The London School of Economics and Political Science

Rule of Law and Human Rights Issues in Social Media Content Moderation

MacKenzie F. Common

A thesis submitted to the Department of Law of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, June 2020.

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 100,824 words.

Abstract

This thesis explores the content moderation process at social media companies. This process is divided into three distinct stages: Creation (the production of terms and conditions), Enforcement (the enforcement of those rules), and Response (the use of both internal and external methods of appeal to enact change). It explains how content moderation occurs and identifies a number of serious issues for both human rights and the rule of law in the current approach. It also proposes a variety of solutions for both small-scale and broader reform and argues for a regulatory approach grounded in procedural rule of law principles and mandatory human rights due diligence.

Table of Contents

Chapter One: Setting the Scene

Social media has become engrained into our society in a very short period of time. Platforms such as Facebook, Twitter, and YouTube have moved from niche offerings aimed at university students to global phenomena that have impacted almost every facet of ordinary life, from politics to entertainment, social relationships to consumer marketing. But even as these companies were refining their technologies, adding more functionalities, and expanding their user base, they were also developing another set of processes: content moderation. Content moderation actually represents a bundle of practices at platforms: creating rules, enforcing them through algorithms and flagging, removing, curating, and categorising content, and responding to appeals or queries from users who feel the wrong moderation decision has been made in their case. Moderation ensures that platforms abide by local laws, avoid negative publicity, and create online environments that users want to access frequently. In fact, despite initially being treated as a secondary concern by platforms, moderation is actually the real commodity platforms offer, an online experience that is "curated, organised, archived, and moderated."[1]

The current approach to content moderation, however, poses problems for lawyers and policymakers. These problems are themes that will appear in a variety of contexts throughout the thesis. The first issue is that these are private companies making decisions that affect human rights. This thesis will use the International Bill of Rights as its source for any substantive discussions of human rights. This is the collective name of the Universal Declaration of Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR,) and the International Covenant on Economic, Social and Cultural Rights (ICESCR). It will also use the UN Guiding Principles and John Ruggie's earlier work on the Protect, Respect, and Remedy framework to define the scope of corporate human rights obligations. The Protect, Respect, and Remedy framework is useful because it articulates the idea that companies may not have the same responsibilities to fulfil human rights as states but they

[1] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media* (New Haven: Yale University Press, 2018).

still have a responsibility to avoid causing or contributing to human rights issues and provide remedies when those situations arise.[2]

Social media platforms can cause or contribute to a wide variety of human rights violations.[3] While it would be beyond the scope of this introduction to enumerate every human rights issue that is relevant to social media, it is possible to provide a brief overview to emphasise the diversity of harms that exist on these platforms. These human rights issues can be divided into three broad categories: physical harm and bodily integrity, civil liberties, and risks to basic needs.[4]

Social media platforms can be used by governments, non-state actors and individuals to advocate, incite, or gather information for the purposes of causing physical harm. For example, Permitting (or even featuring in the case of curated content) content that features war propaganda or incites discrimination or violence against people on the grounds of a protected characteristic is a violation of Article 3 of the UDHR[5] and Article 20 ICCPR.[6] These security rights would also be invoked if platforms hand over information about people (such as human rights activists) to regimes that engage in torture, abuse, or unlawful killings.[7] The rights of the child are also important, with social media companies needing to consider what special safeguards can be used to prevent children from accessing inappropriate content or using social media in a way that puts them at risk of exploitation. Concerns about the specific needs of children should include reference to the Convention on the Rights of the Child,[8] where Article 17(e) states that appropriate guidelines for the protection of children from "information and material injurious to…well-being" should be developed. Article 19 of the

---

[2] Principles 11 and 13, *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework (HR/PUB/11/04)*.

[3] Of course, social media also promotes human rights. It is an important platform for expression, allows individuals and organisations to share information about natural disasters, provides an affordable method for e-commerce and advertisement which can help reduce poverty, and facilitates democratic participation. This section, however, will be discussing the *risks* social media poses to human rights.

[4] These organisational categories are used by BSR (Business for Social Responsibility) when they conduct Human Rights Impact Assessments for companies, including Facebook. See: BSR, *Human Rights Review: Facebook Oversight Board*, 2019). 17.

[5] *The Universal Declaration of Human Rights*. 1948. UN General Assembly Resolution 217 A.

[6] *International Covenant on Civil and Political Rights*. 1966. UN General Assembly Resolution 2200A (XXI)

[7] Which would be a violation of Article 7 ICCPR (torture), Article 6 ICCPR (life), and Article 9 ICCPR (liberty and security of person).

[8] *Convention on the Rights of the Child*. 1989. UN General Assembly Resolution 44/25

Convention on the Rights of the Child also protects children from abuse and exploitation and Article 34 focusses specifically on sexual abuse.

Another category of human rights issues caused or contributed to by social media companies is the risk to civil liberties. Article 2 of the UDHR and Article 3 of the ICCPR are prohibitions against discrimination on the grounds of protected characteristics, and this could be an issue when social media companies fail to address persistent harassment or create and enforce rules that have a discriminatory effect. Privacy issues are covered by Article 12 of the UDHR and Article 17 of the ICCPR, which covers privacy of the individual, family, home, and correspondence, as well as attacks on a person's reputation, and both surveillance and defamation are issues on social media. Fair trial rights (Article 14 ICCPR) could be impeded by the sharing of 'wanted' pictures on social media or by violating prohibitions on pre-trial coverage. Finally, as specified in Articles 19 and 20 of the UDHR and Article 19 of the ICCPR, everyone has the right to freedom of opinion and expression and the right to freedom of peaceful assembly and association. These rights will be a frequent source of discussion throughout the thesis as platforms can both facilitate and inhibit these rights at a very high level through content restrictions, withholding content in certain countries, or featuring and curating material. Most people now rely on social networking platforms as the major outlet for expression, accessing information, and maintaining connections and the implications of this dependence must be investigated. Our lived reality of expression is increasingly moving online and onto platforms controlled by private companies and this could diminish human rights protections. While this thesis will discuss a wide variety of human rights issues relevant to social media content moderation, special attention will be paid to freedom of expression as it is so profoundly connected to the activities platforms carry out.

