APC ISSUE PAPER

ONLINE CONTENT: TO REGULATE OR NOT TO REGULATE — IS THAT THE QUESTION?

By Dr. Mathias Vermeulen

ABSTRACT

Recently there have been a flurry of proposals to "regulate the internet", which in practice boils down to more narrowly regulating online content. These proposals often emerge after high-profile revelations related to the role of online intermediaries in facilitating access to illegal or undefined harmful content. In order to suggest a principles-based approach to regulation, this issue paper highlights positive and negative aspects of some recent initiatives to regulate online content, in one form or another – namely the European Union (EU) code of conduct on hate speech, the EU's reform of its audiovisual media rules, the German NetzDG law, the French approach to co-regulating social media with platforms like Facebook, and the UK online harms paper. It recommends moving towards a process-based co-regulatory approach to online content regulation, which does not make platforms liable for hosting individual pieces of content, but instead imposes a legal obligation on them to fully disclose their self-regulatory efforts to address illegal and harmful content on their services. Such disclosures should be combined with mandatory oversight by an independent regulator in order to allow for independent scrutiny of the necessity, proportionality and effectiveness of these measures.

Dr. Mathias Vermeulen is a Brussels-based independent consultant who has worked on the intersection of tech policy, human rights and new technologies for more than a decade, including for the European Parliament, the United Nations, the Mozilla Foundation and the Belgian Intelligence Oversight Committee. He holds a PhD in European privacy and data protection law. This paper took into account developments until 31 August 2019, and the views expressed are those of its author.

APC
ASSOCIATION FOR
PROGRESSIVE
COMMUNICATIONS

EXECUTIVE SUMMARY

Many proposals to regulate online content run the risk of unintentionally creating more harm than the initial harms they try to combat. Three key principles must be followed to prevent this from happening. Any legitimate intervention must (1) aim for a minimum level of intervention in accordance with the principles of necessity and proportionality in international human rights law, (2) be based on an inclusive consultation process with all relevant stakeholders, and (3) not strengthen the dominant position of the large incumbents.

Self-regulatory approaches to content regulation have turned out to be underwhelming, yet hard regulation has shown to be too prescriptive and rights-intrusive. To solve this conundrum, the EU and some of its main members are looking at co-regulatory approaches to regulate online content. While neither the EU nor European states have a monopoly on proposing new governance mechanisms to regulate online content, they do have an outsized influence on the general debate to regulate the internet. The EU is one of the few global actors that actually has the market power to force dominant tech companies to change their practices. It deliberately aims to set global standards with its tech regulation proposals.

These different initiatives try to different degrees to force social media companies to be more proactive in achieving state-determined public policy objectives. If we cherry pick the best elements from some of these initiatives, what emerges is a governance approach that focuses on the regulation of company processes rather than holding companies liable for failing to swiftly remove individual pieces of content that are illegal or cause undefined harm.

This "ex-ante" approach to regulation gives more teeth to self-regulation efforts while moving beyond harsh punitive measures that penalise unlawful behaviour only after harm has been done. It aims to prevent companies from needing to determine what is acceptable speech and it prevents excessive removal practices. Such an approach also allows for a reframing of the problem of online content regulation. A first priority for governments should be to address the amplification of harmful and illegal content, not merely prevent the appearance of such content on these platforms. Hence, this paper recommends regulating the behaviour of platform-specific architectural amplifiers of illegal or harmful content: recommendation engines, search engine features such as autocomplete, features such as "trending", and other optimisation mechanisms that predict what we want to see next. These are active design choices over which platforms have direct control, and for which they could ultimately be held liable.

This approach reserves the right of platforms to determine how they promote, demote, monetise, demonetise or take any other procedural measure regarding content on their platforms. But this prerogative cannot be unchecked any longer: an independent regulator should be able to assess the effectiveness of these procedural measures against a set of statutory objectives that go beyond simplistic content-related benchmarks such as removal rates and response times.

This governance framework can only work on one condition: it requires total transparency from the platforms vis-à-vis an independent regulator. The regulator should have the power to demand any type of granular information that is necessary for it to fulfil its supervisory tasks, and it should have the power to impose fines or other corrective actions when platforms do not provide that information in a timely manner.

INTRODUCTION

Over the past year there have been an increasing number of proposals to "regulate the internet". French President Emmanuel Macron's speech at the 2018 Internet Governance Forum in Paris called for regulation while positing a binary vision of the internet – an unregulated "California" internet vs. a rigidly regulated "Chinese" internet.[1] Facebook CEO Mark Zuckerberg has also called for government-initiated regulation while at the same time launching a process to create an oversight board.[2] Self-regulation is also being discussed by civil society actors who have proposed a social media council to oversee content moderation practices.[3] Despite the fact that content is already regulated, offline and online, in many jurisdictions through existing defamation, hate speech or counter-terrorism laws, new internet-specific laws are also becoming commonplace, from laws on so-called "revenge porn" to laws addressing hate speech online.

These initiatives make it clear that both the "internet" and "regulation" can mean many things to many different

[1] Macron, E. (2018, 12 November). Speech by M. Emmanuel Macron, President of the Republic at the Internet Governance Forum. https://www.elysee.fr/en/emmanuel-macron/2018/11/12/speech-by-m-emmanuel-macron-president-of-the-republic-at-the-internet-governance-forum

[2] Zuckerberg, M. (2019, 29 March). Mark Zuckerberg: The Internet needs new rules. Let's start in these four areas. *The Washington Post*. 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; see the final charter at https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf

[3] ARTICLE 19. (2019, 11 June). Social Media Councils: Consultation. https://www.article19.org/resources/social-media-councils-consultation

stakeholders. "Hard" regulation has been useful in the past to protect the openness of the open internet's core infrastructure and to promote freedom of speech. Some legislators and states have created binding rules to prevent unjustified interference in this architecture (either by governments or commercial parties) such as net neutrality laws or preventing calls by security agencies to create backdoors in encryption protocols.[4] However, policy makers from around the world are typically not talking about "the internet" as a set of connected networks based on standardised communication protocols that needs to be preserved and protected against governmental overreach.

