European Journal of Law and Technology, Vol. 15 No. 2 (2024)

EJLT European Journal of Law and Technology

A Principled Approach to Infrastructure Moderation

Tristan Goodman[*]

Abstract

At a time when debates on content moderation have largely focused on the role and responsibilities of social media platforms, this article draws attention to the less discussed power that Internet infrastructure providers exert over the way in which content is accessed and shared online. With the rise of less moderated 'free speech' platforms, the role of responsible content moderator is increasingly falling to the companies which keep these platforms online. While many have questioned the legitimacy of so-called 'infrastructure moderation', this article suggests that infrastructure providers should play a limited role in content moderation, which is defined by reference to a regulatory framework based on principles of proportionality, transparency, and procedural fairness.

Keywords: content moderation; infrastructure moderation, internet regulation, free speech.

1. Introduction

On 3 September 2022, the CEO of Cloudflare, a US company that provides content delivery and security services for online applications globally, announced that the company had terminated its services with Kiwi Farms, a trolling website, after 'specific, targeted threats' posing an 'immediate threat to human life' appeared on its website.[1] This announcement followed an intense pressure campaign, led by transgender activist and Twitch streamer, Clara Sorrenti, who was the subject of

[*] Independent.

[1] Matthew Prince, 'Blocking Kiwifarms' (*Cloudflare*, 3 September 2022) <https://blog.cloudflare.com/kiwifarms-blocked/> accessed 8 June 2024.

death threats linked to a discussion thread on Kiwi Farms.[2] After Cloudflare terminated its services with Kiwi Farms, members of the campaign were able to take the website offline, by overwhelming its servers with artificially high levels of internet traffic (known as a 'distributed denial-of-service attack'), until Kiwi Farms found an alternative provider.[3]

While Cloudflare's CEO described this as an 'extraordinary decision' for an Internet infrastructure provider to make,[4] the reality suggests otherwise. Not only has Cloudflare taken the same action against other controversial websites,[5] but there is growing evidence that other Internet infrastructure providers are engaging in content moderation.[6] We are witnessing the rise of 'infrastructure moderation'.[7]

At a time when debates about content moderation tend to focus on the role and responsibilities of a handful of social media platforms,[8] infrastructure moderation

[2] Claire Goforth, 'Kiwi Farms gets back online thanks to the same service that's kept 8kun alive' (*Daily Dot*, 6 September 2022) <https://www.dailydot.com/debug/kiwi-farms-back-online-vanwatech/> accessed 2 October 2023.

[3] Ibid.

[4] Prince (n 1).

[5] Matthew Prince, 'Why We Terminated Daily Stormer' (*Cloudflare*, 16 August 2017) <https://blog.cloudflare.com/why-we-terminated-daily-stormer/> accessed 2 October 2023; Matthew Prince, 'Terminating Service for 8Chan' (*Cloudflare*, 5 August 2019) <https://blog.cloudflare.com/terminating-service-for-8chan/> accessed 2 October 2023.

[6] In this paper, references to 'content moderation' and similar phrases are references to the rules and/or systems which online intermediaries use to determine how user-generated content is treated on their services, where 'online intermediaries' refers to any entity which brings together or facilitates communications and/or transactions between third parties on the Internet. Online intermediaries therefore include user-facing services, such as Facebook and Twitter, but also infrastructure providers, such as Cloudflare and others referred to in this paper.

[7] Jonathan Zittrain, 'The Inexorable Push For Infrastructure Moderation' (*Techdirt*, 24 September 2021) <https://www.techdirt.com/2021/09/24/inexorable-push-infrastructure-moderation/> accessed 2 October 2023.

[8] For example, James Grimmelmann, 'The Virtues of Moderation' (2015) *17 Yale Law Journal of Law & Technology* 42; Jack M Balkin, 'Free speech in the algorithmic society: Big data, private governance, and new school speech regulation' (2017) 51 *UC Davis Law Review* 1149; Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 *Harvard Law Review* 1598; Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 *Harvard Law Review* 526. For notable exceptions, see Joan Donovan, 'Navigating the Tech Stack: When, Where and How Should We Moderate Content?' (*CIGI*, 28 October 2019) <https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content/> accessed 2 October 2023; Tarleton Gillespie, *Custodians of the Internet: platforms content moderation, and the hidden decisions that shape social media* (Yale University Press 2018); Jenna Ruddock and Justin Sherman, 'Widening the Lens on Content Moderation' (2021) Joint PIJIP/TLS Research Paper Series <https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1071&context=research> accessed 2 October 2023; Christoph Busch, 'Regulating the Expanding Content Moderation Universe: A European Perspective on Infrastructure Moderation' (2022) 27 *UCLA Journal of Law & Technology* 32; Prem M Trivedi, 'Content Governance in the Shadows: How Telcos & Other

raises several questions that deserve greater attention. How do different Internet infrastructure providers moderate content? Why are they increasingly doing so? Should they be making content moderation decisions at all? If so, in what scenarios and according to which rules and standards? Without thoughtful answers to these questions, we cannot properly assess what role (if any) infrastructure providers should play in content moderation.

This paper attempts to provide some answers to the questions above in Sections 2 and 3. Section 2 begins by illustrating how different infrastructure providers can, and increasingly do, engage in content moderation. It is suggested that the rise in infrastructure moderation is partly explained by the rise of alternative 'free speech' platforms, which often fail or refuse to engage in meaningful content moderation at the application layer. While some have questioned the legitimacy of infrastructure moderation, Section 3 makes the case for infrastructure providers playing a limited role in content moderation – based on principles of proportionality, transparency and procedural fairness – before concluding with some thoughts on next steps.

2. The Rise of Infrastructure Moderation

2.1 Widening the Scope of Content Moderation

When we use the Internet, we rarely think about how the content which we access and share travels to and from our devices. When someone posts on social media, the only obvious parties involved in that process are themselves, whoever made their device, their (mobile) Internet provider and the social media company. We overlook many other parties whose role in the delivery of content throughout the Internet ecosystem is often essential. Without these other parties, the post may never make its way online. It is important to recognise the full extent of the Internet ecosystem, so that efforts to regulate what we access and share online do not treat the Internet as if it were entirely made up of a handful of social media companies.[9]

The architecture of the Internet is typically depicted as a vertical hierarchy of interrelated layers of technology, commonly referred to as the 'Internet stack'.[10] For the purposes of this paper, it is sufficient to use a simplified version of the Internet stack that distinguishes between the 'application layer' (comprising user-facing services, such as platforms and websites) and the 'infrastructure layer' (comprising everything 'below' the application layer, including web hosting services, content

Internet Infrastructure Companies "Moderate" Online Content' (2023) Joint PIJIP/TLS Research Paper Series <https://digitalcommons.wcl.american.edu/research/90> accessed 2 October 2023.

[9] Mike Masnick, 'The Internet Is Not Just Facebook, Google & Twitter: Creating a 'Test Suite' For Your Grate Idea To Regulate The Internet' (*Techdirt*, 18 March 2021) <https://www.techdirt.com/2021/03/18/internet-is-not-just-facebook-google-twitter-creating-test-suite-your-great-idea-to-regulate-internet/> accessed 2 October 2023.

