustrative example of an indicator that seems to deliver reasonably well on all six factors. Arguably the transparency indicator aligns with many of the core purposes of governance tools for online hate speech, is meaningful and distinguishable from other indicators, has salience for most governance tools across the three different levels of governance, is operationalizable, could be a reasonable proxy for other indicators if necessary, and, finally, no matter one's role, position or perspective on the governance of online hate speech, transparency would seem to be something one could support.

To expand on the last factor, support, clearly different measures or indicators are likely to attract differing levels of support from different kinds of stakeholders. For example, minority rights organisations or NGOs that focus on protecting the rights of a particular community or group in society are likely to place particular importance on protection, redress, victim-sensitivity, and user-friendliness and accessibility, whilst also valuing other indicators of course.[227]

By contrast, civil society organisations that aim to promote and secure human rights including but not limited to the human right to freedom of expression, such as civil liberties organisations, for instance, might instead put more emphasis on human rights, independence, transparency, accountability, due process, legitimacy, necessity, proportionality, publicity and predictability, etc.[228]

Governmental agencies might highlight enforceability, reliability, impact and efficacy, and legitimacy, but also definitional harmonisation, publicity and predictability, and user-friendliness and accessibility, for instance.[229]

For their part, Internet platforms could be more inclined to underscore the special importance of practicality, trust and confidence building, pluralism, and publicity and predictability, but also user-friendliness and accessibility, and mental health and resilience, for example, whilst also valuing other indicators of course.[230]

Nevertheless, the key to support is whether an indicator can attract a strong degree of support not merely in the sense of begrudging acceptance but also active engagement or "buy in". Equally important is whether an indicator can attract support not just from one or a few stakeholders but from a wide and diverse range of stakeholders. Transparency and user-friendliness and accessibility, for example, might be the sorts of indicators that can, and do, attract strong support from a broad range of stakeholders.[231]

[227] 1st consultative meeting, London, 17-18 October, 2019.

[228] Ibid.

[229] Ibid.

[230] Ibid.

[231] Ibid.

IX. CONCLUSIONS AND RECOMMENDATIONS

Based on the above this study draws a series of conclusions and makes a number of practical recommendations covering ten key areas of the governance of online hate speech. The following thoughts, options and recommendations derive from the author's analysis and should be seen as suggestive, rather than definitive and exhaustive.

1. Standardization agenda

1.1 Governmental agencies, Internet platforms and civil society organisations should be aware of, and seek to integrate wherever appropriate, a wide range of different governance tools for online hate speech available at the moderation, oversight and regulatory levels [see sections II, III and IV]. They should also be conscious of the fact that variety in the types of Internet platforms can necessitate pluralism within styles of content moderation, that pluralism within styles of moderation makes appropriate pluralism within systems of oversight of moderation, and that pluralism within systems of oversight of moderation suggests the need for pluralism within regulatory instruments or tools [see sections I.B(iii) and I.D].

1.2 National governments and intergovernmental organisations should recognise that if different standards for the regulation of online hate speech develop across Europe in a patchwork, piecemeal and unpredictable fashion, this could make it not merely more difficult for Internet platforms to operate successfully in Europe but also to achieve regulatory compliance [see section I.A(ii)].

1.3 However, common standards for the regulation of online hate speech in Europe need not mean identical regulatory models or tools. In particular, a new Digital Services Act at the European level could, and should, retain three important forms of decentralisation. First, decentralised regulatory authorities, meaning each country establishes its own national regulator or devolves more powers to existing regulators. Second, a common standard on the responsibility of Internet platforms to remove illegal hate speech content within a specified time frame but with each national regulator applying its own local hate speech laws (which should abide to European standards). Third, a common standard on the responsibility of Internet platforms to remove illegal hate speech content within a specified time frame but with each national regulator designing and implementing slightly different exceptions, exemptions and leniency programmes under this main rubric [see sections I.A(ii), I.D(vi), I.D(vii)].

1.4 Furthermore, national governments and intergovernmental organisations should also be sensitive to the fact of diversity and pluralism among Internet platforms, meaning that a one-size-fits-all approach to the governance of online hate speech is likely to be unsuitable, unworkable and unfair. There is a risk that placing the same demands on all Internet platforms without flexibility could create or maintain an oligopoly of a handful of Internet platforms or could otherwise lead to a flattening of the sorts of Internet platforms that can operate [see sections I.B(iii), I.D].

1.5 That being said, due to the practical difficulties of running parallel regulatory regimes for different kinds of Internet platforms, it is recommended that governance tools themselves incorporate where appropriate the use of exceptions, exemptions and leniency programmes to reflect important differences between Internet platforms [see sections I.D, IV.C].

1.6 In particular, the regulatory model of imposing legal responsibilities on Internet platforms to remove illegal hate speech within specified time frames and levying fines for patterns of failure to comply with this responsibility (Regulatory-C)—a model partly exemplified by the NetzDG Act in Germany and the Avia Bill in France—should be qualified. Governmental authorities should reflect on both country context and diversity among Internet platforms and consider qualifying the model in one or more of the following ways: (i) allowing exceptions for journalistic content [see section IV.C(i)]; (ii) allowing exceptions for Internet platforms that refer grey area cases to competent independent institutions and abide by the decisions [see section IV.C(ii)]; (iii) granting exemptions from regulatory fines for Internet platforms granted “responsible platform” status, where this status depends on *inter alia* platforms devoting reasonable levels of resources to removing illegal hate speech [see section IV.C(iii)]; (iv) giving Internet platforms reductions in fines in return for providing full disclosure about the amounts of illegal hate speech on their platforms (leniency programmes) [see section IV.C(iv)].

1.7 However, governmental authorities should also work to ensure that Internet platforms do not take advantage of or exploit these exceptions, exemptions or leniency programmes in ways that would undermine the core regulatory purposes being pursued. For example, where governmental authorities allow exceptions for Internet platforms that refer grey area cases to competent independent institutions and abide by the decisions [see section IV.C(ii)], the relevant competent independent institutions should be granted the power to select the cases they will hear, so as to prevent Internet platforms from inundating or flooding the institution with cases simply to qualify for exceptions and to avoid fines [see section IV.C(ii)].

