10
Content Moderation: Outline

10.1 Evolution of Practices	185	**10.5 Due Process and Its Criticisms**	193
10.2 Regulation as Division of Powers	188	10.5.1 Why binding rules?	196
10.3 Rule-Setting and Its Legitimacy	189	10.5.2 Can procedural rights help?	196
10.4 User-Empowerment and Customisation	192	10.5.3 Why privatising justice?	197
		10.5.4 What about costs of compliance?	198

10.1 Evolution of Practices

While the term 'content moderation' is relatively novel, it describes a practice that is as old as the internet itself.[1] Probably the oldest content moderation debate concerns the distribution of unsolicited emails, better known as spam.[2] The zero cost of emailing immediately invited abuse. Given that charging per email sent was not viable,[3] the solutions ended up being social, technological, and legal. Email providers started classifying spam sent by people and automated technology and deprioritising it in inboxes.[4] As every email user knows, none works perfectly. Technology gets the classification wrong occasionally, important messages get lost, and laws do not fully deter companies from abusing email. The problem is somewhat under control, but users must exercise care when using email. Despite being based on open protocols, today, Apple Mail and Google's Gmail make up 87.02% of the total email client market.[5] Senders who these two companies wrongly classify have difficulties delivering anything to anyone.

[1] The early cases decided at the dawn of the internet era, such as *Prodigy* in the US, or *CompuServe* in Germany, were about content moderation. They were only framed as intermediary liability problems—see ch 4.

[2] Before emails, spam was observed on Usenet newsgroups. See 'History of Email Spam' (*Wikipedia*, 7 May 2023) <https://en.wikipedia.org/w/index.php?title=History_of_email_spam&oldid=1153712174> accessed 2 September 2023.

[3] In the 1990s, the question was debated under the notion of 'bit tax', see UNDP (ed), *Human Development Report 1999* (OUP 1999) 66 ('There is an urgent need to find the resources to fund the global communications revolution—to ensure that it is truly global. One proposal is a "bit tax"—a very small tax on the amount of data sent through the Internet. The costs for users would be negligible: sending 100 emails a day, each containing a 10-kilobyte document [a very long one], would raise a tax of just 1 cent').

[4] See the famous 2002 essay by Paul Graham, 'A Plan for Spam' (August 2002) <http://www.paulgraham.com/spam.html> accessed 3 September 2023.

[5] Oberlo, 'Most Used Email Clients Worldwide' (July 2023) <https://www.oberlo.com/statistics/most-used-email-clients> accessed 3 September 2023.

Content moderation on digital services, such as social media, suffers many of the same problems. Zero cost of use invites abuse. Again, we observe the implementation of social, technological, and legal solutions. And again, we observe how imperfect they are. What has changed, however, is that the range of content moderation surfaces and types of decisions are much wider (eg labelling, explaining, hiding, geotargeting, demonetising, etc).[6] Unlike with spam, the debate about content moderation today is more visceral because the underlying standards are more political in nature. The decisions are now not only about what content is an unsolicited commercial message but also what content counts as legitimate, bad, or newsworthy. And the scale of decision-making is truly mind-boggling. In just three months, Facebook took down 914,500,000 pieces of content and YouTube 4,496,933 videos.[7] To a large extent, content moderation has become a 'decision factory'.[8]

Content moderation has become an umbrella term for all activities of digital providers that affect users' content. Academics have offered many definitions over the years.[9] Scholars of science and technology studies and law have been documenting and studying content moderation practices and challenges for years.[10] In her influential work, Kate Klonick conceptualised content moderation as a new type of governance,[11] a 'bureaucracy',[12] drawing parallels to administrative and constitutional law.[13] In a model like hers, providers are portrayed as powerful rule-makers and enforcers.[14] The analogy with the administrative state invites us to think about the deficiencies of the process.[15] It sheds light on the legitimacy of the underlying rules,[16] and the extent of the state's pressure on such decision-making.[17]

[6] See extensively, in Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136(2) Harvard Law Review 526.

[7] ibid 537–38.

[8] Robyn Caplan, 'Content or Context Moderation?' (Data & Society 2018) 23 <https://datasociety.net/library/content-or-context-moderation/> accessed 3 September 2023.

[9] eg Grimmelmann defines it as 'the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse', see James Grimmelmann, 'The Virtues of Moderation' (2015) 17 Yale Journal of Law and Technology 42, 47; Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale UP 2018) 4, 6.

[10] Gillespie (n 9); Rebecca MacKinnon, *Consent of the Networked: The Worldwide Struggle for Internet Freedom* (Basic Books 2012); Hannah Bloch-Wehba, 'Automation in Moderation' (2020) 53 Cornell International Law Journal 41, 56; Robert Gorwa, Reuben Binns, and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7(1) Big Data & Society 2053951719897945, 2; Grimmelmann (n 9) 63–70; Danielle Keats Citron, *Hate Crimes in Cyberspace* (Harvard UP 2014); Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598; Nicolas P Suzor, *Lawless: The Secret Rules That Govern Our Digital Lives* (CUP 2019); Sarah T Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (Yale UP 2019); David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (Columbia Global Reports 2019).

[11] Klonick (n 10).

[12] The term is used by Douek to describe what she regards as standard content moderation scholarship, pointing to, among others, the works of influential works of Kate Klonick and Rebecca MacKinnon, see ibid; MacKinnon (n 10) 153–54.

[13] Klonick (n 10) 1163; Douek, 'Content Moderation as Systems Thinking' (n 6) 538.

[14] Klonick (n 10) 1163.

[15] ibid 1664–65; Gillespie (n 9); Suzor (n 10); Roberts (n 10).

[16] Suzor (n 10).