[10] Ulrike Uhlig and others, *How the Internet Really Works: An Illustrated Guide to Protocol, Privacy, Censorship, and Governance* (No Starch Press 2020).

delivery networks, domain name registrars, and Internet service providers (ISPs))[11] – as shown in Figure 1 below. Taking a 'layer-conscious approach' is helpful from a regulatory perspective because content moderation at different layers may require different regulatory norms, particularly because of differences in the technical and economic characteristics of each layer.[12]

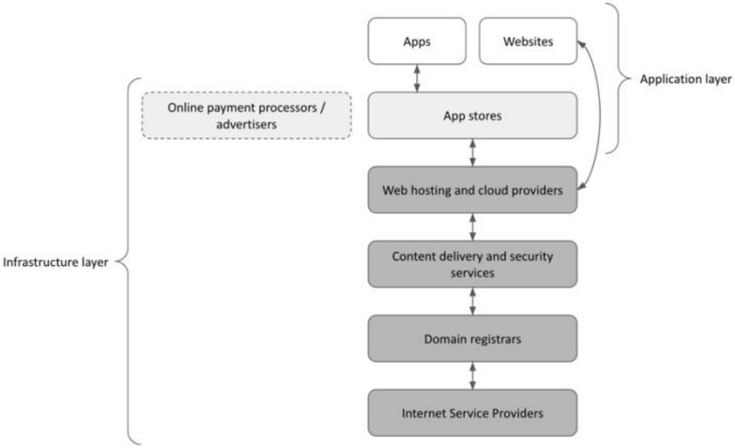
Apps
Websites
Application layer
Online payment processors / advertisers
App stores
Web hosting and cloud providers
Infrastructure layer
Content delivery and security services
Domain registrars
Internet Service Providers

Figure 1. Simplified version of the Internet stack

However, there are a couple of important qualifications to thinking about the Internet in these terms, both reflected in Figure 1 above.[13] First, some online services are both

[11] How the architecture of the Internet should be conceived is a live issue. For example, see Mike Masnick, 'Does An Internet Taxonomy Help Or Hurt?' (*Techdirt*, 6 October 2021) <https://www.techdirt.com/2021/10/06/does-internet-infrastructure-taxonomy-help-hurt/> accessed 2 October 2023; Electronic Frontier Foundation, 'We Need to Talk About Infrastructure' (20 December 2022) <https://www.eff.org/deeplinks/2022/12/we-need-talk-about-infrastructure> accessed 2 October 2023.

[12] Annemarie Bridy, 'Remediating social media: A layer-conscious approach' (2018) 24 *Boston University Journal of Science & Technology Law* 193.

[13] Another qualification, which will not be considered further in this paper, is that some user-facing services can perform similar roles online but in very different ways. For example, although private messaging services such as WhatsApp and Telegram 'can be understood as social media insofar as content sharing among small and large groups, public communication, interpersonal connections, and commercial transactions converge in key features of the app' (Tarleton Gillespie and others, 'Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates' (2020) 9(4) *Internet Policy Review* 1, 7), the content shared on messaging services is often encrypted, unlike social media feeds, making it very difficult or impossible for the service provider to monitor and locate specific pieces of content in the same way as social media platforms can. For the implications of encrypted messaging services for content moderation, see ibid 7–9.

user-facing and infrastructural in nature, depending on who is using the service. App stores, for example, provide essential technical infrastructure for developers wishing to distribute their apps, which, once available on an app store, end users can directly access on their smartphones.[14] Second, some user-facing services are infrastructural in nature. For example, while payment processors and online advertisers do not provide the technical means for content distribution, they do provide the financial means for many platforms and websites to operate. In this way, payment providers and online advertisers form the *financial*, rather than the technical, infrastructure that is often essential for content distribution.[15] Although these more nascent infrastructure services are not traditionally part of the Internet stack, the remainder of this section will illustrate how their role, as well as that of more traditional infrastructure providers, is growing and changing in content moderation.

2.1.1 Technical Infrastructure

App Stores

App stores enable users to access user-facing services, such as social media platforms, by downloading applications specifically designed for accessing and sharing content on mobile devices. In 2022, users downloaded 255 billion mobile applications worldwide, an increase of more than 80 per cent from 140.7 billion downloads in 2016.[16] Mobile applications are therefore becoming a key means by which users access and share content online, meaning platforms are increasingly reliant on app stores as a means of providing their services to end users. In this sense, app stores are located at the interface between the application layer and the infrastructure layer, depending on whose perspective is taken: end user or developer.[17]

Google and Apple dominate the smartphone application platform market, presenting a bottleneck in the distribution of content online: both companies decide which applications are available for download on their app stores and on what terms and conditions.[18] For both the Apple App Store and Google Play Store, these terms and conditions include content moderation rules, which detail not only what content should be moderated but also *how*.[19] For example, Apple requires that applications with user-generated content have certain procedural safeguards built in to ensure

[14] Busch (n 8) 43–44.

[15] Will Duffield, 'A Brief History of Deep Deplatforming' (*Cato Institute*, 22 January 2021) <https://www.cato.org/blog/brief-history-deep-deplatforming> accessed 2 October 2023; Electronic Frontier Foundation (n 11).

[16] Statista, 'Number of mobile app downloads worldwide from 2016 to 2022' (21 April 2023) <https://www.statista.com/statistics/271644/worldwide-free-and-paid-mobile-app-store-downloads/> accessed 2 October 2023.

[17] Busch (n 8) 43.

[18] Nikolas Guggenberger, 'Essential Platforms' (2021) 24 *Stanford Technology Law Review* 237, 262–268.

[19] In terms of *what* content should be moderated, section 1(1) of the App Store Review Guidelines require that applications do not contain 'objectionable content', which includes contents that is 'just plain creepy'.

that objectionable content is moderated, including: '[a] method for filtering objectionable material from being posted to the app; mechanisms to report offensive content and timely responses to concerns; [t]he ability to block abuse users from the service; [and] [p]ublished contact information so users can easily reach you'.[20] Google has very similar requirements for platforms with user-generated content.[21] In this way, Apple and Google set minimum content moderation requirements for third-party applications, through which these companies, in their capacity as infrastructure provider, indirectly exercise control over content moderation systems and rules implemented at the application layer.[22]

The influence of app stores' content moderation requirements on decisions made at the application layer is becoming increasingly apparent. Twitter's 2021 Annual Report states:

> 'Our release of new products, product features and services on mobile devices is dependent upon and can be impacted by digital storefront operators, such as the Apple App Store and Google Play Store review teams, which decide what guidelines applications must operate under and how to enforce such guidelines. Such review processes can be difficult to predict and certain decisions may harm our business.'[23]

According to Twitter's former Head of Trust and Safety, this is an understatement: 'Failure to adhere to Apple and Google's guidelines would be catastrophic, risking Twitter's expulsion from their app stores and making it more difficult for billions of potential users to access Twitter's services'.[24] This risk became a reality for one of Twitter's supposed 'free speech alternatives', Parler, when both Google and Apple swiftly removed the platform from their app stores following the attack on the US Capitol in 2021, after it became apparent that rioters had connected, coordinated and shared live video footage of the attack on the platform.[25] Parler subsequently

[20] Apple, 'App Store Review Guidelines' (last updated 5 June 2023) <https://developer.apple.com/app-store/review/guidelines/> accessed 2 October 2023.

[21] Google, 'Developer Policy Center: User-generated content' <https://support.google.com/googleplay/android-developer/answer/9876937> accessed 20 April 2024.

[22] Busch (n 8) 44.

[23] Twitter, 'Fiscal Year Annual Report 2021' (2021) <https://www.sec.gov/Archives/edgar/data/1418091/000141809121000031/twtr-20201231.htm> accessed 2 October 2023.

[24] Yoel Roth, 'I Was the Head of Trust and Safety at Twitter. This Is What Could Become of It.' (*The New York Times*, 18 November 2022) <https://www.nytimes.com/2022/11/18/opinion/twitter-yoel-roth-elon-musk.html> accessed 2 October 2023.