Instead, the open internet is now often being equated with a small number of online intermediaries – in particular Facebook, and the multiple platforms it owns – which act as closed, walled gardens that are antithetical to the open nature of the internet. A range of high-profile revelations related to the role of these platforms in facilitating access to a diverse range of illegal or undefined harmful content has created a strong appetite for regulating those services. At the same time, many of these proposals remain limited to "virtue signalling" and do not materialise in concrete new governance mechanisms.

A typical example is the Christchurch Call to Action,[5] which was adopted in the wake of the livestreamed Christchurch attacks in March 2019, and signed by 18 governments and eight companies. The call was subsequently endorsed by another 31 governments, as well as the Council of Europe and UNESCO. The non-binding statement of intent includes vague calls for action to "eliminate terrorist and violent extremist content online", including by "enhancing technology".[6] New Zealand Prime Minister Jacinda Ardern stressed the "unprecedented approach" in which tech companies and countries "committed to an action plan to develop new technologies to make our communities safer."[7] The United States refused to sign the accord on the grounds that it violated constitutional free speech protections.[8]

The appetite to adopt actual new regulatory initiatives has been particularly strong in Europe, which is the main geographical focus of this paper. While neither the EU nor European states have a monopoly on proposing new governance mechanisms to regulate online content, they do have an outsized influence on the general debate to "regulate the internet". Hence, European proposals for content regulation deserve particular scrutiny. There are three main reasons for doing this.

Firstly, the EU is one of the few global actors that actually has the market power to force dominant tech companies that operate on a global scale to change their practices, which many individual countries lack. When these changes do occur, they seep into global corporate practices which have an impact on users worldwide.

Secondly, the EU, and some of its largest members, deliberately aim to set global standards with their tech regulation proposals – and they are not shy about it. Frans Timmermans, first vice president of the European Commission, stated in July 2019 that it is "essential for us to shape the global field for the development of the technological revolution."[9] The EU actively promotes the General Data Protection Regulation (GDPR) as the gold standard for other countries' data protection regulations, including – for better or worse – through its trade agreements. France used its chairmanship of the G7 in 2019 to sell its co-regulatory approach towards content regulation to other member states, while the United Kingdom's proposal to counter online harms boldly aims to create a "global coalition of countries" that are all "taking coordinated steps to keep their citizens safe online."[10] In fact, regulation has become Europe's key export product in the tech industry. The disruptive effect of many internet applications took a lot of legislators, officials and traditional industries by surprise. Code effectively became law,[11] which hit a nerve among stakeholders who are used to being in the driving seat when it comes to making regulation. From rule makers they became rule takers; the EU's current plans to develop rules to regulate the deployment of artificial intelligence[12] demonstrate that there is no appetite to make that same mistake again.

[4] Schaake, M., & Vermeulen, M. (2016). Towards a values-based European foreign policy to cybersecurity. *Journal of Cyber Policy*, *1*(1), 75-84 https://www.tandfonline.com/doi/abs/10.1080/23738871.2016.1157617

[5] https://www.christchurchcall.com/christchurch-call.pdf

[6] In September 2019, an advisory network was created to help shape the implementation of the Call in a human rights-protective manner. APC. (2019, 24 September). APC to Christchurch Call leaders: Human rights must be at the core of efforts to combat violent extremist and terrorist content. *APCNews*. https://www.apc.org/en/node/35698

[7] Scott, M., Momtaz, R., & Kayali, L. (2019, 15 May). Macron, Ardern lead call to eliminate online terrorist content. *Politico*. https://www.politico.eu/article/christchurch-call-emmanuel-macron-jacinda-arden-facebook-google-twitter-extreme-harmful-content

[8] Warzel, C. (2019, 16 May). The World Wants to Fight Online Hate. Why Doesn't President Trump? *The New York Times*. https://www.nytimes.com/2019/05/16/opinion/christchurch-online-extremism-trump.html

[9] European Commission. (2019, 24 July). General Data Protection Regulation shows results, but work needs to continue. https://europa.eu/rapid/press-release_IP-19-4449_en.htm

[10] Secretary of State for Digital, Culture, Media and Sport and Secretary of State for the Home Department. (2019). *Online Harms White Paper*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

[11] See Lawrence Lessig's classic *Code and Other Laws of Cyberspace*, originally written in 1999, at http://codev2.cc

[12] Kayali, L. (2019, 18 July). Next European Commission takes aim at AI. *Politico*. https://www.politico.eu/article/ai-data-regulator-rules-next-european-commission-takes-aim

Finally, other actors that used to be influential in setting norms for internet regulation have lost considerable influence. The US Congress has failed to adopt any tech-related big ticket bill in the past five years, and its hands-off approach towards content regulation is seen as untenable by European governments, who are wary that the rules of the game "will evolve in ways determined by, and in the interests of, these companies."[13] Or, as President Macron stated during his speech at the Internet Governance Forum in November 2018, "if we do not regulate Internet [sic], there is the risk that the foundations of democracy will be shaken."[14]

This paper will first highlight how national content regulation laws outside the EU have often focused on criminalising speech itself, with great consequences for the protection of freedom of speech. Then it will describe a small number of self-regulatory approaches that have been considered, and their potential to stave off government-driven regulation. Finally, we will assess both negative and positive new co-regulatory initiatives in the EU to regulate online content, with a focus on the EU code of conduct on hate speech, the EU's reform of its audiovisual media rules, the German NetzDG law, the French approach to co-regulating social media, and the UK online harms paper, in order to distill a principles-based approach to regulate the core issue behind many calls to "regulate the internet": the amplification of harmful or illegal content online.