[17] Jack Goldsmith and Tim Wu, *Who Controls the Internet? Illusions of a Borderless World* (OUP 2006); Evelyn Douek, 'The Rise of Content Cartels' (*Knight First Amendment Institute at Columbia University*, 11 February 2020) <http://knightcolumbia.org/content/the-rise-of-content-cartels> accessed 3 September 2023. Most recently, this topic has been debated with respect to so-called Twitter files, see Adam Rawnsley and Asawin Suebsaeng, 'Twitter

The conceptualisation is also helpful when diagnosing the human rights interferences (Chapter 3).[18]

However, viewing content moderation only as a new governing bureaucracy is a very static view. It directs our attention away from the deeply commercial nature of the practice. It potentially clouds our view of practical solutions that might lie in increasing consumer choice through competition or regulation. Regulators thinking of regulating the assembly line of decisions cannot ignore its intimate relationship with the product design and profit of these digital services. For instance, according to a theory paper in economics, providers with subscription and advertising models have very different incentives in terms of content moderation strategies.[19] The advertising-based services tend to moderate more because they are more concerned about the size of their user base. Content moderation is, therefore, as much 'trust and safety' as it is 'size and profit'. Companies set rules and enforce them, but they primarily do so to attract and retain their user base and advertisers with stricter standards[20] under conditions that also keep their employees happy and enforcement authorities at bay.

Content moderation should be viewed from different angles depending on the context. When it enforces the rules adopted by legislatures as a form of delegated enforcement, it is helpful to think of it as a bureaucracy, a co-regulatory form of governance where the rules are set by the state. When providers cross the legality lines and start ordering lawful content, it is better to think of them as optimisers whose ultimate goal is to make a profit. Because the two contexts are so inextricably mixed in daily practice, ultimately, the best way is to keep thinking about both angles all the time.

The Digital Services Act (DSA) arguably takes both views into account. Whenever it is relevant, the DSA draws the distinction between two sets of decisions: on the illegality or violation of the terms and conditions of services. The regulation confers almost uniform due process rights on all content moderation decisions but also leaves legitimate rulemaking of companies mostly unconstrained. The regulation is more about the standards the companies making these decisions must observe when deploying their solutions. It gives companies constraints within which they can optimise their rule-setting and enforcement without running afoul of the public interest.

Kept Entire "Database" of Republican Requests to Censor Posts' *Rolling Stone* (8 February 2023) <https://www.rollingstone.com/politics/politics-news/elon-trump-twitter-files-collusion-biden-censorship-1234675969/> accessed 3 September 2023. The conservatives reacted by seeking an injunction before US federal Courts. In July 2023, a US federal judge restricted some agencies and officials of the administration of President Joe Biden from meeting and communicating with social media companies to moderate their content, see *State of Missouri et al v Joseph R Biden Jr et al* 3:22-CV-01213 (WD La Mar 20, 2023). For context, see Jon Brodkin, 'Judge Rules White House Pressured Social Networks to "Suppress Free Speech"' *Ars Technica* (5 July 2023) <https://arstechnica.com/tech-policy/2023/07/judge-rules-white-house-pressured-social-networks-to-suppress-free-speech/> accessed 3 September 2023.

[18] Kaye (n 10) 16; Timothy Garton Ash, *Free Speech: Ten Principles for a Connected World* (Yale UP 2016) 369.

[19] Yi Liu, Pinar Yildirim, and Z John Zhang, 'Implications of Revenue Models and Technology for Content Moderation Strategies' [2021] Mack Institute Working Paper <https://marketing.wharton.upenn.edu/wp-content/uploads/2021/09/09.09.2021-Liu-Yi-JMP.pdf> accessed 3 September 2023.

[20] This point is also made by Facebook employees, see Klonick (n 10) 1627.

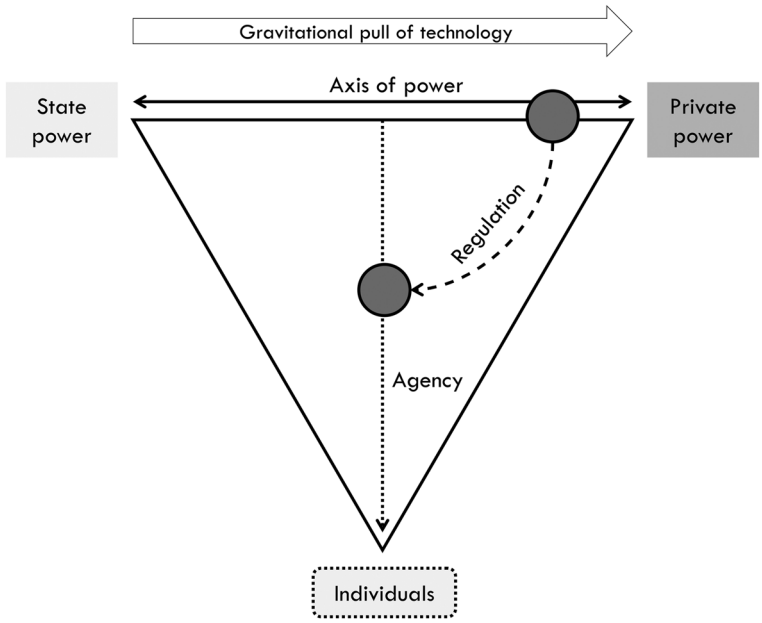

Figure 10.1 Redistribution of power

10.2 Regulation as Division of Powers

In his book, Liberalism and its Discontents, Francis Fukuyama argues that the digital ecosystem creates serious challenges for the classical conception of liberalism. He observes what many internet law scholars have over the last two decades,[21] namely that 'artificial concentrations of power over speech'[22] threaten to undermine our liberties. In the 1990s, techno-libertarians took liberalism to the extreme by arguing against *any* public intervention in cyberspace and supporting limitless self-regulation. This argument is most famously embodied in John Perry Barlow's 1996 Declaration of Independence of Cyberspace.[23] The last 20 years have shown that the period of inaction—where the state shies away from any state intervention—mostly cedes power from public to private hands (see Figure 10.1).