2. Grey area cases

2.1 Governmental agencies, Internet platforms and civil society organisations should be conscious of the need for governance tools for online hate speech that are capable of dealing with not only clear or manifest cases but also grey area cases where it is unclear whether or not a given piece of online speech content (a) is hate speech, (b) contravenes the Internet company's rules on permissible content (i.e. community standards or content policies on hate speech), and (c) is unlawful or illegal based on local hate speech laws (where such laws exist) [see sections I.A(i), I.B(ii), I.B(iv), I.C(ii), I.C(iii), I.E, I.F, VIII.A].

2.2 Some Internet platforms, notably Facebook, have already identified Oversight-D (Referrals of grey area cases to an independent supervisory council, steering committee or oversight board) as an appropriate governance tool for handling grey area cases. However, Facebook's new Oversight Board will not hear cases in "limited circumstances" where its decisions, if acted upon by Facebook, could render Facebook's senior managers liable to criminal liability or could make Facebook as a corporate entity the target of regulatory sanctions [see sections I.F, III.D].

2.3 Regrettably, this limitation on the cases that the Oversight Board will hear could drastically reduce the usefulness of the Board in countries where there is a hostile environment for Internet platforms in terms of their facing criminal liability or regulatory sanctions for a pattern of failure to remove illegal hate speech content (e.g. Germany) [see sections I.F, III.D]. Furthermore, under s. 3(2)3.b) of the NetzDG Act in Germany, the general requirement for Internet platforms to remove unlawful content within 7 days of receiving a report or complaint does not apply *inter alia* in circumstances where the platform refers the case to a competent independent institution within 7 days of receipt, and agrees to accept the decision of that institution. For Facebook to decide that its Oversight Board will not hear cases that potentially raise issues of illegality under the NetzDG regulatory framework arguably constitutes a missed opportunity on the part of Facebook to refer grey area cases of potentially unlawful hate speech to the Board, and in doing so to seek to qualify for the 7 day removal exception set out above. Facebook could attempt to do so by fundamentally changing the Oversight Board, or at least a country subpanel thereof, so as to satisfy the qualifying conditions for the exception. Alternatively, the German government might consider relaxing the high standards for the sort of institution that can be accredited as an "institution of regulated self-governance". Or both parties could work collaboratively to reach a compromise [see sections I.F, III.D].

2.4 It is possible that in the future the new Digital Services Act could require member states throughout Europe to impose fines on Internet platforms for a pattern of failure to remove illegal hate speech content and could also mirror NetzDG by providing for a 7 day removal exception where platforms send grey area cases to competent independent institutions (e.g. oversight boards). In that event, Facebook could no longer treat Germany as a special case, and might need to rethink the nature and function of its Oversight Board for use in Europe as a whole [see sections I.F, III.D].

2.5 More generally, if regulatory fines could create an unwelcome bias or tendency among Internet platforms to remove suspected illegal hate speech content on a "safety first" approach, then referring grey area cases to competent independent institutions (e.g. oversight boards) might help to mitigate that tendency. Thus, the limited circumstances where an Internet platform takes down content because it is potentially illegal but also fears criminal liability or regulatory sanction for not taking it down are precisely the cases where checks and balances are needed the most—the sorts of checks and balances that oversight boards can deliver [see sections I.F, III.D].

2.6 The study concludes that Governmental agencies, Internet platforms and civil society organisations should recognise that some governance tools are more suited to dealing with grey area cases than others. The following governance tools seem especially suitable for grey area cases: Moderation-F (Content management or reducing access to content) [see section II.F], Oversight-D (Referrals of grey area or difficult cases to an independent supervisory council, steering committee or oversight board) [see section III.D]; Oversight-E (Users and Internet platforms agree to avail themselves of a fully independent dispute resolution procedure or mediation process after any internal appeals process has been exhausted) [see section III.E]; Regulatory-C (Impose a legal responsibility on Internet platforms to remove unlawful hate speech content within a specified time frame following notice but also provide exceptions for Internet platforms that refer grey area or difficult cases to competent independent institutions and abide by the decisions) [see section IV.C(ii)].

3. Public opinion

3.1 National governments and intergovernmental organisations should keep in mind that in the UK, France and Germany public opinion surveys show that the general public think it is as important for governmental authorities to impose legal responsibilities on Internet platforms to remove illegal hate speech content as it is for authorities to prosecute the creators of such content [see section VI.D].

3.2 But national governments and intergovernmental organisations should also be aware that the survey results suggest that the general public have a balanced, non-punitive or non-absolutist view on how to regulate Internet platforms. The general public in the UK, France and Germany rate it as not merely important to levy fines on Internet platforms that demonstrate a pattern of failure to remove illegal hate speech content but also important to grant exemptions from such fines if Internet platforms devote reasonable resources to removing illegal hate speech and important to offer Internet platforms reductions in fines in return for them providing full disclosure about the amounts of illegal hate speech on their platforms [see section VI.D].

4. Collaboration

4.1 The study suggests Governmental agencies, Internet platforms and civil society organisations should seek more mutual cooperation and collaboration keeping in mind the many potential benefits of this sort of approach to the governance of online hate speech. The potential benefits include: Increases the influence of less powerful stakeholders; Promotes engagement with governance among stakeholders; Facilitates innovation and creativity in the framing and solving of problems; Promotes mutual understanding and compromise; Encourages sharing of technical knowledge between stakeholders; Produces governance that is more capable of dealing with grey area cases; Results in governance that strikes a better balance among human rights; Enhances public trust in governance measures; Promotes trust among collaborating stakeholders [see section V.A].

4.2 But governmental agencies, Internet platforms and civil society organisations should also be conscious of the pitfalls and challenges in managing collaborative relationships, particularly in terms of dealing with creative differences, power differentials, the maintenance of independence, and sharing of sensitive information [see section V.B].

4.3 Various of these pitfalls and challenges with collaboration can be identified in collaboration between Internet platforms and trusted flaggers [see section V.B(i)]; collaboration between Internet platforms and independent supervisory councils, steering committees or oversight boards [see section V.B(ii)]; and collaboration between Internet platforms and the police and public prosecutors [see section V.B(iii)].