Finally, social media platforms can represent risks to basic needs. The content decisions made could interfere with a person's right to take part in cultural life and benefit from "the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author."[9] The activities of social media companies

[9] *International Covenant on Economic, Social and Cultural Rights*. Article 15(a) and (c).

also affect the content moderators who work for them. The stressful working environments of moderators (especially if no psychological support is available) could be an issue under Article 12 of the ICESCR right to the highest attainable standard of physical and mental health and the Article 7 ICESCR right to just and favourable working conditions. Content moderators would also have rights under labour laws, which includes right to form and join trade unions and initiate collective bargaining.[10] This has been just a brief summary of human rights issues in social media platforms but one can safely conclude that there are many potential problems in the world of content moderation. Businesses have a responsibility to respect all "internationally recognised human rights" because they can have an impact on "virtually the entire spectrum" of rights.[11]

To reiterate, the first issue is that social media companies (which are private platforms) are making decisions that have a significant impact on many human rights. The second issue is that these decisions are being made in a way that would fail any rule of law or due process requirements of accountability, transparency, and access to remedies. These procedural issues exist at every stage in the content moderation process. Accordingly, many of the solutions that will be proposed in this thesis are directed to enhancing procedural protections on social media platforms. These ideas are foundational, creating a solid bedrock for subsequent substantive regulations and helping to address the nuanced situations where companies engage in regulatory initiatives. The final issue is that policymakers are considering various approaches to regulating social media companies but many of their solutions (such as a social media duty of care) fail to address (or may even exacerbate) these rule of law issues. Government reforms can also incentivise censorship as well as the diminishment of other human rights protections in the online environments. This thesis will investigate how these companies are regulating their platforms and why their current approaches to content moderation pose serious problems for the rule of law and human rights principles. It will also consider what reforms could be introduced by the platforms and by UK lawmakers to address the issues identified throughout this thesis.

---

[10] See, for example: Articles 1 and 2, *Right to Organise and Collective Bargaining Convention, 1949, International Labour Organisation*.Article 8 of the *ICESCR*.
[11] Principle 12, *UN Guiding Principles.*

The first topic this thesis must address is that platforms are private companies, who traditionally are not treated as having human rights responsibilities. An orthodoxy of international human rights law is that these are the responsibility of state agencies alone. In many ways, the issues around business and human rights have crystallised in the realm of social media, where companies are being given decision-making powers that affect human rights in an unprecedented way. Therefore, Chapter Two will make the argument that social media companies should be held responsible for human rights issues on their platforms. It will consider a number of theoretical approaches for making this claim such as moral agent, quasi-public, and Business and Human Rights (BHR) ideas and engage with the counterarguments on this issue. This is an important issue because insisting on a strict Westphalian notion of human rights will lead to the impoverishment of human rights protections in a world that is increasingly dominated by privatisation and multinational companies.

The body of the thesis is comprised of three chapters that reflect the three distinct stages of the content moderation process: Creation, Enforcement and Response. Chapter Three covers the Creation stage, which entails the development of rules dictating what content is and is not permissible on the platform. It will argue that the rules are vague, occasionally incoherent, and lack transparency. The consequence for users and activists concerned with human rights is the engendering of a "Kafkaesque uncertainty" online.[12] There is also the issue that the internal guides platforms distribute to moderators to regulate content bear little resemblance to the rules that are publicly available to users. This chapter will explore a range of issues relevant to how the terms and conditions are created by social media companies and argue that greater oversight is needed in order to ensure that the standards are detailed and transparent, and that they broadly reflect rule of law principles.

The Enforcement stage, which is discussed in Chapter Four, concerns how social media companies police their platforms. It will explore the role of the moderator (both human and algorithmic), the process of enforcement, and identify a number of issues related to the inconsistent nature of the enforcement. When rules are created, a tacit promise is

[12] Evgeny Morozov, *The net delusion: the dark side of internet freedom*, 1st ed. (London: Penguin, 2012). 102.

made to the adherents that these are the rules that will be applied to their actions.[13] This continuity allows people to ascertain the law with certainty and adjust their behaviour to avoid sanction, but social media platforms do not always enforce their rules consistently. This chapter will consider why this disparity of application may be occurring and why traditionally marginalised people (such as women and religious minorities) are particularly disadvantaged by inconsistent enforcement.

Chapter Five will discuss the Response stage, which covers both the processes that exist at social networks to handle appeals and the channels that activist groups access when they cannot appeal through the company to initiate change. It argues that the current approach adopted by most platforms is underdeveloped, poses serious human rights issues, and would benefit from a number of reforms. This chapter will explain the principles of an effective appeals system and why alternative channels (such as collective activism) are not an adequate substitute for a robust appeals process at the platform level. This chapter will also discuss the groups who are specially affected by flawed appeals processes such as activists and the growing number of individuals who derive income from their activities on social media.

Currently, there are many possible methods of regulating social media platforms in an attempt to address the issues explored in this thesis. Chapter Six will consider some of the proposals for reform, including the use of platform self-regulation, substantive reform, and a social media duty of care. These solutions are examples of three different regulatory approaches: self-regulation, direct regulation, and co-regulation. This chapter will argue that all these current proposals have significant disadvantages and are incomplete solutions at best.

In Chapter Seven I will identify problems in content regulation and will combine these with the argument that private actors must be held responsible for human rights obligations to suggest a new solution. The thesis will argue that the UK should pass a law mandating human rights due diligence processes for social media platforms (and all businesses more

[13] Lon L. Fuller, *The morality of law*, Storrs lectures on jurisprudence, (New Haven: Yale University Press, 1969). 40.

generally). The objective is to require companies to transition from "generalised commitments" to human rights to the "operationalisation of these commitments-to the rules that give effect to them." [14] This proposal will address the issues of transparency, accountability, and legitimacy that have been raised throughout this thesis and offer a workable solution that still preserves the benefits of the current self-regulatory system but with effective oversight from a business and human rights regulator.