If one accepts that the only way of keeping any power in check is its division in a society,[24] it is inevitable that some form of state regulation is needed to divide the power

[21] Without trying to be exhaustive, some of the early works include: Goldsmith and Wu (n 17); Rebecca Tushnet, 'Power Without Responsibility: Intermediaries and the First Amendment' (2008) 76 The George Washington Law Review 986; Jonathan Zittrain, *The Future of the Internet: And How to Stop It* (Yale UP 2008); Yochai Benkler, 'A Free Irresponsible Press: Wikileaks and the Battle Over the Soul of the Networked Fourth Estate' (2011) 46 Harvard Civil Rights—Civil Liberties Law Review 311.

[22] Francis Fukuyama, *Liberalism and Its Discontents* (Profile Books 2022) 99.

[23] John Perry Barlow, 'A Declaration of the Independence of Cyberspace' (8 February 1996) <https://www.eff.org/cyberspace-independence> accessed 3 September 2023.

[24] For the critical and modern account of the separation of powers, see NW Barber, *The Principles of Constitutionalism* (OUP 2018), ch 3; Jean-Marc Sauvé, 'The Judiciary and the Separation of Powers' (*Conseil*

that markets and technology have allowed private firms to accumulate[25] through their popular products. However, the pendulum can swing too far in the state's corner, too. Proposals that want the state to assume too big a share of the powers over the digital ecosystem threaten to undermine people's liberties all the same, if not worse. In the ideal state, public power must negotiate with private power and people over the outcomes. Each side keeps the other in check (Chapter 1).

The DSA is a creature of European law-making, which might appear too heavy-handed for some jurisdictions.[26] However, it does get right arguably the most important aspect—the division of power between companies, states, and people. This aspect is most visible in how it regulates content moderation. Unlike laws that try to impose 'view-point neutrality',[27] it constrains without taking away all the power of private companies. It does so by recognising the salience of private decision-making about people's digital presence, which it tries to subject to due process procedures while leaving the formulation of policies regarding legal content mostly intact. Although companies lose their monopoly over the interpretation of their contractual policies,[28] they keep the ability to rewrite them. However, they lose their status as the final arbiters of disputes, but they can avoid new ones by clarifying contractual clauses and improving how they solve disputes with users internally.

Thus, the DSA constrains rule-interpretative powers without substantially challenging rule-making powers. The rule-making powers are obviously shared with the state forming rules for a polity through a parliament. The outcome is a greater division of powers: people have individual rights to contest decisions but can only make rules through elective representatives or lobbying their providers—in both cases, they must reconcile their interests with those of others.

10.3 Rule-Setting and Its Legitimacy

The power to make content moderation decisions can be vested with different actors. It can be vested exclusively with providers, content creators and their communities or variously shared among these actors. Thus, content moderation can manifest itself in the following four basic models: top-down, user-centric, community-driven, and mixed models. In addition, each of these models can outsource some content

d'État, 22 June 2011) <https://www.conseil-etat.fr/publications-colloques/discours-et-interventions/the-judiciary-and-the-separation-of-powers> accessed 3 September 2023.

[25] This point is made most repeatedly by Zittrain (n 21).

[26] This is especially the case for the risk mitigation part.

[27] See Florida Social Media Bill, SB 7072, 27th Leg, 1st Reg Sess (Fla 2021); Texas Social Media Bill, HB 20, 87th Leg, 2d CS (Tex 2021). The Polish bill was meant to protect against 'censorship' by prohibiting moderation of legal content, see Maria Wilczek, 'Law to Protect Poles from Social Media "Censorship" Added to Government Agenda' (*Notes From Poland*, 5 October 2021) <https://notesfrompoland.com/2021/10/05/law-to-protect-poles-from-social-media-censorship-added-to-government-agenda/> accessed 3 September 2023.

[28] In theory, courts were always present; in practice, not so much.

moderation to third parties—charities or for-profit firms—that provide tools or people to help with the decision-making.

For instance, a newspaper using social media to distribute its content can hire a company to detect and hide hate speech on its page. A provider can rely on its web hosting providers' tools to detect child abuse images. A short-rental platform might rely on the list of sex offenders compiled by charities. Webmail might rely on blacklists of spammers from some external companies. On most services today, content moderation can rarely be kept entirely in-house. It incorporates various inputs or outsourcing from others. Even a person who runs a personal blog on WordPress will likely use one of the spam plug-ins for the comments section.

The more complex the service and the bigger its content moderation surface, the more likely it is that a company is relying on some external tools. The industry is already developing some standard tools for typical challenges, such as child abuse material or terrorist content.[29] Going forward, as argued in Chapter 20, the DSA is likely only to accelerate standardisation and outsourcing. The size of the market is not negligible. One study estimated that Reddit, a company of 2,000 employees and 52 million daily users, benefits from approximately 170,000 hours of free content moderation labour from its engaged users, which would correspond to a price of $3.4 million a year.[30]

The *locus* of content moderation decision-making influences who can set the underlying rules of engagement—what can be said by whom. In the top-down model, the rules are set by a provider; in the user-centric model, by users; in the community-driven model, by communities of users; and finally, in the mixed model, by several actors acting independently. On social media, the mixed model is usually employed.[31] Providers do some moderation, and so do the users.