5. Mitigating the incentive to over-remove hate speech content

5.1 National governments and intergovernmental organisations should take into account that by imposing legal responsibilities on Internet platforms to remove illegal hate speech content within specified time frames and by levying fines on Internet platforms for a pattern of failure to comply with this responsibility there is a risk that this could create an incentive for Internet platforms to over-remove hate speech content including lawful or legal content [see section I.B(ii), I.C(i), I.C(iii), I.D(iv), I.F, IV.C, IV.D].

5.2 National governments and intergovernmental organisations should also recognise that there are many ways of mitigating the aforementioned incentive. These include: Equivalent fines for over-removal of lawful hate speech content; Reputational damage caused by over-removal of lawful hate speech content; Push-back by Internet platforms based on their corporate values on free speech; Legal obligation on governments to draft Internet laws that safeguard free speech rights; Legal obligation on public prosecutors to enforce Internet laws whilst recognising free speech rights; Legal obligation on administrative courts to interpret Internet laws so as to safeguard free speech rights;[232] Criminal laws targeting individuals for malicious reporting of online hate speech; Qualify fines for under-removal of unlawful hate speech with exceptions or exemptions [see sections IV.C, IV.D, especially Diagram 3].

5.3 That being said, some variants of the alternative interventions listed in paragraph 5.2 above may themselves be incompatible with adopting a victim-sensitive approach. For example, the study recommends that adopting a victim-sensitive approach at the regulatory level involves, amongst many other things, safeguarding [see sections VII.C(iii), IX.7]. One implication is that governments should avoid legislation that creates unnecessary and detrimental friction for persons reporting online hate speech content if they are also the targeted or adversely impacted persons. Now Regulatory-G [see section IV.G] includes, as one possible variant, the introduction of criminal offences relating to the conduct of individuals who maliciously report or flag content as being manifestly unlawful hate speech. Whilst the legislative intent of this governance tool seems legitimate, aiming as it does to protect free speech by combating malicious reporting and flagging of hate speech, there are dangers. If the law is badly framed or overbroad (draconian), or imposes overly severe punishments or is misapplied by the courts (disproportionate), there is a risk that it could discourage honest reporting and flagging of hate speech. Genuine victims might refrain from flagging or reporting for fear of being prosecuted, thus potentially exacerbating an existing problem of under-reporting of online hate speech and blocking access to regulatory remedies.

5.4 In light of this, this study suggests that the introduction of criminal offences for malicious reporting is likely to be suboptimal, partly because it is not the least restrictive alternative, it might deter genuine reporting and otherwise be detrimental to a victim-sensitive approach. Some of the alternative interventions listed in paragraph 5.2 above, including enforce Internet laws whilst recognising free speech rights, are arguably less restrictive but potentially as effective when used in combination or acting together [see section IV.D].

[232] The term "administrative courts" here refers to courts involved in mandating or approving the levying of fines on Internet platforms for patterns of failure to removal unlawful hate speech content.

6. Monitoring voluntary codes of conduct

6.1 Governmental agencies, Internet platforms and civil society organisations should appreciate that Regulatory-B (a voluntary code of practice) can produce improvements in the removal of illegal hate speech content. This can be seen in the European Commission's Code of Conduct on Countering Illegal Hate Speech Online and associated monitoring regime. The first monitoring cycle, conducted in Fall 2016, showed that "[o]ut of 600 notifications, in 169 cases (28.2 percent) the content was removed" (European Commission 2016b: 4). In the fourth monitoring cycle, conducted in Fall 2018, "[o]n average, IT companies are removing 72 percent of the illegal hate speech notified to them" (European Commission 2019: 1) [see section IV.B].

6.2 However, this study also identified a problem that Internet platforms are being made aware of the monitoring period. Based on this problem, it is unclear the extent to which these changes in percentages represent genuine improvements in the removal rate for illegal hate speech throughout the year or in fact reflect improvements in Internet platforms' capacity to game the monitoring process by significantly improving removal rates during the period of monitoring only. That being said, since this seems to have been a weakness in the monitoring system from the start—whereby Internet platforms have been made aware of the period of monitoring through each of the cycles—then it seems safe to conjecture that at least some of the increases in percentages do partly reflect genuine improvement to moderation [see sections IV.B, V.B(i)].

6.3 In addition, some trusted flaggers report that during meetings and training sessions provided by Internet platforms, the latter will typically offer "advertising grants" to trusted flaggers that enable them, for example, to run their own campaigns on the platforms free of charge as a sort of "goodie bag" for participating in the meeting or training session. This sort of close relationship may or may not hinder the independence of organisations working as trusted flaggers but it arguably does reduce the independence, or reduce the appearance of independence, of these organisations as monitoring bodies.

6.4 Reflecting on the above, this study recommends that reforms are made to the monitoring system of voluntary codes of conduct to ensure that Internet platforms are not made aware—deliberately or structurally—of the period of monitoring, such as by extending the monitoring period to 12 months of the year [see sections IV.B, V.B(i)]. It is also recommended that monitoring bodies must publicly declare any "in kind" benefits received from Internet platforms [see sections IV.B, V.B(i)].

7. A victim-sensitive approach

7.1 No approach to the governance of online hate speech would be complete without paying special attention to the needs and experiences of victims [see section VII.A], but particular care should be taken when conceptualising victim-sensitivity [see section VII.B].

7.2 It should also be recognised that within the general category of individuals or groups targeted or adversely affected by online hate speech (or "victims") there might exist individuals or groups who are particularly vulnerable, either because they are subjected to greater amounts of, or more severe forms of, online hate speech or because for some reason they face greater obstacles in using redress mechanisms [see section VII.A].

7.3 It is recommended that when designing, planning, selecting, implementing and delivering governance tools for online hate speech, governmental agencies, Internet platforms and civil society organisations should adopt a victim-sensitive approach to the governance of online hate speech [see section VII]. Moreover, it is recommended that victim-sensitivity should be used by governmental agencies, Internet platforms and civil society organisations, including monitoring bodies, as an indicator or measure of the success or progress of governance tools [see section VIII.A].

