# 16

# Digital Constitutionalism: In Search of a Content Governance Standard

**EDOARDO CELESTE, NICOLA PALLADINO, DENNIS REDEKER AND KINFE MICHEAL YILMA***

## I. Introduction

One of the main issues of global content governance on social media relates to the definition of the rules governing online content moderation.[1] One could think that it would be sufficient for online platforms to refer to existing human rights standards. However, a more careful analysis shows that international law only provides general principles, which do not specifically address the context of online content moderation, and that a single human rights standard does not exist. Even identical provisions and principles are interpreted by courts in different ways across the world. This is one of the reasons why, since their inception, major social media platforms have set their own rules, adopting their own peculiar language, values and parameters. Yet, this normative autonomy too has raised serious concerns. Why should private companies establish the rules governing the biggest public forum for the exchange of ideas? Is it legitimate to depart from minimal human rights standards and impose more stringent rules?

The current situation exposes a dilemma for online content governance. On the one hand, if social media platforms simply adopted international law standards, they would be compelled to operate a choice on which standard to

* This work is the output of a project funded by Facebook Research. The authors conducted their research independently and their findings were subject to external peer reviewing. No review by Facebook Research or its associates took place. We thank the participants to the Conference 'Constitutionalising Social Media' (Dublin, 20–21 May 2021) for their helpful comments on an earlier draft of this chapter as well as the peer reviewers for their constructive criticism.

[1] For a definition of content moderation, see T Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, 2018); S Myers West, 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms' (2018) 20 *New Media & Society* 4366; on platform governance, see R Gorwa, 'What Is Platform Governance?' (2019) 22 *Information, Communication & Society* 854.

follow – for example, between the US freedom of expression-dominated approach or the European standard, which balances freedom of expression with other social values. Moreover, they would also need to put in place a mechanism able to translate, or 'operationalise' such general standards in the context of online content moderation. On the other hand, where social media platforms adopt their own values, rules and terminology to regulate content moderation, thus departing from international law standards, they are accused of censorship or laxity, intrusiveness or negligence.

The core issue and key to solving this dilemma lies in the capacity to define principles and values for online content governance, a task which is part of the broader process of constitutionalising the digital society.[2] This chapter aims to contribute to disentangle this Gordian knot. Firstly, we will clarify to what extent international law standards may provide useful guidance in the context of online content moderation and what their limitations are (section II). Secondly, we will examine a source of normative standards that has been so far neglected by the scholarship: civil society impulses. Over the past few years, a series of initiatives have emerged at societal level, and especially among civil society groups, including NGOs, think tanks, academia, trade unions, and grassroot associations, to articulate rights and principles for the digital age. The output of these efforts mostly consists of non-legally binding declarations, often intentionally adopting a constitutional tone and therefore termed 'Internet bills of rights'.

Our chapter aims to understand what social media platforms' online content governance rules can learn from these civil society initiatives. A peculiarity of these documents lies indeed in their surfacing outside traditional institutionalised constitutional processes. They can be considered as expressing the 'voice' of global communities that struggle to propose an innovative constitutional message within traditional institutional channels: one of the layers of the complex process of constitutionalisation that is pushing towards reconceptualising core constitutional principles in light of the challenges of the digital society in a new form of 'digital constitutionalism'.[3] We have collected a dataset of 40 documents and performed an empirical analysis of their content, looking specifically at how these declarations have articulated the rights and principles related to online content governance (section III). The chapter will then conclude with a case study focusing on Facebook's online content moderation rules, examining to what extent they reflect or depart from international and civil society standards (section IV).

[2] See E Celeste, 'Digital Constitutionalism: A New Systematic Theorisation' (2019) 33 *International Review of Law, Computers & Technology* 76.

[3] See D Redeker, L Gill and U Gasser, 'Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights' (2018) 80 *International Communication Gazette* 302; C Padovani and M Santaniello, 'Digital Constitutionalism: Fundamental Rights and Power Limitation in the Internet Eco-System' (2018) 80 *International Communication Gazette* 295; Celeste (ibid).

# II. International Standards

Recent years have seen a growing attention in the potential of international law in offering normative guidance to address human rights concerns in content governance.[4] Partly in response to pressure from civil society groups, including through the launch of Internet bills of rights that advance progressive content governance standards, social media platforms are also increasingly turning attention to international human rights law.[5] This section considers the extent to which international law offers such guidance to the complex world of platform content governance.

## A. Generic International Law Standards Applicable to Content Governance

The first set of international law standards applicable to content governance are general in scope and formulation. One such generic standard concerns human rights provisions that define the scope and nature of state human rights obligations.[6] The International Covenant on Civil and Political Rights (ICCPR), a human rights treaty widely ratified by states, provides the general framework for any consideration of content governance in international law. One way it does so is by defining the scope of state obligations vis-à-vis Covenant rights. States generally owe two types of obligations under the Covenant: negative and positive obligations.[7] States' negative obligation imposes a duty to 'respect' the enjoyment of rights. As such, it requires states and their organs to refrain from any conduct that would impair the enjoyment of rights guaranteed in the Covenant. States' positive obligations, on the other hand, impose a duty to 'protect' the exercise of rights. This obligation thus concerns state regulation of third parties including private actors to ensure respect for Covenant rights.

Secondly, one finds little-explored norms that would potentially apply to non-state actors, including social media companies directly. One is the Preamble of the Universal Declaration of Human Rights (UDHR), which – at the highest level – states that 'every organ of society' shall strive to promote respect for rights

[4] See M Lwin, 'Applying International Human Rights Law for Use by Facebook' (2020) 38 *Yale Journal on Regulation Online Bulletin* 53.

[5] See Facebook's Corporate Human Rights Policy (2021), available at about.fb.com/wp-content/uploads/2021/04/Facebooks-Corporate-Human-Rights-Policy.pdf. Note that almost all the decisions handed down thus far by Facebook's Oversight Board have drawn upon international human rights standards. See the decisions of the board at oversightboard.com/decision.