The reliance on a particular model of content moderation is especially sensitive when it comes to the regulation of *legal* content. The issue is less pronounced when firms try to implement parliament-mandated rules on illegal content because a user's preference for illegal content is simply an illegal preference. In contrast, purely top-down content moderation is more often blamed for lacking *legitimacy* whenever firms behind services try to redefine what legal content the public should see. Providers contractually limiting disinformation, shocking, vulgar, or self-harm content are criticised for imposing their worldview on their users. According to Pew Research, most Americans think that social media censors what they read, especially political viewpoints.[32] At the same time, academic research shows that the general public does not necessarily see restrictions on disinformation as illegitimate.[33]

[29] Gorwa, Binns, and Katzenbach (n 10); Douek, 'Content Moderation as Systems Thinking' (n 6) 543.

[30] Hanlin Li, Brent Hecht, and Stevie Chancellor, 'Measuring the Monetary Value of Online Volunteer Work' <https://arxiv.org/abs/2205.14528> accessed 3 September 2023. For statistics about the company, see 'Reddit' (*Wikipedia*, 31 August 2023) <https://en.wikipedia.org/w/index.php?title=Reddit&oldid=1173100194> accessed 3 September 2023.

[31] See the analysis of Klonick (n 10); Douek, 'Content Moderation as Systems Thinking' (n 6).

[32] Emily A Vogels, Andrew Perrin, and Monica Anderson, 'Most Americans Think Social Media Sites Censor Political Viewpoints' (Pew Research Center 2020) <https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/> accessed 3 September 2023.

[33] Anastasia Kozyreva and others, 'Resolving Content Moderation Dilemmas between Free Speech and Harmful Misinformation' (2023) 120(7) Proceedings of the National Academy of Sciences of the United States of

The main legitimacy of purely contractual top-down centres is that they help to fund platforms. A free service must be optimised to make money from advertisers.[34] The problem is that the preferences of users and advertisers can grow apart substantially. In such cases, some users might feel they are taken hostage by advertisers who fund the platforms. In the absence of changes to who pays for the services, the only way to satisfy such personal preferences is more diversity through *competition* and more choice for users through *customisation* of their experience.

But even with these two approaches, without a common baseline of what constitutes illegal content across markets, it is hard to avoid some top-down alignment of otherwise legal content. This 'bridging' of various notions of illegality is necessary to allow scalability and uniform application of the same rules across the markets. A lot of Europe's hate speech is free speech in America, but contractual arrangements can bridge the gap.

The legitimacy of rule-setting is especially problematic when markets do not offer choice. If the markets for digital services are competitive, we are usually less concerned about the different standards that various services adopt unilaterally top-down. After all, business partners of providers, including content creators, can vote with their feet and leave for alternatives. For instance, content creators have left Twitch in droves because of its content moderation and advertising policies.[35] This is David Post's idea of a 'market for rules'.[36] However, as noted by Klonick because media services in the same category, such as that of social media, are not perfect substitutes but rather complements, even the existence of several services does not guarantee the existence of such markets.[37]

When market concentration is coupled with extensive top-down content moderation, this increases the risk for society whenever legal content is constrained. Such dominant firms lack the legitimacy of properly elected parliaments and yet can make rules that affect what others can say and share. As argued by Fukuyama and his colleagues, '[p]rivate power faces no checks like popular elections; it can be controlled only by the government (through regulation) or by competition among power holders'.[38]

America 1 ('The majority preferred quashing harmful misinformation over protecting free speech. Respondents were more reluctant to suspend accounts than to remove posts and more likely to do either if the harmful consequences of the misinformation were severe or if sharing it was a repeated offense').

[34] Moreover, as Klonick points out, the companies also face numerous other pressures, such as that of their employees: Klonick (n 10) 1655.

[35] Luci S, 'Why Twitch Streamers Are Moving to YouTube' (*Game Rant*, 12 September 2022) <https://gamerant.com/twitch-streamers-moving-to-youtube-fuslie-faze-swagg-dr-disrespect/> accessed 3 September 2023. This is partly caused by account bans (Jake Selway, 'Every Major Twitch Ban of 2022 So Far, Explained' (*Game Rant*, 26 April 2022) <https://gamerant.com/all-major-twitch-bans-2022-overview-adin-ross-sodapoppin-itssliker/> accessed 3 September 2023) but also by changes to advertising models (Tom Gerken, 'Twitch Scraps Ad Changes after Streamers Leave Platform' *BBC News* (7 June 2023) <https://www.bbc.com/news/technology-65834521> accessed 3 September 2023).

[36] David G Post, 'Anarchy State and the Internet' (1995) Journal of Online Law, art 3 <https://ssrn.com/abstract=943456> accessed 3 September 2023.

[37] Klonick (n 10) 1630.

[38] < >Francis Fukuyama and others, 'Report of the Working Group on Platform Scale' (Stanford Cyber Policy Center 2020) 7 <https://cyber.fsi.stanford.edu/publication/report-working-group-platform-scale> accessed 3 September 2023.

Fukuyama and his colleagues argue that *regulation should guarantee* that users can choose content moderation experience from a new class of middle-ware firms. Their solution is to create competition for such moderation, which should weaken the power of providers who run the digital service.[39] Their paper is not specific regarding the level of integration of middle-ware firms. However, their goal is to reduce the 'unaccountable power that dominant platforms possess' by disciplining it through competition forces and consumer choice.[40] In this sense, the proposal is very close to the DSA and Digital Markets Act (DMA). The DMA opens the private infrastructure to competition by increasing contestability, and the DSA asks companies to grant users agency when moderating content.

10.4 User-Empowerment and Customisation

The DSA's focus on user empowerment also supports the idea of more diversity of digital experiences within the same products. Consider the example of a Slovak start-up that offers its content moderation product to commercial users of social media, such as news organisations. Because large American providers tend to under-invest in small markets with distinct languages, commercial users of social media in those markets must moderate much more content themselves if they want to truly benefit from social media. And this is where companies like *TrollWall* come into the picture.[41] They offer an Artificial Intelligence (AI) tool that helps to detect hate speech at scale. Its customers can tailor the detection and decide how to act, such as whether to hide or remove the content. The tool is not error-free, but it saves news organisations time and effort. It acts as a more sophisticated 'spam filter' for social media. Because it keeps its users in charge, it is also less problematic. The client can decide what rules to set and how to enforce them. The newspaper's readers who dislike the experience can stop following it on social media.