[25] Jay Peters and Kim Lyons, 'Apple removes Parler from the App Store' (*The Verge*, 10 January 2021) <https://www.theverge.com/2021/1/9/22221730/apple-removes-suspends-bans-parler-app-store> accessed 2 October 2023.

returned to both the Apple App Store and Google Play Store, but on the condition that it implemented more robust content moderation policies and tools.[26]

Web Hosting and Cloud Providers

Hosting services are another key component of the Internet's technical infrastructure, often provided by cloud providers such as Amazon Web Services (AWS), Google Cloud, Microsoft Azure and Oracle. These companies effectively leverage the scale of their computing resources to make it cheaper and easier to host content and develop online applications using their technical infrastructure than it would be for a customer to set up and use their own.

The role of web hosting providers in content moderation was highlighted in December 2010 when AWS stopped hosting WikiLeaks shortly after a trove of classified US government documents appeared on its website.[27] More recently, AWS terminated its services with Parler, following the action taken by Google and Apple in the wake of the attack on the US Capitol.[28] In both cases, AWS justified its actions with reference to its Acceptable Use Policy (AUP): both WikiLeak's and Parler's activities had breached AWS's AUP leading to termination of services with both customers, causing them to go offline until they could find another web hosting provider.

Compared to Apple's detailed guidelines for app developers, hosting providers' content moderation rules are brief. For example, AWS's AUP prohibits content which is used to 'threaten, incite, promote or actively encourage violence, terrorism, or other serious harm', but does not specify what constitutes 'serious harm'.[29] The implication, though, is that the bar is high when compared to the terms of other providers such as Oracle, which prohibits content that is used to 'promote in any way

[26] Nico Grant, 'Parler Returns to Google Play Store' (*The New York Times*, 2 September 2022) <https://www.nytimes.com/2022/09/02/technology/parler-google-play.html#:~:text=Parler%2C%20the%20social%20media%20service,for%20content%20that%20incited%20violence> accessed 2 October 2023. Parler ceased to exist in April 2023 shortly after it was sold by its parent company, Parlemant, to Starboard. The CEO of Starboard was quoted as saying: 'No reasonable person believes that a Twitter clone just for conservatives is a viable business any more' (Todd Sprangler, 'Parler Shut Down by New Owner: A "Twitter Clone" for Conservatives Is Not a "Viable Business"' (*Variety*, 6 October 2022) <https://variety.com/2023/digital/news/parler-shut-down-new-owner-starboard-twitter-clone-conservatives-1235583709/> accessed 2 October 2023).

[27] Ewen MacAskill, 'WikiLeaks website pulled by Amazon after US political pressure' (*The Guardian*, 2 December 2010) <https://www.theguardian.com/media/2010/dec/01/wikileaks-website-cables-servers-amazon> accessed 2 October 2023.

[28] Kim Lyons, 'Parler returns to Apple App Store with some content excluded' (*The Verge*, 17 May 2021) <https://www.theverge.com/2021/5/17/22440005/parler-apple-app-store-return-amazon-google-capitol> accessed 2 October 2023.

[29] Amazon, 'AWS Acceptable Use Policy' (last updated 1 July 2021) <https://aws.amazon.com/aup/> accessed 2 October 2023.

… unwelcome or unsociable activities'.[30] Furthermore, none of the major providers specify any content moderation tools which customers should use to deal with harmful content on their website. Instead, these providers seem more focused on selling these tools to user-facing customers: both AWS and Microsoft Azure now offer software that automates and streamlines customers' content moderation processes.[31] In this way, the role of web hosting providers in content moderation at the application layer is growing, capitalising on the fact that content moderation is increasingly becoming an issue of scale.[32]

Content Delivery and Security Services

Among the lesser-known technical infrastructure services are content delivery networks (CDNs), provided by companies such as Cloudflare, Google Cloud CDN and Amazon CloudFront. CDNs are geographically distributed sets of servers which accelerate the delivery of content by storing it on local servers (known as 'caching'), thereby reducing the distance that content must travel. Without this service, the delivery of content can be slowed down to the point where access becomes virtually impossible, making CDNs another point of control over the distribution of content online. In 2020, for example, it was reported that servers used by Facebook in Vietnam were taken offline by state-owned telecommunications companies, rendering the website temporarily unavailable in Vietnam, until Facebook agreed in 2023 to censor significantly more 'anti-state' content.[33]

In addition to accelerating the delivery of content, CDNs often include additional security features, commonly used to prevent distributed denial-of-service (DDoS) attacks, which activists carried out to take Kiwi Farms' website offline.[34] DDoS mitigation services ensure that websites remain online in the face of DDoS attacks on their hosting servers. With over 15 million DDoS attacks predicted for this year – roughly double the number in 2018 – these security services have never been more

[30] Oracle, 'Acceptable Use Policy' (27 October 2020) <https://www.oracle.com/ng/a/ocom/docs/corporate/aconex-acceptable-use-policy.pdf> accessed 2 October 2023.

[31] Namely, Amazon Recognition Content Moderator <https://aws.amazon.com/rekognition/content-moderation/> and Azure Content Moderator <https://azure.microsoft.com/en-gb/products/cognitive-services/content-moderator>.

[32] For example, since the end of August 2023, hosting services have removed or otherwise restricted access to over 16 billion pieces of content. See the DSA Transparency Database <https://transparency.dsa.ec.europa.eu> accessed 20 April 2024. Notably, this does not include all the decisions *not* to remove or otherwise restrict access to content. On the limitations of using automated tools for dealing with the scale of content moderation, see Tarleton Gillespie, 'Content Moderation, AI, and the question of scale' (2020) 7(2) *Big Data & Society* <https://doi.org/10.1177/2053951720943234> accessed 2 October 2023.

[33] James Pearson, 'Exclusive: Facebook agreed to censor posts after Vietnam slowed traffic – sources' (*Reuters*, 21 April 2020) <https://www.reuters.com/article/us-vietnam-facebook-exclusive-idUSKCN2232JX> accessed 2 October 2023.

[34] Goforth (n 2).

important in keeping content online.[35] This is especially true for more controversial websites, which are often the target of DDoS attacks.[36]

Domain Registrars

Further down the Internet stack is the domain name system (DNS), where registrars such as GoDaddy and Tucows operate. Registrars are companies that sell domain names to website operators so that users can easily find their website. In effect, registrars act as intermediaries between the registry operators – the organisations which manage top-level domain names, such '.com' (Verisign) and '.uk' (Nominet) – and the registrant of a domain name. This makes registrars another potential point of control in the distribution of content online, which is as powerful as it is imprecise: the only content moderation tool available to registrars is disabling the *entire* domain name rather than removing abusive parts of a domain name.

For example, in 2017, several registrars (including GoDaddy, Google and Namecheap) stopped providing services to the neo-Nazi website Daily Stormer after it published an article brutalising a victim of the Charlottesville terrorist attack.[37] Similarly, in 2019, Tucows stopped servicing the domain for the anonymous online forum 8chan when the forum was linked to the El Paso Walmart mass-shooting, after allowing the gunman to post a hateful manifesto that remained on its website while others posted their support.[38] In both cases, the entire website was taken offline until it found another willing registrar to service its domain name.

In May 2020, a group of leading domain name registrars (and registry operators) published a 'Framework to Address Abuse' (hereafter the 'DNS Abuse Framework'),[39] which sets out recommended practices for when a registrar should take action in relation to various types of illegal or harmful content online. The DNS Abuse Framework defines five broad categories of harmful activity to which registrars must respond – malware, botnets, phishing, pharming and spam facilitating the delivery of

[35] Cisco, 'Cisco Annual Internet Report (2018-2023) White Paper' (last updated 10 March 2020) <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html> accessed 2 October 2023.