[6] For an analysis of the role of states in social media regulation, see the contributions in Chapter III of this volume.

[7] Art 2(1) ICCPR.

guaranteed in the Declaration.[8] Scholars argue that the reference to 'every organ of society' is said to include the duty of companies to 'respect' human rights.[9] This preambular proviso finds some concrete expression in international human rights law in the form of prohibition of abuse of rights. International law bestows no right upon anyone including 'groups and persons' as well as states to impair or destruct the enjoyment of the rights guaranteed in the Declaration and the Covenant.[10] This abuse of rights prohibition arguably would also apply to social media companies in a sense that their policies and practices, including those relating to content moderation, must not have the effect of impairing or destructing the enjoyment of human rights. In that sense, there is a negative obligation to 'respect' human rights which requires them to refrain from measures that would affect the enjoyment of rights. Indeed, the tendency in these provisions to address private actors directly – albeit generically – appears to be at odds with the state-centred nature of human rights law generally. But the sheer fact that these provisions appear to impose binding duties, regardless of how they would be enforced, would certainly add weight to recent scholarly arguments that international human rights law does, or should, apply to content moderation practices of platforms.[11]

## B. Content Governance Standards in Human Rights and Principles

Content governance standards in international law are also to be found in the catalogue of human rights and principles. First, human rights law prohibits certain types of speech: war propaganda,[12] advocacy for racial, religious and national hatred[13] and racist speech.[14] In outlawing certain types of expression, international human rights law sets forth content governance standards that must be implemented by state parties to the relevant treaties, including in social media platforms. Social media companies are not bound by such international standards, but they may ban or restrict such types of speech to comply with national law or volitionally as they now do in practice.[15] Second, human rights law not

[8] Preamble, para 8 UDHR.

[9] See L Henkin, 'The Universal Declaration at 50 and the Challenges of Global Markets' (1999) 25 *Brooklyn Journal of International Law* 17, 25.

[10] Art 30 UDHR; Art 5(1) ICCPR.

[11] See L McGregor et al, 'International Human Rights Law as a Framework for Algorithmic Accountability' (2020) 68 *International Comparative Law Quarterly* 309.

[12] See Art 20(1) ICCPR.

[13] Art 20(2) ICCPR.

[14] Art 4 of the International Convention on the Elimination of All Forms of Racial Discrimination (21 December 1965).

[15] See M Bickert (Facebook), 'Updating the Values That Inform Our Community Standards' (12 September 2019), available at about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards.

only guarantees the right to freedom of expression but also provides standards for permissible restrictions, namely, legality, necessity and legitimacy.[16]

Third, in addition to freedom of expression, content governance engages a broad range of human rights guaranteed in international law. Common acts of content moderation would normally limit freedom of expression of users. But other human rights and principles such as the right to equality/non-discrimination, right to effective remedy, right to fair hearing, right to honour and reputation and freedom of religion are also impacted by platform content moderation policies and practices. The right to 'enjoy the arts and to share in scientific advancements and its benefits' is another set of socio-economic rights that relates to content governance.[17] This 'right to science and culture' is aimed at enabling all persons who have not taken part in scientific inventions to participate in enjoying its benefits.[18] This provision has barely been invoked in practice, but it arguably would apply to counter aggressive content moderation practices of platforms vis-à-vis copyrighted material. As shall be highlighted in the next section, 'freedom from censorship', including the right not to be subjected to onerous copyright restrictions, is one of the content moderation-related standards proposed by civil society groups.

The 'right to science' has been interpreted to embody the right of individuals to be protected from the adverse effects of scientific inventions and the right to public participation in decision-making about science and its uses.[19] And this, as we will see, comes closer to civil society content governance standards that envisage duties on social media platforms to prevent harm and safeguard vulnerable social groups on social media platforms as well as the need to ensure meaningful participation in the development of policies.

## C. International Soft Law on Content Governance

The UN Guiding Principles on Business and Human Rights (UNGPs, alternatively referred to as the Ruggie Principles) are a potential source of specific international content governance standards. The Ruggie Principles are currently the only international instrument that seeks to address the conduct of businesses and its impact on human rights (with some limits).[20] Not only do they primarily affirm

[16] See Art 19(3) ICCPR.

[17] Art 27(1) UDHR; Art 15(1)(b) of the International Covenant on Economic, Social and Cultural Rights (16 December 1966).

[18] R Adalsteinsson and P Thörhallson, 'Art 27', in G Alfreðsson and A Eide (eds), *The Universal Declaration of Human Rights: A Common Standard of Achievement* (Martinus Nijhoff Publishers, 1999) 575–78.

[19] Report of the Special Rapporteur in the field of Cultural Rights, Farida Shaheed, on the Right to Enjoy the Benefits of Scientific Progress and Its Applications, UN Doc A/HRC/20/26 (14 May 2012) Paras 25, 43–44; see also Committee on Economic, Social and Cultural Rights, General Comment No. 25: On Science and Economic, Social and Cultural Rights, UN Doc E/C.12/GC/25 (30 April 2020) para 74.

[20] OHCHR, Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework, HRC Res 17/4 (16 June 2011).

states as the sole and primary duty-bearers in human rights law[21] but they are also couched in general principles. And this makes them less suited to the complex world of content moderation. It is, however, vital to note that the scope of the Ruggie Principles is defined in a manner to apply to businesses of all types and sectors.[22] This means that the principles would, theoretically, apply to social media platforms.

In a series of reports, the UN Special Rapporteur on Freedom of Expression, David Kaye, has sought to adapt the Ruggie Principles to the social media context.[23] While such elaborative reports provide useful intellectual guidance, they are legally non-binding, meaning that compliance by platforms would be entirely voluntary. Yet, this attests to the fact that a process of adaptation of international law standards, including the Ruggie Principles, to the social media context, is in the process of development.[24] Added to frequent exhortations of civil society groups on the normative value of the Ruggie Principles, emerging attempts at translating the UNGPs to the digital context may potentially contribute to this evolution of international standards on content governance.