In a purely top-down model, only a platform can moderate content; no one else has that capability. While this model of content moderation might still be typical for some services, such as online marketplaces, it does not capture the reality of most digital services. Today, users are often given tools to block or hide, and content creators can remove content or block users. Content creators on YouTube or Facebook set and enforce their own rules for what they accept. While Facebook operates a huge decision factory, what can be seen by the public is equally influenced by numerous individual decisions by page or channel administrators. Providers do some central rule-setting and enforcement, and their content creators engage in their own rule-setting and enforcement in the shadow of the providers' rules. The system, thus already, is not purely top-down.

[39] ibid 33–35.
[40] ibid 36.
[41] 'AI Autopilot For Comment Moderation' (*TrollWall*) <https://trollwall.ai> accessed 3 September 2023.

Most digital services, including social media, already operate mixed models of content moderation with some level of agency given to users. However, providers do not always have business incentives to open all parts of their content moderation practices. Under self-regulation, user empowerment is mostly determined by what is convenient for companies. Some features will be offered by providers because they help the business. For instance, a dating app that does not allow customers to block other users would be quickly resented by those who use it. From the company's perspective, facilitating too much empowerment weakens its ability to set uniform rules across the services and thus attracts some advertisers. For instance, if advertisers ask for bans on nudity and nudity is left to users' personal preferences, the providers might not be able to deliver such bans. Once they empower users to make that choice, they cede the power to users. This is especially sensitive for features, such as recommender systems, that are too close to the bottom line of companies.[42] Thus, there remains a role for regulation to empower users.

The above debate shows why user empowerment is about more than just *who* gets to decide. It is inevitably about how companies earn money and how much competition they will face from new entrants. The ability to customise increases people's agency and autonomy over their information diet. Naturally, it has less space with respect to content where legislatures made their call for something to be illegal. Decentralisation is not a choice of the legislature. If the legislature is wrong, individuals need to use their share of political power to change the content of illegality rules.

The primary benefit of customisation is that instead of patronisingly banning people from consuming unhealthy 'low-quality' or 'dangerous' (yet still legal) information, it asks them to do it by choice. If people become obese due to eating unhealthy food, society does not ask supermarkets to stop selling them unhealthy items; instead they offer better product information, increase awareness about health risks, and provide health care services. This empowering approach cannot solve all problems related to people's unhealthy information diets; however, such are the limits imposed by our commitment to liberalism and human dignity.

The DSA's content moderation rules promote individual agency by giving individual tools and procedural rights. The DSA's risk management rules can also promote empowerment and decentralisation. However, the written rules are so vague that the precise effects will depend on how the European Commission conceptualises its own role. In Chapter 15, I offer a way of looking at the rules that view risk mitigation as an empowerment-first approach for legal content.

10.5 Due Process and Its Criticisms

The DSA regulates content moderation in three ways: by ex ante rules on risk management, due process rights, and transparency obligations. Due process rights give

[42] The recent attempt of Amazon to invalidate some of the DSA regulations of the recommender system is a good example of this—Case T-367/23 *Amazon Services Europe v Commission* (case in progress, lodged 5 July 2023).

(mostly procedural) rights to individuals vis-à-vis providers, and transparency helps reduce the opacity of how companies make decisions. Procedural rights and transparency are largely inspired by a declaration of numerous non-governmental organisations (NGOs) known as the Santa Clara Principles,[43] although it explicitly comes with a note that '[s]tates should not transform the Santa Clara Principles directly into legal mandates'.

Nicolas Suzor made the most powerful case *for* the procedural legal rules adopted by parliaments in his book *Lawless*, where he writes:

> The absence of government regulation is not freedom. This book is called Lawless because so many of the decisions about what we can do and say online are made behind closed doors by private companies. This is the opposite of the standards we expect of legitimate, legal decision-making in a democratic society. Where governments do not set laws to regulate the internet, platforms and other powerful telecommunications providers are constantly making decisions about what types of speech they will carry. The major social media platforms all have rules about the content they deem acceptable, and many of these have expressed limits on hatred and abuse. Without law, though, these rules are not enforced in any way that can be called legitimate. There's no easy way to ensure either that the rules are consistently enforced or that they are enforced in a way that is fair and free from bias.[44]

Suzor speaks in favour of binding procedural rules from the perspective of the rule of law. He points out that the outcome is not legitimate without a fair process. His core message is that the process matters.[45] He positions his proposal around the idea that the private power of technology companies needs to be 'constitutionalised' to 'impose limits on how rules are made and enforced'. In his view, '[c]onstitutionalism is the difference between lawlessness and a system of rules that are fairly, equally, and predictably applied'.[46]

To agree with Suzor's view, one does not also need to accept his or other digital constitutionalists' framework.[47] If the word regulation is substituted for the constitution, Suzor makes a perfectly understandable case for industry regulation that can be understood as consumer or business fairness regulation. While dressing due process rights as

[43] 'Santa Clara Principles on Transparency and Accountability in Content Moderation' (*Santa Clara Principles*) <https://santaclaraprinciples.org> accessed 3 September 2023.

[44] Suzor (n 10) 8.

[45] ibid.

[46] ibid 9.