[36] When Cloudflare terminated its services with neo-Nazi website Daily Stormer after discovering its users were linked to the terrorist attack in Charlottesville (VA), Cloudflare's CEO noted that '[t]he size and scale of the [DDoS] attacks that can now easily be launched online make it such that if you don't have a network like Cloudflare in front of your content, and you upset anyone, you will be knocked offline' (Prince (n 5) (2017)).

[37] Adi Robertson and Andrew Liptak, 'Namecheap has taken down Neo-Nazi site Daily Stormer' (*The Verge*, 20 August 2017) <https://www.theverge.com/2017/8/20/16170370/namecheap-host-take-down-neo-nazi-hate-site-daily-stormer> accessed 2 October 2023.

[38] Sean Keane and Oscar Gonzalez, '8chan's rebranded 8kun site goes offline days after launch' (*CNET*, 25 November 2019) <https://www.theverge.com/2017/8/20/16170370/namecheap-host-take-down-neo-nazi-hate-site-daily-stormer> accessed 2 October 2023.

[39] 'Framework to Address Abuse' (29 May 2020) <https://dnsabuseframework.org/media/files/2020-05-29_DNSAbuseFramework.pdf> accessed 2 October 2023.

any of the former activities – collectively known as 'DNS Abuses'.[40] In contrast to these activities, the DNS Abuse Framework provides that registrars should not respond to harmful or illegal content on a particular website, known as 'Website Content Abuses'.[41] According to the DNS Abuse Framework, 'this distinction is critical in order for the Internet to remain open for free expression', since content moderation at the DNS level is 'in general … a disproportionate remedy that can cause significant collateral damage'.[42]

However, the DNS Abuse Framework provides an exception to this principle for Website Content Abuse which relates to 'the physical and often irreversible threat to human life', such as child sex abuse materials (CSAM), illegal distribution of opioids online, human trafficking, and specific and credible incitement to violence.[43] As the examples above indicate, registrars have often relied on this last exception when terminating their services to websites which have been used to spread and promote hateful violence.[44] However, this does not mean that all registrars will act accordingly. There are hundreds of domain registrars, most of which are not signatories to the Framework and can sometimes function as havens for deplatformed websites.[45]

Internet Service Providers

Finally, at the bottom of the Internet stack are ISPs – such as T-Mobile, Verizon and Comcast – which provide the physical wires, cables, servers, etc. needed for users to access the Internet in the first place. Just as ISPs can provide Internet access, they can also deny it. This is typically in response to government mandates,[46] particularly from authoritarian regimes looking to use ISPs to 'switch off dissent'.[47]

However, ISPs also have other, less binary tools for controlling content online. As well as denying access to a website altogether, ISPs can block access to that website for users in a particular region or block specific content through techniques such as 'deep

[40] Ibid 1–3.

[41] Ibid 3.

[42] Ibid.

[43] Ibid.

[44] For a more recent example, in the wake of the attack on the US Capitol, GoDaddy announced that it had stopped servicing the domain name for AR15.com, an online forum for firearms enthusiasts, for breaching its terms of service, which prohibit customers using its services for activities 'promoting, encouraging or otherwise engaging in violence' (GoDaddy, 'GoDaddy Statement Regarding AR15.com' (15 January 2021) <https://aboutus.godaddy.net/newsroom/news-releases/press-release-details/2021/GoDaddy-Statement-Regarding-AR15.com/default.aspx> accessed 2 October 2023).

[45] Rob Kuznia, Curt Devine and Yahya Abou-Ghazala, 'Epik is a refuge for the deplatformed far right. Here's why its CEO insists on doing it' (*CNN*, 9 December 2021) <https://edition.cnn.com/2021/12/09/business/epik-hack-ceo-rob-monster-invs/index.html> accessed 2 October 2023.

[46] For example, see Internet Society, 'Global Internet Shutdowns' <https://pulse.internetsociety.org/shutdowns> accessed 2 October 2023.

[47] Peter Guest, 'Blackouts: In The Dark' (*Rest of the World*, 22 April 2022) <https://restofworld.org/2022/blackouts/> accessed 2 October 2023.

packet inspection' (a form of filtering content for specific keywords and images).[48] Although using these tools can facilitate more targeted action against harmful content such as CSAM, they can also enable more pervasive surveillance and censorship.[49]

2.1.2 Technical Infrastructure

Accessing and sharing content online relies on both technical *and financial* infrastructure. The reliance of mainstream social media companies on online advertising, and the influence this has on their content moderation policies, is well publicised and documented.[50] However, this 'grand bargain' is less discussed in the context of alternative platforms, despite reports that almost two-thirds of these platforms rely on advertising revenue – roughly twice as much as mainstream websites.[51] When online advertisers, such as Google, have blocked these websites for content that violates their terms of service, there are examples of these websites removing problematic content so that advertising services can be restored.[52]

In other cases, alternative platforms have turned to direct donations from users to stay online, facilitated by payment processors such as PayPal and Patreon. However, content that is problematic for online advertisers can also be problematic for these financial intermediaries, leading to 'financial deplatforming' by other means.[53] In 2018, for example, PayPal and Stripe terminated their online payment services with Gab, another 'free speech Twitter alternative', following reports that the platform hosted anti-Semitic rants by the gunman who killed several people at a synagogue in Pittsburgh, Pennsylvania.[54] Like online advertisers, payment processors can exercise

[48] For a more in-depth discussion on how ISPs exert control over online content, see Trivedi (n 8), particularly 11–12.

[49] Ben Wagner, 'Deep Packet Inspection and Internet Censorship: International Convergence on an "Integrated Technology of Control"' (2009) SSRN <http://dx.doi.org/10.2139/ssrn.2621410> accessed 2 October 2023.

[50] Jack M Balkin, 'Fixing Social Media's Grand Bargain' (2018) Hoover Working Group on National Security Technology and Law, Aegis Series Paper No 1814 <https://www.hoover.org/sites/default/files/research/docs/balkin_webreadypdf.pdf> accessed 2 October 2023; Gillespie (n 8).

[51] Catherine Han, Deepak Kumar and Zakir Durumeric, 'On the Infrastructure Providers That Support Misinformation Websites' (2022) Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (Vol 16) 287 <https://doi.org/10.1609/icwsm.v16i1.1929> accessed 2 October 2023.

[52] In 2019, Google reinstated its advertising services for the site to Zero Hedge, a far-right financial blog, once it had taken down content which breached Google's terms of service, and implemented content moderation policies (Megan Graham, 'Google says Zero Hedge can run Google ads again after removing 'derogatory' comments' (*CNBC*, 14 July 2020) <https://www.cnbc.com/2020/07/14/google-reinstates-zero-hedge-ad-monetization.html> accessed 2 October 2023).

[53] Duffield (n 15).

[54] Andrew Liptak, 'Paypal bans Gab following Pittsburgh shooting' (*The Verge*, 27 October 2018) <https://www.theverge.com/2018/10/27/18032930/paypal-banned-gab-following-pittsburgh-shooting> accessed 2 October 2023.

control over what content is and is not available at the application layer by shaping the content policies of the platforms which rely on them. However, this control is exercised by relatively few compared to online advertising, given how difficult and expensive it is to enter highly regulated financial services markets, meaning that 'the denial of payment processing is one of the most effective levers for controlling speech'.[55]

These levers are not only pulled in respect of content that poses a threat to human life. There are many examples of websites that provide access to lawful sexually explicit content having their ability to receive payments disabled by payment processors, including online booksellers,[56] story archives[57] and alternative social networks.[58] This content conservatism can be explained, in part, by the fact that the custom of any one website or business is of little consequences to a financial services provider, particularly in the face of public and regulatory scrutiny, where it is usually easier and cheaper to sever ties with the problematic customer.[59]

Websites have tried to circumvent financial deplatforming by turning to the decentralised infrastructure of cryptocurrencies. This is what WikiLeaks did in 2011, when mainstream payment processors, including Visa and Mastercard, blocked the website from using their services.[60] However, as the cryptocurrency ecosystem has matured, it has become more centralised. Many users now rely on third-party 'wallets' and exchanges to store and trade their cryptocurrencies. The companies that provide these services are financial intermediaries in another guise, offering new opportunities for financial deplatforming. For example, when Gab began accepting cryptocurrency payments in 2018, those transactions were traded through CoinBase, one of the larger cryptocurrency exchanges, which subsequently removed Gab's account.[61] In response, Gab launched its own payment processor, GabPay, which is

[55] Will Duffield, 'Bankers As Content Moderators' (*Techdirt*, 27 September 2021) <https://www.techdirt.com/2021/09/27/bankers-as-content-moderators/> accessed 2 October 2023.

