CHAPTER 5

Platform Policies Versus Human Rights Standards

Abstract This chapter empirically examines how five social media platforms—Facebook, Instagram, Twitter, TikTok and YouTube—deal with the content governance dilemma and the question of which human rights standard to apply when moderating user content. It builds on previous chapters' analyses of relevant human rights standards in international law and civil society-issued documents to elucidate to what extent substantial and procedural demands are met by the platforms. After an analysis of platform policies—specifically the human rights commitments included in them, the chapter examines substantive content moderation trends in a comparative way. Thereafter, procedural practices of content moderation including transparency reporting and automated content moderation are comparatively discussed. The chapter finds a relatively high degree of convergence among the platforms on a number of practices.

Keywords Platform policies • Human rights standards • Content moderation trends • Transparency reports • Automated content removal

E. Celeste et al., *The Content Governance Dilemma*, Information Technology and Global Governance,
https://doi.org/10.1007/978-3-031-32924-1_5

5.1 Human Rights Commitments as a Window Dressing Strategy?

This chapter empirically examines the content moderation practices of four selected platform companies and five of their social media services—Facebook and Instagram (Meta Inc.), YouTube (Alphabet Inc.), Twitter (Twitter Inc.) and TikTok/Douyin (ByteDance Ltd.). This chapter is based on the analysis of norms in international law presented in Chap. 3 and the findings from the empirical analysis of civil society documents included in Chap. 4. It illustrates how social media platforms deal with the content governance dilemma outlined previously. In Chap. 3, we demonstrated to what extent the content governance of transnational platforms can be regulated or guided by international human rights law (to only a limited extent). The current chapter shows how four globally operating platform companies are dealing with this relative lack of strict guidance, but also with the nonetheless large amount of existing human rights writing and commentary. This chapter also illustrates that a more elaborate human rights standard developed by the international community and put into treaties by states in a multistakeholder process could be desirable, if only to address the gap between those companies that do more to protect human rights in their operations and those that do far less. The chapter also illustrates the value of civil society and multistakeholder charters and declarations that are at times directly cited to be an impetus towards a stronger human rights commitment of platforms. It also compares to what extent substantive and procedural demands raised in these documents are met by different platforms. The chapter only indirectly addresses the legislative codification of international human rights standards into national law and regional rules (e.g. the recently proposed EU Digital Services Act package). Instead, it takes the four platform companies and their services as the locus of the analysis. How these platforms *translate* general human rights commitments into platform policies and practices matters greatly for practical human rights protection online. Observing recent activities of these platforms closely allows us to tease out the different ways in which the content governance dilemma can be addressed, and it helps us to understand *why* platforms address it as they do.

When analysing how these platforms deal with content posted by users in connection to human rights norms, one must take into account both the formal commitments and statements of the platforms and the empirical—or sociological—reality of how platforms incorporate human rights

standards into their processes—or not. Both potentially tell us something about the reasons for adopting, or failing to adopt, a strong human rights stance in content governance. Are human rights commitments a mere window dressing? How do the platforms structure their moderation processes and what moderation outcomes can be observed in relation to how civil society documents frame desirable moderation principles? To address these questions, this chapter is subdivided into three sections. In a first step, we explore the written platform policies on content moderation and show to what extent these documents include human rights language and an explicit commitment to human rights norms or a dedicated policy on human rights. In a second step, we explore how the *substantive* demands by civil society Internet Bills of Rights, focussing on what is seen to be legitimate exceptions from their central freedom of expression claim, are being realised through content moderation at the five platform services. Using the data published by the platforms themselves, we have focused on the human rights relevant moderation practices based on a framing by civil society documents, using the categories developed in Chap. 4. This helps us to connect a discussion about principles adopted in the content policies of platforms with the idea of an emerging convergence—or standard—for content governance. This convergence occurs both on the level of content moderation policy documents and on the level of substantive moderation outcomes. In a third step, we examine the *procedural* category of principles entailed in the Internet Bills of Rights. By looking at two specific principles and respective metrics, (1) the share of moderation decisions taken by automated systems such as AI technologies over time and (2) the increase of transparency reporting of platforms, we focus on some of the key principles demanded by civil society-issued documents. While both procedural principles tell us another story of converging to a standard across platforms, the impact on human rights is less clearly identifiable. The continuous struggle for greater human rights protection is as much needed as are suitable platform policies and moderation practices.

5.2 Platform Policies and Human Rights Commitments

That social media companies concern themselves with the *right* human rights standard for their content moderation operations is a relatively new phenomenon, much like the idea that platforms closely watch what users

publish at all. For a long time, social media platforms happily took on the cloak of 'content intermediaries', which promote free speech and present little in terms of rulebooks to their users. The platforms in fact benefitted from regulations such as Section 230 of the US Communications Decency Act (CDA) and the E-Commerce Directive 2001 in the European Union (Citron and Wittes 2017; Kuczerawy and Ausloos 2015). These regulations allowed them to evade direct liability for content posted by users but obliged them to act when being notified of potential violations and infringing content. This allowed the companies to claim for their social media services the status of neutral tech companies that support (American) First Amendment protections by maximising free speech (and reach) on the Internet. This, almost libertarian, approach mirrors the early Internet ideals most eloquently captured in the Declaration of Independence of Cyberspace (Barlow 1996). This being said, when the four platforms started their operations, there were content limitations such as restrictions on pornographic content, copyrighted materials and spam. Only later would these be coded into their initial content policies.[1] Twitter, in 2009, in its very first iteration of the "Twitter Rules", still stated that

> each user is responsible for the content he or she provides [and thus], we do not actively monitor user's content and will not censor user content, except in limited circumstances described below. (Twitter 2009)

In the same document, Twitter provided a narrow set of exceptions to the focus on freedom of expression, most notably with regard to cases, indeed, referring to spam, pornography, privacy and copyright infringements. Twitter's early platform policies were an important step towards spelling out the rules for speech on the platform but they were less of a concern for policymakers or human rights groups as they are today.

Up until the early 2010s, the scope of platform content policies was relatively limited. Over time, as massive growth of the user base lifted the profile of social media platforms, they became more entangled in political affairs—shaping electoral politics, but also being affected by increasing demand and regulation (Barrett and Kreiss 2019). At the same time, these platforms also represented a viable source of personal data used by national

[1] For an overview of the early platform content policies, see the Platform Governance Archive, https://www.platformgovernancearchive.org/.

intelligence agencies, which also gave rise to a new form of surveillance capitalism (Zuboff 2019). Starting in the mid-2010s, a number of high-profile scandals further put social media platforms into the focus of policymakers in the political capitals of the world. After the 2013 Snowden revelations about spy agencies and their ready access to data from social media companies, 2016 represents another inflexion point, with widespread discussions of misinformation on platforms following the US elections, while the 2018 Cambridge Analytica scandal brought further concerns about privacy and corrupted electoral processes in connection with data collected through Facebook (Hemphill 2019). The "techlash" (Hemphill 2019) that followed the scandals led to greater pressure to design more complex content policies and to innovate with regard to content moderation procedures. Amid this "turn to responsibility" (Katzenbach 2021), platforms moved further away from the notion of platform neutrality in matters of content, which represented a core ingredient of the rise of social media platforms. Instead, today, there exists a "broad consensus that platforms have responsibility for the content and communication dynamics on their services" (Katzenbach 2021, 3). This turn can also be detected through changes in public platform content policies. Twitter, for example, after a major revision of its Rules in June 2019, stated that its

> purpose is to serve the public conversation. Violence, harassment and other similar types of behaviour discourage people from expressing themselves, and ultimately diminish the value of global public conversation. Our rules are to ensure all people can participate in the public conversation freely and safely. (Twitter 2020)

This statement marks a dramatic shift from the initial free-speech absolutism of the platform's early days—and perhaps the days that lie ahead, after the acquisition of the platform by billionaire, Elon Musk.

Twitter is not an exceptional case in this matter. Other platforms have also developed substantial, and elaborate rulesets concerning the kind of content that can be posted on their sites. These platform policies are usually documented on public pages for users to consult. At times, such as in the case of Facebook and Instagram, these rulesets are flanked by transparency centres in the form of websites providing information on enforcement practices. These webpages are ostensibly geared to be of use to policymakers, journalists, members of organised civil society and

academic researchers. This is appropriate because the written policies of large social media companies and their enforcement practices represent a comprehensive and powerful mode of governing communication on the Internet. That online communication is governed in such a way by private actors rather than public entities may seem "lawless" in its current state due to a perceived lack of legitimacy of platforms to rule (Suzor 2019), and it might amount to normative platform authoritarianism as argued in Chap. 2. Notwithstanding the unease many observers perceive, the facticity of "platform law" (Bygrave 2015; Celeste 2022; United Nations 2019) remains, and with it the dominant role intermediaries play in governing the Internet (Suzor 2019). Platform law is at play even where it is not published, where rules are kept secret or otherwise unavailable. There may be a number of reasons why rules are not public, including differential treatment of specific groups (as in the recently revealed separate content moderation for celebrities' content on Meta Inc.'s platforms),[2] the lack of codification (in the case of early platforms) or because content moderation is intertwined with state censorship (as, for instance, in the case of Chinese social media platforms).