[47] The two leading proponents of digital constitutionalism are Suzor (n 10); Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (CUP 2022). This approach has been criticised by Róisín Á Costello, 'Faux Ami? Interrogating the Normative Coherence of "Digital Constitutionalism"' (2023) 12(2) Global Constitutionalism 326 (arguing that while the goals might be noble, however, digital constitutionalism is not constitutionalism at all); Leonid Sirota, 'A Constitution without Constitutionalism: Re-thinking the "Digital Constitutionalism" Debate' (*Verfassungsblog*, 3 July 2023) <https://verfassungsblog.de/a-constitution-without-constitutionalism/> accessed 3 September 2023.

constitutional rights gives them more heft, the truth remains that liberal democracies must convince legislatures to adopt them. Unlike due process rights against the state, these rights are not directly guaranteed by any constitution known to me, even though the underlying idea of procedural fairness might be related.

The DSA's framework of due process rights has three component obligations: to disclose rules, to explain decisions and processes, and to justify decisions. Providers must disclose their rules upfront without hiding them from those bound by them (Article 14(1)). Providers can make rules on the go but cannot apply them retroactively. Once they apply the existing rules, they must explain why they adopt some content moderation decisions (Article 17). Finally, the companies must be open to revisiting the adopted decisions following an internal or external appeal and be able to justify them (Articles 20–21). In addition, they must report publicly on how they handle the decision-making. If companies build their internal decision systems to comply with the above due process rights envisaged by the DSA, they can mostly export the statistics from such systems (Article 15). These obligations will be discussed in detail in the next few chapters. However, before I delve into the details, I consider it important to explain the high-level aspirations and criticisms of the regulatory approach.

Most recognise that solutions to various challenges of content moderation cannot come from one intervention but require 'systemic thinking'.[48] However, what exactly that thinking entails is not always clear. Most likely, the clearest meaning is that content moderation should not be left entirely to firefighting once the fire is already raging but should also involve preventive risk management. Content moderation could be thus described as a combination of ex ante and ex post measures employed by companies. The typical ex post measures are the actions against individual users or their expressions, whether based on the content or their behaviour. The typical ex ante measures are the redesign of the product to curb the abuse. The DSA regulates many ex post measures by prescriptive procedural rules, while ex ante is left regulated by rather general risk management rules.

The following two chapters are devoted to ex post measures. In the literature, many forms of criticism have been levelled against ex post measures, sometimes, I would argue, unfairly so. The criticism of these rights could be clustered as follows:

- they are useful but should not be binding;
- they cannot possibly solve content moderation problems;
- they only further privatise justice; and
- they are going to be too costly and will stifle competition and innovation.

[48] This term was recently used by Evelyn Douek, Douek, 'Content Moderation as Systems Thinking' (n 6). However, it has also attracted criticism for describing something likely commonplace in the literature, Kate Klonick, 'Of Systems Thinking and Straw Men' (2023) 136(6) Harvard Law Review 339.

10.5.1 Why binding rules?

The first criticism makes the point that non-binding due process rights are less likely to be abused by various groups through litigation or state authorities.[49] While that might be true, they are also less likely to be observed by providers who also wish to abuse their users. Any such debate is about how much trust one has in the institutions and the likely market dynamics among providers. The unique mix of both depends on each jurisdiction. If the trust in institutions is low or should be low, and the market is robust, then non-binding rules are preferable. However, if the country has healthy institutions, such as authorities, courts, or civil society, they can filter and deter abuse. Structurally, there is no reason why due process rights arguably should be any more dangerous than any other rights of individuals, such as the right to data protection.

The DSA's design creates separately enforceable due diligence obligations (Chapter 9.4). This feature reduces stakes by ensuring that the observance of such rights is not a life-or-death question for companies. A failure to adopt a user-friendly notification process or issue explanations to users can earn company fines or lawsuits, but the system explicitly prohibits any eye-watering statutory damages.[50] With litigation cost-sharing and proportionality as the ultimate rule for any public fines, there is no obvious reason why obligations should be abused more than any legal instrument can always be.

10.5.2 Can procedural rights help?

The second criticism concerns whether such legislation is not only handwaving that leads to 'procedural theatre'[51] or 'procedural fetish'[52] with little practical impact. Often, it is argued that appeal rates are so low today that we should not rely on individuals to correct mistakes.

It is true that we should not rely *only* on individuals, but we must *also* rely on individuals. The low rates of appeals need to be viewed against the institutional set up that users today face. No remedy, apart from courts, that they had is credible. Why is it such a surprise then that they do not use them? If the remedial set up changes, thanks to external appeals and individual rights, and people internalise these rules, their behaviour can change, too. As I have argued elsewhere, there are many ways in which people's willingness to complain in the right cases can be incentivised.[53] We also need to rely

[49] Eric Goldman, 'How Will the Digital Services Act (DSA) Affect the European Internet?' (*Technology & Marketing Law Blog*, 12 July 2023) <https://blog.ericgoldman.org/archives/2023/07/how-will-the-digital-services-act-dsa-affect-the-european-internet.htm> accessed 3 September 2023.

[50] Art 54 prescribes compensation and thereby also pre-empts punitive damages: see ch 19.

[51] Douek, 'Content Moderation as Systems Thinking' (n 6) 578.

[52] ibid. Most prominently, this point is made by Douek. To support this view, she cites Nicholas Bagley, 'The Procedure Fetish' (2019) 118(3) Michigan Law Review 345.

[53] Lenka Fiala and Martin Husovec, 'Using Experimental Evidence to Improve Delegated Enforcement' (2022) 71 International Review of Law and Economics 106079.

on individuals because if the challenge to the top-down content moderation model is meant to rely on user empowerment, only the users can be in the driving seat of the change. Giving individuals rights against private power is the starting point.