Relatively progressive content moderation-related international standards are emerging through the work of Kaye. He has been part of a coalition of intergovernmental mandates on freedom of expression which, through Joint Declarations, outlined progressive standards on content moderation.[25] The Joint Declarations constitute international soft law that tend to unpack general free speech standards envisaged in the ICCPR as well as regional human rights treaties. The 2019 Resolution, for instance, states in its Preamble that the primary aim of the Joint Declarations is one of 'interpreting' human rights guarantees thereby 'providing guidance' to, inter alia, governments, civil society organisations and the business sector.[26] It further provides that the Joint Declarations have – over the years – 'contributed to the establishment of authoritative standards' on various aspects of free speech.[27]

Intermediary liability is one of the content governance-related themes addressed with some details in the Joint Declarations.[28] Current international hard law on free speech does not address the role of intermediaries, such as social media platforms, in curating and moderating content online. Intermediaries play a key role in the enjoyment of the right to freedom of expression online which makes appropriate regulation of their conduct an imperative. The objective of fair

[21] ibid Part I.

[22] Ruggie Principles (n 20) Part II, para 14.

[23] See Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, UN Doc HRC/38/35 (6 April 2018).

[24] See McGregor et al, 'International Human Rights Law' (2020) 326.

[25] See OSCE, 'Joint Declarations', available at www.osce.org/fom/66176?page=1.

[26] See Twentieth Anniversary Joint Declaration: Challenges to Freedom of Expression in the Next Decade (UN, OSCE, OAS & ACHPR, 10 July 2019) Preamble, para 3.

[27] ibid Preamble, para 4.

[28] See also the Joint Declaration on Freedom of Expression and 'Fake News', Disinformation and Propaganda (UN, OSCE, OAS & ACHPR, 3 March 2017).

intermediary liability, then, is to define the exceptional circumstances where intermediaries would be held liable for problematic content of their users. In filling the normative void in international law, the 2019 Joint Declaration in particular offers some international standards on intermediary liability. On top of the overarching 'responsibility' of intermediaries to 'respect human rights', the Declaration stipulates the principle to put in place clear and predetermined content moderation policies that are adopted after consultation with users, in line with what is called in international human rights law the requirement of legality, the requirement to institute minimum due process guarantees, such as prompt notification to users whose content may be subjected to content action and avenues by which users may challenge impending content actions.

Soft law generally offers authoritative interpretation of high-level principles of international hard law, but the approach in the Joint Declaration raises questions of form and substance in international law. One such question is whether a soft human rights instrument – drawing upon a human rights treaty – can directly address private actors that are not party to the treaty. But this point goes beyond the scope of this chapter.

### D. The Role of International Standards

In conclusion, despite the recent turn to international human rights law for content governance standards, it appears to offer little guidance. It is uncertain, for instance, as to how it would apply to platform content governance. As shown above, this is mainly because international human rights law is – by design – state-centred and hence does not go far in attending to human rights concerns in the private sector. This is exacerbated by the characteristically generic formulation of international human rights standards which – in turn – make them less suited to the world of platform content moderation. The complex, voluminous and routine nature of content moderation requires a rather granular and dynamic system of norms. What is more, the generic international content governance standards have not adequately been unpacked by relevant adjudicative bodies, such as the United Nations Human Rights Committee, to make them fit for purpose to the present realities of content moderation. By and large, content governance jurisprudence at the international level remains thin. However, international law's primary value, thus far, has been to provide the overarching normative framework, on which recent progressive civil society content governance standards build, as we will explore in the next section.

## III. Civil Society Initiatives

This section analyses a selected sample of 40 Internet bills of rights developed by civil society groups, and articulating rights and principles addressing online

content governance.[29] Table 16.1 provides for a synthetic overview of the principles we detected in the corpus, grouped into three major categories. The first one collects all the provisions explicitly concerned with international human rights law compliance. The other two categories distinguish respectively between substantive and procedural standards. 'Substantive standards' refer to people's rights and responsibilities related to the creation and publication of content on the Internet. 'Procedural principles' indicate formal rules and procedures through which substantive rights shall be exercised and enforced, ie the rules through which decisions about users' contents are made, including the rulemaking process itself.

**Table 16.1** Civil society initiatives

| Categories | No of documents | Principles included |
|---|---|---|
| General compliance with human rights standards | 19 | |
| **Substantive principles** | **39** | |
| Freedom of expression | 38 | Freedom from censorship, freedom from copyright restriction, freedom of religion |
| Prevention of harm | 16 | Harassment, cyberbullying, defamation, incitement to violence, cybercrime, human dignity |
| Protection of social groups | 13 | Non-discrimination of marginalised group, discriminating content, hate speech, children rights and protection |
| Public interest | 6 | Public health or morality, public order and national security, fake news and disinformation, protection of infrastructure layer |
| Intermediary liability | 9 | Full immunity, conditional liability, intermediaries are liable in the case of actual knowledge of infringing content, intermediary are liable when fail to comply with adjudicatory order |
| **Procedural principles** | **32** | |
| Rule of law | 24 | Legality, legal certainty, rule of law, judicial oversight, legal remedy, necessity and proportionality |

*(continued)*

[29] See Annex 1, available at bit.ly/3pPmBmr.

**Table 16.1** ***(Continued)***

| Categories | No of documents | Principles included |
|---|---|---|
| Good governance principles | 19 | Transparency, accountability, fairness, participation, multistakeholderism, democratic governance |
| Platform-specific principles | 21 | Notification procedures, human oversight, human rights due diligence, limitations to automated content moderation, informed consent, right to appeal and remedy |

## A. Substantive Principles

Our analysis shows a high degree of consistency among the documents analysed. A consensus emerges on a shared set of principles to be applied to content governance, which, taken together, outline a coherent normative framework. Civil society declarations draw extensively on human rights law. Half of our sample refers explicitly to one of the international human rights law instruments discussed in the previous section (especially ICCPR, HDHR, Ruggie Principles), or more generally advocates for the respect of human rights standards. Moreover, even when not referred to explicitly, the documents we analysed provide for rights and principles drawn from the international human rights literature.