This thesis concludes in Chapter Eight that the challenges in social media content moderation must be addressed in a way that maximises the positive benefits social media offers to society. Instead of allowing platforms free rein to govern their platforms however they wish, governments must create a set of expectations for companies while still maintaining a measure of flexibility. The issues caused by social media companies are an important challenge to 21st century life and it is imperative that we translate human rights and rule of law principles into a workable framework to address these problems.

[14] Emily B. Laidlaw, *Regulating speech in cyberspace: gatekeepers, human rights and corporate responsibility* (Cambridge, UK: Cambridge University Press, 2015). 233.

Chapter Two: Imposing Human Rights Obligations on Social Media Companies

2.1: Introduction

In 1967, Charles L. Black Jr. argued in the *Harvard Law Review* that the state action problem (the notion that private bodies were exempt from constitutional obligations) was the most important problem in American law.[15] His reasoning was based on the fact that the civil rights struggle was unable to confront private discrimination (such as racial discrimination by landlords, employers, and business owners) because of a strict application of this rule. It had even been deployed in 1883 to invalidate the 1875 Civil Rights Act, passed in the wake of the Civil War to remove social barriers for African-Americans.[16] It was only when these laws were relaxed that regulations were passed to address these pernicious problems.[17] The same issue of inaction exists today as social media companies are permitted to govern their platform with little regard for the human rights that are engaged by these practices.

This chapter argues that social media companies should be held responsible for human rights issues on their platforms. This assertion may be supported by a number of different theoretical approaches including the argument that a platform has state-like responsibilities, that it could be considered a moral agent, or should be required to face accountability mechanisms to prevent abuses. It will also outline the orthodox approach to human rights and business and explain why that theory allows some exceptions. Some people may ask why platforms should have human rights obligations as opposed to just being subjected to further forms of regulation. In answer, rights are standards against which we can judge the appropriateness of regulations and offer a "normative vocabulary" to

[15] Charles L. Black, "The Supreme Court, 1966 Term," *Harvard Law Review* 81, no. 1 (1967): 69-70, https://doi.org/10.2307/1339220.
[16] *The Civil Rights Cases* 109 U.S. 3(1883).
[17] "Developments in the Law: State Action and the Public/Private Distinction," *Harvard Law Review* 123, no. 5 (2010): 1258.

identify objectives.[18] This objective should matter whether we are discussing the actions of governments, or of non-state actors who have the power to violate human rights.

2.2: The orthodox approach

2.2.1: Introduction to the orthodox approach

The orthodox approach is that companies (as non-state actors) do not typically have human rights responsibilities. Human rights obligations are traditionally viewed as Westphalian, with a focus on the duties of states. Indeed, a few decades ago "the responsibility of businesses for human rights was, at best, a marginal topic among those concerned with the ethics of business. Some doubted whether business could have any ethical responsibilities at all."[19] There are a number of justifications for this view such as the fact that no businesses (or indeed NGO's) have signed any documents that constitute the International Bill of Rights, which is an indication that there was no expectation that companies be bound by these obligations.[20] Indeed, some of the human rights identified by the UN could only apply to parties that are capable of determining their nationality and immigration status or passing legislation, conduct which is unlikely to be relevant in a commercial setting.[21]

There is a spectrum of different viewpoints that fall under the orthodox approach to business and human rights. At its most liberal conceptualisation, some business scholars have argued that while human rights obligations should not be directly assigned to companies, it is perfectly plausible to develop a firmer approach to complicity, with companies being held accountable for their involvement in human rights violations by

---

[18] Jean Thomas, *Public rights, private relations*, 1st ed. (Oxford: Oxford University Press, 2015), 3.
[19] George G. Brenkert, "Business Ethics and Human Rights: An Overview," *Business and Human Rights Journal* 1, no. 2 (2016): 277-78, https://doi.org/10.1017/bhj.2016.1.
[20] Brenkert, "Business Ethics and Human Rights," 288.
[21] George G. Brenkert, "Google, Human Rights, and Moral Compromise," *Journal of Business Ethics* 85, no. 4 (2009): 455, https://doi.org/10.1007/s10551-008-9783-3.

states.[22] Under this formulation, Facebook could face allegations of being complicit in the human rights abuses perpetrated in Myanmar since the platform has been used to spread hate speech and advocate violence against the Rohingya people.[23] This aligns with the idea that companies have a responsibility to "prevent or mitigate adverse human rights impacts" that are linked to their business even if they have had a passive role in the situation.[24] On the other side of the spectrum is the shareholder primacy theory, which condemns any use of corporate resources on activities not focussed on corporate profits and holds that managers must violate human rights if this would be more profitable than compliance.[25] It should be noted, however, that the shareholder primacy theory founders in a globalised world as it is premised on the assumption that the corporation is operating in a democracy and that the public have therefore been offered an opportunity to weigh in on how corporations are regulated.[26] The shareholder primacy theory also seems at odds with the ethos of social networking, where aspirational language about communities and connections is frequently invoked and claims about improving society are commonplace.