The DSA leans heavily on individuals. If they do not use the tools, little will change in their positions[54] unless they are affected by ex ante risk mitigation interventions. While some might see this as a bug, I personally see this as a feature. The DSA invites civil society to assist individuals in their dispute resolution. They can represent them before providers, in external appeals, or even seek injunctions on their behalf.[55] Thus, numerous credible remedies are created. With different incentives, the behaviour of individuals is also likely to change. They might become less deferential to the providers and more conscious and assertive of their rights. While it is true that some problems, such as unequal distribution of errors,[56] cannot be resolved by simple ex post redress, as it relies on the ability of the affected individuals to defend themselves with a combination of collective structures (see Chapter 19), and ex ante risk management (Chapter 15), procedural rights can still be useful.

A rejection of procedural rights, or private rights more generally, means that we are putting all our faith in the public regulators and giving no agency again to individuals. This is wrong. If companies can rely on imperfect technologies to enforce the law or their contractual clauses, individuals must be given a way to correct them. Making them dependent on the priorities of authorities is problematic. This is why the DSA's mixed enforcement model (Chapter 20), in my view, is essential. Senders and receivers of information cannot become the eternal hostages of imperfect technologies that will continue to govern communication until the providers decide to change them. Individual rights should feed into the calibration of such tools and help to improve them. Even if they are only a *part* of the solution, they have a place.

10.5.3 Why privatising justice?

The third criticism[57] is about how giving procedural rights to private parties, especially allowing them to appeal to private dispute resolution bodies, will lead to fewer disputes being resolved by courts. The assumption is that only courts know how to resolve

[54] Martin Husovec, 'Will the DSA work?: On Money and Effort' (*Verfassungsblog*, 9 November 2022) <https://verfassungsblog.de/dsa-money-effort/> accessed 3 September 2023.

[55] DSA, arts 21, 86, and 90.

[56] Evelyn Douek, 'Governing Online Speech: From "Posts-As-Trumps" to Proportionality And Probability' (2021) 121(3) Columbia Law Review 759.

[57] Daniel Holznagel, 'The Digital Services Act Wants You to "Sue" Facebook over Content Decisions in Private de Facto Courts' [2021] *Verfassungsblog* <https://verfassungsblog.de/dsa-art-21/> accessed 3 September 2023 ('Member States' courts should already possess some of the necessary characteristics for dispute settlement in our context—expertise, impartiality, fair rules of procedure, and the capacity to deliver binding decisions. What may be lacking is speediness and cost-effectiveness. Reform efforts should focus on solving these specific weaknesses, which is less difficult than inventing and implementing a whole new layer of de facto courts from scratch'); Jörg Wimmers, 'The Out-of-Court Dispute Settlement Mechanism in the Digital Services Act' (2021) 12(4) Journal of Intellectual Property, Information Technology and Electronic Commerce Law 421 ('free speech disputes are strictly normative and do not lend themselves to a settlement by private bodies, but are reserved for the judiciary').

disputes, and reducing their agenda is harmful to society at large. The typical concerns are about the fairness of such procedures, their standard-setting nature, or their legitimacy. For some civil society groups, avoiding courts signals that human rights standards will not be respected similarly. The concern is that providers, or specialised dispute settlement bodies, will not be impartial enough.

While I understand that some disputes are so consequential that they should be channelled through courts, it should be stated that the DSA does not limit the jurisdiction of courts. Any party to the dispute can still go to the courts. Some scholars and judges[58] tend to promote judiciary as the only legitimate way to resolve disputes and tend to criticise the privatisation of dispute resolution. However, not all countries have a perfect judiciary. Many judiciaries are slow, less trained in novel issues, or simply expensive. Moreover, the degree of litigiousness differs across countries, and in some, there are serious mental barriers for people to resort to the courts.

If the judicial system is still available—which is the case under the DSA—I fail to see the problem when the law offers a new route to help resolve disputes. If private parties can resolve their differences without state intervention, we should embrace that. For one, society does not have to subsidise such dispute resolution through taxes. The public's finances can be spared from all the grievances that individuals have with platforms. Instead, only providers and users pay for their mutual disputes. The same principles underlie consumer or aviation compensation disputes (Chapter 11), and no one seems to be panicking about the 'disappearance of courts'.

10.5.4 What about costs of compliance?

Finally, the fourth argument concerns the cost of compliance and its impact on competition and innovation.[59] This criticism relates to the first. What if compliance with procedural rights is so costly that it can only be done by bigger players? Isn't the law inviting more than only market concentration? Isn't the DSA going too far in this respect? This criticism is legitimate, as increased regulation can lead to market concentration.[60] Incumbents can and do lobby for increased regulation to stop new entrants or at least slow them down.

However, one does not need to study market developments for too long to see that market concentration is already taking place in the absence of regulation. While it results from many factors unrelated to content moderation, email client concentration provides a unique example. Unlike social media, the email client market is less affected

[58] Holznagel (n 57); Wimmers (n 57).

[59] Goldman (n 49); Douek, 'Content Moderation as Systems Thinking' (n 6) 583 ('mandating a particular kind of process in every case will freeze the current equilibrium in place, … disincentivise innovation, and create barriers to entry given the most well-resourced platforms will find it easiest to comply with such mandates').

[60] Oscar Guinea and Fredrik Erixon, 'Market Concentration, Regulation, and Europe´s Quest for a New Industrial Policy' (*ProMarket*, 7 May 2019) <https://www.promarket.org/2019/05/07/market-concentration-regulation-and-europes-quest-for-a-new-industrial-policy/> accessed 3 September 2023.

by typical 'digital tipping effects', such as direct network effects[61] or data-driven competitive advantages.[62] If anything, indirect network effects are created more by content moderation practices. Nevertheless, the email client market still results in a duopoly of companies controlling the absolute majority of the market. To add insult to injury, email protocol is heavily based on open standards. While one case is hardly conclusive, at least it allows us to study to what extent the cost of content moderation contributes to market concentration because it erects a high barrier to entry.