In this context, it is important to highlight that an increased number of rules for content posted on these services does not necessarily amount to effective human rights protection. Instead, the growth of the number of rules *per se* can also stifle important values and rights such as equality or freedom of expression, which is so central to civil society Internet Bills of Rights. In an environment in which platforms are increasingly pushed to over-moderate to save themselves from legal peril or a public relations disaster, a multitude of rules that allow for speech to be removed may be an outright risk to freedom of speech. Ideally, at least in the framework of this book, platform rules directly refer to the human rights document, whose implementation they ought to support. However, when examining the four platform provider's content moderation rules, such immediate references cannot be found. This may well be due to the difficulty to simply copy and paste the content of international human rights documents, as discussed in Chap. 3. However, human rights are usually referred to by the platform services in some way. For our purposes and the

[2] A 2021 leak, the so-called Facebook Files, entailed information about Meta's XCheck program, which in late 2020 shielded at least 5.8 million important and celebrity users from the content moderation procedures applied to other users (see Horwitz 2021).

remainder of this section, we are most interested in *how* human rights are included in the platform policies of each of the four platform companies.

5.2.1 Meta

Meta Inc. is the parent company of two major social media services—Facebook and Instagram, which share a common set of policy documents. The popularity of the platforms—just under 3 billion people use Facebook every month and just under 1.5 billion use Instagram—is the foundation for the company's place among the largest companies globally by valuation (Statista 2022). WhatsApp, a messenger service, is another popular service owned by Meta, which is however not examined in this chapter. The content posted on Meta's platforms is governed by its "Community Standards", which were first published in 2007 (Facebook 2007). The document has greatly expanded from a little over 700 words in its first iteration to more than 19,700 words spread across several sub-pages as of late 2022, now including many explanations and examples (Meta 2022a). The Community Standards also apply to content posted on Instagram. Due to these services' enormous number of users, the rules may very well be one of the most effective tools to affect the enjoyment of human rights worldwide, true constitutional instruments of these online spaces (Celeste 2019). Hence, the pressure on the company by civil society activists and by the former UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, to adopt *a* human rights standard for its content moderation has been immense (Helfer and Land 2022; United Nations 2018, 2019). Starting in 2018 and 2019, Meta (then still known as Facebook Inc.) started to adopt human rights references in communications about content moderation by top management (Allan 2018; Zuckerberg 2019). This talk was then followed up with the creation of the Meta Oversight Board, which has its own charter and by-laws that explicitly put the board on a path to negotiate between the platform's Community Standards on the one hand and, external, human rights standards on the other hand, as we will discuss below. Legally speaking, it still holds true what the company communicated in 2018, that is, "we're not bound by international human rights laws that countries have signed on to" (Allan 2018). Nonetheless, in 2021, Meta gave itself its own corporate human rights policy, stressing both a commitment to the non-binding Ruggie Principles and other international law instruments:

> We are committed to respecting human rights as set out in the United Nations Guiding Principles on Business and Human Rights (UNGPs). This commitment encompasses internationally recognized human rights as defined by the International Bill of Human Rights—which consists of the Universal Declaration of Human Rights; the International Covenant on Civil and Political Rights; and the International Covenant on Economic, Social and Cultural Rights—as well as the International Labour Organization Declaration on Fundamental Principles and Rights at Work. (Meta 2022c)

As discussed in Chap. 3, the UNGPs are international soft-law standards that systematically address businesses, albeit indirectly. For a platform company like Meta to commit specifically to the Ruggie Principles should be a minimum standard, or as the former UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression recommended:

> The Guiding Principles on Business and Human Rights, along with industry-specific guidelines developed by civil society, intergovernmental bodies, the Global Network Initiative and others, provide baseline approaches that all Internet companies should adopt. (United Nations 2018)

To what extent these listed commitments to international human and labour rights standards result in a coherent human rights standard, particularly considering the strong competing role of Meta's other values, is discussed further below.

5.2.2 Twitter

Twitter Inc.'s platform may be the go-to place for politicians, journalists, academics and others to communicate political messages, advertise their own latest products or publications and engage in shoulder-rubbing by other means. However, importantly, many more users engage in everyday conversations about their life, leisure and politics. Twitter has also developed into a tool for human rights defenders to speak to members of the media to create awareness for human rights abuses by governments and companies, allowing for messages unfiltered by the press, governments or non-governmental organisations. For instance, in 2011, in Cairo's Tahrir Square, activists and ordinary citizens connected to one another and to a global audience through their 'tweets', at the very least *supporting* the

2011 Egyptian Revolution. In early 2022, Twitter counted more than 400 million active monthly users globally (Statista 2022), and not all of them were bots (Milmo 2022). Twitter had been the favourite outlet of thoughts and supposed policy formulations by former US President Donald Trump. Like Facebook, Twitter faced a decision on how to continue the relationship with the (then-sitting) president after the storming of the US Capitol on January 6, 2021. Twitter decided to permanently suspend Trump's account (Guo 2021). Generally, what Twitter users can and cannot post online is regulated by Twitter Rules and additional documents. A dedicated human rights policy was not publicly available as of September 2022. Twitter claims that its commitment to user rights is based on a commitment to both the US Constitution and the European Convention on Human Rights (Twitter 2022a). It also refers to the fact that its content moderation is "informed (…) by works such as United Nations Principles on Business and Human Rights" (ibid.). However, unlike Meta, the company does not specify particular rights or further international human rights documents that could amount to a binding policy or standard.

5.2.3 TikTok

TikTok has been *the* recent quick starter among social media platforms globally. The platform's focus on short video clips as a core format harnesses the increased access to fast mobile data connections and suitable mobile phones to record scenes. The number of monthly active users more than doubled between the second quarter of 2020 and the second quarter of 2022, to an estimated 1.46 billion, according to the trade website Business of Apps (Iqbal 2022). In China, the mobile app and social media service is known as Douyin. While there had been pressure by the US government to spin-off its operations in the United States to a local joint venture, this did not directly occur. An attempted ban of TikTok by the previous administration was revoked by US President Biden (Kelly 2021). Being pressured to increase the protection of US data, TikTok's mother company ByteDance Ltd. arranged a deal with Oracle to store American user data within the country exclusively (The Guardian 2022). As of September 2022, TikTok's content policy, the "Community Guidelines" do not make any reference to human rights or international legal norms (TikTok 2022b). Nonetheless, in its transparency centre, after pointing to the fact that "as a global entertainment platform, TikTok

spans most major markets except China, where ByteDance offers a different short-form video app called Douyin", the company published the following human rights statement:

> Technology is an essential gateway to the exercise of human rights. (…) Responsibility for upholding human rights is shared: while governments have the responsibility to protect human rights, TikTok and other businesses have a responsibility to respect those human rights. Respecting human rights is essential for TikTok to build and sustain trust among our employees, creators, advertisers, and others who engage with our company. Our philosophy is informed by the International Bill of Human Rights (which includes the Universal Declaration of Human Rights and the International Labour Organisation's Declaration on Fundamental Principles and Rights at Work) and the United Nations Guiding Principles on Business and Human Rights. As part of our commitment, we will strive to respect human rights throughout our business and will comply with applicable laws and regulations intended to promote human rights where we conduct business globally. We will continuously evaluate our operations to identify, assess, and address salient human rights risks; engage key stakeholders; and prioritise key areas where we have the greatest opportunity to have a positive impact. (TikTok 2022a)

The statement generally follows the Ruggie Principles but with a caveat that the platform would prioritise actions avoiding human rights based on where such actions would have the greatest benefit, rather than striving for an overall protection of human rights. Such a utilitarian approach to the balancing of rights and other objectives is interesting, particularly with regard to the processes and actors involved in such decisions. Based on the policy itself, little is to be expected in terms of aligning with the substantial and procedural demands by civil society voiced in various Internet Bills of Rights examined in Chap. 4.