The orthodox approach is more subtly embedded in schemes that encourage corporate social responsibility [CSR] or which focus on the "business case for human rights." CSR typically relies on corporate voluntarism and the notion that corporations should act as "voluntary and affirmative contributors to human rights realisation."[27] CSR began in the aftermath of World War Two, when leading business scholars began to conceptualise the responsibilities of an ethical corporation.[28] It is often linked to self-regulatory schemes,

---

[22] Nien-hê Hsieh, "Should Business Have Human Rights Obligations?," *Journal of Human Rights* 14, no. 2 (2015): 229, https://doi.org/10.1080/14754835.2015.1007223.
[23] This will be discussed in detail at 6.2.3. For background, see: *Report of the independent international fact-finding mission on Myanmar (A/HRC/42/50)* (Geneva: United Nations, 2019), 12.
[24] Principle 13(b) *UN Guiding Principles.*
[25] Milton Friedman, "The Social Responsibility of Business Is to Increase Its Profits," in *Corporate Ethics and Corporate Governance*, ed. Walther Ch Zimmerli, Markus Holzinger, and Klaus Richter (Berlin: Springer, 2007). For an overview (but not an endorsement) of this school of thought see: Denis G. Arnold, "Corporations and Human Rights Obligations," *Business and Human Rights Journal* 1, no. 2 (2016): 268, https://doi.org/10.1017/bhj.2016.19.
[26] Arnold, "Corporations and Human Rights Obligations," 268.
[27] Anita Ramasastry, "Corporate Social Responsibility Versus Business and Human Rights: Bridging the Gap Between Responsibility and Accountability," *Journal of Human Rights* 14, no. 2 (2015): 237-38, https://doi.org/10.1080/14754835.2015.1037953.
[28] The two seminal works on the development of CSR were both published in 1949. See: Donald K. David, "Business Responsibilities in an Uncertain World," *Harvard Business Review* 27, no. supplement (1949): 1-8; Bernard Dempsey, "Roots of Business Responsibility," *Harvard Business Review* 27 (1949): 393-404.

which will be discussed throughout this thesis. CSR has recently been the target of criticism from Business and Human Rights scholars (at 2.5) because CSR fails to prioritise remedies for victims of corporate human rights abuses.[29] Companies should not be permitted to acquire wealth through human rights violations (such as violations of labour rights, cooperation with autocratic regimes, or maintaining unsafe working conditions) and then off-set those negative images through corporate philanthropy. It is admirable, for example, that Jeff Bezos has announced a ten billion dollar donation to fight climate change but it would be even more admirable if he also addressed the myriad of human rights challenges at Amazon (including threatening to fire employees who have spoken out about Amazon's poor environmental practices).[30]

The business case for human rights (or the enlightened self-interest approach) seeks to appeal to businesses by arguing that it is profitable for corporations to respect human rights and has helped to inform John Ruggie's work on the UN Guiding Principles (UNGP's).[31] The UNGP's were created as a reaction to earlier, failed attempts to create international corporate human rights obligations such as the UN Norms project.[32] The UNGP's chose a soft law approach and prioritised consensus-building and consultation with industry in the hope that there would be wider acceptance of rules that companies helped to create.[33] It is true that compliance with soft laws on human rights (such as the UNGP's) offers some important benefits to companies. These benefits are discussed at 6.2.1 but in short they include: positive public image, the prevention of situations that would disrupt commercial activities such as labour unrest, a reduction in the risk of litigation, and an enhanced ability to appeal to investment partners concerned with ethical investments. The issue is, however, that there

---

[29] Ramasastry, "Bridging the Gap," 247.
[30] Richard Luscombe, "Amazon's Jeff Bezos pledges $10bn to save Earth's environment," The Guardian, last modified February 17, 2020, https://www.theguardian.com/technology/2020/feb/17/amazon-jeff-bezos-pledge-10bn-fight-climate-crisis.. For a selection of Amazon's human rights issues, see: "Amazon.com stories," Business and Human Rights Resource Centre, accessed March 5, 2020, https://www.business-humanrights.org/en/amazoncom.
[31] Louise J. Obara, "'What Does This Mean?': How UK Companies Make Sense of Human Rights," *Business and Human Rights Journal* 2, no. 2 (2017): 254, https://doi.org/10.1017/bhj.2017.7.
[32] John Gerard Ruggie, "Global Governance and 'New Governance Theory': Lessons from Business and Human Rights," *Global Governance* 20, no. 1 (2014), https://doi.org/10.1163/19426720-02001002. 9.
[33] United Nations Economic and Social Council, *Interim Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises (E/CN.4/2006/97)* (Geneva: United Nations, 2006).

may not always be a business case for human rights. There may be many jurisdictions where the company would suffer no major losses in reputation or harmony from violating human rights because the country experiences frequent human rights violations.[34] Take, for example, a social media company cooperating with the government to remove any content the government terms "fake news" but which may actually only be critical of the party line. If this were to occur in a country that also has a heavily censored media, then the public may not expect anything better from the platform. There may also be situations where the business case is strongly in favour of non-compliance (particularly in countries with weak governance and limited forms of accountability) and by transforming human rights into a cost-benefit analysis, theorists implicitly allow that non-compliance is appropriate. A social network, for example, may hand over information about activists to a government because a refusal would result in their exclusion from that market. Cragg concludes that the business case is therefore flawed because "enlightened self-interest is not capable of sustaining the human rights agenda against competing business imperatives."[35]

2.2.2: Social media companies: a possible exception to the orthodox approach?

An orthodox view of social media companies is that they are entitled to govern their platforms in whatever way they consider appropriate, removing any content that they wish as "there is no such thing as private censorship."[36] Of course, the platform would also be able to keep up content that constitutes hate speech, violates privacy or harasses individuals. This idea is best exemplified by Clay Shirky's famous quote that the "Internet is not a public sphere. It is a private sphere that tolerates public speech."[37] This argument might have been less problematic in the earlier days of the internet, where users were spread across an endless number of online enclaves. Then users were able to "vote with their feet" if they felt their rights were being disrespected. However, now that a small number of tech companies

---

[34] Wesley Cragg, "Ethics, Enlightened Self-Interest, and the Corporate Responsibility to Respect Human Rights: A Critical Look at the Justificatory Foundations of the UN Framework," *Business Ethics Quarterly* 22, no. 1 (2012): 14, https://doi.org/10.5840/beq20122213.
[35] Cragg, "Ethics, Enlightened Self-Interest,..." 10.
[36] Tom Bowden, "Blacklists are not Censorship," Ayn Rand Institute, last modified March 23, 1999, https://ari.aynrand.org/issues/government-and-business/individual-rights/blacklists-are-not-censorship/.
[37] Erica Newland et al., *Account Deactivation and Content Removal: Guiding Principles and Practices for Companies and Users* (Cambridge, MA: Berkman Centre for Internet and Society, 2011), 5.