[56] Rainey Reitman, 'Free Speech Coalition Calls on Paypal to Back Off Misguided Book Censorship Policy' (*Electronic Frontier Foundation*, 7 March 2012) <https://www.eff.org/tl/deeplinks/2012/03/free-speech-coalition-calls-paypal-back-misguided-book-censorship-policy> accessed 2 October 2023.

[57] Kurt Opsahl and Rainey Reitman, 'Payment Provider Stripe Upholds Free Speech, Reactivates Nifty Archives' (*Electronic Frontier Foundation*, 2 November 2012) <https://www.eff.org/deeplinks/2012/11/payment-provider-stripe-upholds-free-speech-reactivates-nifty-archives> accessed 2 October 2023.

[58] Paige Collings, 'No Nudity Allowed: Censoring Naked Yoga' (*Electronic Frontier Foundation*, 19 December 2022) <https://www.eff.org/deeplinks/2022/12/no-nudity-allowed-censoring-naked-yoga> accessed 2 October 2023.

[59] Duffield (n 15).

[60] Roger Huang, 'How Bitcoin And WikiLeaks Saved Each Other' (*Forbes*, 26 April 2019) <https://www.forbes.com/sites/rogerhuang/2019/04/26/how-bitcoin-and-wikileaks-saved-each-other/?sh=18e453ff74a5> accessed 2 October 2023.

[61] Erin Carson, 'Gab says it was kicked off Coinbase' (*CNET*, 7 January 2019) <https://www.cnet.com/culture/gab-says-it-was-kicked-off-coinbase-again/> accessed 2 October 2023.

promoted as a 'PayPal alternative' providing a 'free speech-friendly financial infrastructure'.[62] According to Gab's CEO, 'God had a plan and getting banned only led us to build GabPay'.[63] However, it seems that God's plan included a long list of content-related exceptions to using GabPay, as provided in its extensive terms and conditions, which bear a striking resemblance to PayPal's.[64]

2.2 The Push for Infrastructure Moderation (and its Pushbacks)

The foregoing suggests that content moderation at the infrastructure layer of the Internet is not uncommon, but this has not always been the case. Until recent years, infrastructure providers have left the role of content moderator to actors at the application layer, concentrating instead on their primary responsibility of delivering content and facilitating transactions over a secure and efficient network. This division of labour is not inevitable; it is by design. The architecture of the Internet is primarily characterised by the 'end-to-end' principle, which calls for a 'stupid network' with 'smart applications': infrastructure providers which constitute the network should simply transmit content, without discriminating between different content generated at the application layer.[65] For a long time, the argument was that 'actors who are responsible for the pipes of the Internet should not concern themselves with the kind of water that runs through them'.[66]

However, this division of labour has become strained as application layer providers increasingly fail or refuse to uphold their end of the bargain. As we have already seen, alternative 'free speech' platforms such as Gab and Parler have emerged, characterised by less moderation than their mainstream equivalents.[67] This has pushed many infrastructure providers to take on the role of content moderator

[62] Andrew Torba, 'An Alternative To Paypal is Coming' (*Gab News*, 1 August 2021) <https://news.gab.com/2021/08/an-alternative-to-paypal-is-coming/> accessed 2 October 2023.

[63] Ibid.

[64] For GabPay's terms and conditions, see GabPay, 'End User Licence Agreement' <https://gabpay.live/TermsOfService/> accessed 2 October 2023. For PayPals' terms and conditions, see PayPal, 'Acceptable Use Policy' (last updated 29 October 2022) <https://www.paypal.com/us/legalhub/acceptableuse-full> accessed 2 October 2023.

[65] Lawrence B Solum and Minn Chung, 'The Layers Principle: Internet Architecture and the Law' (2004) 79(3) *Notre Dame Law Review* 815, 829.

[66] Konstantinos Komaitis, 'Infrastructure And Content Moderation: Challenges And Opportunities' (*Techdirt*, 4 October 2021) <https://www.techdirt.com/2021/10/04/infrastructure-content-moderation-challenges-opportunities/> accessed 2 October 2023.

[67] Despite claiming allegiance to free speech, almost all alternative social media platforms moderate content beyond legal requirements to do so (Galen Stocking and others, 'The Role of Alternative Social Media in the News and Information Environment' (*Pew Research Center*, 6 October 2022) <https://www.pewresearch.org/journalism/2022/10/06/the-role-of-alternative-social-media-in-the-news-and-information-environment/> accessed 2 October 2023). Arguably, platforms must, in some form or another, engage in content moderation to run a functioning platform that appeals to both users and advertisers (Gillespie (n 8)).

instead, albeit reluctantly.[68] As Internet policy expert Konstantinos Komaitis notes, 'It really comes down to a simple equation: if user-generated platforms don't do their job, infrastructure providers will have to do it for them'.[69]

However, it is important to recognise that the way in which infrastructure providers play the role of content moderator is different from user-facing services such as platforms. Because infrastructure providers generally do not have the technical ability to differentiate between different pieces of content, any moderation they do undertake is, by design, generally application-wide rather than content-specific: infrastructure moderation tends to result in the entire platform being taken offline, not just the harmful content. This means that infrastructure providers are generally not concerned with specific pieces of content – that is the proper concern of the application layer provider – but rather with whether user-facing services are undertaking meaningful content moderation. For example, when Cloudflare terminated its services with 8chan after it was linked to with the El Paso mass-shooting, its CEO announced:

> 'The rationale is simple: [8chan] have proven themselves to be lawless and that lawlessness has caused multiple tragic deaths. Even if 8chan may have not violated the letter of the law in refusing to moderate their hate-filled community, they have created an environment that revels in violating its spirit.'[70]

In this way, infrastructure providers tend to play the role of 'meta-moderator', taking a systemic approach to content moderation: if an application layer provider fails to meet minimum content moderation standards set out in the infrastructure provider's terms of service, the infrastructure layer provider will take action instead.[71] What these minimum standards should be is debateable, although Apple's App Store Review Guidelines provide a helpful example.[72]

But the more fundamental question is whether this kind of content moderation by infrastructure providers is legitimate at all. Many have called for a return to the longstanding *status quo*, where infrastructure providers adhere to a policy of content agnosticism.[73] Three key arguments are often used to support this position. First,

[68] For example, Cloudflare's CEO referred to the company's decisions to withhold services to customers for content-related reasons as 'dangerous' (Prince (n 1) and (n 5) (2017)) and 'incredibly uncomfortable' (Prince (n 5) (2019)).

[69] Komaitis (n 66).

[70] Prince (n 5) (2019).

[71] Busch (n 8) 33.

[72] As discussed in Section 2.1.1.