7.4 Governmental agencies, Internet platforms and civil society organisations are recommended to implement the general guidelines for victim sensitivity as proposed in section VII and the practical recommendations at each of the three levels of governance of online hate speech [see section VII.C], as follows.

> 7.4.1 At the moderation level key practical recommendations include: notification of moderation decisions must be sent to the victim; notifications sent to the victim should go beyond pro forma communications and standardized explanations to provide messages that contain at least some personalised or semi-personalised content of a suitable form, where feasible and appropriate; reporting mechanisms should use plain language and should be made available in multiple languages and formats; reporting forms should enable bulk reporting, such as by allowing victims to highlight and report several bits of content at the same time; Internet platforms should be proactive in identifying and removing "identical", "equivalent" (e.g. language translations), or "very similar" content; wherever feasible moderation should empower the victim, or put them back in control, such as by giving them power to select between moderation outcomes [see section VII.C(i)]
>
> 7.4.2 At the oversight level key practical recommendations include: oversight mechanisms should seek to make oversight part of the process of healing or recovery for victims, such as by giving victims "their day in court" or access to some form of restorative justice; victims should not merely be informed of oversight decisions but should also be provided with an account of the reasoning behind the oversight decision; an independent supervisory council, steering committee or oversight board involved in oversight (e.g. Facebook's Oversight Board) should be willing to receive statements or "amicus briefs" setting out the merits of hearing certain cases [see section VII.C(ii)].
>
> 7.4.3 At the regulatory level key practical recommendations include: notification of regulatory decisions in particular cases must be sent to the victim; any regulatory mechanisms that require victims to submit information should avoid any demands that

would run a significant risk of retraumatisation; any regulatory mechanisms should not impose unnecessary and detrimental friction for victims, such as by avoiding draconian laws or disproportionate sanctions against malicious reporting; complaints bodies should be willing to act as a “one stop shop” for complaints, such as by enabling victims to launch group or “class action” complaints and to launch simultaneous complaints against multiple Internet platforms; victims should be empowered by regulatory processes, such as by enabling victims to play an active role in the case as it progresses through the relevant legal or administrative processes including by giving evidence or testifying where appropriate based on consent and with necessary legal and psychological support [see section VII.C(iii)].

7.5 Governmental agencies, Internet platforms and civil society organisations should avoid taking a softer or weaker approach to the governance of online hate speech when it is posted or shared by political figures—such as by adopting content management rather than content removal. This may run counter to principles of victim-sensitivity despite the rationales offered for this approach (e.g. public interest). Consider Internet platforms that attach warning labels to online hate speech posted by political figures rather than removing the content (e.g. Twitter) [see section II.F]. By prioritising the free speech of political figures such Internet platforms risk downgrading the importance of the experiences and needs of victims of online hate speech especially in terms of empowerment [see section VII.C(i)] in comparison to speech of non political figures.

8. Proactive use of text extraction and machine learning tools or algorithms

8.1 Internet platforms' legal compliance teams should be proactive in identifying unlawful or illegal hate speech content, such as by more closely monitoring accounts after one bona fide instance of illegal content has been discovered on those account, or by using text extraction and machine learning tools or algorithms to search for illegal content [see section II.A]. This would also reflect principles of victim-sensitivity [see section VII.C(i)].

8.3 It is recommended that legal compliance teams supply programmers with sample relevant court decisions (summaries) in the relevant countries (i.e. cases where local hate speech laws have been applied by the courts to bits of content) [see section II.A]. Legal compliance teams furnishing programmers with cases that the legal compliance teams have reached definitive and accurate judgments about will improve the quality of the "training data" or "benchmark data sets" used in the programming of machine learning tools or algorithms.

8.4 It is also recommended that governmental authorities, following recommendations of intergovernmental organisations, should place legal responsibilities on Internet platforms to be proactive in identifying unlawful or illegal hate speech content [see section IV.C].

8.5 National governments and intergovernmental organisations should bear in mind that the legal responsibility on Internet platforms to be proactive in identifying unlawful or illegal hate speech content can be specified in stronger or weaker ways [see section IV.C, especially diagram 1]. However, the precise extent of that responsibility, the methods of enforcement and the use of any exceptions, exemptions or leniency programs should also reflect country context [see sections I.A(ii), I.D(vi), I.D(vii), IV.C].

9. Indicators of success in the governance of online hate speech

9.1 Governmental agencies, Internet platforms and civil society organisations, including monitoring bodies, should be aware of the fact that comparing the relative success or progress of different governance tools for online hate speech is likely to be easier for different tools within each of the three levels of governance than for different tools across the levels (moderation, oversight, regulatory) [see section VIII.B].

9.2 Further monitoring, reporting, and research is needed to assess the success or progress of actual or real world governance regimes or instruments for online hate speech that are currently being implemented or administered by governmental agencies, Internet platforms and civil society organisations across Europe. In undertaking this endeavour monitoring bodies, reporting organisations and researchers may find it useful to apply and/or adapt the ideal types or models of governance tools outlined in the study [see sections II, III, IV], particularly those best suited to address grey area cases as outlined in point 2.6 above.

9.3 Monitoring bodies, reporting organisations, and researchers should also recognise that there are a large number of potentially valid and important indicators or measures that could be used when assessing the success or progress of governance tools for online hate speech. This study identified the following indicators: Protection; Redress; Victim-sensitivity; Human rights; Equity; Transparency; Independence; Due process; Validity; Accuracy and context; Reliability; Enforceability; Practicality; Impact and efficacy; Legitimacy; Necessity; Proportionality; Publicity and predictability; User-friendliness and accessibility; Representativeness; Collaborativeness; Definitional harmonisation; Facts of pluralism and diversity within the sector; Facts of technological advancement; Democratization; Accountability; Trust and confidence building; Mental health and resilience; Avoidance of structural bias; Compatibility [see section VIII.A].