NATIONAL CONTENT REGULATION LAWS

A number of countries are using legitimate concerns about the proliferation of online misinformation and hate speech to deepen their control over their citizens. These legislative initiatives have some shared similarities: they give discretionary powers to executive bodies to decide whether a piece of content is false or misleading, and give these bodies the power to issue fines, corrections or even hand out prison sentences for creating, publishing or disseminating individual pieces of content. In these cases, it is the creators, disseminators and publishers of disinformation who are the main targets of regulation, not online intermediaries. Such regulatory initiatives often limit the essence of the right to freedom of expression to such a degree that the right itself is in jeopardy.

In Egypt, ordinary citizens are treated as publishers in the eyes of the law when they post disinformation or spread hate speech. Article 19 of the 2018 Egyptian Media and Press Law grants the Supreme Media Council the authority to "suspend any personal website, blog, or social media account that has 5,000 followers or more if it posts fake news, promotes violence, or spreads hateful views."[15] Bloggers can be subjected to prosecution for publishing false news or incitement to break the law. Egypt's public prosecutor reportedly even set up a hotline for citizens to report "fake news and rumours".[16]

Malaysia adopted the Anti-Fake News Law in 2018, which provides prison sentences for up to 10 years for knowingly creating, distributing or publishing "fake news", defined to include "news, information, data and reports" that are "wholly or partly false." The law also applies to individuals or organisations operating outside of the country if the "fake news" concerns Malaysia or affects Malaysian citizens.[17] Human Rights Watch claims that the bill was deliberately defined vaguely to allow maximum discretion for the government to target critics of the ruling party and the government.[18]

In June 2019, Singapore adopted its Protection from Online Falsehoods and Manipulation Act, which empowers any Singaporean government minister to issue a range of corrective directions, including fines and prison sentences, against online "falsehoods" deemed to be against the public interest.[19] Government ministers will also be able to issue "general correction orders" to online intermediaries to remove or "correct" content that the government disagrees with. The bill states further that false statements cannot be transmitted to users in Singapore through the internet, or through systems "that enable the transmission through a

[13] House of Lords Select Committee on Communications. (2019). *Regulating in a digital world*. https://publications.parliament.uk/pa/ld201719/ldselect/ldcomuni/299/299.pdf

[14] Macron, E. (2018, 12 November). Op. cit.

[15] Sadek, G. (2018, 6 August). Egypt: Parliament Passes Amendments to Media and Press Law. *The Library of Congress*. https://www.loc.gov/law/foreign-news/article/egypt-parliament-passes-amendments-to-media-and-press-law

[16] Michaelson, R. (2018, 27 July). 'Fake news' becomes tool of repression after Egypt passes new law. *The Guardian*. https://www.theguardian.com/global-development/2018/jul/27/fake-news-becomes-tool-of-repression-after-egypt-passes-new-law

[17] In April 2018, a Danish citizen was fined and sentenced to one week in prison for posting a video criticising the police's response to a targeted killing in Kuala Lumpur. Human Rights Watch. (2018, 29 March). Malaysia: Drop Proposed 'Fake News' Law. https://www.hrw.org/news/2018/03/29/malaysia-drop-proposed-fake-news-law

[18] https://www.hrw.org/world-report/2019/country-chapters/malaysia

[19] In Singapore, "ministers presented their approach as being of a kind with moves in Europe." Guest, P. (2019, 19 July). Singapore Says It's Fighting 'Fake News.' Journalists See a Ruse. *The Atlantic*. https://www.theatlantic.com/international/archive/2019/07/singapore-press-freedom/592039

mobile network" of text and multimedia messages.[20] This potentially gives the government power to police encrypted chat apps like Signal and WhatsApp, although it is unclear how the government would get access to such content.

Sri Lanka is the latest nation in this region that is contemplating the adoption of amendments to the penal and criminal procedure codes to criminalise the dissemination of "false news" where it is deemed to affect "communal harmony" or "state security".[21]

These criminalisation efforts could be seen as one of the "worst threats" for freedom of speech, since they do not limit government discretion in a manner that distinguishes between lawful and unlawful expression.[22] History is filled with examples of regimes that apply criminal provisions to quash dissent and criticism, including against journalists and human rights defenders. As a reminder, Article 19(2) of the International Covenant on Civil and Political Rights establishes states parties' obligations to respect and ensure the right "to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice." Any restriction to this right needs to pass a three-part test. The restriction must be provided through a law that is "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly."[23] Restrictions must only be imposed to protect legitimate aims, which can include the protection of other people's right to freedom of speech. And the measure must be necessary to achieve that aim.

Given the civil liberty issues at stake, any state-driven regulatory intervention to regulate online content should be subject to additional precautions. Any intervention must be based on an inclusive consultation process with all relevant stakeholders and not strengthen the dominant position of the large incumbents by creating barriers for new entrants to the market.

One approach to regulation of online content that has been rejected on this basis by courts, UN bodies and civil society groups is the general monitoring of content. States should not impose a general obligation on companies to indiscriminately monitor information that they transmit or store – both from a fundamental rights and a competition perspective. Firstly, such an obligation would have a significant impact on the right to privacy and the right to data protection of the users of that service, as such monitoring would involve "the identification, systematic analysis and processing of information connected with the profiles created on the social network by its users."[24] Secondly, such a system could potentially undermine freedom of information, since it might not distinguish adequately between unlawful content and lawful content, with the result that its introduction could lead to the blocking of lawful communications. Finally, such an obligation would also have an impact on a company's freedom to conduct its business, since it would require that a provider install a potentially costly monitoring system at its own expense.