Within this framework, freedom of expression emerges as the core principle and main concern of civil society when dealing with content moderation in the context of digital constitutionalism initiatives. Not only is it the most quoted principle, but most of the other rights and interests emerge within discussions about free speech and its limitations, and are clearly presented as subordinate to freedom of expression. Three categories of substantive principles, respectively related to the prevention of online and offline harm, the protection of social groups and public interest, set the boundaries of what civil society considers acceptable derogation to the principle of freedom of expression, and which in turn justifies content removal. Sixteen documents refer to the protection of individuals from potential harms including harassment, cyberbullying, incitement to violence, damage to reputation and dignity. Thirteen documents call into question principles aiming at protecting minorities or vulnerable groups from troubling content, or content that constitute incitement to hatred, violence or discrimination. Six documents refer to some articulations of public interest, such as protection of national security, public health, morals, or more recently fake news and disinformation. It is worth noting that these kinds of permissible restrictions reflect those foreseen by international law, and in particular Article 19 of UDHR and ICCPR.

## B. Procedural Principles

In relation to substantive principles, civil society initiatives limit themselves to reiterate general international human rights standards in the context of social media platforms. Conversely, as regards procedural principles, Internet bills of rights perform a more articulated 'generalisation and respecification' of international standards in light of the peculiar nature of the social media environment.[30] The procedural principles advanced by civil society can be divided into three subcategories. Firstly, a series of principles address states and online platforms, and can be collected under the label of 'rule of law' principles. Civil society organisations require, on the one hand, governments to establish a legal framework providing certainty and predictability for content moderation, and on the other hand, platforms to execute content removal requests from states only when provided by law. As a minimum, civil society require states to:

(1) define freedom of expression restrictions and what constitutes illegal content through democratic processes and according to international human rights law standards, adopting in particular the tripartite test of legality, necessity and proportionality
(2) clearly establish under which conditions intermediaries are deemed responsible for user-generated content, and which kind of actions they must undertake
(3) guarantee appropriate judicial oversight over content removal and the right to legal remedy.

The main concern here is that constitutional safeguards and rule of law are circumvented by outsourcing online content moderation adjudication and enforcement to private entities. Furthermore, civil society groups in particular claim that states must not impose a 'general monitoring obligation' to intermediaries.[31] Human rights defenders fear that encouraging a 'proactive' content moderation will lead to 'over-removal of content or outright censorship'.[32] In doing so, civil society groups recall the warnings advanced by the UN Special Rapporteur David Kaye in his 2018 report on the promotion and protection of the right to freedom of opinion and expression.[33]

[30] See E Celeste, 'Terms of Service and Bills of Rights: New Mechanisms of Constitutionalisation in the Social Media Environment?' (2019) 33 *International Review of Law, Computers & Technology* 122; see also G Teubner, *Constitutional Fragments: Societal Constitutionalism and Globalization* (Oxford University Press, 2012).

[31] Access Now, '26 Recommendations on Content Governance' (2020). See also The Manila Principles (2015).

[32] Access Now (2020) 13. See also, among others, Association for Progressive Communication, 'Content Regulation in the Digital Age' (2018); Art 19 The Universal Declaration of Digital Rights (2017); EDRI, 'The Charter of Digital Rights' (2014); Declaration on Freedom of Expression In response to the adoption of the Network Enforcement Law (2017).

[33] D Kaye, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression – A/HRC/38/35' (June 2019), available at documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf?OpenElement.

The second category refers to a series of 'good governance' standards that apply both to states and private companies. They include the principles of transparency, accountability, fairness and participatory decision-making. However, the most interesting group of procedural principles is the third one, which we termed 'platforms-specific principles'. It is indeed here that civil society contextualises and adapts general international human rights standards into more granular norms and rules to be implemented in the platform environment.

This last group includes six procedural principles:

(1) *Certainty and predictability*. As discussed above in relation to transparency, platforms are asked to provide certainty and predictability through accessible terms of service and community standards, including well-defined and transparent decision-making processes. These principles articulate rule of law standards with specific reference to social media platforms.
(2) *Appeal and remedy*. Fourteen civil society documents affirm a right to appeal and remedy. This principle partially overlaps with the right to legal remedy mentioned in the previous category, but in this case specifically focuses on private companies' practices. According to the Santa Clara Principles, 'Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension', then describing what the 'minimum standards for a meaningful appeal' are.[34] Companies are also requested to provide remedies, such as 'restoring eliminated content in case of an illegitimate or erroneous removal', 'providing a right to reply', 'issuing apologies or corrections' or 'providing economic compensation'.[35]
(3) *Notification procedures*. Social media companies should provide notice to each user whose contents have been subject to content moderation decisions. This notification must include at least all relevant details about the content removed, which provision of the Terms of Service breached, how it was detected and an explanation of the user's possibilities to appeal the decision.[36]
(4) *Limitations to automated content moderation*. Companies should limit the use of automated content moderation systems to well-defined manifestly illegal content. Platforms should provide clear and transparent policies for automated content moderation. Users should be informed about the usage of automated systems. Companies should provide for human oversight or human review of automated decisions and should adopt an approach to minimize human rights risks by design.
(5) *Human rights due diligence or impact assessment*. Social media platforms should scrutinise on an ongoing basis their policies, products, services with the consultation of third-party human rights experts in order to evaluate their

[34] Santa Clara Principles on Transparency and Accountability in Content Moderation (2018).
[35] 'Access Now' (2020) 40.
[36] On notification requirements see also Access Now (n 31); and Association for Progressive Communication, 'Content Regulation' (2018).

impact on human rights. Companies are also called to share information and data with researchers and civil society organisations and to support independent research.