5.2.4 YouTube

YouTube is a global video platform owned by Google Inc., which itself is a subsidiary of Alphabet Inc. Google ran with the slogan "don't be evil", which could also be found in the company's Code of Conduct—at least until it was removed from there in 2018 (Conger 2018). YouTube counts more than 2.5 billion monthly active users globally, which places it within the top four social media services, together with three services offered by

Meta Inc. (Statista 2022). Like the other platforms featured in this book, YouTube's very existence—particularly in countries with restricted public discourse—can be seen as a contribution towards enhancing freedom of expression and the right to information. Thus, when the platform was blocked by governments, human rights courts have repeatedly found that this blocking amounted to a human rights violation of the platform's users (Deutsche Welle 2015). However, YouTube has also been subject to allegations that it does not do enough to fight human rights violations (AccessNow 2020). Interestingly, as of September 2022, no explicit commitment to human rights can be found in YouTube's "Community Guidelines", which govern the kind of content that can be posted on the platform (YouTube 2022a). However, Google Inc. has a human rights policy that also applies to YouTube. Specifically, Google's policy asserts that the company finds orientation in internationally recognised human rights standards "in everything it does", adding a commitment to respect the rights included in the Universal Declaration of Human Rights and related treaties (Google 2022). The statement also specifically mentions the Ruggie Principles and, interestingly, the principles of the Global Network Initiative (GNI). The human rights policy also includes information on how Google and YouTube aim to implement these commitments, and thus translate the principles into moderation practices.

When observing how the four platforms discuss human rights in their policy documents, some stark similarities but also differences can already be made out. All five platforms (or their parent companies) include references to human rights, including specifically to the Ruggie Principles. Thus, on paper, one might say that a strong convergence on committing platforms to human rights standards has developed in the field, even if the scope of applicable human rights documents differs and differences in approach can be made out (e.g. TikTok's decidedly utilitarian approach). However, importantly, these human rights policies are distinct from content governance policies. The former are likely not integrated into the latter, in part due to the challenges that are posed by the application of any *one* human rights standard to content governance. As argued above, digital constitutionalism and, specifically, civil society Internet Bills of Rights are a potential catalyst to solving the content governance dilemma. Civil society advocates show platforms the way by balancing, in their documents at least, various human rights and good governance principles against each other. Consequently, the findings from Chap. 4, particularly if taken in their aggregate, can inform efforts to apply human rights standards to

platform content moderation. Consequently, the next two sections investigate how the content governance practices of the featured platforms perform against the background of these civil society demands. The next section focuses on substantial demands by civil society and the quantitative outcomes of platform content moderation.

5.3 Substance Matters! Platform Moderation Outcomes Versus Civil Society Demands

There are many who rightly emphasise the importance of process when it comes to evaluating the content governance of social media platforms (Kettemann and Schulz 2020; Klonick 2018; Suzor et al. 2018). However, as we show below, there is value in gauging to what extent comparative substantive enforcement outcomes relate to the civil society demands, if we assume that these demands, particularly in aggregate, are an important interpretation of how a human right-based platform governance regime should be designed. Addressing the three categories of "prevention of harm", "protection of social groups" and "public interest" as outlined in Chap. 4, this section considers which of these categories is most often used to justify limitations on the chief principle of freedom of expression. First, we categorised the substantive principles by which the four platforms organise content moderation into the three derived categories. The principles represent only the reported substantive principles for which data is available for analysis. Copyright infringements and related moderation principles are excluded from the analysis, due to differential reporting of data by platforms.[3] Table 5.1 shows which substantive principles can be found within the three mentioned categories, for all four platform companies, the information is limited to their reported data for 2021 (Meta 2022b; TikTok 2021a; TikTok 2021b; TikTok 2022c; TikTok 2022d; Twitter 2022b; YouTube 2022b).

Following the categorisation from Chap. 4, this chapter aims to show where platforms' respective focus lies. This framework is here first applied to Meta's two platforms and data for 2021, the last complete year for which data is available. In the following, the same framework is applied to Twitter, TikTok and YouTube.

[3] A recent report offers insights into substantive data on copyright moderation and copyright actions reporting by major platforms over time (Quintais et al. 2022).

Table 5.1 Substantive content moderation principles and categories from civil society documents

Categories	*Principles reported on in transparency reports (2021)*
Meta	
Prevention of harm	Bullying and harassment, suicide and self-injury, dangerous organisations: organised hate, dangerous organisations: terrorism, violence and incitement
Protection of social groups	Child endangerment: nudity and physical abuse, child endangerment: sexual exploitation, child nudity and sexual exploitation, adult nudity and sexual activity, violent and graphic content, hate speech
Public interest	Regulated goods: firearms, regulated goods: drugs
Twitter	
Prevention of harm	Abuse/harassment, hacked materials, impersonation, non-consensual nudity, private information, promoting suicide or self-harm, terrorism/violent extremism, violence
Protection of social groups	Child sexual exploitation, hateful conduct, sensitive media
Public interest	Civic integrity, COVID-19 misleading information, illegal or certain regulated goods or services, manipulated media
TikTok	
Prevention of harm	Harassment and bullying, suicide, self-harm and dangerous acts, violent and graphic content, violent extremism
Protection of social groups	Adult nudity and sexual activities, hateful behaviour, minor safety
Public interest	Illegal activities and regulated goods, integrity and authenticity
YouTube	
Prevention of harm	Promotion of violence and violent extremism, harmful or dangerous content, harassment and cyberbullying
Protection of social groups	Child safety, nudity or sexual content, violent or graphic content, hateful or abusive content
Public interest	Spam, misleading content, scams

Meta's Community Standards apply to the platforms Facebook and Instagram. Their content policy is informed by the organisation's self-proclaimed "core values", which emphasise their aim to create "a place for expression and giving people voice" (Meta 2022a). 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 Chap. 3. The Community Standards are structured into six chapters, of which the first four outline restrictions on

content based on Facebook's core values; the fifth chapter affirms intellectual property and its protection, whereas the sixth chapter outlines which user requests Meta complies with, including those to protect children and youth. The first four chapters currently entail a total of 21 principles defining content that must not be posted on the platform (Meta 2022a). Each principle is described in some detail, some with bullet-pointed lists of what constitutes an offence to be removed.

For the year 2021, Meta issued content moderation transparency reports covering 15 categories of content that can cause an action to delete content (apart from copyright-related actions and actions based on national legal requirements). These categories do not neatly fit the 21 principles of the Community Standards. For instance, the principle not to share "Restricted Goods and Services" includes goods such as weapons, drugs, blood, endangered animals, weight loss products, historical artefacts or hazardous goods and materials. Reporting, however, is only done for drugs and weapons. Two reporting categories—prohibitions on fake accounts and on spam—are arguably not as closely associated with the three most important categories of civil society demands; they will not be discussed in this analysis. In addition, reporting on these is only available for Facebook, lowering the number of reporting categories for Instagram to 13. Data for 2021 is available separately for Facebook and Instagram (downloaded through Meta's Transparency Center, see Meta 2022b). Furthermore, data for "Child nudity and sexual exploitation" is only reported for the first quarter of 2021. Starting from the following quarter, Instagram and Facebook's data differentiate between "Child endangerment: Nudity and physical abuse" and "Child endangerment: Sexual exploitation" when reporting moderation actions in this area. Table 5.2 shows each category from the civil society documents and the corresponding principles from the Community Standards on which transparency reporting occurs. For some quarters, no data is reported for one or the other platform.

Differences between Facebook and Instagram in terms of relative share among the justifications to limit the core value of 'voice' are distinctive. Instagram users have been moderated more commonly based on justifications of preventing bullying, suicide and self-injury, adult nudity and the advertisement of drugs. On Facebook, there are relative shares of moderation due to child sexual exploitation, in relation to terrorist organisations and hate groups, and due to gun offerings.

Table 5.2 Moderation outcomes and civil society categories, Facebook and Instagram (2021)

Category/principle	*Content actions (FB)*	*Content actions (IG)*	*Share of total (FB)*[a]	*Share of total (IG)*
Prevention of harm	**150,700,000**	**51,086,900**	**25.74%**	**31.48%**
Bullying and harassment	34,100,000	24,500,000	5.83%	15.10%
Suicide and self-injury	36,600,000	17,000,000	6.25%	10.48%
Dangerous organisations: organised hate	19,600,000	1,330,100	3.35%	0.82%
Dangerous organisations: terrorism	34,400,000	2,356,800	5.88%	1.45%
Violence and incitement	26,000,000	5,900,000	4.44%	3.64%
Protection of social groups	**416,500,000**	**105,580,600**	**71.15%**	**65.06%**
Child endangerment: nudity and physical abuse	5,900,000	1,968,200	1.01%	1.21%
Child endangerment: sexual exploitation	66,600,000	5,600,000	11.38%	3.45%
Child nudity and sexual exploitation	5,000,000	812,400	0.85%	0.50%
Adult nudity and sexual activity	126,700,000	42,000,000	21.64%	25.88%
Violent and graphic content	115,900,000	29,300,000	19.80%	18.05%
Hate speech	96,400,000	25,900,000	16.47%	15.96%
Public interest	**18,200,000**	**5,616,100**	**3.11%**	**3.46%**
Regulated goods: firearms	6,000,000	516,100	1.02%	0.32%
Regulated goods: drugs	12,200,000	5,100,000	2.08%	3.14%
Not categorised	**10,190,500,000**	**N/A**	–	**N/A**
Fake accounts	6,500,000,000	N/A	–	**N/A**
Spam	3,690,500,000	N/A	–	**N/A**