Unfortunately, we are witnessing a worrying trend in the EU that goes against this direction, by leaving platforms no choice but to install so-called "upload filters" to implement a legal provision, which would amount to pre-publication censorship. Article 17 of the EU's new copyright directive[25] makes platforms directly liable for the content that is uploaded by their users, if they cannot or will not be able to pay licensing fees from rights owners. The law states that the application of the article "shall not lead to any general monitoring obligation," but in practice it is hard to imagine in what other way service providers can ensure copyrighted works are not made available without proper licensing.[26] Similarly, the

[20] https://www.parliament.gov.sg/docs/default-source/default-document-library/protection-from-online-falsehoods-and-manipulation-bill10-2019.pdf

[21] AFP. (2019, 5 June). Sri Lanka proposes new law on fake news after Easter attacks. *France 24*. https://www.france24.com/en/20190605-sri-lanka-proposes-new-law-fake-news-after-easter-attacks

[22] In general, see the Report of the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35 at https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23218&LangID=E; another approach is the French law against the "manipulation of information", which created a legal injunction allowing the "circulation of fake news to be swiftly halted." A judge would interpret the definition of "fake news" on the basis of three criteria: the "fake news" must (1) be manifest, (2) be disseminated deliberately on a massive scale, and (3) lead to a disturbance of the peace or compromise the outcome of an election. See https://www.gouvernement.fr/en/combating-the-manipulation-of-information

[23] UN Human Rights Committee. (2011). General comment No. 34, article 19: Freedoms of opinion and expression, CCPR/C/GC/34. https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf

[24] Court of Justice of the European Union, C-360/10, Sabam v. Netlog (2012).

[25] https://eur-lex.europa.eu/eli/dir/2019/790/oj

[26] As Aleksandra Kuczerawy summarises: "To effectively recognize infringing content, a technological tool must be used to examine all newly uploaded content on the platform and comparing it with an existing database. This amounts to installing upload filters by the service providers and systematic monitoring of the entirety of the users' content. Despite repeated attempts to convince the broad public that the Copyright in DSM Directive was not meant to introduce upload filters, several officials admitted, soon after the vote, that the filters are unavoidable." Kuczerawy, A. (2019, 10 July). To Monitor or Not to Monitor? The Uncertain Future of Article 15 of the E-Commerce Directive. *Centre for IT and IP Law*. https://www.law.kuleuven.be/citip/blog/to-monitor-or-not-to-monitor-the-uncertain-future-of-article-15-of-the-e-commerce-directive

EU is currently discussing the adoption of a regulation to prevent the "dissemination of terrorist content online". According to Article 6 of the proposal, service providers are expected to take "proactive measures", including using "reliable technical tools" to identify terrorist content. The proposal still needs to be further negotiated between the European institutions.

SELF-REGULATORY ALTERNATIVES TO CONTENT REGULATION LAWS

Given the dangers of national content regulation laws outlined above, some civil society advocates have promoted self-regulatory alternatives. Self-regulatory approaches are voluntary arrangements initiated and undertaken by a particular company, sector or industry, which are not formally sanctioned or endorsed by governmental actors. Companies usually take voluntary action to redress a policy concern or stave off more onerous government regulation. These actions typically rely on self-reporting, and the only penalty that is often available is exclusion from an association or other professional body that has developed the standards. Self-regulation has an immediate effect on the behaviour of the regulated company, while prescriptive regulation can take years to materialise and implement (especially in the EU).

A typical example of self-regulation is press councils, which evaluate complaints of unethical or wrongful media behaviour or journalistic reporting. Press councils are typically set up by an association and governed by a code of ethics for journalists, and cases are considered by juries that consist of journalists. In the same vein, the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has expressed support[27] for ARTICLE 19's idea to create an independent social media council that would enable industry-wide complaint mechanisms for content moderation decisions and remedies for improper removals that violate freedom of expression standards. These initiatives aim to prevent excessive censorship rather than prevent exposure to harmful or illegal content.[28] Questions around scope, jurisdiction and funding of such a council are still being scrutinised, but its proponents stress that signing up to such a mechanism would not create legal obligations, making it self-regulatory in nature.[29]

Unsurprisingly, the online intermediaries themselves have been looking for self-regulatory solutions. Facebook has announced the creation of an oversight board for content decisions, which will review Facebook's "most challenging content decisions". The board will be able to reverse Facebook decisions about whether to allow or remove certain posts on its platforms based on a set of values which will include "concepts like voice, safety, equity, dignity, equality and privacy". Importantly, the board will not decide cases where reversing Facebook's decision would violate the law.[30] These self-regulatory initiatives are laudable, but they can only function in a broader co-regulatory framework that provides a "transparency-forcing function",[31] and which has the power to impose sanctions if these self-regulatory initiatives do not live up to their expectations. In general, leaving content decisions entirely in the hands of private companies leaves us with the worst of both worlds: it incentivises companies to remove excessive amounts of content, and it leaves our fundamental rights in the hands of private companies.

However, governments are increasingly critical about the potential of self-regulatory initiatives as a mechanism to regulate the activities of online intermediaries. The UK parliament stated that "self-regulation by online platforms which host user-generated content, including social media platforms, is failing."[32] When Germany was dissatisfied with the outcomes of a code of conduct on hate speech, it created new legal obligations to achieve the same goal (see below).

As a result, governments are increasingly looking to new co-regulatory responses to these content-related challenges. These initiatives try – to different degrees – to force social media companies to be more proactive in achieving state-determined public policy objectives. This "ex-ante" approach to regulation seeks to give more teeth to self-regulatory efforts, while simultaneously moving beyond harsh punitive measures that penalise unlawful behaviour only after harm has been done. These approaches generally also aim to prevent companies from determining what is acceptable speech as well as to prevent excessive removal practices.

[27] Kaye, D. (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35. https://freedex.org/wp-content/blogs.dir/2015/files/2018/05/G1809672.pdf

[28] See also https://cddrl.fsi.stanford.edu/global-digital-policy-incubator//content/social-media-councils-concept-reality-conference-report

[29] ARTICLE 19. (2019, 11 June). Op. cit.