(6) *Independent self-regulation bodies*. Some civil society organizations call for the establishment of independent self-regulatory bodies, following the example of press councils or ethics committees. The most developed attempt in this regard is provided by the NGO Article19, which outlines the model of a Social Media Council.[37]

## IV. A Comparison with Facebook's Community Standards

Facebook is a global social media platform, offering its services across jurisdictions. Its transnational nature implies that content moderation on Facebook generates a variety of normative conflicts, among users and between users and state actors. Since, in principle, no single cultural or legal standard can be applied to decide how to reconcile these normative conflicts, Facebook resorted to a solution in its scale 'unique to communication practices in the digital age':[38] it set its own rules for a communication space used by almost 1.8 billion daily users.[39] The norms that guide Facebook's content moderation practices – its Community Standards – are perhaps the most important example of 'platform law' today.[40] The Community Standards are described as a 'living set of guidelines'[41] governing what is allowed and prohibited on the platform. These guidelines are made and remade by the company's Policy Team in a process that brings in academics and other experts in the field, but that failed to be grounded in a 'popular' vote by the platform's users.[42]

Following the articulation of the previous two sections which examined substantive and procedural principles of international law and civil society declarations in the context of online content moderation, this section assesses how

[37] ARTICLE 19, 'Self-Regulation and "Hate Speech" on Social Media Platforms' (2018), available at www.article19.org/wp-content/uploads/2018/03/Self-regulation-and-'hate-speech'-on-social-media-platforms_March2018.pdf; cf also UK Government, 'Online Harms White Paper' (2019) 46, available at assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/973939/Online_Harms_White_Paper_V2.pdf.

[38] See MC Kettemann and W Schulz, 'Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study' (January 2020), available at www.hans-bredow-institut.de/uploads/media/default/cms/media/k0gjxdi_AP_WiP001InsideFacebook.pdf.

[39] See Statista, 'Number of Daily Active Facebook Users Worldwide as of 4th Quarter 2020', available at www.statista.com/statistics/346167/facebook-global-dau.

[40] See LA Bygrave, *Internet Governance by Contract* (Oxford University Press, 2015); Kaye, 'Report of the Special Rapporteur' (2019).

[41] See Facebook, 'Writing Facebook's Rulebook', available at about.fb.com/news/2019/04/insidefeed-community-standards-development-process.

[42] See Kettemann and Schulz, 'Setting Rules for 2.7 Billion' (2020); N Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4 *Social Media + Society* 1.

Facebook's content governance standards measure up in these two dimensions, starting with a discussion of substantive principles.

## A. Substantive Principles Entailed in the Community Standards

Facebook's Community Standards are informed by the organisation's self-proclaimed 'core values', which emphasise creating 'a place for expression and giving people voice'.[43] What limits free expression are four values – authenticity, safety, privacy and dignity – and the application of copyright rules and national law. The focus on freedom of expression is absolutely consistent with the emphasis on this principle by civil society, as we observed in section III of this chapter.

The Community Standards are structured into six chapters, of which the first four outline restrictions on content based on Facebook's core values; chapter V affirms intellectual property and its protection, whereas chapter VI outlines procedural questions and refers to the Oversight Board. The first four chapters currently entail a total of 23 principles defining content that must not be posted on the platform.[44] These 23 substantive principles align rather well with the principles we found in civil society documents. Specifically, addressing the three categories of 'prevention of harm', 'protection of social groups' and 'public interest' as outlined above, this section considers which of these categories is most often used to justify limitations on the chief principle of freedom of expression. Facebook issues content moderation transparency reports[45] covering 12 of the 23 principles associated with these three categories of limitations of freedom of expression.[46] Two principles – prohibitions on fake accounts and on spam – are not associated with the three most important categories of civil society demands. Table 16.2 indicates each category from the civil society documents and the corresponding Community Standards principles on which transparency reporting occurs.[47] It also shows the number of content takedowns in 2020 and the share of those takedowns appealed by users.[48] Finally, the table shows the shares of content takedowns triggered by

[43] See Facebook, 'Community Standards', available at www.facebook.com/communitystandards.

[44] If content is posted that violates these principles, it is deleted (unless a successful appeal is lodged). Recently, academics have proposed to keep copies of violating content in a 'poison cabinet'. See J Bowers, E Sedenberg and J Zittrain, 'Platform Accountability Through Digital "Poison Cabinets"' (Knight First Amendment Institute at Columbia University, 13 April 2021), available at knightcolumbia.org/content/platform-accountability-through-digital-poison-cabinets.

[45] See Facebook, 'Community Standards Enforcement Report', available at transparency.facebook.com/community-standards-enforcement.

[46] Facebook provides separate transparency reports for its actions concerning content that is covered by copyright rules (although also referenced in the Community Standards), See Facebook, 'Intellectual Property', available at transparency.facebook.com/intellectual-property, and, concerning legal requests by states, see Facebook, 'Content Restrictions Based on Local Law', available at transparency.facebook.com/content-restrictions.

[47] As of early May 2021.

[48] This data is for Facebook only: Instagram has its own, very similar but not identical, transparency reporting based on the same standards ('Facebook and Instagram share content policies. This means

automatic recognition, as opposed to flagging done by humans, in the last quarter of 2020.