[a]Excluding the moderation categories of fake accounts and spam, for which there is no data from Instagram and which dwarf the remainder of the categories in the case of Facebook (94.6% of total content actions of the 15 reported on categories)

Table 5.3 shows data for Twitter for the entire year of 2021 (Twitter 2022b). The platform's transparency reporting differentiates content policy-related sanctions into 'content deletions' and 'account suspensions', both adding up to 'account actions'. Across different categories derived from the Internet Bills of Rights, account suspension shares differ. 'Public interest'-related enforcements and account suspensions make up

Table 5.3 Moderation outcomes and civil society categories, Twitter (2021)

Category/principle	*Account actions*	*Account suspensions*	*Share of total (actions)*	*Share of total (suspensions)*
Prevention of harm	**3,338,114**	**670,047**	**34.13%**	**26.19%**
Abuse/harassment	1,984,204	182,536	20.29%	7.13%
Hacked materials	143	0	0.00%	0.00%
Impersonation	398,490	368,625	4.07%	14.41%
Non-consensual nudity	58,471	15,660	0.60%	0.61%
Private information	64,895	5741	0.66%	0.22%
Promoting suicide or self-harm	753,243	18,818	7.70%	0.74%
Terrorism/violent extremism	78,668	78,667	0.80%	3.07%
Protection of social groups	**5,990,781**	**1,679,348**	**61.25%**	**65.64%**
Child sexual exploitation	1,055,669	1,050,751	10.79%	41.07%
Hateful conduct	2,010,891	238,150	20.56%	9.31%
Sensitive media	2,773,618	282,616	28.36%	11.05%
Violence	150,603	107,831	1.54%	4.21%
Public interest	**452,747**	**209,058**	**4.63%**	**8.17%**
Civic integrity	674	27	0.01%	0.00%
COVID-19 misleading information	51,947	1993	0.53%	0.08%
Illegal or certain regulated goods or services	399,983	207,038	4.09%	8.09%
Manipulated media	25	0	0.00%	0.00%

46% of actions; the rate is significantly lower for the 'prevention of harm' (20%) and 'protection of social groups' (28%) categories. The high 'public interest' share is due to a relatively more hard-line approach to moderation of displays and offers of 'regulated goods or services', as part of which accounts are more often suspended. Notably, almost all violations of the principles against 'terrorism/violent extremism' and 'child sexual exploitation' led to account suspensions rather than mere content removal.

The data for Twitter shows that more than 60% of account actions occurred to "protect social groups". Within that category "child sexual exploitation" makes up the by far largest reason for moderation actions.

Table 5.4 shows data for TikTok for all four quarters of 2021. Data availability is for "video removals" in these time intervals, rather than "content actions" (Meta) or "account actions" and "account suspension"

Table 5.4 Moderation outcomes and civil society categories, TikTok (2021)

Category/principle	*Share video removals*			
	Q1	*Q2*	*Q3*	*Q4*
Prevention of harm	**14.20%**	**13.10%**	**11.90%**	**13.90%**
Harassment and bullying	8.00%	6.80%	5.30%	5.70%
Suicide, self-harm and dangerous acts	5.70%	5.30%	5.70%	7.40%
Violent extremism	0.50%	1.00%	0.90%	0.80%
Protection of social groups	**62.70%**	**65.20%**	**70.90%**	**66.00%**
Adult nudity and sexual activities	15.60%	14.00%	11.10%	10.90%
Hateful behaviour	2.30%	2.20%	1.50%	1.50%
Minor safety	36.80%	41.30%	51.00%	45.10%
Violent and graphic content	8.00%	7.70%	7.40%	8.50%
Public interest	**23.1%**	**21.70%**	**17.10%**	**21.10%**
Illegal activities and regulated goods	21.1%	20.90%	16.60%	19.50%
Integrity and authenticity	2.00%	0.80%	0.50%	0.60%

(Twitter). Differences of metrics reported here are based on differences in platform reporting. In addition, TikTok only provides quarterly figures in their Transparency Center. In general, the degree of detail is relatively low for TikTok. However, the platform does offer data on content actions by countries, at least for a small number of countries, potentially useful information only few other platforms (such as YouTube) report on in detail. TikTok's video removal data relates to only nine moderation principles for posted content. Not included in the table is more detailed information about spam and fake accounts and engagement, not reported in this detail by other platforms. For instance, in the last quarter of 2021, TikTok "prevented" more than 152 million spam accounts, removed more than 46 million spam videos as well as 442 million fake followers, 11.9 billion fake likes and more than 2.7 billion fake follow requests (TikTok 2022d).

The data for TikTok shows that, like the previous three platform services, protection of social groups makes up the largest share of the three categories derived from civil society demands. The included limitations on freedom of speech are mostly justified with "minor safety"; here this single moderation principle amounts to more than half of all deleted videos (at least during the period July and September 2021). It appears that this justification is often used, perhaps due to the current character of TikTok used by younger users.

YouTube's Community Guidelines entail 21 moderation principles or sub-guidelines on its website (YouTube 2022a). One of these, however, is itself a list of other guidelines pertaining to four reasons for moderation that appear on the face of it to be too small to be an entire principle alongside the others. Interestingly, the guideline category of misinformation entails three sub-guidelines (or principles) prohibiting general misinformation, misinformation related to elections and medical misinformation related to the COVID-19 pandemic. This, perhaps once again, illustrates the effect world events have on the policies themselves, if not also their enforcement. YouTube's report on "YouTube Community Guidelines enforcement" entails only data on eight principles (YouTube 2022b). Table 5.5 shows how the substantive moderation decisions the platform reports on for 2021. Like the other platforms examined here, YouTube's reporting shows a strong—or even stronger—quantitative emphasis on removing videos that may (be used to) hurt (the sensibilities of) certain groups, including children and protected groups. Spam video removal is included in the data presentation here, because of its bundling up with other moderation principles such as misleading content and scams.

There is a lack of comparative research into substantive content moderation outcomes of Platforms. We ventured to conduct a comparison utilising a broad framework developed from civil society demands for one year of reported data. These demands, we argue, can help platforms understand what rights and principles should be considered when limiting

Table 5.5 Moderation outcomes and civil society categories, YouTube (2021)

Category/principles	*Share video removals*			
	Q1	*Q2*	*Q3*	*Q4*
Prevention of harm	**4.74%**	**15.40%**	**13.33%**	**18.62%**
Promotion of violence and violent extremism	0.91%	6.9%	4.07%	1.90%
Harmful or dangerous content	2.22%	4.80%	4.58%	8.11%
Harassment and cyberbullying	1.61%	3.70%	4.68%	8.61%
Protection of social groups	**87.30%**	**70.40%**	**76.70%**	**72.27%**
Child safety	54.03%	29.90%	32.45%	31.53%
Nudity or sexual content	16.63%	22.40%	18.72%	18.42%
Violent or graphic content	15.73%	16.80%	23.70%	19.92%
Hateful or abusive content	0.91%	1.40%	1.83%	2.40%
Public interest	**7.96%**	**14.10%**	**9.97%**	**9.11%**
Spam, misleading content, scams	7.96%	14.10%	9.97%	9.11%

speech on their platforms. The clear downsides of such an analysis are, first, the differences in how many moderation principles are actually reported on—relative to the number of principles entailed in the platforms' content policies, and, second, the differences of what metrics are reported on. Regarding the latter, we see removals of videos (TikTok and YouTube), "content actions" (Facebook, Instagram, Twitter), and "account actions" (Twitter) as the dominant metrics in this space. Some platforms report on additional metrics whose use would, however, have made comparison even less viable.

With these caveats in mind, we find that, surprisingly, the shares between the three categories and for all five platform services are relatively similar. Table 5.6 shows that around two-thirds of all reported (non-spam, non-fake account) moderation actions are associated with the protection of social groups (range: 61.25% to 72.27%). Between 13.90% and 34.13% of moderation actions occurred to prevent harm, while between 4.63% and 21.10% of reported moderation decisions are categorised to be in the public interest. The deviations between platform services are certainly less than we would have expected. This is likely the case for two possible reasons. First, users' behaviour could be assumed to be relatively similar across platforms. This would mean that social media users globally conduct themselves on social media platforms in such a way as to require moderation in similar ways, say on TikTok as on Instagram. Some users, independently of which platform they are on, harm each other and post videos and other media that are deemed to be inappropriate for certain viewers, or they engage in behaviour that is regulated such as the sale of drugs. The other possible reason or the similarity observed has more to do with the reactions of the platforms to user behaviour. Political and market forces, including the recent techlash, have apparently impacted the content of the platform policies and the moderation processes in such a way that substantial moderation foci are relatively similar across platforms.