[30] This paper was drafted before Facebook's final charter came out. See https://fbnewsroomus.files.wordpress.com/2019/01/draft-charter-oversight-board-for-content-decisions-2.pdf

[31] Douek, E. (2019, 18 May). Two Calls for Tech Regulation: The French Government Report and the Christchurch Call. *Lawfare*. https://www.lawfareblog.com/two-calls-tech-regulation-french-government-report-and-christchurch-call

[32] House of Lords Select Committee on Communications. (2019). Op. cit.

NEW APPROACHES TO CO-REGULATION

Under a system of co-regulation, the regulatory role is traditionally shared between government and industry. Typically, a representative group of industry representatives formulate a code of practice or a code of conduct in consultation with a governmental actor, with breaches of the code usually punishable in some way. Depending on the model, these sanctions can be imposed by a professional industry body or a governmental regulator. In theory, this approach allows an industry to take the lead in the regulation of its members by setting standards and encouraging greater responsibility for performance. It could lead to significantly greater levels of compliance, as industries become co-monitors, while it also encourages participants to see good industry-wide performance as a common goal. However, in order to overcome potential risks of regulatory capture and anti-competitive activities (as regulatory barriers to entry can be developed by incumbents), a proper regulatory design that focuses on transparency and follows specified regulatory principles is needed to guide the development of the codes.[33]

EU CODE OF CONDUCT ON HATE SPEECH

In December 2015, the European Commission launched the EU Internet Forum, which brought large internet platforms, Commission officials and civil society together to develop "a joint, voluntary approach to detect and address harmful material online."[34] After terrorist attacks in Brussels in March 2016, the Internet Forum was transformed into the creation of a Code of Conduct against hate speech online.[35] The Code was presented as a "voluntary" commitment, and did not include legal sanctions for non-compliance, which appears to make it a pure self-regulation instrument. However, the parties did commit themselves to review the majority of valid notifications for removal of illegal hate speech "in less than 24 hours" and to remove or disable access to such content. Furthermore, the companies agreed to produce reports that would mainly assess the implementation of this public commitment on a regular basis. Since the creation of the Code, four "naming-and-shaming" evaluations took place which use two simple benchmarks for success: the amount of notifications that were assessed within 24 hours, and the removal rates of the companies.[36]

Civil society organisations were particularly critical about three aspects of the Code.[37] Platforms are encouraged to interpret "illegal" hate speech in a uniform manner across all EU member states, yet there is a wide variety in national hate speech laws. Secondly, the code lacks any due process requirements and does not include any safeguards against potential abuse of the notice procedure, nor does it offer remedies for wrongful removals. Ironically, the Code of Conduct would never pass the EU's own best-practice Principles for Better Self- and Co-Regulation.[38] Non-governmental organisations such as EDRi and Access Now were dismayed that they were "systematically excluded from the negotiations."[39]

GERMAN NETZDG LAW

In the midst of the refugee crisis in 2015, Germany witnessed an increasing amount of hate speech online against religious minorities, including defamation of religious institutions and public incitement to violence, which are all illegal under German law. This trend prompted the German Ministry of Justice and Consumer Protection to set up a task force with the largest social media providers and a number of civil society organisations. The task force resulted in a code of conduct wherein the platforms committed to improve their moderation procedures and remove hate speech quickly upon notice.

However, the German government was not satisfied with the first results of the code of conduct, claiming that "too much illegal content" remained nevertheless available on the sites.[40] The government quickly introduced the Network Enforcement Act – also known as the NetzDG law. The law requires providers of large social networks

[33] See in general Brown, I., & Marsden, C. T. (2013). *Regulating Code: Good Governance and Better Regulation in the Information Age*. MIT Press.

[34] European Commission. (2015, 3 December). EU Internet Forum: Bringing together governments, Europol and technology companies to counter terrorist content and hate speech online. https://europa.eu/rapid/press-release_IP-15-6243_en.htm

[35] European Commission. (2016, 31 May). European Commission and IT Companies announce Code of Conduct on illegal online hate speech. https://europa.eu/rapid/press-release_IP-16-1937_en.htm

[36] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en

[37] ARTICLE 19. (2016). *EU: European Commission's Code of Conduct for Countering Illegal Hate Speech Online and the Framework Decision*. https://www.article19.org/data/files/medialibrary/38430/EU-Code-of-conduct-analysis-FINAL.pdf; McNamee, J. (2016, 3 June). Guide to the Code of Conduct on Hate Speech. *EDRi*. https://edri.org/guide-code-conduct-hate-speech; Bukovská, B. (2019). *The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An analysis of freedom of expression implications*. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. https://www.ivir.nl/publicaties/download/Bukovska.pdf

[38] https://ec.europa.eu/digital-single-market/sites/digital-agenda/files/CoP%20-%20Principles%20for%20better%20self-%20and%20co-regulation.pdf

[39] EDRi. (2016, 31 May). EDRi and Access Now withdraw from the EU Commission IT Forum discussions. https://edri.org/edri-access-now-withdraw-eu-commission-forum-discussions

[40] https://ec.europa.eu/growth/tools-databases/tris/en/search/?trisaction=search.detail&year=2017&num=127

(with more than two million users located in Germany) to set up an effective and transparent procedure for handling user complaints about content that violates existing provisions of German criminal law. Essentially, platforms must determine whether these complaints are valid or not. "Blatantly illegal content" must be removed within 24 hours of notice, while "other illegal content" requires deletion within a week of notice. This aspect of the law attracted widespread criticism because the government was seen as outsourcing the traditional judicial authority for determining criminality to the private sector. Given the stringent time requirement, and the steep fine for lack of compliance (up to EUR 50 million), the law was said to incentivise platforms to err on the side of taking down flagged content even if that content is not manifestly criminal.[41]