**Table 16.2 Content moderation by category derived from civil society documents, 2020**[49]

| Category/principle | Number of content actions | % appealed | % of automation (Q4, 2020) |
|---|---|---|---|
| **Prevention of harm** | **73.3m** | **0.2%** | – |
| – Bullying and harassment | 14.5m | 0.3% | 48.8% |
| – Suicide and self-injury | 6.4m | 0.8% | 92.8% |
| – Organised hate | 19.1m | 0.2% | 98.3% |
| – Terrorism | 33.3m | 0.1% | 99.8% |
| **Protection of social groups** | **332.3m** | **1.1%** | – |
| – Child nudity and sexual exploitation of children | 35.9m | 0.1% | 98,8% |
| – Adult nudity and sexual activity | 139.9m | 1.7% | 98.1% |
| – Violent and graphic content | 75.5m | 0.1% | 99.5% |
| – Hate speech | 81m | 1.5% | 97.1% |
| **Public interest** | **27.8m** | **1.2%** | – |
| – Regulated goods: firearms | 5.1m | 5.5% | 92.2% |
| – Regulated goods: drugs | 22.7m | 0.2% | 97.3% |
| **Not categorised** | **12.0bn** | **0,0%** | – |
| – Fake accounts | 5.8bn | – | 99.6% |
| – Spam | 6.2m | 0.0% | 99.8% |

The data reveals that, in 2020, most removal actions occurred to safeguard specific groups (around 332 million). More than 40 per cent of these actions relate to adult nudity and sexual activity, ostensibly protecting minors (Facebook generally allows accounts for anyone over 13 years old) and those who 'may be sensitive to this type of content'.[50] The category 'prevention of harm' counted for more than 73 million content removal actions in 2020, of which the removal of terrorist content makes up almost half (45.4 per cent). An association of categories from the previous section with the 12 principles Facebook reports on appears difficult with regard to the category of terrorism, which could also have been coded as public interest issue (specifically referring to the civil society principle of

if content is considered violating on Facebook, it is also considered violating on Instagram'.) See Facebook, 'Community Standards Enforcement Report' (n 45).

[49] See Facebook, 'Community Standards Enforcement Report' (n 45).

[50] See Facebook, 'Adult Nudity and Sexual Activity', available at www.facebook.com/communitystandards/adult_nudity_sexual_activity.

'public order and national security'), rather than a matter of harm prevention. However, such is the case with all illegal activities that bring in the state as an interested person *qua* the nature of criminal law. Content moderation based on the 'public interest' category identified above made up only about 28 million cases in 2020, with a clear concentration on identifying posts designed to 'purchase, sell or trade non-medical drugs, pharmaceutical drugs and marijuana' compared to a lower number of actions against content related to the 'purchase, sale, gifting, exchange and transfer of firearms'[51] (only accounting for about 18 per cent of this category). A great number of content actions are not associated with the demands that we found in civil society declarations to limit free speech (see the category 'not categorised' above): in 2020, 12 billion actions were taken by Facebook to remove fake accounts or spam.[52]

Overall, the substantive standards found in the 40 civil society documents analysed in the previous section appear well captured by the Community Standards. However, on the level of individual principles, ie below the broader categories, some substantial differences can be identified. Freedom of expression, the most-often cited value, is at times extended to the freedom from copyright restrictions when advocated for by civil society. Facebook's Community Standards, being less aspirational – and encompassing binding copyright rules – explicitly emphasise the protection of intellectual property and dedicate a separate chapter to it. The company states that protection of intellectual property is 'important to promoting expression, creativity and innovation in our community'.[53]

With regard to standards in international law, Facebook's Community Standards appear, once again, particularly conducive toward enabling and protecting free expression, based on its relatively limited list of exceptions. Its preamble stresses this focus on the provision of 'voice'. At the same time, considering the billions of content removals that occurred on the platform, one clearly sees that Facebook takes on the role of making and enacting rules to protect other important rights and principles, too. Some of these principles cover norms articulated in the UDHR and the ICCPR: for instance, with regard to prohibitions on advocacy for racial, religious and national hatred, the Community Standards expressly prohibit content that constitutes 'a direct attack against people on the basis of … protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease'.[54] However, while international law includes protections for groups, not merely individuals, Facebook's Community Standards, in practice, focus on protecting individuals

[51] See Facebook, 'Regulated Goods', available at www.facebook.com/communitystandards/regulated_goods.

[52] Here, fake accounts relate to the lack of an authentic person or organisation setting them up, unrelated to whether these accounts are engaged in spreading fake news/misinformation.

[53] See Facebook, 'Respecting Intellectual Property', available at m.facebook.com/communitystandards#!/communitystandards/respecting_intellectual_property; see L Lessig, 'Copyright's First Amendment Melville B. Nimmer Memorial Lecture' (2001) 48 *UCLA Law Review* 1057.

[54] See Facebook, 'Hate Speech', available at www.facebook.com/communitystandards/hate_speech.

with these attributes rather than the respective group. More recently, Facebook's moderation practices shifted toward a stronger emphasis on groups that are particularly vulnerable – in contrast to protecting all groups equally, including, for instance, white male Americans.[55]

In 2021, Facebook created a corporate human rights policy to demonstrate that it is 'committed to respecting human rights in [their] business operations, product development, policies and programming'.[56] In this document the company specifically refers to the UNGPs and vows to regularly report on how it addresses human rights. The company sees its human rights commitment as an extension of its work as part of the Global Network Initiative (GNI), the multistakeholder group behind the GNI Principles on Freedom of Expression and Privacy, included in the dataset of civil society declaration in the previous section. As part of the policy, next to public reporting on its human rights due diligence and the establishment of the independent Oversight Board, Facebook aims to alter 'key content policies, including creating a new policy to remove verified misinformation and unverifiable rumours that may put people at risk for imminent physical harm'.[57] The latter is a reaction to its assessment of human rights impacts in Sri Lanka, based on the framework of the UNGPs.[58]

## B. Procedural Principles in the Community Standards

Without a way to enforce the Community Standards, they would be moot. Facebook's content moderation practices entail both automated moderation and human moderation teams, the latter consisting of 15,000 content moderators working in 50 languages.[59] As can be seen from Table 16.2, a large majority of content takedowns are not reviewed by humans at all. These 'proactive' content removals are based on algorithms that automatically (and probabilistically) detect supposed infractions of the Community Standards in audio-visual and textual user content. Reviews against these decisions can be requested by affected users.

[55] E Dwoskin, N Tiku and H Kelly, 'Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show' *Washington Post* (3 December 2020), available at www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech.