Table 5.6 Overall share of reported moderation actions by category, all five platform services (2021)

Category	*Facebook*	*Instagram*	*Twitter*	*TikTok (Q4)*	*YouTube (Q4)*
Prevention of harm	25.74%	31.48%	34.13%	13.90%	18.62%
Protection of social groups	71.15%	65.06%	61.25%	66.00%	72.27%
Public interest	3.11%	3.46%	4.63%	21.10%	9.11%

We suggest that the degree of similarity is indeed explained not just by coincidence. Such a suggestion arguably requires the assumption of relatively similar behaviour of users across platforms. It should also be noted that, where geographic differences in usership exist, these might have a slight impact on the overall trends. These effects notwithstanding, we argue that platforms converge in their global content moderation around a standard affected by public pressure.

To illustrate convergence, which describes an ongoing process, the moderation principle relating to the promotion of violence can be explored in some detail. Even before 2021, platforms usually had policies in place that would outlaw incitement to engage in violence. However, paying respect to the right of freedom of expression, platforms have been relatively less strict in their enforcement. This changed dramatically in January 2021 and shortly afterwards. There is evidence that the January 6 US Capitol attack strongly affected platform policies and practices. On that day, violence erupted around the building that houses both chambers of the US parliament, right when parliamentarians were to certify the results of the November 2020 general election. Both Twitter and Meta banned accounts of the then-US President Trump, who was identified as inciting and condoning the violence by way of his posts during and in the aftermath of the attack on Capitol Hill in Washington. Other platforms were quick to react rhetorically, with YouTube announcing that "due to the disturbing events that transpired yesterday, and given that the election results have now been certified, starting today any channels posting new videos with false claims in violation of our policies will now receive a strike" (Ha 2021). The adaptation of content moderation principles took a bit longer then, often pushed by external actors. For instance, in May 2021 the ban on Trump's accounts was in principle confirmed by Meta's Oversight Board, which took the case and decided that the decision made by the company was to be upheld. However, the Oversight Board argued that it was "not appropriate for Facebook to impose the indeterminate and standardless penalty of indefinite suspension" (Oversight Board 2021). Thereafter, the platform's Community Standards were substantially revised. The new version of the content policy included thinly veiled references to the riot at the Capitol and the role Trump played in the incitement of violence. Specifically, the late January 2021 version of the Community Standards prohibits content that makes "implicit statements of intent or advocacy, calls to action, or aspirational or conditional statements to bring armaments to locations, including but not limited to places

of worship, educational facilities, polling places, or locations used to count votes or administer an election (or encouraging others to do the same)" (Facebook 2021). Less elaborate changes occurred faster. The term "incitement" was added to the Twitter Rules in January 2021, now stating that "content that wishes, hopes, promotes, *incites*, or expresses a desire for death, serious and lasting bodily harm, or serious disease against an entire protected category and/or individuals who may be members of that category" (Twitter 2021, emphasis added). This change, however, mimics the language of the day without being as directly related to the January 6 Capitol attack.

By mid-2021, all platform services included in our analysis entailed some reference to the principle of "incitement to violence", as is attested by Table 5.7. For Facebook and Instagram, data on the new principle is only available for the second half of the year 2021, quickly making up between 5 and 10% of overall moderation actions on the two platform services. The data further shows how the prohibition on incitement of violence was relatively more often invoked as a reason to take a content moderation action for some of the platforms studied here. This increase occurred on a low level in the case of TikTok and, slowly but strongly, in the case of YouTube, with a fall-off in terms of relative share of moderation actions in the fourth quarter of the year. In the case of Twitter, there was no significant change between the first and second half of 2021.

Changes in policies and moderation outcomes for one specific principle illustrate how a further convergence towards a *common standard* adopted by platforms can occur. These changes transpire due to public and political pressure to secure certain human rights—here, the right to life and the right to democratic elections. Civil society documents, including the GNI Principles and the Santa Clara Principles, have an impact when the platforms grasp for solutions to their policy and enforcement woes (as

Table 5.7 Share of reported moderation actions for incitement of violence, all five platform services (2021)

Platform	*Principle*	*Q1*	*Q2*	*Q3*	*Q4*
Facebook	Violence and incitement	N/A	N/A	8.81%	9.32%
Instagram	Violence and incitement	N/A	N/A	6.98%	5.93%
Twitter	Terrorism/violent extremism	0.86%		0.74%	
TikTok	Violent extremism	0.50%	1.00%	0.90%	0.80%
YouTube	Promotion of violence and violent extremism	0.91%	6.9%	4.07%	1.90%

indicated by them being cited in relation to human rights policies). However, not all global constituents have the same influence on shaping this standard; great tragedies could remain without an impact on policies if it was not for strong advocacy organisations to engage in reporting about platform failings. As examples from India and Myanmar show, platforms have been slow to adopt effective human rights-respecting policies and to conduct impact assessments (Al Ghussain 2022; Amnesty International 2022). Importantly, substantive moderation outcomes are not just affected by changes in policies. In fact, process matters quite a bit for the enjoyment of human rights, attested by the 40 civil society documents analysed previously. The next section examines two of these six core demands entailed in the Internet Bills of Rights in some more detail.

5.4 Process Matters! Platform Moderation Processes Versus Civil Society Demands

The substantive moderation outcomes discussed above tell us only as much about how civil society demands, which we understand as in their aggregate as a reasonable approach to how human rights-based content governance ought to be implemented, are actually met in practice. While the principles demanded, such as protection from hate and the protection of democratic elections create an important foundation for a human-rights-respecting moderation system, a suitable process is required to enable effectiveness, fair treatment and transparency in content moderation. Chapter 3 identified 32 procedural principles in 3 categories representing civil society demands with regard to the process of content moderation. In order to examine to what extent the empirical moderation practices of the four platform services adhere to demands by civil society, and to see whether they again converge, this section focuses on two of these principles: the *limitation of automated content moderation* and *transparency*. This focus is aimed to allow for more in-depth analysis.

5.4.1 Curbing Automated Content Moderation?

A relatively high number of civil society documents entail demands for limitations to automated content moderation. These demands are likely driven by the principled idea that every user and their posts should be evaluated by another thinking human being and not by a cold machine

that perhaps does not really understand the point of the joke or the circumstance of the post in the first place. Examples of such false positives include the breast cancer awareness post in Brazil removed by automated systems for infringement on the company policy against nudity, even though it was clearly stated that the nude female breasts were shown for that exact, and permitted, purpose (Oversight Board 2020). In another case, a human moderator penalised a user for posting an Iranian protest slogan amid the 2022 protests against the Iranian government. The user, appealing to the initial decision by Meta, did not receive a decision by another human moderator, but an automated system closing the appeal (Oversight Board 2022). Civil society documents reviewed by us demonstrate that there is a hope that limitations on 'automated', 'proactive' or 'AI-based' moderation may help to reduce false positives, thereby strengthening freedom of expression.

All five platform services studied for this chapter rely on automation in their content governance systems. However, Twitter, at least until late 2022, did not report the number or the share of content removals triggered by automated detection of a policy violation. Demands of civil society concerning automated moderation, as seen in Chap. 3, are diverse. The Global Forum for Media Development's stance against any kind of automated content moderation, and the demands found in the "Charter of Digital Fundamental Rights of the European Union" drafted by members of the German civil society both amount to a demand for a right not to have decisions over humans be made by algorithmic systems. Such a demand is certainly difficult to reconcile with the strikingly pervasive use of automated moderation systems in content moderation. The results of the analysis are displayed in Table 5.8. The data shows that the four reporting services heavily rely on automated moderation to remove content from their platform. The striking exception is TikTok, which only removes about half of the videos it deems to violate its content policies upon a

Table 5.8 Share of automated moderation actions of total actions

Platform	*Framing*	*Share of automation*	*Reference period*
Facebook	Proactive detection	94.20%	Q3/2022
Instagram	Proactive detection	94.70%	Q3/2022
Twitter	N/A	N/A	N/A
TikTok	Videos removed by automation	48.02%	Q3/2022
YouTube	Automated flagging	94.52%	Q3/2022

prompt by an automated system. It can thus be concluded that automated systems take over a large share of the moderation workload, and they take (or at least took) on tasks even when a user appealed to a decision.

On the other hand, since human moderators would take longer to react to (automatically) flagged content, proactive moderation means that less material presumed to infringe platform policies will be viewed by users. Here, automated moderation and the call for it or a rejection become a balancing act between competing rights. Chiefly among others, the desire not to be over-moderated (as in the cases in Brazil and Iran) and to exercise freedom of expression. On the other hand, the 'right' to not see violent, hateful, sexual or privacy-infringing content and to be protected from online incitement of violence. Under-moderation, this is the tenor of the past several years, can have grave consequences for individuals and entire communities (Amnesty International 2022). Table 5.9 shows data from YouTube, illustrating the effects of automated moderation on the views of potentially policy-infringing content. Shown below is data for YouTube across a period of three years (2020–2022). Data is displayed for the third quarter of the year each and then annual intervals of data back to the earliest third quarter data available (YouTube 2022b).