exert a powerful dominance over most online activity it becomes more troubling. In fact, this concentration of power has become a source of concern for a number of academics such as Cass Sunstein and Eli Pariser, who have written extensively on the balkanisation of the internet.[38] This privatisation of the public sphere represents a fundamental shift in the availability of opportunities for expression and one that necessitates a re-evaluation of the orthodox perspective. The orthodox view of social media companies, therefore, seems to have been transported wholesale from a time when the President of the United States didn't announce his policies first on Twitter[39] and when abstaining from social media didn't hinder one's employment and social prospects.[40]

Even if one were to adopt the orthodox approach, one might nevertheless be able to make the argument that platforms have publicly aligned themselves with human rights and have directly benefitted from their association with human rights. In light of those circumstances, it could be reasonable to infer that social media companies have accepted responsibility for the human rights impacts that occur on their platforms. Social media companies benefit from their affiliation with civic life and freedom of expression, "appropriating" this allusion "to salvage the virtues of the corporate sphere."[41] Because of this affiliation with human rights, social media companies are also offered a measure of privilege and protection. Take, for example, the case of *Packingham v North Carolina*, which will also be discussed at 2.2.3.[42] In 2017, the US Supreme Court struck down a law that prohibited registered sex offenders from using social media. In the judgement, Kennedy J (providing the opinion of the majority) stated that social media platforms are "websites integral to the fabric of our modern society

[38] Cass Sunstein, however, thought the balkanisation would come from users growing increasingly polarised but his work is very important in identifying the dangers of fragmentation. Eli Pariser, however, identified the growing power of certain websites (Google and Facebook are singled out) as driving the fragmentation process. Eli Pariser, *The Filter Bubble: What the Internet is Hiding from You,* (London: Penguin, 2011), 9. Cass Sunstein, *Republic.com 2.0* (Princeton: Princeton University Press, 2007), 72.
[39] Jessica Estepa, "We're all atwitter: 3 times President Trump made major announcements via tweets," USA Today, last modified December 15, 2019, https://eu.usatoday.com/story/news/politics/onpolitics/2018/03/13/were-all-atwitter-3-times-president-trump-made-major-announcements-via-tweets/420085002/.
[40] Danielle Keats Citron argues that social media has significant implications for a person's reputation, career, social circle, and romantic life and therefore opting-out of online participation can be detrimental. Danielle Keats Citron, *Hate crimes in cyberspace* (Cambridge, MA: Harvard University Press, 2014), 7-10.
[41] José van Dijck, *The culture of connectivity: a critical history of social media* (Oxford: Oxford University Press, 2013), 16.
[42] *Packingham v. North Carolina*, 137 S. Ct. 1730(2017).

and culture"[43] and that they "provide the most powerful mechanisms available to a private citizen to make his or her voice heard."[44] It was a ringing endorsement of social media platforms as places where First Amendment rights could be exercised and it was clear that the Court would be strictly scrutinising any laws seeking to inhibit social media use.[45] The *Packingham* decision, however, leads to a fascinating dichotomy, a schism of reason. These platforms are being affirmed as the pre-eminent place for the exercise of the right to free expression (as well as other democratic virtues), a resource so important that states should not prevent citizens from using them, and yet these platforms are treated as having no real human rights obligations, as not being capable of infringing the very acts of expression that the Court is seeking to protect. It seems very clear that *Packingham* affords these platforms a privileged status as essential forums for human rights. It therefore does not seem radical to argue that such protected status should come with some obligations as well.

This idea that platforms have accepted responsibility for human rights could be linked to a point that business ethicists make about the role of corporations. In direct contradiction to the shareholder primacy theory, ethicists argue that corporations are creatures of statute and that with the privileges afforded to these companies, there should be some responsibility to further social goals.[46] This argument can lapse into hyperbole (that corporations are designed to further human rights specifically[47]) but its specific application appears valid: that companies should assume "some responsibility" to ensure that their business activities

[43] *Packingham v. North Carolina*, 137 S. Ct., at 10.
[44] *Packingham v. North Carolina*, 137 S. Ct., at 8.
[45] The fact that this case concerned registered sex offenders also shows the high level of protection the Supreme Court was affording social media companies as this is a group that society often expects will be heavily restricted in their activities.
[46] David Bilchitz, "Corporate Obligations and a Treaty on Business and Human Rights: A Constitutional Law Model?," in *Building a treaty on business and human rights: context and contours*, ed. Surya Deva and David Bilchitz (Cambridge, UK: Cambridge University Press, 2018), 203.
[47] Indeed, Bilchitz makes this argument when he writes "In creating laws, legislatures cannot themselves be motivated simply to enrich private individual interests; that would be an illegitimate exercise of their power. The normative basis for the exercise of legitimate legislative power must be founded in a commitment to advance the interests of all members of society and, to do so, in a manner that demonstrates the equal importance of every individual in the society. The protection of human rights is a core purpose that flows from this principle of equal importance: in performing their tasks, legislators must thus have the realisation of such rights as one of their core goals." This seems like a stretch as many legislators would perceive value in laws that encourage property rights and economic freedom. Bilchitz, "A Constitutional Law Model?," 203.

results in the attainment of objectives that the public expects from these laws.[48] Cragg uses the example of pharmaceutical companies justifying patent laws on the public benefits they generate (innovation, resources for further study etc) and argues that they have a responsibility to integrate human rights into their practices to encourage further public benefit.[49] The notion that platforms have accepted responsibilities for human rights (or at the very least, civil rights) is not the most assertive argument for imposing corporate human rights obligations but it is one possible avenue that could be explored. This chapter will show, however, that the orthodox approach represents only one of many different approaches to corporate human rights obligations and arguably one that is losing ground as the public becomes more concerned about the actions of private actors such as social media companies.