9.4 When selecting indicators or measures to be used in assessing the success or progress of actual or real world governance regimes or instruments, the relevant monitoring bodies, reporting organisations, and researchers should bear in mind six factors identified in this study: alignment with the core purposes of governance tools for online hate speech; meaningfulness and distinguishability; salience; operationalizability; proxies; support [see section VIII.C].

9.5 In order to be reliable and effective, the monitoring bodies, reporting organisations, and researchers will need access to data, both quantitative and qualitative, from the relevant governmental authorities, Internet platforms and civil society organisations [see section I.G]. For example, transparency should mean providing access not merely to numbers of governance decisions but also to the substantive findings, reasoning and processes involved in those decisions. Going forward policy planning around the governance of online hate speech in Europe should be based on evidence-based assessments of the success or progress of real world or actual governance regimes and instruments and not assumptions about what might happen.

9.6 That being said, in the absence of complete or perfect data, sometimes a precautionary approach to the governance of online hate speech can be rational. Precautionism can justify policies designed to mitigate the risk of over-removal of lawful content as well as policies that aim to mitigate the risk of under-removal of unlawful hate speech [see section IV.C]. For example, national governments and intergovernmental organisations should recognise that by imposing legal responsibilities on Internet platforms to remove illegal hate speech within specified time frames and by levying fines on Internet platforms for patterns of failure to comply with this responsibility there is a risk that this could create an incentive for Internet platforms to over-remove hate speech content including lawful or legal content [see sections I.B(ii), I.C(i), I.C(iii), I.D(iv), I.F, IV.C, IV.D, IX.5]. It is recommended that a precautionary approach is taken to the risk, in the light of incomplete or imperfect evidence [see section IV.C]. Governmental authorities can, and should, consider various different steps to mitigate the risk of this potential incentive [see sections IV.C, IV.D, IX.5].

10. Equitable sharing in the governance of online hate speech

10.1 It is right for governmental agencies, Internet platforms and civil organisations to argue for, and accept, equitable sharing in the practical burden of, and legal responsibility for, tackling online hate speech [see section I.B(ii)].

10.2 One way to understand this idea of equitable sharing of burden and responsibility is in terms of fair proportion: for example, that Internet platforms should share a burden and responsibility that is in fair proportionate to their role in (unintentionally) facilitating online hate speech and/or in fair proportion to their capacity in tackling online hate speech [see section I.B(ii)].

10.3 It is wrong to think that the only way for Internet platforms to meet their equitable share of burden and responsibility for tackling online hate speech is by means of content removal. Content management is also an important tool, especially in grey area cases [see section II.F].

10.4 A truly responsible Internet platform is one that, on occasion and where appropriate, is willing to defend in courts its decisions not to remove content, on the grounds of promoting and protecting the human right to freedom of expression [see section I.B(ii)].

10.5 Governmental agencies, Internet platforms and civil society organisations should also be conscious of the fact that the governance tools for online hate speech examined in this study are only part of the story in tackling online hate speech. Other important policy interventions are outlined in various of the key documents at the European level discussed in this study [see section I.B]. ECRI's General Policy Recommendation (GPR) No. 15, for example, sets out a wide range of measures including but not limited to providing education to children, as well as guidance and training to political figures and media professionals, about the harms of hate speech, and supporting organisations involved in distributing counter-narratives to hate speech.[233] Governmental authorities should also implement these recommendations.

[233] CRI(2016)15, Strasbourg, 8 December 2015. Available at: https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01 [last accessed 7 October, 2019].

LIST OF ORGANISATIONS WHO PARTICIPATED, WERE CONSULTED OR WERE ENGAGED DURING THE STUDY

Aequitas
American University, Washington College of Law
Article 19
Belgian Police, Online hate crime and hate speech special unit
Berkman Klein Center for Internet & Society, Harvard University
Centre for Communication Governance, National Law University in Delhi
Centre for Democracy and Technology
Centre for Internet Studies, Bangalore, India
Centre for Studies on Freedom of Expression and Access to Information, University of Palermo in Buenos Aires, Argentina
Change.org
Council of Europe, Anti-Discrimination Department, ECRI
Community Security Trust (CST)
Dailymotion
The Danish Institute for Human Rights
Derechos Digitales, Santiago, Chile
ePanstwo Foundation
EQUINET
European Commission, DG Justice
European Schoolnet
Facebook
French National Assembly, Laetitia Avia
German government, Federal Ministry of Justice and Consumer Protection
Gitanos
Global Network Initiative (GNI)
Global Partners Digital
Google
Hatter
Human Rights House Zagreb
Human Rights Watch
INACH
International Network Against Cyber Hate
International Society
Millicom
No Hate Speech Movement
Ofcom
Open Technology Institute at the New America Foundation in Washington
Oxford Brookes University
Pink Armenia
Snapchat
Spanish government, Ministry of the Interior
Stanford Global Digital Policy Incubator
Twitter
UK government, Department for Digital, Culture, Media & Sport
UK government, Home Office
UK police, Online Hate Crime Hub
Unia
University of East Anglia
Verizon Media
Victim Support Europe
Wikimedia Foundation
YouTube

REFERENCES

Alkiviadou, N. (2016) ‘Regulating Internet Hate: A Flying Pig?’, *Journal of Intellectual Property, Information Technology and E-Commerce Law* 7: 216-228.

Article 19 (2016) *EU: European Commission’s Code of Conduct on Countering Illegal Hate Speech Online and the Framework Decision*, June. Available at: https://www.article19.org/data/files/medialibrary/38430/EU-Code-of-conduct-analysis-FINAL.pdf [last accessed 1 October, 2019].

Article 19 (2017) *Germany: The Act to Improve Enforcement of the Law in Social Networks*, August. Available at: https://www.article19.org/wp-content/uploads/2017/09/170901-Legal-Analysis-German-NetzDG-Act.pdf [last accessed 1 October, 2019].

Article 19 (2019) *France: Analysis of Draft Hate Speech Bill*, 3 July. Available at: https://www.article19.org/resources/france-analysis-of-draft-hate-speech-bill/ [last accessed 4 October, 2019].

Avram, D. (2019) ‘Towards an enhanced responsibility of online platforms: the EU Digital Services Act’, *Inline*, 31 Jul 2019. Available at: https://www.inlinepolicy.com/blog/towards-an-enhanced-responsibility-of-online-platforms-the-eu-digital-services-act [last accessed 1 December 2019].