NetzDG also introduced some transparency mechanisms. It imposes an obligation on platforms that receive more than 100 complaints to publish semi-annual reports that detail their content moderation procedures, including statistical information about the number of complaints, the number of content removal decisions and the amount of personnel and other resources that were dedicated to content moderation. Unfortunately, the first transparency reports that came out have been criticised for their "low informative value",[42] which in turn does not allow for a proper evaluation of the effect of the law. Interestingly, in July 2019, Facebook was fined precisely because it did not meet the transparency requirements of the NetzDG law. The government claimed that Facebook made it easier for users to complain that a post violated the platform's community standards as opposed to making a complaint under NetzDG. Germany's Federal Office of Justice said that by tallying only certain categories of complaints, Facebook "had created a skewed picture of the extent of violations on its platform."[43] The NetzDG experiment in Germany makes it clear that legislators need to carefully think through how an effective transparency mechanism could work in practice. Access to granular data and standardised reporting formats are crucial to meaningfully assess the effect of any co-regulatory response.[44]

EU REFORM OF RULES ON AUDIOVISUAL MEDIA

The EU's Audiovisual Media Services Directive (AVMSD) was originally designed for broadcasters. It included rules to protect minors from viewing any "harmful" content that can disrupt their physical, mental and moral development as well as rules that protect all citizens from illegal content such as child sexual abuse material, terrorist content and hate speech. It only applied to media services providers that had editorial responsibilities, meaning those who have effective control over the selection and organisation of programmes on their channels.

When the AVMSD was reformed in 2018, regulators decided to include video-sharing services such as YouTube, where professional and non-professional users alike upload a wide variety of content. YouTube does not exercise editorial control in the traditional meaning of the word; its key difference from traditional media is precisely that users can upload what they see fit. Is this a typical example of regulators trying to retrofit a regulatory model that worked in one sector (TV) and apply it to a different sector with its own rules and (technical) specificities? It is[45] – but with a twist.

Article 28b of the AVMSD is based on the following logic: services like YouTube are responsible for the organisation of the content on their services[46] and hence

[41] See for instance Human Rights Watch. (2018, 14 February). Germany: Flawed Social Media Law. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law or Kuczerawy, A. (2017, 30 November). Phantom Safeguards? Analysis of the German law on hate speech NetzDG. *Centre for IT and IP Law*. https://www.law.kuleuven.be/citip/blog/phantom-safeguards-analysis-of-the-german-law-on-hate-speech-netzdg

[42] Heldt, A. (2019, 12 June). Reading between the lines and the numbers: an analysis of the first NetzDG reports. *Internet Policy Review*. https://policyreview.info/articles/analysis/reading-between-lines-and-numbers-analysis-first-netzdg-reports

[43] Escritt, T. (2019, 2 July). Germany fines Facebook for under-reporting complaints. *Reuters*. https://www.reuters.com/article/us-facebook-germany-fine/germany-fines-facebook-for-under-reporting-complaints-idUSKCN1TX1IC

[44] For a good analysis see Tworek, H., & Leerssen, P. (2019). *An Analysis of Germany's NetzDG Law*. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

[45] Before the conclusions of the negotiations, seven EU member states warned that "it seems obvious that many of these services cannot be expected by an end-user to be regulated similarly to audiovisual media services." These countries noted: "The impracticability of regulating any small platform, of regulating livestreaming and of determining whether a platform or social media service carry a significant proportion of audiovisual content is obvious and seems to border the realm of the impossible." Joint non-paper from the Czech, Danish, Finnish, Irish, Luxembourg, Dutch and Swedish delegations on the scope of the Audiovisual Media Services Directive – Revised Presidency compromise text amending Directive 20110/13/EU (AVMS), April 2017.

[46] While the text acknowledges that a "significant share of the content provided on video-sharing platform services is not under the editorial responsibility of the video-sharing platform provider, [...] those providers typically determine the organisation of the content, namely programmes, user-generated videos and audiovisual commercial communications, including by automatic means or algorithms." See also the French white paper: "Yet through their ordering of published content and moderation policies, social networks have the ability to take direct action against the worst abuses to prevent or respond to them and thereby limit the damage to social cohesion." Republic of France. (2019). *Creating a French framework to make social media platforms more accountable: Acting in France with a European vision*. https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf

they need to protect their consumers through proactively adopting in their terms of service a number of rules that protect 1) minors from harmful content, and 2) all users from hate speech, terrorist content and other forms of illegal content. Users need to be able to flag or report such content to the video-sharing service. The service needs to implement age verification or rating and control systems, establish "transparent, easy-to-use and effective" procedures to resolve user complaints, and provide media literacy tools.[47]

The controversial[48] legislation was critically received, with most commentators focusing on the fact that the legislator has basically outsourced its regulating role to video-sharing services, which goes against a basic principle of constitutional democracies. Juan Barata from Stanford University stated that the AVSMD "assigns to private actors a new de facto role as interpreters and enforcers of the most sensitive and impactful rules affecting freedom of expression in European member States' legal systems."[49] The new legislation indeed allowed services like YouTube to take stricter measures on a voluntary basis.[50] In practice, this will boil down to video service providers interpreting how issues such as "incitement to hatred" should be understood "to the appropriate extent" within the meaning of the EU's legal framework on hate speech.[51] It remains to be seen whether such a legal construction will pass a fundamental rights test, since the restriction of speech is neither accessible nor predictable.[52]

The text also forces member states to make out-of-court dispute settlements available between users and video-sharing platforms, and ultimately allows users to assert their rights before a court in relation to decisions taken by video-sharing platform providers. In the EU, a national (media) regulator is responsible for overseeing the processes that video-sharing platforms will set up. In the best case scenario, as Lubos Kuklis, chair of the European Regulators Group for Audiovisual Media Services said, this may "create an environment where users are not only protected from the harmful content of other users, but also from overbearing or arbitrary intrusions by the platform itself."[53] However, the problem with this approach is that the platform can always say it removed content on the basis of a violation of its terms of service, rather than because it violated the new AVMSD.