[56] See Facebook, 'Our Commitment to Human Rights', available at about.fb.com/news/2021/03/our-commitment-to-human-rights.

[57] See Facebook, 'Our Commitment to Human Rights', available at about.fb.com/news/2021/03/our-commitment-to-human-rights.

[58] In addition to the UNGPs, Facebook's new human rights policy includes specific international law instruments including the International Convention on the Elimination of All Forms of Racial Discrimination, the Convention on the Elimination of All Forms of Discrimination Against Women, the Convention on the Rights of the Child, the Convention on the Rights of Persons with Disabilities, the Charter of Fundamental Rights of the European Union and the American Convention on Human Rights (see Facebook, 'Corporate Human Rights Policy', available at about.fb.com/wp-content/uploads/2021/04/Facebooks-Corporate-Human-Rights-Policy.pdf).

[59] See Facebook, 'Understanding the Community Standards Enforcement Report', available at transparency.facebook.com/community-standards-enforcement/guide.

If a human review remains unsuccessful, an appeal to the newly created Oversight Board can be lodged. The nuts and bolts of the process are indeed what civil society groups are particularly concerned with.

The previous section identified six procedural categories of principles representing civil society demands with regard to the process of content moderation: certainty and predictability; appeal and remedy; procedure; limitations to automated content moderation; rights due diligence and impact assessment; and self-regulatory bodies. The existence of Community Standards makes Facebook's content governance relatively *certain and predictable*. Its policies also state both the consequences of violating posts, ie removal of the content and – depending on repeated infractions – changes to the ability to post or to use the account altogether, and the decision-making procedures (in general terms) and pathways, including regular transparency reporting.

Opportunities to *appeal and remedy* can be readily found by users on the platform itself. While the percentage of content decisions for which a human review was requested is generally low, as shown in Table 16.2, there exists some variance. Content decisions based on the above-defined category of 'prevention of harm' are rather unlikely to be appealed by users (0.2 per cent overall), whereas the rate is significantly higher for the 'protection of social groups' and 'public interest' categories (at 1.1 per cent and 1.2 per cent respectively). Users are most likely to request a review for content that has been deemed to violate the policy to not sell firearms (5.5 per cent). The remedy is usually for Facebook to restore the content and, possibly, account functionality. The outcome of the review procedure is then communicated to users. However, the rate of successful reversal of an initial content removal differs between principles of the Community Standards. For instance, in the last quarter of 2020, concerning the principle on the sale and promotion of drugs, a total of 80,200 reviews were requested and 58,600 decisions reversed (73 per cent).[60] In contrast, concerning the principle on bullying and harassment, 443,000 initial decisions were appealed to, but only 41,000 content restorations upon appeal were reported (9 per cent).[61] At least theoretically, this variance points to a differing ability of Facebook's algorithms to detect content correctly and not to over-censor it. It should be noted that Facebook's appeal procedures provide users with limited opportunities to argue against alleged violation of community standards. Notifications of violations and reviews do not give explanations of the reasoning that informed the decision and, in the case of review, they do not usually refer to users' replies. Usually, they consist of standardised sentences.

The *notification procedure* employed by Facebook appears only partly in line with the demands by civil society. Users do not always receive sufficient notification about the reason for a content or account action as well as information about

[60] See Facebook, 'Community Standards Enforcement Report – Regulated Goods: Drugs and Firearms', available at transparency.fb.com/data/community-standards-enforcement/regulated-goods/facebook.

[61] See Facebook, 'Community Standards Enforcement Report – Bullying and Harassment', available at transparency.fb.com/data/community-standards-enforcement/bullying-and-harassment/facebook.

how to appeal the decision. In cases of copyright content takedowns, contact details about the alleged owner of the intellectual property are provided to allow direct interaction.[62] Users are not informed whether the decision has been made by a human or a machine, a demand entailed in the civil society documents. While the notification process with regard to content takedown and account actions appears to be at least fairly consistent with demands by civil society, there is an important difference between this system and the News Feed ranking process. The downranking of individual posts does not in principle make it impossible to discover a post. However, in practice, a low ranking score in the Feed means that a post will be viewed by others less frequently. Assuming that communication of content to be viewed by others is the primary purpose of posting on Facebook, a low rank represents a severe limitation.[63] While there are reasons to rank a post lower or higher, eg depending on if a post is made up of text, a picture or a video, increasingly, content is in the focus here, too. Specifically, in early 2021, Facebook changed its ranking to downrank website links considered more borderline to appear less often on users' Feeds.[64] Similar mechanisms exist with regard to misinformation and clickbait.[65] The processes concerning borderline content have been criticised for their lack of notification.[66] This process is entirely unrelated to the Community Standards, and yet it creates questions concerning the civil society demand of appropriate notification. The lack of notification about certain posts being 'punished' by the algorithm or the clear statement of the criteria for such down-ranking mean that users are unaware and unable to appeal to the decision or alter their behaviour. Whereas the Community Standards represent a written code with an elaborate appeals procedure, the News Feed ranking is more akin to computer code involved in content decisions without these safeguards.[67]

Relatedly, a relatively high number of civil society documents entail demands for *limitations to automated content moderation*. Limitations on so-called 'proactive' moderation may help to reduce false positives, thereby strengthening freedom of expression. On the other hand, since human moderators would take

[62] See Facebook, 'Content that I posted on Facebook was removed because it was reported for intellectual property infringement. What are my next steps?', available at www.facebook.com/help/365111110185763.

[63] This does not relate to the part of Facebook's ranking algorithm where it is based on people's prior interests, but to criteria that go beyond the individual. See Facebook, 'How machine learning powers Facebook's News Feed ranking algorithm', available at engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking.

[64] See E Dreyfuss and I Lapowsky, 'Facebook is Changing News Feed (Again) to Stop Fake News' (Wired, 4 October 2019), available at www.wired.com/story/facebook-click-gap-news-feed-changes.

[65] See Facebook, 'Working to Stop Misinformation and False News', available at www.facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news.