Bakalis, C. (2015) *Cyberhate: An issue of continued concern for the Council of Europe’s Anti-Racism Commission*, November. Available at: https://edoc.coe.int/en/cybercrime/6883-cyberhate-an-issue-of-continued-concern-for-the-council-of-europe-s-anti-racism-commission.html [last accessed 3 December, 2019].

Banks, J. (2011) ‘European regulation of cross-border hate speech in cyberspace: The limits of legislation’, *European Journal of Crime, Criminal Law and Criminal Justice* 19: 1-13.

Benesch, S. and Matias, J. N. (2018) ‘Launching today: new collaborative study to diminish abuse on Twitter’, *Medium*, 6 April. Available at: https://medium.com/@susanbenesch/launching-today-new-collaborative-study-to-diminish-abuse-on-twitter-2b91837668cc [last accessed 19/11/2019].

Brown, A. (2009) *Personal Responsibility: Why it Matters*. London: Continuum Press.

Brown, A. (2015) *Hate Speech Law: A Philosophical Examination*. London: Routledge.

Brown, A. (2016) ‘The “Who?” Question in the Hate Speech Debate: Part 1: Consistency, Practical, and Formal Approaches’, *Canadian Journal of Law & Jurisprudence* 29: 275-320

Brown, A. (2017a) ‘What is Hate Speech? Part 1: The Myth of Hate’, *Law and Philosophy* 36: 419-468.

Brown, A. (2017b) ‘What is Hate Speech? Part 2: Family Resemblances’, *Law and Philosophy* 36: 561-613.

Brown, A. (2017c) 'The "Who?" Question in the Hate Speech Debate: Part 2: Functional and Democratic Approaches', *Canadian Journal of Law & Jurisprudence* 30: 23-55.

Brown, A. (2017d) 'Hate Speech Laws, Legitimacy, and Precaution: Reply to Weinstein', *Constitutional Commentary* 32: 599-617.

Brown, A. (2017e) 'Averting Your Eyes in the Information Age: Hate Speech, the Internet, and the Captive Audience Doctrine', *Charleston Law Review* 12: 1–54.

Brown, A. (2017f) 'New Evidence Shows Public Supports Banning Hate Speech Against People with Disabilities', The Conversation, March 1. Available at: https://theconversation.com/newevidence-shows-public-supports-banning-hate-speech-against-people-with-disabilities-73807 [last accessed 4 December, 2019].

Brown, A. (2018a) 'What is so Special About Online (as Compared to Offline) Hate Speech? Internet Companies, Community Standards and the Extragovernmental Regulation of Cyberhate', *Ethnicities* 18: 297–326.

Brown, A. (2018b) 'Retheorizing Actionable Injuries in Civil Lawsuits Involving Targeted Hate Speech: Hate Speech as Degradation and Humiliation', *Alabama Civil Rights & Civil Liberties Law Review* 9: 1–56.

Brown, A. (2019a) 'The Meaning of Silence in Cyberspace: The Authority Problem and Online Hate Speech' in K. Gelber and S. Brison (eds.) *Free Speech in the Digital Age* (Oxford University Press).

Brown, A. (2019b) 'New Evidence Shows Increasing Public Support for Hate Speech Laws post-Brexit', Eastminster, June 18. Available at: http://www.ueapolitics.org/2019/06/18/newevidence-shows-increasing-public-support-for-hate-speech-laws-post-brexit/ (last accessed 4 December, 2019].

Brown, A. and Sinclair, A. (2019) *The Politics of Hate Speech Laws*. London: Routledge.

Butler, J. (1997) *Excitable Speech: A Politics of the Performative*. New York: Routledge.

Bychawska-Siniarska, D. (2017) *Protecting the Right to Freedom of Expression Under the European Convention On Human Rights: A Handbook for Legal Practitioners*, Council of Europe, July. Available at: https://rm.coe.int/handbook-freedom-of-expression-eng/1680732814 [last accessed 10 December 2019].

Caplan, R. (2018) 'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches', *Data & Society*. Available at: https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf.

Carnegie UK Trust (2019) *The Online Harms White Paper: a summary response from the Carnegie UK Trust*, 18 June. Available at: https://www.carnegieuktrust.org.uk/blog/online-harms-response-cukt/ [last accessed 9 October, 2019].

Citron D. K. (2014) *Hate Crimes in Cyberspace*. Harvard, MA: Harvard University Press.

Citron, D. K. and Norton, H. (2011) 'Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age', *Boston University Law Review*, 91: 1435-1484.

Cohen-Almagor R. (2015) *Confronting the Internet's Dark Side: Moral and Social Responsibility on the Free Highway*. Cambridge: Cambridge University Press.

Delgado, R. and Stefancic, J. (2004) *Understanding Words That Wound*. Boulder, CO: Westview Press.

Delgado, R. and Stefancic, J. (2014) 'Hate Speech in Cyberspace', *Wake Forest Law Review* 49: 319–343.

Dworkin, R. (2011) *Justice for Hedgehogs*. Cambridge, MA: Harvard University Press.

ECRI (2017) ECRI Report on Luxembourg, Fifth Monitoring Cycle, 28 February. Available at: https://rm.coe.int/fifth-report-on-luxembourg/16808b589b [last accessed 29 April 2020].

ECRI (2018) ECRI Report on Spain, Fifth Monitoring Cycle, 27 February. Available at: https://rm.coe.int/fifth-report-on-spain/16808b56c9 [last accessed 29 April 2020].

ECRI (2019) *ECRI Report on Romania, Fifth Monitoring Cycle*, 5 June. Available at: https://rm.coe.int/fifth-report-on-romania/168094c9e5 [last accessed 3 December 2019].

ECRI (2020) *ECRI Report on Germany, Sixth Monitoring Cycle*, 17 March. Available at: https://rm.coe.int/ecri-report-on-germany-sixth-monitoring-cycle-/16809ce4be [last accessed 13 April 2020].