The essential problem with the orthodox theory is that it offers no real answer to how to respond to corporate activities that would be considered human rights abuses if committed by state authorities. There are some suggestions of course, such as encouraging voluntarism, appealing to the self-interest of the business, and enlarging complicity rules, but these solutions seem unsatisfactory when the scale of the problem is so large. When one contemplates recent scandals such as the BP oil spill, the Rana Plaza collapse, and the electoral manipulations perpetrated by Cambridge Analytica, adages about shareholders and profits seem outdated, perhaps even toxic. There may be ways to moderate the orthodox theory, making exceptions for companies that assume responsibility for human rights or benefit from their affiliation with human rights values. There should also be greater expectations of how companies conduct their activities even if some theorists refuse to make human rights a priority. At the minimum, for example, the public should be guaranteed some measure of transparency and fairness in how social media platforms operate even if orthodox theorists would not support a fuller approach.

Ultimately the orthodox theory should be supplanted by approaches that respond to the new challenges of multinational corporations and weak governance regimes by arguing that

---

[48] Wesley Cragg, "Human Rights, Globalisation and the Modern Shareholder Owned Corporation," in *Human Rights and the Moral Responsibilities of Corporate and Public Sector Organisations*, ed. Tom Campbell and Seumas Miller (Dordrecht: Springer, 2005), 118.
[49] Cragg, "Human Rights..." 121.

corporate human rights obligations should exist. Orthodox theorists might condemn the introduction of such laws but it should be noted that "almost no management, finance, or economics scholars explicitly defend the idea that companies should violate the law in the interest of additional wealth creation."[50] Therefore, if these obligations were introduced then orthodox theorists would conclude that the appropriate response must be compliance. The following sections will consider theories on why human rights obligations should be imposed on social media companies.

2.3: Quasi-Public Spaces

2.3.1: Introduction

It might be argued that social media companies fulfil a state-like function and that it is therefore appropriate to expect human rights protections from them as if they were states, or at least quasi-states. The terminology may differ, but a number of different approaches have emerged to describe what is essentially the same situation, at least in relation to social media. There is the notion that social networks are performing activities that are usually the responsibility of governments and they should therefore be constrained by a similar set of rules. A related idea is that platforms resemble traditional public spaces (or public forums) to such an extent that people are entitled to exercise their rights (most notably free expression) as if it were a public space.[51] These spaces are sometimes termed "pseudo-public spaces" as there is no legal authority designating them as public spaces even if there is some form of "geographical/cultural/media" understanding of them as public.[52] A similar

[50] Arnold, "Corporations and Human Rights Obligations," 259.
[51] A recent article by the Bishop of Chelmsford takes this idea further, arguing that the Internet should indeed be treated like a public space but as public spaces are open to all ages (including children), the Internet should also be regulated as a safe space for children. It is an interesting idea but it disregards other important priorities in regulating the Internet, such as creating a space for free expression, access to information (some of which is not appropriate for children), and democratic participation. No child should be shown, for example, a video of a peaceful protester being shot by the police but that does not mean it should sanitised from the Internet. "The internet must be made safe for children," accessed 31st January 2019, https://www.chelmsford.anglican.org/news/article/the-internet-must-be-made-safe-for-children.
[52] Daithí Mac Síthigh, "Virtual walls? The law of pseudo-public spaces," *International Journal of Law in Context* 8, no. 3 (2012), https://doi.org/10.1017/S1744552312000262. 396.

approach may be found in the UK Human Rights Act (HRA), which uses the term "hybrid public authority" to designate certain "functions of a public nature" committed by private organisations as falling within the scope of the HRA.[53] There is a certain logic to these approaches and that intuitive sense has informed the development of a pragmatic body of case-law and theoretical discussion. This school of thought moderates the orthodox approach by acknowledging that private companies may be treated as having human rights obligations *in certain circumstances*. This section will attempt to describe the themes that cut across this area and show why the unique status of social media platforms warrants the imposition of human rights obligations.[54]

2.3.2: The vanishing town square

The first theme that can be identified in this field is that the public sphere is being supplanted by the private sector. This could be more specifically construed as the notion that the town square (as a symbolic embodiment of civil society)[55] is vanishing and that civic life is increasingly shifting into privately owned spaces in both the physical world (such as shopping malls) and virtual space (such as social media companies).[56] Jørgensen argues that while private domains have always had a place in public life, such as in coffee shops or through newspapers, the current situation is different in scope as the "vast majority of social interactions" online occur on platforms provided by private companies.[57] This migration

---

[53] *Human Rights Act 1998*, c. 42 (Eng. and Wales).Section 6(3) b and Section 6(5).

[54] These themes are somewhat similar to the three principles identified by the court in Marsh v Alabama but these themes have been adjusted to reflect a broader area of scholarship and the contemporary challenges of technology. The principles identified in the case are the company town was no different from any other municipality other than the fact it belonged to a private corporation so it could violate human rights just as easily, the town was accessible to everyone so it was serving a public function, and holding the town constitutionally accountable would be beneficial to the public. *Marsh v. Alabama*, 326 U.S. 501, 503-06 (1946).

[55] Hunter argues that the two archetypal town squares that are frequently evoked in academic literature and jurisprudence is the Athenian Senate and Hyde Park Speaker's Corner, and the "myth of their influence and importance is hard to dispel." Dan Hunter, "Cyberspace as Place and the Tragedy of the Digital Anticommons," *California Law Review* 91, no. 2 (2003): 488, https://doi.org/10.2307/3481336.

[56] Jackson argues, however, that social media platforms are even more essential than spaces like shopping malls, because malls are primarily focused on commerce and "whose ability to serve as a forum for speech is merely incidental" whereas platforms are designated specifically for expression. Benjamin F. Jackson, "Censorship and freedom of expression in the age of Facebook," *New Mexico Law Review* 44 (2014): 146.