The data suggests—not surprisingly—that automated detection decreases the share of videos removed for content policy violations ever seen by users. Videos picked up by other detection sources (for YouTube this means users, organisations or governments flagging content) are usually seen by people. In the case of many classes of content, such immediacy has great value. Abhorrent violence, pornography and terrorist propaganda may arguably not be suitable for young users. To wait for moderators to pick up the lead may well mean thousands or millions view content that will eventually be removed for policy violations. Still lacking explanation, the share of videos never seen by users decreased over time, from the third quarter of 2020 to the third quarter of 2022. In any case, much depends on the quality of the automated detection, which will likely matter when

Table 5.9 Share of removed videos not viewed, for automated and other detection, YouTube (2020–2022)

Detection type	*Q3/2020*	*Q3/2021*	*Q3/2022*
Automated detection	45.2%	38.7%	38.3%
All other detection sources	2.7%	0.8%	3.6%

balancing between over-moderation and under-moderation of platforms. This in turn depends on the quality of training data stemming from human moderators, which may be biased in a number of ways (Binns et al. 2017).

The over-time comparison of automation rates is an interesting indicator to observe trends and their stability of the algorithmic moderation. Arguably, such comparison can show how algorithmic moderation 'learns', taking over a larger share of the initial detection work from users and other actors. Table 5.10 shows data for the platform Facebook across a period of five years (2018–2022). Data is displayed for the third quarter of the year each and then annual intervals of data back to the earliest third quarter data available.

The data shows that the level of automated moderation is generally very high throughout the principles of Meta's Community Standards reported on. The overall rate of automation has increased over the last five

Table 5.10 Share of proactive detection by category and year, Facebook (2018–2022)

Category/principles	*Q3/2018*	*Q3/2019*	*Q3/2020*	*Q3/2021*	*Q3/2022*
Prevention of harm					
Bullying and harassment	14.8%	16.2%	31.0%	59.4%	67.8%
Suicide and self-injury	N/A	96.8%	95.7%	99.0%	98.6%
Dangerous organisations: organised hate	N/A	N/A	97.8%	96.4%	94.3%
Dangerous organisations: terrorism	99.3%	98.5%	99.8%	97.9%	99.1%
Violence and incitement	N/A	N/A	N/A	96.7%	94.3%
Protection of social groups					
Child endangerment: nudity and physical abuse	N/A	N/A	N/A	97.1%	97.5%
Child endangerment: sexual exploitation	N/A	N/A	N/A	99.1%	99.5%
Child nudity and sexual exploitation	99.1%	99.5%	99.5%	N/A	N/A
Adult nudity and sexual activity	97.3%	98.8%	98.2%	98.8%	96.9%
Violent and graphic content	96.7%	99.0%	99.5%	99.4%	99.1%
Hate speech	52.9%	80.6%	94.8%	96.5%	90.2%
Public interest					
Regulated goods: firearms	N/A	97.6%	96.2%	96.7%	98.3%
Regulated goods: drugs	N/A	93.8%	91.7%	94.1%	94.8%
Not categorised					
Fake accounts	99.6%	99.6%	99.4%	99.8%	99.6%
Spam	99.7%	99.9%	99.9%	99.6%	98.5%

years. For principles for which the automation rate has been relatively low still in 2018, such as principles against bullying and harassment, as well as hate speech, the automation shares have swiftly risen (from 14.8% to 67.8% for the former, and from 52.2% to 90.2% for the latter). These two principles illustrate how technically challenging the detection of hate speech, bullying and harassment are, given such expressions' contextual character. For other principles, the rate of automation has been consistently above 96–99% over the same period, amounting to the overall automated moderation of 94.2% referred to in Table 5.8. The data reported on a principle basis 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 (and perhaps spam). In addition, as pointed out above, although demanded by some civil society documents, not all automated content decisions are being reviewed by a human.

Far from a limitation of automated moderation, the platforms studied here have extended their automated detection mechanisms and scaled them up. While data for Twitter is not readily available in their transparency report, it can be assumed that the service does not differ from the others on this indicator. We see once again that platforms become more similar and converge on the notion of near-complete automation of moderation, with moderators taking care of appeals (if at all). TikTok lags behind relatively speaking, but this might merely be a snapshot. The platform's overall automation rate has increased from 33.91% in the third quarter of 2021 to 48.02% in the third quarter of 2022 (TikTok 2022e). Whether this affords TikTok more appreciation by civil society is doubtful. The demands of civil society, as discussed in Chap. 3, are clearly not met. Neither is only "manifestly illegal" content being automatically detected, for instance, through a hash procedure as often done with copyrighted and terrorist content (Gorwa et al. 2020). Instead, increasingly, automation dominates moderation across content categories. A number of cases, in which Meta's Oversight Board has ordered the company to improve automated detection, shows that there are still regular and decisive failings of automated moderation even where, arguably, the most extensive set of training data should be available (Oversight Board 2020, 2022). As this subsection shows, being able to judge platforms on their self-reported data is key to understanding empirical developments and how they relate to any standard for content moderation extrapolated from civil society-authored Internet Bills of Rights. The following subsection shows how

transparency reporting has also converged on a relatively extensive standard.

5.4.2 Transparency Reporting: Which Standard to Adopt?

Transparency is a core principle demanded in 16 of the civil society charters. For a platform to be transparent about its content moderation allows for others to scrutinise it, including but not limited to the question of whether the platform promotes and protects human rights. Why—apart from their human rights commitments, do platforms engage in activities that foster transparency and thus accountability? The goal of platforms when engaging in transparency-increasing measures—such as the creation of transparency reports and transparency microsites (transparency centres) that bring together various metrics and by engaging researchers and others—is to gain legitimacy. Transparency reports have become a key tool "to cultivate legitimacy with users and civil society organizations" (Suzor et al. 2018, 393). Legitimacy relates to the "right to govern" in the eyes of the users (the governed) but also, as a response or pre-emptive measure to public regulation, in the eyes of politically powerful stakeholders. Indeed, increasingly, regulators prescribe how platforms are required to report about their content moderation practices. India's Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules of 2021 require larger platforms to produce monthly reports about complaints and actions taken (Tewari 2022). The Digital Services Act (DSA) and the Platform Accountability and Transparency Act (PATA) are respectively a recently adopted EU regulation and a US legislative proposal that would increase transparency requirements for platforms significantly. Transparency can further be enhanced by providing data on content moderation to academic researchers. Consequently, regulators increasingly perceive "access to data for empirical research (…) as a necessary step in ensuring transparency and accountability" (Nonnecke and Carlton 2022, 610). The DSA specifically "seeks a new level of granularity in transparency surrounding content moderation practices", surpassing previous national transparency reporting requirements such as the bi-annual requirement of the German NetzDG and India's transparency rules (Tewari 2022). Less in the focus of public attention yet already codified are transparency reporting standards for platforms towards their business partners as part of the EU's Platform-to-business Regulation of 2019 (European Union 2019). Based on this, Meta now regularly reports to

their advertisers not only the number of complaints lodged against decisions and the type of complaint, but also the average time to process such appeals.

With regard to copyright-related notice-and-takedowns, additional voluntary transparency practices exist. For instance, the Lumen project at Harvard's Berkman Klein Center for Internet & Society collects and makes available DMCA takedown notices from those who receive them. This allows researchers and others to gain an understanding of individual practices and overarching trends. As of late 2021, Lumen included more than 18 million notices, most of them copyright-related, from companies such as Wikipedia, Google (including YouTube) and Twitter (Lumen 2022). Pending the passage of some of the more stringent legislative proposals, what is the level of transparency if platforms are being compared? Until 2019, the Electronic Frontier Foundation produced an annual report in which the content moderation practices of 16 online platforms were compared based on six overarching categories such as transparency about government takedown requests, transparency about content removal based on the platform's policies, transparency about appeals and even endorsement of one of the civil society-issued documents, the Santa Clara Principles (Crocker et al. 2019). In 2019, for the last iteration of the report, Reddit was able to receive a star in all six categories with Facebook, Instagram, Vimeo and Dailymotion performing particularly poorly.

Ranking Digital Rights produces an annual Big Tech Scorecard, which evaluates the corporate accountability of 14 (2022a) large digital platforms from the United States, China, South Korea and Russia, subdivided by offered services, such as Facebook and Instagram for Meta Inc. (Ranking Digital Tech 2022a). The report includes indicators on content moderation transparency reporting in its section on freedom of expression, such as the reporting of "data about government demands to restrict content or accounts", and data about platform policy enforcement. Overall, in that section, the report finds that Twitter "took the top spot, for its detailed content policies and public data about moderation of user-generated content" (Ranking Digital Tech 2022b). Table 5.11 shows an excerpt of the results for the subcategory of algorithmic transparency, also relevant for the proceeding subsection of this chapter.

The data in Table 5.11 suggests generally low scores in algorithmic transparency across the tech sector, with platforms doing relatively well. TikTok is not included in the ranking. There are various other projects

Table 5.11 Ranking 'algorithmic transparency', Big Tech Scorecard 2022 by Ranking Digital Rights (2022a)

Rank (out of 14)	*Platform*	*Score*
1	Meta	22%
4	Twitter	20%
6	YouTube (Google)	14%

Table 5.12 Reporting of content policy-based moderation data (2016–2022)

Platform	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*	*2022*
Facebook	No	Yes	Yes	Yes	Yes	Yes	Yes
Instagram	No	No	No	Yes	Yes	Yes	Yes
Twitter	No	No	Yes	Yes	Yes	Yes	Yes
TikTok	No	No	No	Yes	Yes	Yes	Yes
YouTube	No	Yes	Yes	Yes	Yes	Yes	Yes

examining platform moderation transparency. New America's Transparency Report Tracking Tool is a continuously updated project that curates data from transparency reports of six services of five platform companies (Singh and Doty 2021). The tracking tool allows readers to find in one place the categories of transparency reporting included in transparency reports of Facebook, Instagram, Reddit, TikTok, Twitter and YouTube. The tracking tool also allows an over-time view of when certain reporting categories have been added or dropped by the services. What is not included is any attempt to find common categories of transparency reporting that would allow to compare changes over time between the different platforms. On a general level, it is worth examining when the five platforms' services have started to disclose transparency reports concerning moderation actions based on their content policies (as opposed to government requests, etc.). Such longitudinal data is presented in Table 5.12.[4]

As Table 5.12 shows, with regard to content moderation action based on conflict with platform content policies, there appears to be a degree of isomorphism across platforms. The evolution of this common practice is interesting, though. Content policy is actually not one of the first categories of content removal that was introduced into transparency reporting

[4] The underlying data is derived from Quintais et al. (2022).

(Quintais et al. 2022). The amount of content moderated was first shared by Facebook and YouTube in 2017. Only subsequently in 2018, Twitter started to disclose the data for content removed due to its Twitter Rules. Instagram and TikTok started to reveal the data for such platform-policy-based moderation of content in 2019. However, the quality of reporting also matters greatly. On the one hand, what is crucially lacking is a common standard by which data is reported, even if the reporting slowly converges towards common criteria. It remains difficult to make data actually comparable. On the other hand, the protection of human rights requires an in-depth understanding by the public and by policymakers regarding the processes at play, especially concerning the harms that platforms may have data on.

Whether such data is relevant to content moderation can often only be seen once additional reasons to restrict specific content are established. The Facebook Files relate to a recent whistleblowing and succeeding scandal, in which one shortcoming of Meta received a particularly high degree of media attention. The leaks demonstrated that the company had long known the impact of the use of its platforms on the mental-health of young adults, specifically that "Instagram is harmful for a sizable percentage [young users], most notably teenage girls" (Wells et al. 2021). Not being transparent where internal data suggests major issues is highly problematic. For instance, Leightley et al. (2022) argue that access to platform data could be used to better understand the mental-health implications, suggesting that "limited data access by researchers is preventing such advances from being made". Whether content governance would be a tool to tackle these challenges would have to be established. This demonstrates that transparency is required in a serious and comprehensive way in order to protect human rights, rather than being a mere exercise of counting and publishing high-level data.

Overall, this chapter demonstrates a number of noteworthy trends when it comes to integrating human rights claims into platform policies and with regard to both substantial outcomes and procedural content moderation practices. It becomes apparent that the platforms are using the language of human rights but often not in their content policies. With regard to the latter, it can be said that—as far as reported—content moderation outcomes can be better understood through the lenses of the Internet Bills of Rights introduced in Chap. 4. Such a perspective also allows us to observe that the five studied platforms appear to largely converge on a number of indicators. Convergence on a common standard

is also a useful narrative to understand practices related to automated content moderation and transparency reporting, even if—particularly with regard to automation of content moderation—there are substantial deviations from civil society demands. Importantly, even a standard of practices on which platforms converge, while referencing human rights at least in name, does not suffice to fully solve the content governance dilemma platforms face. Deep engagement with human rights standards and continuous exchange with those who defend them are needed to ensure human rights are indeed realised. The ongoing process by UNESCO for "Guidelines for Regulating Digital Platforms" (2022) might be an additional way forward, as may be general moves towards so-called platform councils that bring together different stakeholders to counsel platform policy teams (Tworek 2019).

References

AccessNow. 2020. *Open Letter to Facebook, Twitter, and YouTube: Stop Silencing Critical Voices from the Middle East and North Africa, December 2020*. Open Letter. https://www.accessnow.org/facebook-twitter-youtube-stop-silencing-critical-voices-mena/. Accessed December 21, 2022.

Al Ghussain, Alia. 2022. *Meta's Human Rights Report Ignores the Real Threat the Company Poses to Human Rights Worldwide*. Amnesty International. https://www.amnesty.org/en/latest/campaigns/2022/07/metas-human-rights-report-ignores-the-real-threat-the-company-poses-to-human-rights-worldwide/. Accessed February 20, 2023.

Allan, Richard. 2018. Hard Questions: Where Do We Draw the Line on Free Expression? *about.fb.com*. https://about.fb.com/news/2018/08/hard-questions-free-expression/. Accessed September 24, 2022.

Amnesty International. 2022. Myanmar: Facebook's Systems Promoted Violence Against Rohingya; Meta Owes Reparations. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/. Accessed February 20, 2023.

Barlow, John Perry. 1996. A Declaration of Independence of Cyberspace. *eff.org*. https://www.eff.org/cyberspace-independence. Accessed October 28, 2022.

Barrett, Bridget, and Daniel Kreiss. 2019. Platform Transience: Changes in Facebook's Policies, Procedures, and Affordances in Global Electoral Politics. *Internet Policy Review* 8 (4): 1–22.

Binns, Reubes, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In *Social Informatics 9th International Conference, SocInfo 2017, Oxford, UK,*

September 13–15, 2017, Proceedings, Part II, ed. Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, 405–415. Oxford: Springer.

Bygrave, Lee A. 2015. *Internet Governance by Contract*. Oxford: Oxford University Press.

Celeste, Edoardo. 2019. Terms of Service and Bills of Rights: New Mechanisms of Constitutionalisation in the Social Media Environment? *International Review of Law, Computers & Technology* 33 (2): 122–138. https://doi.org/10.1080/13600869.2018.1475898.

———. 2022. *Digital Constitutionalism: The Role of Internet Bills of Rights*. London: Routledge.

Citron, Danielle Keats, and Benjamin Wittes. 2017. The Internet Will Not Break: Denying Bad Samaritans Sec. 230 Immunity. *Fordham Law Review* 86 (2).

Conger, Kate. 2018. Google Removes 'Don't Be Evil' Clause from Its Code of Conduct. *gizmodo.com*. https://gizmodo.com/google-removes-nearly-all-mentions-of-dont-be-evil-from-1826153393. Accessed September 22, 2022.

Crocker, Andrew, Gennie Gebhart, Aaron Mackey, Kurt Opsahl, Hayley Tsukayama, Jamie Lee Williams, and Jillian C. York. 2019. Who Has Your Back? *eff.org*. https://www.eff.org/wp/who-has-your-back-2019. Accessed October 30, 2022.

Deutsche Welle. 2015. European Court of Human Rights Rules Turkey's Ban on YouTube Violated Rights. https://www.dw.com/en/european-court-of-human-rights-rules-turkeys-ban-on-youtube-violated-rights/a-18886693. Accessed December 21, 2022.

European Union. 2019. Regulation (EU) 2019/1150. 20 June 2019. *Promoting Fairness and Transparency for Business Users of Online Intermediation Services*. https://eur-lex.europa.eu/eli/reg/2019/1150/oj.

Facebook. 2007. Facebook Community Standards (Version of 30 August 2007). Platform Governance Archive. https://github.com/PlatformGovernanceArchive/pga-corpus/tree/main/Versions/PDF/Facebook. Accessed June 16, 2023.

———. 2021. Facebook Community Standards (Version of 30 January 2021). Platform Governance Archive. https://github.com/PlatformGovernanceArchive/pga-corpus/tree/main/Versions/PDF/Facebook. Accessed June 16, 2023.

Google. 2022. Human rights. *Google.com*. https://about.google/human-rights/. Accessed September 24, 2022.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data & Society* 7 (1).

Guo, Eileen. 2021. Deplatforming Trump Will Work, Even if it Won't Solve Everything. *MIT Technology Review*. https://www.technologyreview.

com/2021/01/08/1015956/twitter-bans-trump-deplatforming/. Accessed October 20, 2022.

Ha, Anthony. 2021. YouTube Will Start Penalizing Channels that Post Election Misinformation. *TechCrunch*. https://techcrunch.com/2021/01/07/youtube-election-strikes/. Accessed January 20, 2023.

Hemphill, T. A. 2019. 'Techlash', responsible innovation, and the self-regulatory organization. *Journal of Responsible Innovation* 6 (2), 240–247.

Helfer, Laurence R., and Molly K. Land. 2022. The Facebook Oversight Board's Human Rights Future. *Duke Law School Public Law & Legal Theory and Research Paper Series No. 2022-47* 44 (6). https://doi.org/10.2139/ssrn.4197107.

Horwitz, Jeff. 2021. Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353. Accessed December 21, 2022.

Iqbal, Mansoor. 2022. TikTok Revenue and Usage Statistics (2022). *Businessofapps.com*. https://www.businessofapps.com/data/tik-tok-statistics/. Accessed September 24, 2022.

Katzenbach, Christian. 2021. AI Will Fix This—The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society* 8 (2). https://doi.org/10.1177/20539517211046182.

Kelly, Makena. 2021. Biden Revokes and Replaces Trump Orders Banning TikTok and WeChat. *The Verge*. https://www.theverge.com/2021/6/9/22525953/biden-tiktok-wechat-trump-bans-revoked-alipay. Accessed September 24, 2022.

Kettemann, Matthias, and Wolfgang Schulz. 2020. Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study. *Working Papers of the Hans-Bredow-Institut, Works in Progress# 1*. Hamburg: Hans-Bredow-Institut.

Klonick, K. 2018. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131 (6): 1598–1670.

Kuczerawy, Aleksandra, and Jef Ausloos. 2015. From Notice-and-Takedown to Notice-and-Delist: Implementing Google Spain. *Colorado Technology Law Journal* 14 (2): 219–258.

Leightley, Daniel, Amanda Bye, Ben Carter, Kylee Trevillion, Stella Branthonne-Foster, Maria Liakata, Anthony Wood, Dennis Ougrin, Amy Orben, Tamsin Ford, and Rina Dutta. 2022. Maximising the Positive and Minimising the Negative: Social Media Data to Study Youth Mental Health with Informed Consent. *Frontiers in Psychiatry* 13.

Lumen. 2022. About us. *lumendatabase.org*. https://lumendatabase.org/pages/about. Accessed October 28, 2022.

Meta. 2022a. Facebook Community Standards. *transparency.fb.com*. https://transparency.fb.com/policies/community-standards/. Accessed December 21, 2022.

———. 2022b. Community Standards Enforcement Report. *transparency.fb.com*. https://transparency.fb.com/data/community-standards-enforcement/. Accessed December 21, 2022.

———. 2022c. Corporate Human Rights Policy. https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf. Accessed December 14, 2022.

Milmo, Dan. 2022. Twitter Says it Suspends 1m Spam Users a Day as Elon Musk Row Deepens. https://www.theguardian.com/technology/2022/jul/07/twitter-says-it-suspends-1m-spam-users-a-day-as-elon-musk-dispute-deepens. Accessed September 22, 2022.

Nonnecke, Brandie, and Camille Carlton. 2022. EU and US Legislation Seek to Open up Digital Platform Data. *Science* 375 (6581): 610–612.

Oversight Board. 2020. Breast Cancer Symptoms and Nudity. Case 2020-004-IG-UA. https://www.oversightboard.com/decision/IG-7THR3SI1

———. 2021. Former President Trump's Suspension. Case 2021-001-FB-FBR. https://www.oversightboard.com/decision/FB-691QAMHJ/

———. 2022. Iran Protest Slogan. Case 2022-013-FB-UA. https://www.oversightboard.com/decision/FB-ZT6AJS4X/

Quintais, João Pedro, Péter Mezei, István Harkai, João Carlos Magalhaes, Christian Katzenbach, Sebastian Felix Schwemer, and Thomas Riis. 2022. *Copyright Content Moderation in the EU: An Interdisciplinary Mapping Analysis*. reCreating Europe Report. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4210278.

Ranking Digital Tech. 2022a. The Big Tech Scorecard 2022. *Rankingdigitalrights.org*. https://rankingdigitalrights.org/index2022/. Accessed October 31, 2022.

———. 2022b. Key Findings from the 2022 RDR Big Tech Scorecard. *Rankingdigitalrights.org*. https://rankingdigitalrights.org/bts22/. Accessed October 5, 2022.

Singh, Spandana, and Doty, Leila. 2021. The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules. *Newamerica.org*. https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/. Accessed October 28, 2022.

Statista. 2022. Most Popular Social Networks Worldwide as of January 2022, Ranked by Number of Monthly Active Users. *statista.com*. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Accessed September 23, 2022.

Suzor, Nicolas. 2019. *Lawless: The Secret Rules that Govern Our Digital Lives*. Cambridge: Cambridge University Press.

Suzor, Nicolas, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the Legitimacy of Platform Governance: A Review of Research and a Shared Research Agenda. *International Communication Gazette* 80 (4): 385–400. https://doi.org/10.1177/1748048518757142.

Tewari, Shreya. 2022. *Transparency Initiatives in the DSA: An Exciting Step Forward in Transparency Reporting. Lumen Project.* https://www.lumendatabase.org/blog_entries/transparency-initiatives-in-the-dsa-an-exciting-step-forward-in-transparency-reporting. Accessed October 28, 2022.

The Guardian. 2022. TikTok Moves to Ease Fears Amid Report Workers in China Accessed US Users' Data. https://www.theguardian.com/technology/2022/jun/17/tiktok-us-user-data-china-bytedance. Accessed September 22, 2022.

TikTok. 2021a. Community Guidelines Enforcement Report, January 1, 2021–March 31, 2021. *tiktok.com*. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2021-1/. Accessed January 12, 2023.

———. 2021b. Community Guidelines Enforcement Report, April 1, 2021–June 30, 2021. *tiktok.com*. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2021-2/. Accessed January 12, 2023.

———. 2022a. Upholding Human Rights. *tiktok.com*. https://www.tiktok.com/transparency/en-au/upholding-human-rights/. Accessed November 24, 2022.

———. 2022b. Community Guidelines. *tiktok.com*. https://www.tiktok.com/community-guidelines?lang=en. Accessed November 24, 2022.

———. 2022c. Community Guidelines Enforcement Report, July 1, 2021–September 30, 2021. *tiktok.com*. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2021-3/. Accessed January 12, 2023.

———. 2022d. Community Guidelines Enforcement Report, October 1, 2021–December 31, 2021. *tiktok.com*. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2021-4/. Accessed January 12, 2023.

———. 2022e. Community Guidelines Enforcement Report, July 1, 2022–September 30, 2022. *tiktok.com*. https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2022-3/. Accessed January 13, 2023.

Twitter. 2009. The Twitter Rules (Version of 18 January 2009). Platform Governance Archive. https://github.com/PlatformGovernanceArchive/pga-corpus/tree/main/Versions/PDF/Twitter. Accessed June 16, 2023.

———. 2020. The Twitter Rules (Version of 28 October 2020). Platform Governance Archive. https://github.com/PlatformGovernanceArchive/pga-corpus/tree/main/Versions/PDF/Twitter. Accessed June 16, 2023.

———. 2021. The Twitter Rules (Version of 27 January 2021). Platform Governance Archive. https://github.com/PlatformGovernanceArchive/pga-corpus/tree/main/Versions/PDF/Twitter. Accessed June 16, 2023.

———. 2022a. Defending and Respecting the Rights of People Using Our Service. *twitter.com*. https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice. Accessed September 18, 2022.

———. 2022b. Rules Enforcement. *Transparency.twitter.com*. https://transparency.twitter.com/en/reports/rules-enforcement.html. Accessed January 13, 2023.

Tworek, Heidi. 2019. Social Media Councils. In *Models for Platform Governance—A CIGI Essay Series*, 97–102. Waterloo: Centre for International Governance Innovation.

UNESCO. 2022. *Guidelines for Regulating Digital Platforms: A Multistakeholder Approach to Safeguarding Freedom of Expression and Access to Information*. CI-FEJ/FOEO/3 Rev. https://unesdoc.unesco.org/ark:/48223/pf0000384031

United Nations. 2018. *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye*. UN Doc A/73/348.

———. 2019. *Report of Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye*. UN Doc A/47/486.

Wells, Georgia, Jeff Horwitz, and Deepa Seetharaman. 2021. Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?mod=hp_lead_pos7&mod=article_inline. Accessed September 22, 2022.

YouTube. 2022a. Community Guidelines. *youtube.com*. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/. Accessed September 24, 2022.

———. 2022b. Transparency Report. *youtube.com*. https://transparencyreport.google.com/youtube-policy/. Accessed January 14, 2022.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

Zuckerberg, Mark. 2019. *Facebook's Commitment to the Oversight Board, September 2019*. Open Letter. https://about.fb.com/wp-content/uploads/2019/09/letter-from-mark-zuckerberg-on-oversight-board-charter.pdf. Accessed September 24, 2022.