Eberwine, E. T. (2004) 'Sound and Fury Signifying Nothing? Jürgen Büssow's Battle Against Hate-Speech on the Internet', *New York Law School Law Review* 49: 353-410.

Equinet (2018) *Extending the Agenda: Equality Bodies Addressing Hate Speech*, December. Available at: http://equineteurope.org/wp-content/uploads/2019/05/hate_speech_perspective_-_web.pdf [last accessed 6 December, 2019].

European Commission (2016a) *Report on the progress made in the fight against trafficking in human beings*, 19 May, Brussels. Available at: https://ec.europa.eu/anti-trafficking/sites/antitrafficking/files/report_on_the_progress_made_in_the_fight_against_trafficking_in_human_beings_2016.pdf [last accessed 2 October, 2019].

European Commission (2016b) *Code of Conduct on countering illegal hate speech online: First results on implementation, Factsheet*, December. Available at: https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=54300 [last accessed 10 October 2019].

European Commission (2016c) *Strategic Engagement for Gender Equality 2016-2019*, 2 June. Available at: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality/gender-equality-strategy_en [last accessed 4 December, 2019].

European Commission (2017) *Code of Conduct on countering illegal hate speech online: One year after, Factsheet*, June. Available at: https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=71674 [last accessed 10 October 2019].

European Commission (2018) *Code of Conduct on countering illegal hate speech online: Results of the 3rd monitoring exercise, Factsheet*, January. Available at: https://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=612086 [last accessed 10 October 2019].

European Commission (2019) *Code of Conduct on countering illegal hate speech online: Fourth evaluation confirms self-regulation works, Factsheet*, February. Available at: https://ec.europa.eu/info/sites/info/files/code_of_conduct_factsheet_7_web.pdf [last accessed 9 October, 2019].

Facebook (2019) *Global Feedback and Input on the Facebook Oversight Board for Content Decisions*, 27 June. Available at: https://fbnewsroomus.files.wordpress.com/2019/06/oversight-board-consultation-report-2.pdf [last accessed 1 October, 2019].

Frydman, B. and Rorive, I. (2002) 'Regulating Internet Content Through Intermediaries in Europe and the USA', *Zeitschrift für Rechtssoziologie* 23: 41-59.

Gao, C. (2017) 'China Fines Its Top 3 Internet Giants for Violating Cybersecurity Law', *The Diplomat*, September 26. Available at: https://thediplomat.com/2017/09/china-fines-its-top-3-internet-giants-for-violating-cybersecurity-law/ [last accessed 26 April 2020].

Gelber, K. (2017) 'Hate Speech—Definitions and Empirical Evidence', *Constitutional Commentary* 32: 619-629.

Gelber, K. (2019) 'Differentiating Hate Speech: A Systemic Discrimination Approach', *Critical Review of International Social and Political Philosophy* [Online First]. Available at: www.tandfonline.com/doi/full/10.1080/13698230.2019.1576006 [last accessed 05/02/2019].

Gelber, K. and McNamara, L. J. (2016) 'Evidencing the Harms of Hate Speech', *Social Identities*, 22: 324-341.

Gillespie, T. (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Global Network Initiative (GNI) (2017) 'Proposed German Legislation Threatens Free Expression Around the World', *GNI*, 20 April. Available at: https://globalnetworkinitiative.org/proposed-german-legislation-threatens-free-expression-around-the-world/ [last accessed 1 December 2019].

Global Partners Digital (2019) *An Oversight Board to Review Facebook's Content Decisions: Global Partners Digital Response*, May. Available at: https://www.gp-digital.org/wp-content/uploads/2019/05/Facebook's-Draft-Charter-An-Oversight-Board-for-Content-Decisions-GPD-Submission___.pdf [last accessed 1 October, 2019].

Greene, D. (2018) 'EFF to Court: Remedy For Bad Content Moderation Isn't to Give Government More Power to Control Speech', eff.org, 26 November. Available at:

https://www.eff.org/deeplinks/2018/11/eff-court-remedy-bad-content-moderation-isnt-give-government-more-power-control [last accessed 6 December 2019].

Heinze, E. (2006) ‘Viewpoint Absolutism and Hate Speech’, *Modern Law Review*, 69: 543–82.

Hurst, B. (2019) ‘The Online Harms White Paper: Duty of Care in the Context of Online Harms’, *Media Writes*, 17 June. Available at: https://mediawrites.law/social-media-online-harms-white-paper-series-uty-of-care-and-online-harms-duty-of-care-and-online-harms/ [last accessed 9 October, 2019].

INACH (2017) *Strategic Paper—Policy Recommendations to Combat Cyber Hate*, 30 November. Available at: http://www.inach.net/policy-recommendations-to-combat-cyber-hate/ [last accessed 22 October, 2019].

INACH (2019) The State of Policy on Cyber Hate in the EU: Bringing the Online in Line with Human Rights. Available at: http://www.inach.net/wp-content/uploads/The_State_of-_Policy_on_Cyber_Hate_in_the_EU_full_final.pdf [last accessed 1 December 2019].

ISD (2019) *Online Harms White Paper: The Duty of Care in Our Democracy*, July. Available at: https://www.isdglobal.org/wp-content/uploads/2019/07/Online-Harms-White-Paper.pdf [last accessed 6 December, 2019].

Kees, S. J. et al. (2016) Hate Crime Victim Support in Europe: A Practical Guide, RAA Saxony – Counselling Services for Victims of Hate Crimes, RAA Sachsen. Available at: https://www.equalrightstrust.org/sites/default/files/ertdocs/2016_RAA_Saxony-Hate_Crime_Victim_Support_2016_Vers.final_.pdf [last accessed 3 October, 2019].

Mackinnon, R. (2012) ‘Consent of the Networked: How can digital technology be structured and governed to maximize the good and minimize the evil?’, Slate, 30 January. Available at: https://slate.com/technology/2012/01/consent-of-the-networked-rebecca-mackinnon-explains-why-we-must-assert-our-rights-as-citizens-of-the-internet.html [last accessed 6 December, 2019].