[57] Rikke Frank Jørgensen, "Framing human rights: exploring storytelling within internet companies," *Information, Communication and Society* 21, no. 3 (2018): 340, https://doi.org/10.1080/1369118X.2017.1289233.

into the private sphere has led to greater interest in how these spaces are governed. By the 1990's, the traditional delineation between the responsibilities of public and private bodies began to be challenged. The border between the two categories became more porous as privatisation, globalisation, and the growth of transnational corporations forced many scholars, judges, and activists to question whether their rights really stopped at the entrance of private property. This fragmentation of political authority and blurring of the line between public and private spheres has led to a re-evaluation of the duties of non-state actors.[58] The increasing power of corporations has oxygenated the quasi-state argument, with many academics pointing to this newfound power in the international political system as evidence that "political authority should imply public responsibility."[59]

There is, therefore, an instinctive sense that the power and control that a private company wields can approach a certain threshold. After this threshold is reached, justice demands that these companies be given commensurate obligations. An early case that exemplifies this situation is *Marsh v Alabama (1946)*.[60] This case concerned a company town (a town which is owned by a private company that provides the majority of jobs, housing, and amenities) that prohibited Jehovah's Witness members from distributing religious literature on the town's sidewalks. The American Supreme Court treated the town like a state actor, holding that that the private property rights of the company did not "justify the State's permitting a corporation to govern a community of citizens so as to restrict their fundamental liberties." [61] This functional equivalency approach has only grown more appealing as our concerns changed from company towns to multinational corporations.[62]

---

[58] Stephen J. Kobrin, "Private Political Authority and Public Responsibility: Transnational Politics, Transnational Firms, and Human Rights," *Business Ethics Quarterly* 19, no. 3 (2009): 353, https://doi.org/10.5840/beq200919321.
[59] Kobrin, "Private Political Authority and Public Responsibility: Transnational Politics, Transnational Firms, and Human Rights," 350.
[60] *Marsh v. Alabama*, 326 U.S.
[61] *Marsh v. Alabama*, 326 U.S., 509.
[62] One academic tried to argue that *Marsh v Alabama* should be applied to the virtual roleplaying game Second Life. This argument seems dated now as Second Life never achieved the cultural embeddedness of platforms like Facebook and Twitter. Jason S. Zack, "The Ultimate Company Town: Wading in the Digital Marsh of Second Life," *University of Pennsylvania journal of Constitutional Law* 10 (2007).

Social media platforms are a clear example of how a privately owned space can become essential to free expression, as well as facilitating accountability of the state through the sharing of information. In the first American Supreme Court case that considered internet regulation, (*ACLU v Reno*) the Court stated that "through the use of chat rooms, any person with a phone line can become a town crier with a voice that resonates farther than it could from any soapbox."[63] While this is an optimistic assertion (and similar assertions are made in *Packingham*) the concern is that by moving civic life onto private domains, human rights protections may be diminished.

2.3.3: The impacts of corporate activities

The second theme is that as these companies grow more powerful, their activities begin to impact on the public in such significant ways. In addition, the scale and severity of these impacts is comparable to the state, the "dynamic that characterises the relation between individuals and state has begun to appear in non-state relations."[64] This societal shift means that now, "most major decisions about our lives are made in the private sector, not by a state bureaucracy."[65] This idea is applicable to current social media companies as these platforms have become essential to expression because they are free to users, offer the opportunity to connect to large audiences, and are open to anyone with the requisite technology (which is becoming more affordable every year). Platform governance, therefore, can have serious impacts on people because deciding which content to permit or which users should be allowed to remain on the platform will affect people's ability to express themselves, access information, protect their privacy, and take part in cultural life. These impacts are comparable to the effects that public authority decisions could have on a person.[66] The power companies exert, therefore, can move from influence to de facto

[63] *Reno v. ACLU*, 521 U.S. 844, 870 (1997).
[64] Thomas, *Public rights, private relations*, 188.
[65] Frank Pasquale, *Black box society: the secret algorithms that control money and information* (Cambridge, MA: Harvard University Press, 2015), 17.
[66] Of course, it would depend on the decision in question. An individual who uses YouTube to watch music videos would suffer considerably less if their account was blocked than an individual who uses YouTube to create monetised content and is dependent on the platform for their primary source of income. In the same way, a decision that an individual is no longer eligible for housing benefits will have a much more serious impact than the decision that an individual must remove the building materials from their front yard.

authority as stakeholders begin to comply with corporate requirements and accord them some measure of legitimacy through their compliance.[67]

There have been a number of cases that have concerned tech companies and the impact they have on the public. All the cases already discussed in this chapter and most of those to be discussed next are American cases. As so many major social media companies are headquartered in US Federal jurisdictions, it has been the natural forum for legal challenges. These cases, however, are interesting and relevant beyond America because they offer a chronological account of how technological change has transformed from a niche, secondary concern into a sector that has fundamentally altered society. One early (and failed) attempt to argue that tech companies had reached a comparable level to states is the case of *Cyber Promotions Inc. v. America Online Inc.*[68] which concerned a spamming company claiming that AOL had violated its First Amendment rights by banning them from its network. In order to be successful, Cyber Promotions had to convince the court that AOL was acting like a state and was therefore capable of violating human rights. They argued that AOL acts like a government from the perspective of the user and has created a virtual town square where "public discourse, conversations and commercial transactions can and do take place."[69] The court rejected this claim, concluding that "AOL is merely one of many private online companies which allow its members access to the Internet through its e-mail system where they can exchange information with the general public. The State has absolutely no interest in, and does not regulate, this exchange of information... around the world."[70]

Cyber Promotions was likely unsuccessful because AOL did not seem powerful enough to be comparable to a state.[71] Since then, however, technology has become more integrated into society and power has concentrated into a handful of tech companies. This has resulted in a number of cases from multiple jurisdictions affirming the public role of tech

---

[67] Florian Wettstein, *Multinational corporations and global justice: human rights obligations of a quasi-governmental institution* (Stanford, CA: Stanford Business Books, 2009), 210-11.
[68] *Cyber Promotions Inc. v. America Online Inc*, 948 F. Supp. 436(E.D. Pa. 1996).
[69] *Cyber Promotions Inc. v. America Online Inc*, 948 F. Supp., 441-42.
[70] *Cyber Promotions Inc. v. America Online Inc*, 948 F. Supp., 442.
[71] It is also possible that the unappealing activities of Cyber Promotions (spam) influenced the court. Then again, Packingham concerned the rights of registered sex offenders to use social media so these two cases could stand as a testament of how much technology and society had changed in 21 years.