[66] A Heldt, 'Borderline speech: caught in a free speech limbo?' (15 October 2020) *Internet Policy Review*, available at policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510.

[67] cf L Lessig, *Code: And Other Laws of Cyberspace, Version 2.0* (Basic Books, 2006).

longer to react to (automatically) flagged content, proactive moderation means that less material presumed to infringe the Community Standards will be viewed by users. The degree to which posts that are removed from the platform are either proactively found, mostly through algorithms, or flagged by users depends on which principle is concerned, as shown in Table 16.2. For instance, in the last quarter of 2020, the rate of automation has been high for most infractions such as offering or promoting drugs (97.3 per cent), terrorism (99.8 per cent) and adult nudity and sexual activity (98.1 per cent). An outlier is the principle on bullying and harassment: only 48.8 per cent of posts sanctioned under this provision of the Community Standards were proactive – the remainder had to be flagged by users. This shows how some rules require more contextual knowledge and human understanding of a situation to be applied effectively.[68]

The rate of automation has increased over the last few years, even in those areas that can be considered more difficult to judge compared to others. While the proactive detection of hate speech made up only 23.6 per cent of all content actions in the last quarter of 2017, the rate rose to 60.7 per cent in the same quarter of 2018, on to 80.9 per cent a year later and further increased to 97.1 per cent in the last quarter of 2020.[69] For other principles entailed in the Community Standards, the rate of automation has been consistently above 96–99 per cent over the same period.[70] This clearly shows that there is even a tendency away from the demand by civil society documents that automated content moderation should be limited to 'manifestly illegal content'. In addition, as pointed out above, although demanded by some civil society documents, not all automated content decisions are being reviewed by a human.

The demand by civil society documents that social media platforms conduct *human rights due diligence or impact assessments* has been met through such assessments on the country level, such as recently in Sri Lanka, Indonesia and Cambodia.[71] In order to be a more legitimate judge over what content or accounts can remain on the platform, in 2018 Facebook created an independent Oversight Board tasked to pick, deliberate and decide concerning the decision of the platform's content moderation system.[72] It does so based on the Community Standards, and its governing documents and internal rulebook have also been reviewed from a human rights due diligence perspective by an outside non-profit organisation, ensuring its grounding 'in human rights principles, including the rights

[68] See R Gorwa, R Binns and C Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 *Big Data & Society* 1.

[69] See Facebook, 'Community Standards Enforcement Report – Hate Speech', available at transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook.

[70] See Facebook, 'Community Standards Enforcement Report' (n 45).

[71] See Facebook, 'An Update on Facebook's Human Rights Work in Asia and Around the World', available at about.fb.com/news/2020/05/human-rights-work-in-asia.

[72] See ch 14 (Schulz) in this volume.

to freedom of expression, privacy and remedy'.[73] The board's power extends to making binding decisions on content removal, based on the appealed cases before it. In addition to these decisions, the Oversight Board issued broader 'policy advisory statements' with its initial seven decisions, in which it asks for policy changes, including substantive clarifications and procedural improvements.[74] However, at present, it remains to be seen how the policy teams at Facebook will implement these recommendations and how sustainable the Oversight Board proves to be.[75]

## V. Conclusion

Global online content governance is currently facing a problem which is not novel in its essence. Determining which principles govern global spaces is an issue that characterised all phenomena related to globalisation and has affected the Internet since its origin. In his seminal book, *Code 2.0*, Lessig schematised this dilemma as being the choice between a 'no law', 'one law' and 'many laws' worlds.[76] In the social media environment, the decision of private platforms to adopt their own internal rules to bypass the legal pluralism that characterise national and international law has been accused of arbitrariness and lack of accountability, being even associated with a 'no law' scenario.[77] However, this strong critique is gradually pushing online platforms to rethink their internal content moderation rules and procedures in a way that better reflects the relevance of the virtual space they offer vis-à-vis fundamental rights. A process of constitutionalisation of this environment is currently underway. Core principles of contemporary constitutionalism are rearticulated to address the challenges of the social media environment.

Despite the evocative image that 'constitutionalisation' brings to mind, there are no founding fathers sitting in the same room for days to define the constitution of social media.[78] The process of constitutionalisation of this environment reflects the complex, global and pluralist scenario in which social media operate. It is characterised by a high level of fluidity and informality. Social media online content governance rules are being fertilised by a multistakeholder constitutional input, which may also potentially contribute to enhancing the legitimacy of social media rules from a global perspective. Indeed, an aerial view on this phenomenon

[73] See Facebook, 'An Update on Building a Global Oversight Board', available at about.fb.com/news/2019/12/oversight-board-update.

[74] See Oversight Board, 'Board Decisions', available at oversightboard.com/decision.

[75] For a detailed discussion of the creation and the limits of the Oversight Board, see *K Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2019) 129 Yale Law Journal* 2418.

[76] See Lessig, *Version 2.0* (2006) ch 15.

[77] See N Suzor, *Lawless. The Secret Rules That Govern Our Digital Lives* (Cambridge University Press, 2019).

[78] Celeste, 'Terms of Service and Bills of Rights' (2019); KM Yilma, 'Digital Privacy and Virtues of Multilateral Digital Constitutionalism – Preliminary Thoughts' (2017) 25 *International Journal of Law and Information Technology* 115.

witnesses multiple, simultaneous processes of 'parallel' or 'collateral' constitutionalisation that are currently ongoing.[79] In this chapter, we mapped the contributions of international law, civil society impulses, and the norms developed by social media platforms themselves. Future research in the field is encouraged to take in account the complexity of this multi-layered constitutional landscape, investigating how these normative sources are mutually complementing each other.

[79] See E Celeste, 'The Constitutionalisation of the Digital Ecosystem: Lessons from International Law' (2021) Max Planck Institute for Comparative Public Law and International Law (MPIL) Research Paper No 2021-16, available at papers.ssrn.com/abstract=3872818.