If the cost is significant, then the question is whether the DSA is not only adding fuel to the fire by increasing such costs for new entrants but potentially helping reduce the fire. This is ultimately an empirical question. The DSA has two countervailing effects: first, the new regulatory requirements initially increase the regulatory costs of operating content moderation and thus market entry; second, the regulation can facilitate more standardisation and market entry in content moderation solutions. If a bigger industry needs cheap and scalable solutions that comply with certain standardised legal metrics, it can be expected that other companies jump on the opportunity and start offering them. The industry's steps in this direction are already visible. Cloudflare is offering a tool to catch child-sexual-abuse material as part of its Content Delivery Network services.[63] Amazon's Amazon Web Services (AWS) is offering various content moderation systems for hosted services.[64]

Arguably, the DSA softens the regulatory costs through its tiered approach, where bigger players attract more responsibility to give effect to the procedural rights of individuals. The main obligations apply to companies with at least 50 employees and a €10 million turnover. Only providers with over 45 million monthly active users in the EU are subject to a more elaborate ex ante risk management system requiring more in-house adjusting and customisation. One might argue about these thresholds as being low or high. However, while doing so, we should not forget that the DSA also helps companies do business across the border by lowering their cost of compliance with different rules. Thus, as explained below, there are potential substantial savings from harmonisation per se.

The online platform tier effectively limits the more expensive compliance to fewer companies because of the company size. Undoubtedly, in this tier, there could be companies that are merely local and thus cannot immediately reap the cross-border benefits. However, even such companies benefit whenever they seek expansion to the other EU markets. For the big tech, the threshold corresponds to roughly 10% of the EU population. Only a handful of single-language services can cross this threshold—namely,

[61] Matthew T Clements, 'Direct and Indirect Network Effects: Are They Equivalent?' (2004) 22(5) International Journal of Industrial Organization 633.

[62] Jens Prufer and Christoph Schottmüller, 'Competing with Big Data' [2017] TILEC Discussion Paper No 2017-006, CentER Discussion Paper 2017-007 <https://papers.ssrn.com/abstract=2918726> accessed 3 September 2023.

[63] Justin Paine and John Graham-Cumming, 'Announcing the CSAM Scanning Tool, Free for All Cloudflare Customers' (*The Cloudflare Blog*, 18 December 2019) <http://blog.cloudflare.com/the-csam-scanning-tool/> accessed 3 September 2023.

[64] Amazon Web Services, 'Content Moderation | Machine Learning' (*Amazon Web Services, Inc.*) <https://aws.amazon.com/machine-learning/ml-use-cases/content-moderation/> accessed 3 September 2023.

those offered in German, French, and Italian, and counting second language speakers, English, Spanish, and Polish.[65] Complying with the VLOP/VLOSE rules effectively means doing business across the border.

It is underappreciated that the DSA creates huge benefits for non-EU companies that provide businesses in the EU. Any provider not established in the EU is *not* shielded by the country-of-origin principle.[66] This meant that such companies pre-DSA might have dealt with hundreds of authorities from 27 countries speaking 24 different languages. Under the ECD, there was no system where a company would simply choose a legal representative to reduce that number to one, as it is now under Article 13 DSA for rules *harmonised by the DSA*.[67] According to ECD, foreign companies had to be 'established' in the EU to benefit from the country-of-origin principle. This involves 'the actual pursuit of an economic activity through a fixed establishment for an indefinite period' (Recital 19 ECD). The DSA thus offers non-EU companies a simple way to comply with one set of rules, and if they move their establishment later, those rules will remain the same because they are harmonised.

Therefore, to answer the original empirical question of whether the DSA compliance costs harm competition, one needs to look at the countervailing effects: the beneficial effects caused by standardisation and improved cross-border entry and the potential negative effects of lower entry caused by such regulation. The proposition that more rules only equal higher costs and less competition is far from immediately true for the above reasons.

Furthermore, deregulation is not necessarily socially cheaper. Content moderation of *illegal content* regulation has true distributive costs. If a platform is asked to moderate illegal content, its failure to do so means more work for users of such digital spaces (eg administrators of pages on social media dealing with illegal hate speech). A provider's underinvestment raises costs for its users. A newspaper, as a content creator, needs to invest in people and technology to detect and moderate such content at speed and scale. A failure of platforms to moderate illegal content clearly creates externalities for its users. Regulating such underinvestment thus does not mean creating costs out of thin air but reallocates costs to the platforms asked to internalise the externalities of their design. Since users cannot change the design of the platforms (eg what content is incentivised), having them bear the high cost of externalities seems questionable. If we accept that the regulation of digital services should respect the principle of shared burden for societal costs (Chapter 21), each side should shoulder part of the burden.

I hasten to add that the situation is different, however, for content moderation of *legal content*—information or behaviour that is not prohibited by parliaments. There, the platforms and their users can have different views of what they want to see. If users

[65] TNS Opinion & Social, 'Europeans and Their Languages' (European Commission 2012) Special Eurobarometer 386 <https://web.archive.org/web/20160106183351/http:/ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf> accessed 3 September 2023.

[66] This remains the same under the DSA, as emphasised by art 13(5) DSA.

[67] For the rules that the DSA does not harmonise, the pre-DSA situation applies—ie any national regulator can expect its rules to be enforced against such companies and potentially has jurisdiction.

fail to convince providers to change their content standards, they can still lobby for tools that allow them to enforce different standards in their own digital spaces. Thus, a Christian newspaper using social media can decide to restrict all content that is not aligned with its values or worldview. They can use external tools to do so at the required speed or scale. While such preferences on the individual level can be justified, they have serious consequences if rolled out for entire services. If individuals are unhappy with the lax default existing standards of content that are lawful, they can direct their attention to parliaments or competitors. Parliaments not only have the highest legitimacy to change what is prohibited, but they are also properly accountable to the public for their rulemaking in a public process, which is ultimately overseen by courts as the ultimate guardians.