Mehreen Khan, M. and Murgia, M. (2019) ‘EU draws up sweeping rules to curb illegal online content’, Financial Times, 24 July. Available at: https://www.ft.com/content/e9aa1ed4-ad35-11e9-8030-530adfa879c2 [last accessed 1 December 2019].

Klonick, K. (2017) ‘The New Governors of Speech: The People, Rules, and Processes Governing Online Speech’, *Harvard Law Review* 131: 1598-1620.

Langton, R. (2012) ‘Beyond Belief: Pragmatics in Hate Speech and Pornography’, in I. Maitra and M. McGowan (eds.) *Speech and Harm*: *Controversies Over Free Speech*. Oxford: Oxford University Press.

Langton, R. et al. (2012) ‘Language and Race’, in G. Russell and D. Graff Fara (eds.) *Routledge Companion to the Philosophy of Language*. London: Routledge.

Lawrence, C. (1990) ‘If He Hollers Let Him Go: Regulating Racist Speech on Campus’, *Duke Law Journal* 1990: 431-483.

Lawrence, C. (1992) ‘Cross Burning and the Sound of Silence: Anti-Subordination Theory and the First Amendment’, *Villanova Law Review* 37: 787-804.

Leyen, von der, U. (2019) *A Union that Strives for More: My Agenda for Europe*. Available at: https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf [last accessed 1 October, 2019].

Lomas, N. (2017) ‘Facebook, Twitter Still Failing on Hate Speech in Germany as New Law Proposed’, techcrunch.com, 14 March. Available at: https://techcrunch.com/2017/03/14/facebook-twitter-still-failing-on-hate-speech-in-germany/ [last accessed 6 October, 2019].

Massanari, A. (2017) ‘#Gamergate and the Fappening: How Reddit’s Algorithm, Governance, and Culture Support Toxic Technocultures’, *New Media & Society* 9: 329-346.

Matsakis, L. (2019) ‘Twitter Trust and Safety Advisers Say They’re Being Ignored’, *Wired*, 23 August. Available at: https://www.wired.com/story/twitter-trust-and-safety-council-letter/ [last accessed 6 October, 2019].

Matsuda, M. (1989b) ‘Public Response to Racist Speech: Considering the Victim’s Story’, *Michigan Law Review* 87: 2320-2381.

Miller, D. (2007) *National Responsibility and Global Justice*. Oxford: Oxford University Press.

Neller, J. (2018) ‘The Need for New Tools to Break the Silos: Identity Categories in Hate Speech Legislation’, *International Journal for Crime, Justice and Social Democracy* 7: 75–90.

Nielsen, L. B. (2002) ‘Subtle, Pervasive, Harmful: Racist and Sexist Remarks in Public as Hate Speech’, *Journal of Social Issues* 58: 265-280.

No Hate Speech Movement (2014) Starting Points for Combating Hate Speech Online: Three studies about online hate speech and ways to address it. Council of Europe. Available at: https://rm.coe.int/1680665ba7 [last accessed 29.11.2019].

Ofcom (2019) Use Of AI in Online Content Moderation. Cambridge Consultants. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf [last accessed 2 December 2019].

Packer, H. (1968) *The Limits of the Criminal Sanction*. Stanford, CA: Stanford University Press.

Parekh, B. (2005–6) ‘Hate Speech: Is There a Case for Banning?’, *Public Policy Research* 12: 213-223.

Parekh, B. (2012) ‘Is There a Case for Banning Hate Speech?’, in M. Herz and P. Molnar (eds.) *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge: Cambridge University Press.

Pollitt, C. (2013) ‘The Logics of Performance Management’, *Evaluation* 19: 346-363.

Radu, R. (2019) *Negotiating Internet Governance*. Oxford: Oxford University Press.

Saccardo, N. (2016) 'How to Stand Up to Hate Speech', *Vice*, 13 December 2016. Available at: https://www.vice.com/en_au/article/vd8y7m/confronting-online-hate-speech-share-some-good [last accessed 4 October, 2019].

Schauer, F. (2009) 'Is It Better to Be Safe than Sorry?: Free Speech and the Precautionary Principle', *Pepperdine Law Review* 36: 301-315.

Segarra Crespo, D. M. J. (2018) *Memoria Elevada Al Gobierno De S. M. Presentada Al Inicio Del Año Judicial Por El Fiscal General Del Estado, Volumen 1*.

Tsesis, A. (2002) *Destructive Messages: How Hate Speech Paves the Way for Harmful Social Movements*. New York, NY: New York University Press.

Tsesis, A. (2017) 'Campus Speech and Harassment', *Minnesota Law Review* 101: 1863-917.

Watts, K. A. (2011) 'Constraining Certiorari Using Administrative Law Principles', *University of Pennsylvania Law Review* 160: 1-68.

Wenar, L. (2007) 'Responsibility and Severe Poverty' in T. Pogge (ed), *Freedom from Poverty as a Human Right*. Oxford: Oxford University Press.

Zuckerberg, M. (2019) 'The Internet needs new rules. Let's start in these four areas', *Washington Post*, 30 March. Available at: https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html [last accessed 1, December 2019].

For some users the current Internet epoch can be considered the Internet of Hate which poses serious human rights concerns. Reflecting the scale and seriousness of the problem, innovations in governance tools for online hate have been initiated by national governments, intergovernmental organisations and Internet intermediaries across Europe in past years.

This study maps, explains and critically evaluates these emerging innovations covering three levels: moderation, oversight, and regulatory level. It reviews whether and how these innovations deliver a victim sensitive approach; uphold human rights including freedom of expression and prohibition of discrimination; and fulfil goals, aims, values and expectations of governmental agencies, Internet platforms, civil society organisations and the general public when it comes to the governance of online hate speech.

The study identifies 30 indicators that could assess the success or progress of different governance tools for online hate speech and makes many practical recommendations covering ten key areas.

PREMS 057220

ENG

www.coe.int

The Council of Europe is the continent's leading human rights organisation. It comprises 47 member states, including all members of the European Union. All Council of Europe member states have signed up to the European Convention on Human Rights, a treaty designed to protect human rights, democracy and the rule of law. The European Court of Human Rights oversees the implementation of the Convention in the member states.

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE