

10. Unpacking content moderation: The rise of social media platforms as online civil courts

Catalina Goanta and Pietro Ortolani

10.1 INTRODUCTION

In today's data-driven economy, social media platforms enable an ever-growing range of interactions. As a consequence, users have been increasingly generating 'content', such as tweets, Facebook posts, or Instagram images. Unavoidably, such interaction (which is often global in scale) also generates disputes. For example, social media content can infringe copyright, constitute an unfair commercial practice, or violate privacy. Social media platforms allow users to 'report' content, i.e., to file complaints. Through these reporting mechanisms, users request the platforms to take action, and 'moderate' the content. Moderation typically amounts to removing the content, making it inaccessible, and/or suspending or terminating the accounts of those who posted the content. The purpose of this chapter is to analyse content moderation in relation to the dispute resolution architectures currently being built by social media platforms. By moderating content, social media platforms operate as veritable online courts.

In the past years, content governance has found itself high on the agenda of many social media platforms. Platforms have been developing private infrastructures intended to enhance self-regulation (e.g., Facebook Oversight Board; Twitch Safety Council),¹ or to ensure compliance with mandatory norms (e.g., the EU Audiovisual Media Services Directive).² Through these initiatives, platforms attempt to show their willingness to create or adhere to

¹ 'Facebook Oversight Board' <https://oversightboard.com> accessed 29 April 2021; 'Introducing the Twitch Safety Advisory Council' (*Twitch*, 14 May 2020) www.blog.twitch.tv/en/2020/05/14/introducing-the-twitch-safety-advisory-council/ accessed 29 April 2021.

² Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or

rules that ought to protect public interest. However, these very platforms are essentially private entities, not designed to uphold fundamental rights in the same way as sovereign States. To a certain extent, this explains why platforms such as Facebook, Instagram, TikTok or Twitter prioritize certain forms of content moderation over others. In other words, platforms decide which rights should be protected (either algorithmically, or through human moderation), and which rights remain without a remedy. As an illustration, copyright or terrorist activity on social media is often algorithmically screened and taken down in case of infringement, whereas content affecting consumers (e.g., the non-disclosure of advertising)³ beyond criminal activities (e.g., illegal sales) is not. These preliminary observations show how uneven and fragmented the current landscape of content moderation is. As a consequence, it is necessary to scrutinize the phenomenon, assessing how, and to what extent, social media platforms contribute to the protection of different types of rights.

As already mentioned, when users have problems with the content they encounter, they can report it to the platform. In limited instances, the platform operates as a mediator; for instance, to comply with the US Digital Millennium Copyright Act,⁴ Youtube allows users to negotiate remedies such as sharing revenue in the case of so-called Content ID claims. In the vast majority of cases, however, platforms act as adjudicators. On the one hand, this phenomenon can enhance access to justice, allowing an efficient resolution of low-value, high-frequency disputes that cannot be realistically dealt with by individual courts. On the other hand, however, the way in which platforms currently perform these functions raises significant worries; at the moment, access to justice depends on private, largely automated decision-making processes established by platforms, lacking transparency and legitimacy.

The remainder of the chapter proceeds as follows. Section 10.2 lays down the theoretical framework, explaining why content moderation should be understood as a form of digital dispute resolution, and how social media platforms have progressively embraced their role as private adjudicators. This section includes an empirical overview of the reporting mechanisms of four social media platforms (Facebook, TikTok, Twitch and Twitter). In section 10.3, we explore the current shortcomings of content moderation, pointing out three problems which currently affect the transparency and legitimacy

administrative action in Member States concerning the provision of audiovisual media services [2010] OJ L 95 (Audiovisual Media Services Directive).

³ The use of ‘advertorials’ constitutes an unfair commercial practice, pursuant to point 11 of Annex I to Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market, [2005] OJ L 149 (Unfair Commercial Practices Directive).

⁴ Digital Millennium Copyright Act of 1998.

of the dispute resolution activities carried out by platforms through content moderation. In section 10.4, we analyse Facebook's recent decision to institute a quasi-judicial body, the Oversight Board, to deal with delicate and significant content moderation disputes. Section 10.5 explores the possible future regulation of content moderation, with specific reference to the European Union and the proposal for a Digital Services Act (hereinafter, DSA). Finally, section 10.6 presents some conclusions.

10.2 CONTENT MODERATION AS CONFLICT RESOLUTION

Social media platforms have undergone numerous iterations, which resulted in fundamental changes as to user demographics, business models, and range of interactions. If, in its early days, Facebook had the aim of connecting people studying at the same university,⁵ its goals and uses gradually shifted towards content monetization. In 2021, Facebook also displays a marketplace connecting buyers and sellers,⁶ an 'ad library' allowing businesses to promote sponsored content into users' news feeds,⁷ as well as considerable volumes of undisclosed advertisement taking the form of goods or services endorsed by influencers,⁸ just to give a few examples of some defining characteristics. The same trends can be observed for other social media platforms as well.

In this landscape, an increasing variety of behaviours and interactions is channelled into content. Given the increasing complexity of the range of activities that users carry out through platforms, the past years were marked by mounting concern regarding platform governance, with a particular focus on content moderation policies. What should be allowed on platforms, and what should be taken down? Legal scholarship focusing on freedom of expression, intermediary liability, hate speech and democracy, among others, has thoroughly reflected upon the different facets of content moderation, at the crossroads between private ordering and public regulation.⁹

⁵ C McFadden, 'A Brief History of Facebook, Its Major Milestones' (*Interesting Engineering*, 2020) www.interestingengineering.com/history-of-facebook accessed 29 April 2021.

⁶ 'Facebook Marketplace' www.facebook.com/marketplace/ accessed 29 April 2021.

⁷ 'Facebook Ad Archive' www.facebook.com/ads/library/ accessed 29 April 2021.

⁸ See for instance C Goanta and S Ranchordás (eds), *The Regulation of Social Media Influencers* (Edward Elgar 2020).

⁹ J M Balkin, 'The Future of Free Expression in a Digital Age Free Speech and Press in the Digital Age' (2008) 36 *Pepperdine Law Review* 427; A Heldt, 'Borderline Speech: Caught in a Free Speech Limbo?' (2020) *Internet Policy Review* www

Social media are, to a certain extent, a breeding ground for conflict. Research done at the intersection of social science and computer science reveals that user-defined communities will inevitably experience conflict. To be sure, this is not an exclusive feature of the online realm; however, empirical research confirms that online communities typically entail a significant degree of conflict and, as such, generate the need for conflict resolution. In a study looking at 40 months of Reddit comments and posts, Kumar et al. show that ‘1% of all communities initiate 74% of all conflicts’ (‘wars’ or ‘raids’) on this social media platform.¹⁰ As a consequence, when social media platforms make binding decisions on content moderation, they unavoidably profile themselves as engaging in private dispute resolution, deciding on a wide range of delicate disputes where free speech must be balanced against other rights and values.¹¹ For instance, the Facebook Oversight Board has recently decided on whether former US President Donald Trump should be allowed back to the platform.¹² In light of this, it is necessary to take a closer look at social media content moderation through the ‘lens’ of dispute resolution.

10.2.1 Dealing with Conflicts on Social Media Platforms

The role of platforms as dispute resolution service providers has been inspired by eBay’s early model of online dispute resolution, which relied on a combination of reputation mechanisms, evidence submission and automated remedies,¹³ which in turn encouraged mediation experiments such as those

.policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510 accessed 29 April 2021.

¹⁰ S Kumar, W L Hamilton, J Leskovec and D Jurafsky, ‘Community Interaction and Conflict on the Web’ *The Web Conference* (2018) <https://snap.stanford.edu/conflict/> accessed 29 April 2021.

¹¹ K Klonick, ‘The New Governors: The People, Rules, and Processes Governing Online Speech’ (2018) 131 *Harvard Law Review* 1598–1670; O Pollicino and M Bassini, ‘Free Speech, Defamation and the Limits to Freedom of Expression in the EU: A Comparative Analysis’ in A Savin and J Trzaskowski (eds), *Research Handbook on EU Internet Law* (Edward Elgar 2014) 508–542.

¹² Oversight Board, Case decision 2021-001-FB-FBR. See also L Gradoni, ‘Constitutional Review via Facebook’s Oversight Board: How platform governance had its Marbury v Madison’ (*Verfassungsblog*, 2021) <https://verfassungsblog.de/fob-marbury-v-madison/> accessed 25 June 2021; K Klonick, ‘The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression’ (2020) 129 *Yale Law Journal* 2318–2499; E Douek, ‘Facebook’s Oversight Board: Move Fast with Stable Infrastructure and Humility’ (2019) 21 *North Carolina Journal of Law & Technology* 1.

¹³ L Mommers, ‘Visualization of Dispute Resolution: Establishing Trust by Recycling Reputation’ (2006) 15 *Information and Communications Technology Law*

conducted by Katsh et al.¹⁴ Other platforms, such as Wikipedia, developed an equally acclaimed dispute resolution approach combining mediation, arbitration, and negotiation.¹⁵ It would be beyond the scope of this chapter to assess what the added value of these dispute resolution mechanisms may be, with reference to the specific platform within which they are embedded, and the type of activities carried out on each platform. For the purposes of this chapter, suffice it to note that approaches to dispute resolution vary significantly across platforms. In fact, not all platforms deemed it necessary to design dispute resolution mechanisms; sharing economy platforms, in particular, have for a long time refused to resolve disputes among users. This has been the case for ride-sharing platform Uber, which at the time of writing still does not feature any formal dispute resolution centre. Airbnb initially adopted the same approach, but then introduced a 'Resolution Center' (in the words of the platform itself) which uses existing reservations logged by users as the starting point for two particular actions: requesting and sending money.¹⁶ Overall, according to the legal scholarship on ODR and the sharing economy, the dispute resolution role of platforms such as Uber or Airbnb remains fluid and a 'work in progress'.¹⁷ By contrast, social media platforms have recently adopted a more explicit role as adjudicators, most likely in the light of all the attention received by content governance in public policy around the world.

An example of the progressive 'judicialization' of social media platforms is the aforementioned Facebook Oversight Board, whose structures clearly resemble that of a judicial authority. This body will be analysed more thoroughly in section 4. A variation of this model is the Twitch Safety Advisory

175, 180. See also N Solovay and C K Reed, *The Internet and Dispute Resolution: Untangling the Web* (Law Journal Press 2003); R Koulu, *Law, Technology and Dispute Resolution* (Routledge 2018); L Edwards and C Wilson, 'Redress and Alternative Dispute Resolution in EU Cross-Border E-Commerce Transactions' (2007) 21 *International Review of Law Computers & Technology* 315, 326.

¹⁴ E Katsh, J Rifkin and A Gaitenby, 'E-Commerce, E-Disputes, and E-Dispute Resolution: In the Shadow of eBay Law' (2000) 15 *Ohio State Journal on Dispute Resolution* 705.

¹⁵ D A Hoffman and S K Mehra, 'Wikitruth through Wikiorder' (2009) 59 *Emory LJ* 151; O Rabinovich-Einy and E Katsh, 'Digital Justice: Reshaping Boundaries in an Online Dispute Resolution Environment' (2014) 1 *International Journal of Online Dispute Resolution* 5.

¹⁶ Airbnb Resolutions, <https://www.airbnb.com/resolutions> accessed 29 April 2021.

¹⁷ H Scheiwe Kulp and A L Kool, 'You Help Me, He Helps You: Dispute Systems Design in the Sharing Economy' (2015) 48 *Washington University Journal of Law & Policy Introduction* 179. See also M Cantero Gamito, 'Regulation.com. Self-Regulation and Contract Governance in the Platform Economy: A Research Agenda' (2017) 9 *European Journal of Legal Studies* 53.

Council, looking to advise Twitch on content moderation policies. Moreover, even in platforms (like Twitter) that have not instituted such a court-like body (yet), there is an increasing acknowledgement that content moderation requires conflict-resolving decisions, with important repercussions on freedom of expression. These decisions are normally taken through a combination of algorithms, which detect certain categories of content, and human moderators.

These, however, are rather new developments. In an initial phase, platforms refrained from developing dispute resolution infrastructures. As earlier iterations of social networks were generally user-defined, moderation of speech was left up to the communities built around the platforms, as an inherent characteristic of new, digital public spaces. More recently, platform governance has increasingly eroded the power of users in shaping their own participatory space, given the continuous automation of processes that will be discussed in further detail in the following section.¹⁸

Before looking at how platforms currently deal with users and conflicts, it is important to sketch a basic taxonomy of the possible conflicts that may come into play, in the context of content moderation. Departing from a broad understanding of ‘conflict’, as an expression of disagreement which may have legal repercussions, it is useful to distinguish among three main categories of conflicts:

- (i) *Conflicts between users and the platform*: users disagree with platform actions (e.g., taking down their content) or policies (e.g., nudity policies).
- (ii) *Conflicts between different users*: users disagree between themselves, whether bilaterally (e.g., making disparaging or misleading statements about a business competitor, posting content that violates another user’s trademarks or other IP rights, or violating another user’s consumer rights)¹⁹ or multilaterally (e.g., ‘cancel culture’ practices, whereby an ad-hoc swarm of social media users deliberately attack another user’s reputation).²⁰

¹⁸ C Goanta and J Spanakis, ‘Influencers and Social Media Recommender Systems: Unfair Commercial Practices in EU and US Law’ (TTLF Working Papers No. 54/2020) www.law.stanford.edu/publications/no-54-influencers-and-social-media-recommender-systems-unfair-commercial-practices-in-eu-and-us-law/ accessed 29 April 2021; J G Webster, ‘User Information Regimes: How Social Media Shape Patterns of Consumption’ (2010) 104 *Northwestern University Law Review* 593, 604.

¹⁹ Paolo Cavaliere, ‘Glawischnig-Piesczek v. Facebook on the Expanding Scope of Internet Service Providers’ Monitoring Obligations’ (2019) 5 *European Data Protection Law Review* 573.

²⁰ N K Carr, ‘How Can We End #CancelCulture - Tort Liability or Thumper's Rule?’ (2020) 28 *Catholic University Journal of Law and Technology* 133.

- (iii) *Conflicts involving third parties*: content posted on social media platforms may infringe the rights of non-users, for reasons analogous to those considered under (ii) above. Non-users, however, find themselves in a different situation, as they may be unable to activate the content moderation procedures of the platform.

While the mapping of stakeholders, interests and potential clashes which may escalate into conflicts can always be undertaken more systematically, this basic taxonomy fleshes out a crucial characteristic of this ecosystem: platforms are often parties to conflicts arising out of their private ordering. This raises considerable concerns relating to transparency and impartiality. From this perspective, social media platforms are the regulators, judges and enforcers of content moderation. Users present their case through procedures designed by the platform, limited by the goals of the platform, and often inaccessible to non-users.²¹

In sum, content moderation has an intrinsic dispute resolution dimension. On the one hand, platform-administered content moderation procedures are not the only avenue of redress: in the examples mentioned above, disputes arising out of violations of copyright, consumer or competition law may be brought before the competent national court(s), or resolved through a variety of 'traditional' ADR mechanisms. On the other hand, however, this is often unlikely to happen in practice. Many parties will not pursue any of those avenues, due to costs and other bottlenecks. For many categories of disputes, especially when the economic value is relatively limited, plaintiffs will remain inactive.²² Content moderation procedures, by contrast, are relatively simple to initiate and, most importantly, virtually costless. In practice, thus, social media platforms may be the only authority before which these disputes are brought for resolution. To be sure, it would be misleading to claim that content moderation on social media platforms should be exclusively understood as a form of private dispute resolution; the governance aspect of the phenomenon is obviously crucial and should not be disregarded. However, dispute resolution is one of the facets of content moderation, and a non-negligible one.

²¹ For example, Facebook does not provide customer support via telephone: S John, 'How to contact Facebook for problems with your account and other issues' (*Business Insider*, 2019) www.businessinsider.nl/how-to-contact-facebook-problems-with-account-other-issues?international=true&r=US accessed 29 April 2021.

²² F Weber, "'A chain reaction'" or the Necessity of Collective Actions for Consumers in Cartel Cases' (2018) 25 *Maastricht Journal of European and Comparative Law* 208; M Ioannidou, 'Compensatory Collective Redress for Low Value Consumer Claims in the EU: A Reality Check' (2019) 27 *European Review of Private Law* 1367.

Understanding content moderation as a form of dispute resolution may be useful both at the descriptive and at the normative level. Descriptively, it is possible to highlight similarities and differences between content moderation and other dispute resolution mechanisms (such as court litigation, or arbitration), which procedural lawyers are more familiar with. Such an approach can help understand to what extent current content moderation procedures offer sufficient due process guarantees, and how they perform these dispute resolution functions. Prescriptively, a procedural viewpoint can help determine whether content moderation is in need of reform. As already mentioned, for many categories of parties, platforms may be the only viable dispute resolution forum. Therefore, if platforms do not offer an avenue of redress, those parties will not be able to obtain the enforcement of the rights which, on paper, they would be entitled to. In certain cases, the question whether the ‘law in the books’ coincides with the ‘law in action’ may largely depend on whether social media platforms offer an effective possibility of redress through content moderation. Inasmuch as content moderation procedures constitute a sort of informal procedural law-making, parties left without redress may in practice have no right at all (*nullum jus sine actione*).

10.2.2 An Empirical Incursion into Social Media Reporting Systems

So far, we explored the topic of content moderation from a broad perspective of conflict resolution, in a theoretical sense. To further flesh out the issues that content moderation poses to platform governance and legal compliance, this section reports the findings of an empirical exercise, focusing on what we identify as ‘actionable content’ on the mobile applications of four social media platforms namely, Facebook, TikTok, Twitch, and Twitter. These platforms were selected due to a combination of factors based on their size, novelty and platform affordances.

Actionable content reflects the procedural grounds on which platforms allow users to report content. They can equally be referred to as ‘platform affordances’ in the context of social science research.²³ We systematically explored actionable content on the four selected platforms, in order to make an overview of existing labels used in these environments. We use observations gathered during November 2020²⁴ through a combination of digital ethnogra-

²³ See for instance B Rieder and J Hofmann, ‘Towards Platform Observability’ (2020) 9(4) *Internet Policy Review* <https://policyreview.info/articles/analysis/towards-platform-observability> accessed 22 June 2021.

²⁴ As platform affordances are constantly upgraded, we decided to focus on a given moment in time, instead of systematically observing changes across a more extensive period.

phy and walk-through methods, extensively used in communication studies and sociology, among other fields.²⁵ To report on these observations, in what follows, we will focus on two main points: (i) the availability of reporting affordances; (ii) the reporting labels attributed by platforms to actionable content. While platforms may also add relevant references to content moderation in community guidelines, it is important to note that our analysis focuses on describing the reporting mechanisms, which is an understudied aspect of content moderation by social media platforms.

10.2.2.1 The availability of reporting affordances

On their mobile applications, all four platforms provide reporting support, which can be engaged through as little as one click (or tap). Twitter and Facebook users can report posts by clicking the 'More options' button, which is placed on the upper right corner of each post to indicate additional actions which can be taken to engage with that content. TikTok offers the report feature under its 'Share' button. Unlike the other three platforms, which allow users to report specific posts made by another user, Twitch focuses exclusively on reporting users instead of posts, although the options of doing so indirectly reflect content reports as well.²⁶ To do so, users must engage with the 'Options' button found on the top right corner of a live stream.

While report buttons can be engaged with only one action, the way in which they are designed raises two questions. First, there is an issue of access: are these buttons easy to find for users who want to report potential content conflicts? This is a matter for further behavioural research at the intersection of computer-human interaction and website design. Second, there is a question of perception: platform interfaces shape the way in which users perceive available actions. If sharing and reporting content are placed under the same menu, the way in which users relate to and use reporting mechanisms may be affected. For instance, TikTok's 'Share' button is sometimes visible as an arrow, and sometimes as the WhatsApp icon, in order to communicate the desirability of sharing that content on other media. In other words, TikTok has adopted a much different approach in content dissemination than other platforms, as it facilitates the sharing of content for instance on WhatsApp. Users can thus send TikTok links to one another in order to increase traffic to the (web)app. If reports are available under this particular button, but users are nudged more towards sharing than towards filing complaints, it can be argued

²⁵ S Bishop, 'Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias' (2021) *Social Media + Society*.

²⁶ This is possible on Facebook, Twitter and TikTok as well, but with slight variations which are not covered by this empirical exercise.

that the design of this button is a dark pattern. A dark pattern is an interface design choice that may result in nudging users into behaviour patterns which are against their interests.²⁷ This conclusion, unfortunately, cannot be clearly drawn without additional behavioural research, which is necessary to reveal whether interface design is used in a way that is conducive to a constructive use of these reporting mechanisms.

10.2.2.2 The reporting labels attributed by platforms to actionable content

The content labels used by the four social media platforms under scrutiny for reporting purposes differ widely in granularity. We report on the main categories of actionable content, which we clustered according to very broad legal features. This resulted into six main clusters, as indicated in Table 10.1 below. Several observations can be made on the basis of this overview. First, it is noteworthy that most of the content which the four platforms allow users to report is labelled around content restrictions which seem to have a criminal nature, and which is likely to be prohibited under the law of most countries where the platforms operate. Examples include content relating to terrorism or jeopardizing minor safety. Hate speech is featured in this cluster as well, although national legal systems will most likely translate hate speech into a broader array of crimes, such as incitement to hatred, insult and/or defamation. Second, three out of the four platforms have separate support centres for intellectual property infringements. This often entails submitting reports which allow users to provide more evidence (e.g., screenshots), but which may also require a certain capacity to submit the complaint (e.g., being the holder of an intellectual property right). Third, when comparing the different clusters of actionable content in the light of platform affordances, Twitch stands out with its approach of focusing on reporting users, rather than content. From this perspective, Twitch does not police individual content, but user behaviour. If this behaviour is reported on the grounds allowed for by the platform, this can have repercussions for the user, which can include temporary or permanent banning.

All in all, reporting mechanisms as the ones briefly investigated in this section show that platforms have an inherent hierarchy of actionable content, which is determined according to internal policies relating to legal compliance.

²⁷ K B Cornelius, ‘Zombie Contracts, Dark Patterns of Design, and “documentation”’ (2019) 8(2) *Internet Policy Review* <https://policyreview.info/articles/analysis/zombie-contracts-dark-patterns-design-and-documentation> accessed 22 June 2021. See also A Mathur, M Kshirsagar and J Mayer, ‘What Makes a Dark Pattern... Dark?: Design Attributes, Normative Considerations, and Measurement Methods’ (2021) *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* 1–18.

Table 10.1 Actionable content on Facebook, TikTok, Twitch and Twitter

Actionable content cluster	Facebook	TikTok	Twitch	Twitter
Criminal	<ul style="list-style-type: none"> • Nudity • Violence • Hate speech • Harassment • Terrorism • Mocking victims • Bullying • Child abuse • Promoting drug use • Non-consensual intimate images 	<ul style="list-style-type: none"> • Dangerous organizations and individuals • Harassment or bullying • Minor safety • Violent and graphic content • Animal cruelty • Hate speech • Pornography and nudity 	<ul style="list-style-type: none"> • Violence or threats to harm • Harassment • Impersonation • Sexually explicit or sexually suggestive content • Sexual violence or exploitation • Extreme violence, gore, or other obscene content • Terrorism or acts of mass violence 	<ul style="list-style-type: none"> • Suspicious or spam • Abusive or harmful • It displays a sensitive photo or video
Consumer	<ul style="list-style-type: none"> • Fraud or scam • Unauthorized sales 	<ul style="list-style-type: none"> • Illegal activities and regulated goods • Misleading information 	<ul style="list-style-type: none"> • Spam, scam/malicious content • Prohibited game 	
Intellectual property	<ul style="list-style-type: none"> • Intellectual property 	<ul style="list-style-type: none"> • Separate form 	<ul style="list-style-type: none"> • Separate form 	<ul style="list-style-type: none"> • Separate form
Personhood			<ul style="list-style-type: none"> • Underaged user 	
Procedural			<ul style="list-style-type: none"> • Bits acceptable use policy violation^a • Other terms of service violation • Site suspension evasion • Chat ban evasion 	
Other	<ul style="list-style-type: none"> • False news • Spam • Suicide or self-injury 	<ul style="list-style-type: none"> • Suicide, self-harm and dangerous acts • Spam • Others 	<ul style="list-style-type: none"> • Self-harm • Offensive username • Miscategorized content or other category violation • Cheating in online game 	<ul style="list-style-type: none"> • Suspicious or spam • Self-harm or suicide

Note: ^a 'Bits' are virtual tokens which can be purchased on Twitch and spent on live streams by transferring them to favourite streamers.

By emphasizing intellectual property infringements, or content of a potentially criminal nature, platforms inadvertently show what mandatory legal norms they choose to prioritize in their content moderation approaches. Conversely, platforms choose not to engage with mandatory rules which may come from other fields of law, such as consumer protection. While in the European Union it is generally accepted that users who monetize content on social media must disclose sponsored posts,²⁸ and content monetization is skyrocketing, none of the investigated platforms had any actionable content label for non-disclosed advertising. In addition, the architecture of reporting mechanisms can also shed light on the reason why these mechanisms do or do not exist. On Twitch, user reporting, coupled with sanctions relating to account access if violations are determined to have occurred, supports the platform in policing content which goes against policy guidelines. We argue that reporting mechanisms are not primarily used to give users access to platform justice, but rather to channel the policy areas on which platforms want to take measures. One benefit of doing so is to use reporting mechanisms to crowdsource content labelling, which can be further used in the various content recognition models deployed internally.

10.3 THREE ISSUES WITH CONTENT MODERATION

As illustrated in the previous section, social media platforms allow users to file complaints about the content that is present on the platform. In doing so, platforms can potentially contribute to the effective protection of rights, and the enforcement of the law. However, in practice, these goals are pursued in an uneven and imbalanced fashion. While platforms generally provide an avenue of redress against content that would be qualified in several legal systems as criminally relevant, the same does not hold true for other fields of law, such as consumer protection law, as our empirical analysis shows. Just to give a practical example, on many social media platforms, content is ‘monetized’ according to a business model where famous users (‘influencers’) advertise products, without disclosing that they are indeed conducting advertisement. As already mentioned, this practice is incompatible with EU consumer law; for this reason, if platforms allowed users to report this type of content, they could meaningfully contribute to the enforcement of consumer law, making good on their own legal obligations to engage in such disclosures of advertisement.²⁹ However, as illustrated, many platforms do not offer this possibility to users. Platforms, hence, seem to ignore this type of business-to-consumer (B2C)

²⁸ C Riefa and L Clausen, ‘Towards Fairness in Digital Influencers’ Marketing Practices’ (2019) 8 *Journal of European Consumer and Market Law* 64.

²⁹ *Ibid.*

disputes, in spite of a growing body of transnational standardization, soft law and jurisprudence, all pointing in the same direction – that disclosing advertising on social media is a matter of consumer protection, and therefore (in some jurisdictions) a hard, mandatory limit to the contractual freedom enjoyed within a platform's private ordering.

Furthermore, platform infrastructure – in other words, the code on which each platform rests – creates semi-automated procedures which may limit the usefulness (or even fairness or legality) of reporting mechanisms. In many social media platforms, if a user reports content with the wrong label (i.e., choosing the wrong option out of the menu that the platform offers), this will lead to an unsuccessful report. What is more, the general terms and conditions of many social media platforms make no disclosure in this respect. Such a strict and opaque approach has a significant negative impact on the overall fairness of content moderation as a form of dispute resolution. To start with, the procedural frameworks used by social media platforms already limit the choice of a user in reporting content, which must be fitted within a closed number of often ill-designed categories. In addition, if the existing categories are not used properly, this deprives the user of any access to a potential remedy. It is crucial to consider the consequence of such automatism, in order to grasp the shortcomings of current content moderation practices from the perspective of dispute resolution. While the latter is supposed to be user-centric, and focused on remedying harms, content moderation in its current state is not even a matter of bilateral B2C customer care. Instead, what transpires when looking at the categories of actionable content is that reporting may very well be a techno-social process of crowdsourcing algorithmic progress, in such a way as to facilitate the evasion of liability by platforms as intermediaries. In other words, there is a significant risk that, instead of valuing the users' input in identifying unlawful content and providing effective redress, platforms mainly use the input of users to train their algorithms, and at the same time to preserve their immunity as mere hosts of user-generated information.³⁰ This leads to a situation where there is an appearance that platforms are trying to solve user conflicts, when in fact, other factors may play a role in the practice of implementing content moderation approaches.

Lastly, an emerging issue is to what extent reporting mechanisms enabled by platforms are still based on the 'notice and take-down' dynamic, and to what extent they are shifting towards a 'notice and other action' model, in

³⁰ See Art 14 of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L 178 (E-Commerce Directive).

the light of new European reforms. This will be further explored in section 10.5. The question, here, is what type of measure social media platforms should take, and whether the ‘take-down’ of the content is always an adequate remedy. In its general terms, Facebook makes the following specification: ‘We also can remove or restrict access to your content, services or information if we determine that doing so is reasonably necessary to avoid or mitigate adverse legal or regulatory impacts to Facebook.’³¹ While the removal of content is a rather clear measure, the same does not hold true for the restriction of access. When visiting a social media platform, the content we see is selected and curated by a so-called ‘recommender system’, which should present us with the most relevant information, given our data and past behaviour on the platform.³² Recommender systems, hence, are the backbone of information diffusion on social media platforms. An interesting question, then, is whether platforms should also allow different forms of redress, by adapting their recommender systems, so that for instance a certain piece of content does not circulate to the point of becoming ‘viral’. In principle, such a gradation of remedies could play an important role, allowing platforms to strike an adequate balance between freedom of expression and the protection of other rights and values.³³ However, recommender systems are opaque, and operate according to parameters that remain generally undisclosed, and are practically impossible to reverse-engineer. As a result, the current reality of content moderation lacks transparency and accountability. For instance, users on social media platforms may receive a so-called ‘shadow ban’, i.e., a measure whereby their profile or content is made less visible on the platform, but the affected user receives no notification or warning that they have been banned.

In the current landscape of content moderation, platforms are faced with increasing pressure to control content generated by its users. In doing so, the three issues of content moderation explored in this section (selective reporting mechanisms, their limited scope and algorithmic opacity) raise concerns relating to the actions platforms undertake to give users access to viable private remedies and procedures that do not affect their access to justice. Needless

³¹ ‘Facebook Terms’ <https://www.facebook.com/terms.php> accessed 29 April 2021.

³² P Covington, J Adams and E Sargin, ‘Deep Neural Networks for YouTube Recommendations’ (*Proceedings of the 10th ACM Conference on Recommender Systems* 2016) www.dl.acm.org/citation.cfm?id=2959190 accessed 29 April 2021; N Tintarev and J Masthoff, ‘Explaining Recommendations: Design and Evaluation’ in F Ricci, L Rokach and B Shapira (eds), *Recommender Systems Handbook* (Springer 2015) 382.

³³ E Goldman, ‘Content Moderation Remedies’ (2021) *Michigan Technology Law Review*, forthcoming.

to say, such an untransparent way of dealing with content-related disputes is incompatible with legitimate and accountable dispute resolution.

10.4 PLATFORMS EMBRACING THEIR ADJUDICATIVE FUNCTION: THE CASE OF THE OVERSIGHT BOARD

Social media platforms are embracing the dispute resolution dimension of content moderation in their public relations strategies, acknowledging (with varying degrees of openness) how content moderation constitutes a form of private adjudication. The most explicit example is the one of Facebook, with the institution of the Oversight Board, since no other social media platform has gone so far in mimicking an independent, precedent-creating institution which is supposed to rule on content moderation matters. In November 2018, Mark Zuckerberg announced Facebook's intention to 'create a new way for people to appeal content decisions to an independent body, whose decisions would be transparent and binding'.³⁴ Against that background, Facebook launched a public consultation process with a variety of stakeholders.³⁵ The results of the consultation fed into the design of this new body,³⁶ which was officially announced in September 2019 with the publication of a 'Charter', setting out inter alia the composition of the Board, its procedures, and its governance. This document, starting from its very title, explicitly confirms Facebook's ambition to create a court-like structure, borrowing from the experience (and the symbolic power) of both national and international courts. The Oversight Board Charter (hereafter 'Charter') has been rightly defined as a 'constitution-like document'³⁷ and can be understood as one of the steps of 'self-constitution-alization' of the platform.³⁸ Observing the phenomenon through the lens of dispute resolution, the Board can be understood as an example of transnational procedural law-making, whereby Facebook expressly acknowledges the adjudicative dimension of content moderation, and the need for adequate due process guarantees. For decades, procedural lawyers have highlighted how, by

³⁴ 'A Blueprint for Content Governance and Enforcement' www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/?hc_location=ufi accessed 29 April 2021.

³⁵ Brent Harris, 'Getting Input on an Oversight Board' (Facebook, 1 April 2019) <http://about.fb.com/news/2019/04/input-on-an-oversight-board/> accessed 29 April 2021.

³⁶ Klonick (n 12).

³⁷ *Ibid.*, 2457–8.

³⁸ G Teubner, 'Self-Constitutionalizing TNCs? On the Linkage of "Private" and "Public" Corporate Codes of Conduct' (2011) 18 *Indiana Journal of Global Legal Studies* 617.

looking at the procedures of the courts of a given State, we can gain insights as to how that State organizes its own authority, and what 'ideal of officialdom' it embraces.³⁹ Also from this point of view, then, a close look at the procedural dimension of content moderation can be meaningful; inasmuch as social media platforms emerge as repositories of power and as quasi-sovereign entities, the procedures they create can help us grasp their understanding of their own authority.

The Oversight Board, which is currently composed of 20 Members, can be seized in two main ways. First, the Board can review Facebook's content moderation decisions, on the request of one or more users. If a certain content has been taken down by Facebook, the user that posted the content may ask the Oversight Board to review the decision, and order that the content be reinstated on the platform. Alongside the initiative of users, Facebook itself may submit requests for review to the Board.⁴⁰ In both instances, the basis of decision making is supposed to be 'Facebook's content policies and values'.⁴¹ Drawing an analogy with litigation in civil and commercial matters, the users' initiative may be understood as a form of private enforcement of such 'policies and values', while Facebook's own referrals constitute an example of public enforcement. Needless to say, the substantive rules to be enforced here would not be provisions of national law, but 'policies and values' that legally bind the users inasmuch as they are enshrined in Facebook's terms and conditions. In its first decisions, however, the Oversight Board has interpreted its mandate broadly, and made references to international human rights law as a normative standard against which the legality of these policies and values should be assessed.⁴² To date, hence, it remains unclear to what extent content moderation at the Oversight Board will remain confined within the boundaries of Facebook's own internal rules and guidelines.

According to the Charter, 'the board has the discretion to choose which requests it will review and decide upon',⁴³ and should select the cases that have 'the greatest potential to guide future decisions and policies'.⁴⁴ This mechanism of case selection resembles the certiorari process, whereby the US Supreme Court selects a small number of cases out of all petitions, generally

³⁹ M Damaška, *The Faces of Justice and State Authority: A Comparative Approach to the Legal Process* (Yale University Press 1986).

⁴⁰ Art 2(1) Oversight Board Charter.

⁴¹ Art 2(2) Oversight Board Charter.

⁴² See Gradoni (n 12) with reference to Case Decision 2020-004-IG-UA, where the Oversight Board relied on the International Covenant on Civil and Political Rights (ICCPR) and the UN Guiding Principles on Business and Human Rights (UNGPs).

⁴³ Art 2(1) Oversight Board Charter.

⁴⁴ *Ibid.*

focusing on disputes which raise important legal questions, and can contribute to the development of a certain area of law. Furthermore, Facebook's ability to request from the Board guidance as to 'future decisions and policies'⁴⁵ echoes the role played by certain domestic and international courts in the context of preliminary ruling procedures. Such a discretion also entails that, as already mentioned, certain parties will obtain adequate redress through content moderation, while others will face the choice between bringing their case elsewhere, or factually giving up on their legal entitlements.

Each selected case is assigned to a panel of five members, assisted by a case manager. The procedure is rather inquisitorial: the Board receives information from Facebook itself, and can 'gather additional information, including through subject matter experts, research requests or translation services'.⁴⁶ The involvement of private parties, by contrast, is limited: according to the Charter, 'the posting person or the reporting person will have the opportunity to submit relevant and informed written statements to the Board'.⁴⁷ In other words, the party seeking review will be allowed to make written submissions and to present documentary evidence, but will generally not be invited to an oral hearing. In a nutshell, the Board procedure resembles that of many continental European constitutional courts, with a prevalence of written over oral submission, a limited role for the disputing parties, and an emphasis on the development of the law for the future (rather than the resolution of a specific dispute).

The Board's decisions 'include a determination on the content, as well as a corresponding plain language explanation of the board's rationale'. The Board may also decide to include in the decision a 'policy advisory statement, which will be taken into consideration by Facebook to guide its future policy development'.⁴⁸ This reinforces the conclusion that one of Facebook's primary reasons-for-action, when instituting the Board, was the need to develop a consistent body of precedent, to guide and justify future content moderation decisions and policies.⁴⁹ It remains to be seen to what extent Facebook will maintain its willingness to adhere to that case-law and whether the Board ends up reshaping the platform's policies in a substantial and invasive way. For the time being, the cases decided by the Board are not sufficient to make long-term predictions as to the Board's standpoint vis-à-vis Facebook.

The Board's decisions are public, and archived in a publicly accessible database, so as to ensure a strong persuasive (if not outright binding) precedential

⁴⁵ Ibid.

⁴⁶ Art 3(3) Oversight Board Charter.

⁴⁷ Art 3(3) Oversight Board Charter.

⁴⁸ Art 3(4) Oversight Board Charter.

⁴⁹ Klonick (n 12).

value.⁵⁰ The Charter declares the binding nature of the Board's decisions, and Facebook's obligation to implement them promptly.⁵¹

10.5 INJECTING LEGITIMACY FROM WITHOUT: THE DIGITAL SERVICES ACT (DSA) PROPOSAL

In its current state, content moderation on social media platforms can be analysed from the perspective of dispute resolution, existing at the crossroads between two normative tensions. The first tension concerns the decision-making basis, i.e., the substantive standards that platforms use, to determine whether a certain content should be taken down, or otherwise restricted. On the one hand, the relationship between platforms and users remains, to date, mainly a matter of contract law, so that platforms will normally seek a decision-making basis in their own terms and conditions, claiming the faculty to moderate content which does not comply with the obligations that the users contractually undertake. On the other hand, however, contract law does not exist in a vacuum, but within a broader framework of (national and supranational) law. The latter may sometimes restrict freedom of contract, typically to ensure the enforcement of mandatory provision of law, and to avoid violations of public policy. In the case of content moderation, this entails that the law may occasionally require the removal of illegal content, despite the fact that the content is not incompatible with the platform's terms and conditions. Conversely, the law may sometimes regard a content moderation decision as an excessive limitation of a user's fundamental rights (such as freedom of expression), despite the fact that the decision was taken in accordance with contractual terms and conditions that the user agreed to.

The second tension happens at the boundaries between platform self-regulation, and regulation at the national and supranational level. On the one hand, developments such as Facebook's decision to create the Oversight Board suggest a willingness on the side of the platforms to introduce further self-regulation and procedural guarantees in the field of content moderation (although the reasons for such willingness remain, to date, unclear).⁵² On the other hand, however, calls for more public regulation of content moderation are intensifying on both sides of the North Atlantic. This tension is, in and of itself, nothing new or specific to content moderation: whenever a form of private, out-of-court dispute resolution emerges and gains traction, the question unavoidably arises whether and to what extent sovereign entities should

⁵⁰ Art 3(6) Oversight Board Charter.

⁵¹ Art 4 Oversight Board Charter.

⁵² Klönick (n 12).

impose a ‘judicialization’ of that dispute resolution mechanism. This typically results in the injection of due process guarantees from without, through mandatory provisions of procedural law. In this vein, van Loo proposes that the procedural regulation of platforms should build on past experiences, namely on the regulation of other private dispute resolution schemes, such as credit card chargebacks.⁵³

Understanding the two aforementioned tensions (between contract and mandatory law, and between self- and public regulation) is useful to contextualize the EU proposal for a new Digital Services Act (hereinafter, DSA). The DSA is the most important reform in platform governance which will shape the future of European regulation on the matter during the next decades. Moreover, as it will be explored in this section, the DSA proposes extensive dispute resolution obligations for digital platforms.

On 15 December 2020, the European Commission published a proposal for a new ‘Regulation on a Single Market For Digital Services’,⁵⁴ commonly referred to as DSA. The proposed instrument, which forms part of European Digital Strategy *Shaping Europe’s Digital Future*, aims to overhaul the EU legislative framework concerning digital services and online communications, largely developed before the emergence of online platforms and social media. If the DSA proposal will be adopted, EU law will set forth a comprehensive set of rules on content moderation, applicable to social media platforms as well as other online intermediaries.

10.5.1 Content Moderation under the DSA: Distinctions and Purposes

The DSA proposal acknowledges that content moderation may serve two distinct purposes: the enforcement of a platform’s terms and conditions (when the content is incompatible with the users’ contractual obligations), or the enforcement of (national and/or EU) law (when the content is unlawful). The DSA proposal also acknowledges that content moderation can interfere with the principle of freedom of expression, and that sufficient guarantees must be put in place to ensure that the latter is respected.⁵⁵

The DSA proposal is based on two key distinctions:

- (a) The distinction between the moderation of ‘unlawful’ content (i.e., content that is incompatible with mandatory provisions of national or EU

⁵³ R Van Loo, ‘Federal Rules of Platform Procedure’ (2020) 88 *University of Chicago Law Review* 1.

⁵⁴ COM(2020) 825 final, 2020/0361(COD).

⁵⁵ Recital 22 and Art 12(2) DSA.

law), and the moderation of content that is legal but in violation of the platform's terms and conditions;

- (b) The distinction between moderation performed on the platform's own initiative, and moderation requested by a user through one of the notification systems described above (labelled by the DSA proposal as 'notice and action mechanisms').⁵⁶

The DSA does not regulate the procedures of the national judicial or administrative authorities that may be seized with a content moderation case: in accordance with the principle of procedural autonomy of the Member States, it is up to the latter to provide procedures that enable complainants (in our example, the consumer) to seek the removal or restriction of illegal content, as long as those procedures comply with the general principles of equivalence and effectiveness. By contrast, if the complainant triggers the available notice and action mechanism, the DSA sets forth an articulate body of procedural rules and guarantees.

10.5.2 Notice and Action Mechanisms

Pursuant to Article 14 of the DSA proposal, platforms must 'put mechanisms in place to allow any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content'. Moreover, Article 14 requires that the mechanisms at hand be available to 'any individual or entity', thus including non-users (i.e., parties that have no account on the platform). Nevertheless, not all notices are 'born equal' under the DSA: according to Article 19, certain entities can be granted the status of 'trusted flaggers', and the platform will then be obliged to process and decide upon the notice of trusted flaggers 'with priority and without delay'.⁵⁷

Furthermore, Article 14 requires that these mechanisms 'be easy to access, user-friendly, and allow for the submission of notices exclusively by electronic means'. While compliance with the last requirement is easy to assess, criteria such as ease of access and user-friendliness seem to be intrinsically subjective, and not easily enforceable. At first glance, therefore, there seems to be a risk that different platforms will interpret these requirements in different ways, ultimately resulting in procedural fragmentation. Article 34 of the DSA proposal, however, tries to limit this risk, by encouraging voluntary standardization across platforms. This approach seems to strike an adequate balance between

⁵⁶ Art 14 DSA.

⁵⁷ Art 19(1) DSA.

the desirability that platforms shape their own procedures and update them without unnecessary constraints, on the one hand, and the need for procedural uniformity and predictability, on the other hand. While it is obviously too early to predict whether such standardization will indeed occur, it is worth mentioning that encouraging precedents exist: in the field of international arbitration, for instance, arbitral institutions around the world have progressively converged towards a certain procedural model, borrowing from each other's experiences until a relatively uniform and predictable procedure emerged, while at the same time preserving sufficient space for flexibility.

Article 14(2) streamlines the submission of notices, specifying which elements should be included. The requirement that notices be 'precise' and 'adequately substantiated' ensures that the platform will have sufficient information to determine whether a certain content is unlawful and will be able to comply with the obligations that derive from the notice. After the notice, the platform must send the complainant a notification of receipt, and process the notice 'in a timely, diligent and objective manner'.⁵⁸ At this stage, the platform is free to determine whether the notice should be processed by a human, or by a machine; however, if the platform uses 'automated means', this must be disclosed to the complainant.⁵⁹ The platform must notify its decision to the individual or entity that has made the notice. That party must also be informed on the redress possibility in respect of the decision.⁶⁰

The DSA imposes a set of procedural guarantees on the platform, in cases where the latter decides to remove or disable access to a certain content, irrespective of whether the decision was initiated through a notice and action mechanism or not, and of whether the platform deemed the content to be unlawful, or incompatible with its own terms and conditions. The platform is obliged to provide a 'clear and specific statement of reasons'⁶¹ to the user that posted the content. Pursuant to Article 15, the statement must explain whether the content has been removed, or access to it has been disabled. In the latter case, the platform must also disclose 'the territorial scope of the disabling of access'. Second, the platform must indicate 'the facts and circumstances relied on in taking the decision'. Among these, the platform must disclose whether the decision originated from a notice, within the meaning of Article 14. Third, the platform must disclose whether the decision was made through 'automated means', including whether a machine detected or identified the content. Fourth, the platform must indicate what legal or contractual ground

⁵⁸ Art 14(6) DSA.

⁵⁹ Art 14(6) DSA.

⁶⁰ Art 14(5) DSA.

⁶¹ Art 15 DSA.

it relies on. If the content was deemed to be illegal, the platform must explain why the content is considered to run counter to the legal ground identified by the platform. Conversely, if the content is deemed to be incompatible with the platform's terms and conditions, the platform must indicate the contractual ground it relies upon and explain why the content is considered to be incompatible with that ground. Fifth, the statement must indicate the redress possibilities available to the user. In this respect, three main avenues are possible, the platform's internal complaint-handling mechanisms, out-of-court dispute settlement, and judicial redress. The next section will focus on the first avenue, as it directly concerns the adjudicative role of platforms. First, however, a further obligation must be pointed out. Article 15(4) requires that the platform to 'publish the decisions and the statements of reasons, referred to in paragraph 1, in a publicly accessible database managed by the Commission'. Not only does this enhance platform accountability by enabling external scrutiny, but in addition, the consolidation of a body of precedent encourages consistency not only within each platform, but also across them. From this point of view, the policy goals of the DSA are akin to those of Facebook with the institution of the Oversight Board, which also aims at ensuring more predictability and consistency of outcomes.

10.5.3 Internal Complaint Handling

Article 17 obliges platforms to put in place an 'effective internal complaint-handling system'. This system should be available to users for at least six months, following a decision finding that a certain content is illegal, or incompatible with the platform's terms and conditions. More specifically, the platform may have decided not only to remove or disable access to the content, but also to suspend or terminate the provision of the service to the user, and either suspend or terminate the user's account, or both. In all of these cases, the user must be able to access a 'user-friendly' complaint-handling mechanism, which must 'enable and facilitate the submission of sufficiently precise and adequately substantiated complaints'.⁶² Article 17 contains some binding provisions, dictating how the complaint-handling system should be organized by the platform. First, decision-making at this stage cannot be entirely automated; while the initial content moderation decision could be made by a machine (as long as it complies with the requirements of Art 15), this appellate stage necessarily requires human intervention. Furthermore, the provision requires that complaints be handled 'in a timely, diligent and objec-

⁶² Art 17(2) DSA.

tive manner'.⁶³ The DSA proposal does not expound on the standard of review that the platform should adopt; however, the decision should be 'reversed', if the complaint 'contains sufficient grounds' to conclude that the content was not illegal or incompatible with the platform's terms and conditions. In other words, whether the decision will be reversed depends on how convincing and well-illustrated the complaint is, while the platform does not have a duty to carry out any independent fact-finding. As a consequence, the platform may refuse to reverse a content moderation decision because the complaint does not sufficiently elaborate on the grounds it relies upon. The complaint-handling mechanism of Article 17, hence, should be seen not as a *de novo* assessment of the case, but as a review, which the platform will only carry out if and inasmuch as an adequately substantiated complaint has been brought. Conversely, the platform may refuse to reconsider whether the content is illegal or incompatible with its own terms and conditions, if the complainant failed to specify while he/she filed the complaint.

In light of the above, it is in the interest of a complainant to provide a sufficiently detailed overview of the grounds on which the complaint is based. In practice, the platform's interface will play a crucial role, taking into account the fact that the complainant is unlikely to have legal representation at this stage. As described above, the interface can facilitate or hinder the filing of the complaint, thus ensuring 'access to justice by design'. Ease of use, transparency, predictability and granularity should be key goals for platforms, when designing their complaint-handling systems. Standardization would be a significant added value: if different platforms offered comparable complaint-handling mechanisms, this could significantly facilitate the filing of sufficiently substantiated complaints. From this point of view, Article 34 comes across as a missed opportunity, as it encourages the standardization of the submission of notices under Article 14, but not of the submission of complaints under Article 17. This is all the truer, considering that the industry itself is currently attempting such standardization. The Oversight Board, for instance, has been designed as to be potentially adopted and used by platforms other than Facebook and Instagram in the future. As noted by Klonick, this is reflected by the language of the Oversight Board Charter, which is intentionally vague, so as allow other platforms to submit themselves to the Board's decision-making.⁶⁴

⁶³ Art 17(3) DSA.

⁶⁴ Klonick (n 12) 2475-6.

10.5.4 Beyond Platform Procedure: Out-of-court Dispute Settlement and Court Litigation

Complainants have the right to bring their case outside platforms. In this respect, two main avenues must be considered, out-of-court dispute settlement and court litigation. The DSA partially regulates both avenues.

The DSA mainly conceives of out-of-court dispute settlement as an extrajudicial appellate mechanism, available to parties whose complaints have been declined by the platform's internal complaint-handling system. Article 18 enables the national Digital Services Coordinators of each Member State to certify dispute settlement bodies established on the territory of the Member State in question. Additionally, Member States also have the possibility of directly setting up a dispute settlement body, which would then operate alongside the other certified bodies. A complainant can select any certified body, whose decision will be binding on the platform. Platforms have a duty to cooperate with the body in good faith, and promptly implement the body's decisions.

As for court litigation, in accordance with the principle of procedural autonomy of the Member States, the DSA does not regulate the procedures through which national courts resolve content moderation-related disputes. However, the DSA does regulate certain aspects concerning the enforcement of court decisions in the field of content moderation. More specifically, pursuant to Article 8(1), Member State courts can issue 'orders to act against illegal content'. Article 8 indicates the minimum elements that such orders must contain, so that platforms can understand what implementing measures are required, and act accordingly.

A particularly interesting aspect is the territorial scope of court orders. According to Article 8(2)(b), the territorial scope of the order should 'not exceed what is strictly necessary to achieve (the order's) objective'. Therefore, while violations of EU law would generally warrant an EU-wide removal of illegal content, content that is declared incompatible with national provisions of law will generally only be removed within the territory of the relevant Member State. The rationale of Article 8(2)(b) is to ensure that national and EU law will not be applied extraterritorially. Member State courts, hence, are indirectly discouraged from requiring worldwide action against content that may not be illegal outside the EU, or outside a single Member State. At the same time, this regulatory framework can lead to striking practical results. While illegal content may remain visible and accessible outside the territory of a certain Member State (or of the EU), content that is incompatible with the platform's terms and conditions will normally be removed by the platform with worldwide effects.

10.6 CONCLUSION

In this chapter, we have analysed the dispute resolution dimension of content moderation on social media. We have observed how social media platforms unavoidably perform certain dispute resolution functions, when taking decisions with respect to content. We also offered empirical insights from a study on the actionable content mechanisms offered by four social media platforms. From this point of view, content moderation can effectively contribute to the enforcement of substantive law, making sure that the parties' rights are actually translated into practice. We have observed how social media platforms have initially resisted such a dispute resolution role but have recently come to accept it. Despite this progressive acceptance, however, the reality of content moderation is far from satisfactory. Despite the current shortcomings, it is interesting to notice how social media platforms increasingly accept their role as adjudicators, and sometimes even set up court-like structures, such as the Oversight Board. In the future, due process guarantees may be injected by way of regulation. In the EU, the current DSA proposal has the potential to reshape content moderation as a form of digital dispute resolution.