It remains to be seen how European member states – and companies – will implement the new rules. They have until September 2020 to decide, and the text allows them to impose stricter rules to achieve the goals of the legislation.[54] This can go both ways: countries can demand more take-downs of content, with regulators enforcing sanctions when companies fail to take action against dubiously legal content; or companies can face sanctions if they do not provide granular transparency on content moderation decisions, including on individual cases.

THE FRENCH APPROACH TO CO-REGULATING SOCIAL MEDIA

In the first experiment of its kind, 10 French officials from various ministries started to inspect Facebook's internal processes for managing content in January and February 2019. Without receiving access to "truly confidential information", the embedded regulators looked at how flagging works, how Facebook identifies problematic content, how Facebook decides if such content is problematic or not, and what happens when Facebook takes down a post, a video or an image.

After the experiment, a non-binding report commissioned by the French government summarised the main findings

[47] Article 28b3, AVMS. https://eur-lex.europa.eu/eli/dir/2018/1808/oj

[48] Plucinska, J. (2017, 18 May). Parliamentary committee's tech power grab. *Politico*. https://www.politico.eu/article/parliament-committee-power-grab-crowds-out-most-meps

[49] Barata, J. (2018, 24 October). The new Audiovisual Media Services Directive: Turning video hosting platforms into private media regulatory bodies. *Center for Internet and Society*. https://cyberlaw.stanford.edu/blog/2018/10/new-audiovisual-media-services-directive-turning-video-hosting-platforms-private-media

[50] Recital 49, AVMS. https://eur-lex.europa.eu/eli/dir/2018/1808/oj

[51] Recital 17, AVMS. https://eur-lex.europa.eu/eli/dir/2018/1808/oj

[52] EDRi stated that this reform can "lead to censorship of perfectly legal material online." See also EDRi. (2016). *Position Paper: Proposed revision of the Audio-Visual Media Services Directive*. https://edri.org/files/AVMSD/edrianalysis_20160713.pdf and https://edri.org/avmsd-reform-document-pool

[53] Kuklis, L. (2018, 29 November). European regulation of video-sharing platforms: What's new, and will it work? *LSE Media Policy Project*. http://blogs.lse.ac.uk/mediapolicyproject/2018/11/29/european-regulation-of-video-sharing-platforms-whats-new-and-will-it-work; for a more critical assessment, see Barata, J. (2019, 18 February). New EU rules on video-sharing platforms: Will they really work? *Center for Internet and Society*. https://cyberlaw.stanford.edu/blog/2019/02/new-eu-rules-video-sharing-platforms-will-they-really-work

[54] Article 28a6. The Irish regulator has published a first attempt at transposing the rules in a national framework; see Broadcasting Authority of Ireland. (2019, 24 June). BAI publishes submission on regulation of harmful online content / implementation of new Audiovisual Media Services Directive. https://www.bai.ie/en/bai-publishes-submission-on-regulation-of-harmful-online-content-implementation-of-new-audiovisual-media-services-directive

of the experiment. Refreshingly, the report focuses on the infrastructural issues that underpin content-related problems. It states that "the inadequacy of [Facebook's] moderation systems and the lack of transparency of their platforms' operation justify intervention by the public authorities," while recognising that "any state intervention must be strictly necessary, proportionate and transparent whenever it affects public freedoms that are as important as the freedom of expression and freedom of communication." It further points out that a pure self-regulatory approach is neither adequate nor credible, largely because of "the extreme asymmetry of information" between social networks on the one hand, and public authorities and civil society on the other.[55]

The report recommends a co-regulatory approach to content regulation that takes inspiration from banking supervisory authorities. These authorities do not punish a bank when it becomes clear that a customer used their account for unlawful purposes such as money laundering or financing terrorism, but they do intervene when a bank has not taken sufficient due diligence measures to prevent their infrastructure from being used for illegal purposes.[56] Applied to large online intermediaries, this approach does not call for penalties over individual failures to police content by the platforms. Instead, it urges much more regular audits and more transparency on internal processes for handling illegal, and perhaps even harmful, content by an independent supervisor. Macron called this at the time an "innovative experiment" which he "would like us to spread."[57]

THE UK ONLINE HARMS WHITE PAPER

In April 2019, the UK's Department for Digital, Culture, Media and Sport (DCMS) and Home Office jointly launched an ambitious proposal for tackling "online harms".[58] The UK government plans to establish a mandatory duty of care "to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services." In practice, this means that a wide range of actors (from social media companies and public discussion forums to retailers that allow users to review products online, along with non-profit organisations, file-sharing sites and cloud hosting providers) would be under a duty to take "reasonable steps" to keep their users safe. The obligations that this overarching duty would impose would be further developed by an independent regulator that could produce codes of practice which should "outline the systems, procedures, technologies and investment, including in staffing, training and support of human moderators, that companies need to adopt to help demonstrate that they have fulfilled their duty of care to their users."[59] The independent regulator would be able to issue fines, block access to websites, and even impose liability on individual members of the companies' senior management.