[73] Bridy (n 12); Jack M Balkin, 'How to regulate (and not regulate) social media' (2021) 1 *Journal of Free Speech Law* 71; Corynne McSherry, India McKinney and Jillian C York, 'Content Moderation Is A Losing Battle. Infrastructure Companies Should Refuse to Join the Fight' (*Electronic Frontier Foundation,* 1 April 2021) <https://www.eff.org/deeplinks/2021/04/content-moderation-losing-battle-infrastructure-companies-should-refuse-join-fight> accessed 2 October 2023.

because infrastructure moderation tends to be application-wide rather than content-specific, there will usually be significant collateral effects on other content that deserves protection.[74] It is very unlikely that *all* content on a given platform or website will be in breach of the infrastructure provider's terms of service. Second, content moderation at the infrastructure layer is generally less transparent than at the application layer. Recent research has found that, unlike online platforms, virtually no infrastructure providers disclose anything about their content moderation decisions beyond occasional announcements and reports by journalists.[75] Third, the more that infrastructure providers engage in content moderation, the more vulnerable they are to government pressure to block or filter content without public scrutiny, particularly from more authoritarian regimes.[76] For example, after terminating its services with Daily Stormer and 8chan, Cloudflare reported that it 'saw a dramatic increase in authoritarian regimes attempting to have us terminate security services for human rights organisations'.[77]

While these are reasonable arguments against infrastructure moderation as it is now, they should be treated as reasons for regulation rather than prohibition. It is important to recognise that as the Internet has evolved, so has the relationship between infrastructure providers and content.[78] Strict adherence to content agnosticism no longer seems viable when content moderation has become 'an expansive socio-technical phenomenon, one that functions in many contexts and takes different forms'.[79] The challenge for content moderation is to ensure that the role and responsibilities of infrastructure providers are appropriately scoped. Section 3 makes the case for infrastructure providers playing a limited role in content moderation, based on principles which have informed content moderation policy at the application layer: proportionality, transparency and procedural fairness.

3. Defining the Limits of Infrastructure Moderation

To determine what role infrastructure providers should play in content moderation, a good starting point is to consider applicable principles on which there is broad consensus. Some have pointed to the 'Santa Clara Principles On Transparency and Accountability in Content Moderation', which were initially developed by a group of academics and civil society organisations in 2018, and have since been endorsed by

[74] Jack M Balkin, 'Free Speech is a Triangle' (2018) 118 *Columbia Law Review* 2011.

[75] Katie Stoughton and Paul Rosenzweig, 'Toward Greater Content Moderation Transparency Reporting' (*Lawfare*, 6 October 2022) <https://www.lawfareblog.com/toward-greater-content-moderation-transparency-reporting> accessed 2 October 2023.

[76] Balkin (n 74).

[77] Matthew Prince and Alissa Starzak, 'Cloudflare's abuse policies & approach' (*Cloudflare*, 31 August 2022 <https://blog.cloudflare.com/cloudflares-abuse-policies-and-approach/> accessed 2 October 2023.

[78] Komaitis (n 66).

[79] Gillespie and others (n 13) 9.

platform providers, including Apple, Meta and Google.[80] The Santa Clara Principles describe best practices for 'how best to obtain meaningful transparency and accountability around Internet platforms' increasingly aggressive moderation around user-generated content'.[81] This includes specific requirements regarding transparency and procedural due process for content removals and account suspensions.

While the Santa Clara Principles may serve as a helpful reference point for infrastructure providers engaging in content moderation, there are two important limitations to these principles. First, they are principally designed for entities operating at the application layer, as reflected by the majority of endorsements coming from these entities.[82] Second, the authors are explicit that the Principles 'are not designed to provide a template for regulation'.[83]

In light of these limitations, this article suggests that greater attention should be paid to another set of principles: the 'the 'Manila Principles on Intermediary Liability', a series of reform proposals developed by civil society organisations in 2015.[84] According to the Background Paper for the Manila Principles, they are 'directed at laws, policies, norms, practices, and private terms of service that relate to content restriction, including removal, blocking or filtering by intermediaries'.[85] Here, 'intermediaries' are understood to be parties that 'bring together or facilitate transactions between third parties on the Internet', which include both application

[80] 'The Santa Clara Principles On Transparency and Accountability in Content Moderation' <https://santaclaraprinciples.org/> accessed 2 October 2023 (hereafter the 'Santa Clara Principles'). Busch (n 8) and Komaitis (n 65) both suggest that the Santa Clara Principles should inform the basis of any regulatory framework for infrastructure moderation. Trivedi (n 8) appears to be the only commentator to have endorsed the applicability of the Manila Principles (see n 84) to infrastructure moderation – although Trivedi also suggests the Santa Clara Principles are application, without suggesting that one set of principles is preferable over the other in this context.

[81] Ibid (Santa Clara Principles).

[82] Gennie Gebhart, 'Who Has Your Back? Censorship Edition 2019 (*Electronic Frontier Foundation*, 12 June 2019) <https://www.eff.org/wp/who-has-your-back-2019> accessed 20 April 2024.

[83] Santa Clara Principles (n 80).

[84] 'Manila Principles on Intermediary Liability' (24 March 2015) <https://www.eff.org/files/2015/10/31/manila_principles_1.0.pdf> accessed 2 October 2023 (hereafter the 'Manila Principles'). Busch (n 8) claims that there is no set of principles for infrastructure moderation that is equivalent to 'The Santa Clara Principles On Transparency and Accountability in Content Moderation' <https://santaclaraprinciples.org/> (hereafter the 'Santa Clara Principles'), which were initially developed by a group of academics and civil society organisations and first published in 2018. The Manila Principles, which were also developed by civil society organisations from around the world, undermine this claim. Trivedi (n 8) appears to be the only commentator to have endorsed the applicability of the Manila Principles to infrastructure moderation.

[85] Manila Principles on Intermediary Liability Background Paper (30 May 2015) <https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf> accessed 2 October 2023 (hereafter the 'Background Paper') 4.

and infrastructure layer providers.[86] The Manila Principles therefore provide a more suitable guide for those involved in infrastructure moderation and for informing the development of any existing or future regulatory framework.

The Manila Principles focus on proportionality, transparency and procedural fairness. The Background Paper elaborates on each principle, explaining the rationale behind them and breaking each down into sub-principles with examples of best practices. Building on the analysis in Section 2, this Section will draw out the relevance of the Manila Principles to infrastructure moderation and how they might be applied in practice.

3.1 Proportionality

Principle IV of the Manila Principles provides that content moderation practices and policies must be necessary and proportionate. This Principle addresses one of the key concerns about infrastructure moderation: the likelihood of significant collateral effects on legitimate speech.[87] The Background Paper explains that the least restrictive technical means must be used, which should take into account the harm caused (or likely to be caused) by the content, the likely harm that would be caused by the intermediary taking action, and the proximity of the intermediary to the content.[88]

The key practical implication is that application layer providers, which can take targeted action against *specific* pieces of harmful content, should, by default, make content moderation decisions, rather than infrastructure providers.[89] This is not only because infrastructure providers typically lack the technical means to take targeted action, but also because they are further removed from the context of the content. As a result, it is very difficult for infrastructure providers to assess the likely harm (to be) caused by the content in question. Therefore, when an infrastructure provider becomes aware that its services are being used to distribute harmful content, it should first notify those closest to the content, who can make a more informed assessment and take targeted action if deemed appropriate. In most cases, this will be the operator of the website or platform. For example, before AWS terminated its services with Parler, it had repeatedly notified the platform that it was in breach of the parties' cloud-hosting agreement and requested that it remove content which threatened public safety.[90] However, as this case shows, application layer providers can be uncooperative, meaning infrastructure providers might need to take action

[86] Ibid 6.

[87] Balkin (n 73).

[88] Background Paper (n 39) 36.

[89] Solum and Chung (n 66) refer to this as the 'Layers Principle'.