The white paper deserves credit for starting a conversation about the right way forward to potentially address online harms. If implemented properly, its co-regulatory structure could be effective. However, its current lack of detail, and the extraordinarily wide scope of the "online harms" concept, give pause for thought. According to the proposal, "online harms" includes clearly illegal content such as child sexual abuse material, but also a wide category of content "that may not cross the criminal threshold but can be particularly damaging to children or other vulnerable users"[60] including activities such as "intimidation" or "cyberbullying". Ultimately, it might even include content that "threatens our way of life in the UK, either by undermining national security, or by reducing trust and undermining our shared rights, responsibilities and opportunities to foster integration."[61]

CONCLUSION AND RECOMMENDATIONS

Regulating "the internet" in a smart way boils down to two rules of thumb. Rule one: governments need to preserve the integrity of the core technical infrastructure of the internet by ensuring the protection of net neutrality and encryption, which allows people to communicate with each other in a secure way. Rule two: governments should not be afraid to step in and force online intermediaries, as users of this infrastructure, to live up to universal standards of human rights law. Regulation can be done right if the basic reasons for liberal democracies to adopt legislation are kept in mind: the need to uphold the rule of law online as well as offline, the need to ensure fair, open and competitive markets, and the need to ensure that universal human rights are upheld.

Given the human rights and social and political freedoms issues at stake, any state-driven intervention to regulate content online should be subject to particular precautions. Any legitimate intervention must (1) aim for a minimum level of intervention in accordance with the principles of necessity and proportionality, (2) be based on an inclusive

[55] Republic of France. (2019). Op. cit.

[56] Ibid.

[57] Macron, E. (2018, 12 November). Op. cit.

[58] Secretary of State for Digital, Culture, Media and Sport and Secretary of State for the Home Department. (2019). Op. cit.

[59] Ibid.

[60] Ibid.

[61] Ibid.

consultation process with all relevant stakeholders, and (3) not strengthen the dominant position of the large incumbents by creating barriers for new entrants to the market. Many regulatory proposals to counter online harms run the risk of unintentionally creating more harm than the initial harms they try to combat.

Self-regulatory approaches have turned out to be underwhelming, yet hard regulation has shown itself to be too prescriptive and restrictive of rights, and at the same time ill suited to the technologies it is meant to regulate. Hence, European governments are increasingly looking at co-regulatory approaches which contain some promising developments, but which also have the potential to have a detrimental effect on freedom of expression and related rights – within the EU and globally.

If we cherry pick the best elements from some of the approaches outlined above, what emerges is a governance approach that focuses on the regulation of company processes rather than content. This governance structure requires platforms to develop processes that ensure a systematic approach to controlling and minimising well-defined risks, and which take as a starting point the precautions outlined above. Shifting scrutiny towards these processes would help address some of the causal factors that give rise to illegal and harmful content online, without unduly interfering with individuals' rights and the open architecture of the internet.

This approach steers us away from the disproportionate attention that is given to the removal of individual pieces of content. Rather than trying to regulate the impossible, i.e. removal of individual pieces of content that are illegal or cause undefined harm, we need to focus on regulating the behaviour of platform-specific architectural amplifiers of such content: recommendation engines, search engine features such as autocomplete, features such as "trending", and other mechanisms that predict what we want to see next. These are active design choices over which platforms have direct control, and for which they could ultimately be held liable. Regulating these architectural elements is a more proportionate response than regulating content as such. Renee DiResta has summarised succinctly the differences between both approaches: "free speech does not mean free reach. There is no right to algorithmic amplification."[62] The European Court of Human Rights might agree with this position, as it has earlier suggested that restricting the scale, extent and quantity of dissemination of lawful speech is justifiable.[63]

This approach reserves the right of platforms to determine how they promote, demote, demonetise – or take any other procedural measure[64] – regarding content on their platforms. Digital service providers are likely to prove more effective in identifying hazards and developing technical solutions than a central regulatory authority. But this prerogative cannot be unchecked any longer: an independent regulator should be able to assess the effectiveness of these procedural measures against a set of statutory objectives that go beyond simplistic content-related benchmarks such as removal rates and response times. The regulator should have a foundational mandate to respect and indeed vindicate individuals' rights online, and platforms should always have recourse to judicial review to challenge disproportionate regulatory action.

Also, this governance framework can only work on one condition: it requires total transparency from the platforms vis-à-vis the regulator. The regulator should have the power to demand any type of granular information that is necessary for it to fulfil its supervisory tasks, and it should have the power to impose fines or other corrective actions when platforms do not provide that information in a timely manner. The French approach towards transparency in its white paper is a good first step in the right direction.

As Peter Pomerantsev has pointed out:

> Get the regulatory approach right and it will help formulate rights and democracy in a digital age; get it wrong and it will exacerbate the very problems it is trying to solve, and play into the games of authoritarian regimes all agog to impose censorship and curb the free flow of actual information across borders.[65]

A clearly defined, principles-based approach that is aligned with the co-regulatory approach discussed above would ensure that platforms address online harms as a key operational objective, but in a way which is reflective of their reach, their technical architecture, their resources, and the risk such content is likely to pose.

[62] DiResta, R. (2018, 30 August). Free Speech Is Not the Same As Free Reach. *Wired*. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach

[63] European Court of Human Rights, Satakunnan Markkinapörssi Oy and Satamedia Oy v. Finland (App. no. 931/13), 27 June 2017.

[64] This includes, for example, notice and action processes, and processes to provide remedies for wrongful take-downs.

[65] Pomerantsev, P. (2019, 10 June). How (Not) to Regulate the Internet. *The American Interest*. https://www.the-american-interest.com/2019/06/10/how-not-to-regulate-the-internet

APC
ASSOCIATION FOR
PROGRESSIVE
COMMUNICATIONS

Internet and ICTs for social justice and development

APC is an international network of civil society organisations founded in 1990 dedicated to empowering and supporting people working for peace, human rights, development and protection of the environment, through the strategice use of information and communication technologies (ICTs).

We work to build a world in which all people have easy, equal and affordable access to the creative potential of ICTs to improve their lives and create more democratic and egalitarian societies.

www.apc.org info@apc.org

ONLINE CONTENT: TO REGULATE OR NOT TO REGULATE – IS THAT THE QUESTION?
OCTOBER 2019

ISBN 978-92-95113-14-5 APC-201910-GAPS-IP-EN-DIGITAL-303