[90] Alian Selyukh, 'Amazon Says Parler Systematically Unwilling To Remove Violent Content' (*NPR*, 13 January 2021) <https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/13/956362434/amazon-says-parler-systematically-unwilling-to-remove-violent-content> accessed 2 October 2023.

instead. Whether that action would also be *proportionate* will depend on several factors which the infrastructure provider should consider.

First, what is the least restrictive action that will remedy the problem? In AWS's case, for example, it could have suspended its services to Parler to add weight to its warning. Other intermediate measures could be taken, depending on the technical features of the infrastructure provider. For example, ISPs can temporarily throttle the speed of certain services as part of a graduated response to terms of service violations.[91] Similarly, payment processers could display a warning notice to a user who is about to make a donation to a hate speech platform. Infrastructure providers should only withhold services from a user if intermediate steps like these prove ineffective.

Second, what is the likely level of harm (to be) caused by the content in question? Since infrastructure providers are usually unable to assess the context of specific pieces of content, any action should be reserved for content where little context is required to determine the level of harm, such as specific and credible incitements to violence. Content moderation decisions by infrastructure providers are vulnerable to miscalculation when the potential harm of a specific piece of content is highly context-dependent, such as nudity or hate speech.[92] This could be an argument against financial intermediaries (or any other infrastructure providers) withholding services to customers providing access to lawful sexually explicit content.[93]

Third, does the assessed level of harm (to be) caused by the content outweigh the likely level of harm that would be caused by taking the least restrictive but effective action? This is another difficult assessment for an infrastructure provider to make, but one factor that is often overlooked is the dominance of the provider in its service market. For example, if Google or Apple decide to remove a platform from their app stores, the impact is likely to be greater than if GoDaddy decides to stop servicing a platform's domain name, as there are many more alternative domain registrars than there are app store providers.[94] Just as the stakes are very different when moderating at the infrastructure layer than when moderating at the application layer, so too are the stakes when different types of infrastructure providers take action.

3.2 Transparency

[91] Eric Goldman, 'Content Moderation Remedies' (2021) 18 *Michigan Law Review* 1, 17–19.

[92] This is not to say that context is unproblematic for content moderation at the application layer – it is – but the fact that application layer providers have access to more information surrounding a specific piece of content means they are still better placed to make determinations about whether that content breaches their terms of service.

[93] As discussed in Section 2.2.

[94] Accordingly, the Electronic Frontier Foundation (n 11) distinguishes between 'essential infrastructure' and all other 'infrastructure', arguing that only 'essential infrastructure' – which, according to the EFF, includes app stores but not content delivery and security services – should not engage in content moderation.

Transparency is essential in order to assess whether content moderation decisions made by infrastructure providers are necessary and proportionate. Without transparency, we cannot make such assessments and hold the decision-makers to account.[95] As the private power of social media platforms over the public right to freedom of expression has become more evident, calls for greater transparency have grown,[96] leading many platforms to implement a range of policies which fall broadly into one of three dimensions of transparency: disclosure to users; public reporting; and granting researchers access to internal company data.

Despite the influence that infrastructure providers also have on users' right to freedom of expression online, only a small minority disclose anything about their content moderation decisions.[97] One explanation for this may be that public and regulatory expectations of transparency have not developed in the same way as for platforms, since infrastructure providers have traditionally avoided making content moderation decisions. Nevertheless, as the relationship between infrastructure providers and content changes, so should expectations of transparency.

But what exactly should our expectations be? Clearly, greater transparency is needed, but should transparency requirements around content moderation decisions be the same at the infrastructure layer as at the application layer? Much will depend on what is technically feasible at each layer. A layer-conscious approach is important to ensure that high-level principles such as transparency are effectively implemented. For example, while platforms can estimate how many users have seen harmful content on their platform,[98] infrastructure providers do not have automatic access to this information.

However, there are many ways in which transparency at both layers can be broadly similar. Principles VI(c) and (e) of the Manila Principles require *all* intermediaries to publish, or otherwise disclose, any terms of service or policies which inform their content moderation decisions, how those terms of service and policies are applied,

[95] In this way, transparency should be understood as an instrumental good, which matters to the extent it is required to further the aims of intrinsic goods, such as accountability. More transparency is not always desirable in the context of content moderation. It is not difficult to imagine situations in which transparency would be unhelpful (e.g., disclosing source code of content moderation software to users) or even detrimental (e.g., publicly disclosing personal information).

[96] Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129 *Yale Law Journal* 2418.

[97] Stoughton and Rosenzwieg (n 75).

[98] In recent years, several platforms have converged on using exposure-based metrics to increase platform accountability for content moderation decisions. For example, Meta measures 'Prevalence' (Meta, 'Transparency Center: Prevalence' (last updated 18 November 2022) <https://transparency.fb.com/en-gb/policies/improving/prevalence-metric/> accessed 2 October 2023), while YouTube and Snap measure 'Violative View Rate' (Google, 'Transparency Report: Views' <https://transparencyreport.google.com/youtube-policy/views> accessed 2 October 2023; Snap, 'Transparency Report (1 July 2022–31 December 2022)' (last updated 15 June 2023) <https://values.snap.com/en-GB/privacy/transparency> accessed 2 October 2023).

and when and why action has been taken. There is no ostensible reason why infrastructure providers could not disclose this type of information in the same way as platforms, subject to legal restrictions such as data protection laws.

In terms of *how* this information should be disclosed, there is also significant overlap between the infrastructure and application layers. Principle VI(c) of the Manila Principles states that all intermediaries should publish their content moderation policies online, in clear language and accessible formats. This is a transparency measure that is in fact widely adopted by both infrastructure providers and platforms.[99] What is much less common at the infrastructure layer, however, is transparency reporting on content moderation, despite increased levels of reporting at the application layer.[100] This is another form of disclosure required by Principle VI(e) of the Manila Principles for *all* intermediaries, which infrastructure providers should address, not least because content moderation transparency is now a legal requirement for infrastructure providers under EU's Digital Services Act (DSA).[101]

Notably, the Manila Principles are silent on whether infrastructure providers (or other online intermediaries) should give researchers access to internal company data in order to conduct external assessments of infrastructure providers' content moderation practices. Some have argued that, of all the dimensions of transparency, this is the most important in the context of application layer content moderation.[102] This seems to have resonated with policymakers around the world: in the US, members of Congress have introduced or published several bills or discussion drafts which include provisions compelling access to data held by platforms with high numbers of monthly active users to vetted researchers;[103] in the EU, the DSA is now the first major piece of legislation to enact such requirements;[104] and in the UK, the Online Safety Act (OSA) takes a more incremental approach, by requiring Ofcom, the UK's communications regulator, to produce a report within 18 months of the OSA coming into force assessing the extent to which greater access to data is needed to inform research into online safety.[105] However, none of these provisions, as currently drafted, would apply to infrastructure providers.

A reasonable justification for not requiring infrastructure providers to give researchers access to data about their content moderation practices is that it would be disproportionate to do so. This is explicitly recognised in recital 57 of the DSA,

[99] Stoughton and Rosenzweig (n 75).

[100] Ibid.

[101] Council Regulation (EC) 2022/2065 on a Single Market for Digital Services and amending Directive 2000/31/EC [2022] OJ L277 (DSA), Article 15.

[102] Mark MacCarthy, 'Transparency is essential for effective social media regulation' (*Brookings*, 1 November 2022) <https://www.brookings.edu/articles/transparency-is-essential-for-effective-social-media-regulation/> accessed 2 October 2023.

[103] For example, see the Platform Accountability and Transparency Act, S. 5339, 118th Cong. (2023); the Digital Services Oversight and Safety Act, H. R. 6796, 117th Cong. (2022); the Social Media DATA Act, H. R. 3451, 117th Cong. (2021).

[104] See Article 40 of the DSA, which came into effect on 14 February 2024.

[105] OSA, s 162.

which considers that the additional transparency obligations on 'very large online platforms' are appropriate because they have 'a larger reach and greater impact in influencing how recipients of the service obtain information and communicate online' than other online intermediaries. This may be true when comparing the impact of content moderation decisions between very large and smaller platforms, but the analysis is not as straightforward when comparing the impact between the application layer and the infrastructure layer. Although infrastructure providers will make far fewer content moderation decisions than very large online platforms, the impact of those decisions is usually far greater: infrastructure moderation often removes *the ability* to obtain information and communicate online in the first place.[106]

To make an informed assessment of whether additional transparency obligations, such as data access for researchers, should be imposed on infrastructure providers, policymakers need more information on the impact of infrastructure moderation. Transparency reporting by infrastructure providers could provide this information. Policymakers outside the EU should consider whether there is a need to introduce a mandatory reporting regime for infrastructure providers, as provided under the DSA. For example, section 77 of the OSA introduces a mandatory transparency reporting regime for online platforms, which could be extended to infrastructure providers or replicated in separate legislation.

3.3 Procedural Fairness

Transparency alone is not enough to ensure that infrastructure moderation is carried out in a fair and responsible manner. Just as at the application layer, infrastructure providers should build procedural fairness into their content moderation practices in the form of safeguards.[107] In this context, procedural safeguards not only protect users from misinformed or misjudged content moderation decisions by infrastructure providers, but also limit government pressure on infrastructure providers to restrict content voluntarily, outside the rule of law.

The design of these safeguards should recognise that content moderation decisions at the infrastructure layer are, most of the time, qualitatively more impactful than at the application layer and prone to miscalculation. It may therefore be insufficient for infrastructure providers to have procedural safeguards allowing a user to contest a blocking decision after it has taken effect.[108] In light of this, Principle V(a) of the Manila Principles requires infrastructure providers to give users an effective right to be heard *before* a decision is taken, unless doing so would be impossible or impractical (in which case, the infrastructure provider should notify the user and review the decision as soon as possible). In practice, this would involve the infrastructure provider

[106] Ben Thompson, 'A Framework for Moderation' (*Stratechery*, 7 August 2019) <https://stratechery.com/2019/a-framework-for-moderation/> accessed 2 October 2023.

[107] Balkin (n 73); Rory Van Loo, 'Federal rules of platform procedure' (2021) 88(4) *The University of Chicago Law Review* 829.

[108] Busch (n 8) 77.

notifying the user that their website is hosting content that the provider has determined to be in breach of their services agreement with the user, providing a clear explanation of how that determination was reached and how the user can contest it both before (if applicable) and after any final decision is made.

However, disputes will inevitably remain after decisions take effect, and users should have effective recourse to appeal. As at the application layer, this is likely to be achieved with several complementary mechanisms. First, like all major platforms, infrastructure providers could introduce in-house appeal mechanisms where users can appeal content decisions by requesting that the provider review its decision. These proceedings would be cost-free for the user and could prove effective – there is evidence that a considerable number of appeals are successful at the application layer.[109] However, it would still be the infrastructure provider deciding the appeal, with no external oversight. This could justify infrastructure providers setting up an independent, self-regulatory body, similar to Facebook's Oversight Board.[110] While such a body would introduce greater accountability in what would otherwise be an unchecked appeals process, infrastructure providers would be disincentivised to relinquish their discretion over which appeals are heard before such a body when they are voluntarily footing the bill.[111] The DSA attempts to address this at the *application* layer by encouraging EU Member States to establish low-cost, private dispute resolution bodies, which would hear proceedings in which *platforms* would have to participate, so long as the user's claim is brought in good faith.[112] How realistic or proportionate these ambitions are is questionable, and they may well be unnecessary – users can, and have, successfully sued platforms to restore content removals by platforms in their national courts.[113] At least in theory, the same would apply for infrastructure moderation.

[109] For example, Meta reported that about 198,000 content decisions on the grounds of hate speech alone were successfully appealed in Q4 of 2022 (Meta, 'Transparency Center: Hate Speech' <https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook> last accessed 2 October 2023).

[110] Konstantinos Komaitis, 'An (Im)perfect Way Forward On Infrastructure Moderation' (*Techdirt*, 13 October 2022) <https://www.techdirt.com/2022/10/13/an-imperfect-way-forward-on-infrastructure-moderation/> accessed 2 October 2023. Of course, lessons would need to be learned from the exercise of setting up and running Facebook's Oversight Board (Klonick (n 92)). In particular, many have voiced scepticism about its purported independence, which has resulted in leading social media researchers and experts to create the Real Facebook Oversight Board ('The Real Oversight Board' <https://the-citizens.com/campaign/real-facebook-oversight-board/> accessed 2 October 2023).

[111] Daniel Holznagel, 'The Digital Services Act wants you to "sue" Facebook over content moderation decisions in private de facto courts' (*Verfassungsblog*, 24 June 2021) <https://verfassungsblog.de/dsa-art-21/> accessed 2 October 2023.

[112] DSA, Article 21 and recital 59. Article 21 applies to 'online platforms' only, which, as defined under Article 3(i), do *not* include the infrastructure providers mentioned here.

[113] Holznagel (n 107); Daniel Holznagel, 'Enforcing the Rule of Law in Online Content Moderation: How European High Court decisions might invite reinterpretation of CDA §230' (*Business Law Today*, 9 December 2021) <https://businesslawtoday.org/2021/12/rule-of-law-in-

4. Conclusion

This paper has attempted to widen content moderation debates to include the infrastructure providers which keep user-facing platforms, such as Facebook and Twitter, online. It is evident that infrastructure providers can and do engage in content moderation. We must ask ourselves how they are doing so and whether they should be.

Section 2 highlighted that infrastructure moderation is usually not about specific pieces of content but rather ensuring that meaningful content moderation is carried out at the application layer. In this way, infrastructure providers tend to play the role of meta-moderator: if an application layer provider does not meet minimum content moderation standards set by the infrastructure provider, the infrastructure provider will attempt to enforce them by withholding their services from the application layer provider. Withholding services will often take an application provider offline (at least temporarily), meaning the impact of content moderation at the infrastructure layer is qualitatively far greater than at the application layer: infrastructure moderation is about controlling not only how content is treated online but who can be online in the first place. This may be a power which infrastructure providers do not wish to have, but they nevertheless do, and such power is increasingly called upon when harmful content which is (or would be) moderated on mainstream platforms appears on alternative 'free speech' platforms.

In response, Section 3 argued that content agnosticism should not be abandoned at the infrastructure layer, but limited exceptions to this general rule should be recognised and implemented through a framework based on principles of proportionality, transparency and procedural fairness. Putting principles into practice will require regulatory intervention in some form, which we are already seeing in the EU with the introduction of content moderation transparency reporting obligations for infrastructure providers under the DSA. While the DSA could certainly go further – for example, detailed rules for procedural safeguards such as internal complaints-handling systems and external out-of-court settlement apply to platforms only – transparency is an essential first step. To properly scope the role and responsibilities of infrastructure providers in content moderation, regulators should introduce mandatory transparency reporting regimes for infrastructure providers to better understand the frequency, impacts and justifications of their content moderation decisions. Otherwise, the superficial view of content moderation that has so far dominated debates will continue to result in equally superficial responses from policymakers, leaving infrastructure providers to make difficult value judgements with few guiding norms or rules to reference. This is not a position that these providers should be in, as many of them, including Cloudflare, have publicly said.

online-content-moderation-european-high-court-decisions-reinterpretation-cda-section-230/> accessed 2 October 2023.