companies. In the 2012 European Court of Human Rights case *Yildirim v Turkey*, the Court held that blocking access to Google was a breach of Article 10 of the ECHR and that the Internet "has now become one of the principal means by which individuals exercise their right to freedom of expression…"[72] The implications of this case seem to be that some Internet companies are so important to the public that denying access to them is a violation of human rights. This argument is hard to reconcile with assertions that Internet companies do not occupy a special place in society. Another case that is pertinent to this point was already introduced (at 2.2.2), *Packingham v North Carolina*.[73] This case came to a similar conclusion as Yildirim on the issue of prohibiting registered sex offenders from using social media. The importance of these cases is that if platforms are being treated as essential conduits for human activities then one must conclude that any decision that impacts a user's ability to participate in social media could trigger legal action.

2.3.4: The appropriate course of action: treat them as quasi-states

Advocates of the quasi-state approach argue that the only way to address the influence of certain private sector companies and the power they have over the public is to treat them as if they were states and hold them accountable for human rights violations. This conclusion is based on the idea that we have a legitimate interest in how these governance decisions of private actors are being made and that they should not have unbounded discretion in how they order their activities. Wettstein in particular argues that transnational corporations have become quasi-state entities and that it is therefore appropriate to attribute human rights responsibilities to them.[74] This is important because the current approach to non-state actors does not treat them as having "political responsibility towards a particular community" or any obligations to respond to the people their activities affect.[75]

Despite the failure of *Cyber Promotions*, recently there has been a renewed interest in interpreting the actions of platforms (and the actions of users on those platforms) as quasi-

[72] *Ahmet Yildirim v. Turkey (application no. 3111/10), ECHR 458, 54 (2012).*
[73] *Packingham v. North Carolina*, 137 S. Ct.
[74] Wettstein, *Multinational corporations and global justice*, 18.
[75] Dalia Palombo, *Business and human rights: the obligations of the European home states* (London: Hart, 2020). 4.

state activities. This started with *Packingham*, which seemed to signal a new era of affirming the significance of social media usage in everyday life and subjecting laws that exclude people from these platforms to strict scrutiny. While Packingham is a fascinating case, it will likely be the precursor to more important challenges in the future. Klonick argues that future cases might use Packingham's contention that "access to private online platforms" is a First Amendment right to argue that these platforms perform "quasi-municipal functions."[76]

Another case, *Knight First Amendment Institute v. Donald J. Trump*[77] appears to illustrate this expansion. This case argued that blocking them from viewing President Trump's Twitter account because they disagreed with his political views was a violation of their First Amendment rights. The plaintiffs argued that this account was a public forum and that they were being prevented from accessing information about the many government decisions announced first on Twitter. Both the lower court and appeals court held that that Trump's Twitter account was indeed a public forum and it was unconstitutional to block users because of their political views. There have since been a number of other cases that also apply this rule to government officials at a lower level.[78]

Knight is more progressive than *Packingham*, advancing the notion of human rights being enforceable on platforms. Of course, it should be noted that *Knight* and *Packingham* did not concern the enforcement of human rights against the platforms directly but rather against a state that had passed a law excluding certain individuals from social media and against public figures (such as the President of the United States) who block people from Twitter. Still, these cases contribute to a climate of human rights on the platform,[79] a sense of enhanced accountability that will likely result in a case against a social media company for infringing human rights being brought in the near future. It should also be noted that these cases are not about protecting social media companies, they are about protecting users from

---

[76] Kate Klonick, "New Governors: The People, Rules, and Processes Governing Online Speech," *Harvard Law Review* 131 (2017): 1611.
[77] *Knight First Amendment Institute v. Donald J. Trump*, 928 F.3d 226(2nd Cir. 2019).
[78] See: *Davison v. Randall*, 912 F.3d 666, 680 (4th Cir. 2019); *Robinson v. Hunt City, TX*, 921 F.3d 440, 447 (5th Cir. 2019).
[79] Another recent case held that clicking the "like" button on Facebook is protected speech so human rights are being increasingly integrated into our perceptions of platform activities. See: *Bland v. Roberts*, 730 F.3d 368, 385-86 (4th Cir. 2013).

interference with their ability to carry out activities on social media so it does not seem unreasonable to predict that a future case will consider how social media platforms themselves can interfere with the activities of users and why this would need to be regulated.

The challenge of treating social media companies as quasi-states (for the purposes of human rights) is that these cases tend to be fact-specific.[80] Even senior British judges have struggled to decide whether private companies are behaving like public authorities, with Lord Neuberger once commenting that the words "functions of a public nature" are "so imprecise in their meaning that one searches for a policy as an aid to interpretation."[81] Which companies would merit this treatment is often difficult to predict. It seems easier to conclude, for example, that Google and Facebook have the requisite power and influence to be treated as exercising public functions but what about Pinterest or Snapchat? This approach is also problematic when it comes to new and emergent social media companies as during their emergent phase they might adopt policies and engage in activities that will later be identified as in conflict with human rights responsibilities once they become successful. This may give an unfair advantage to smaller companies, and could result in controversial practices that may become embedded in their business culture. Another complication of applying notions of the quasi-public is that it results in social media companies being subject to both "public and private content controls spanning multiple jurisdictions and differing social mores."[82] This is of less concern, however, because social media companies are already subject to a range of different laws around the world and harmonisation of these regulations is unlikely to be achieved anytime soon. Ultimately, the quasi-public idea (and the other related ideas detailed above) is very important as it identifies themes that are very relevant to regulating social media companies. One might, however, achieve a more uniform and predictable response by adopting a process that

[80] Jackson argues that this approach might offer less certainty to users than leaving the decisions with social media platforms, who do have accessible rules. How detailed and accessible these rules really are, however, will be explored in the next chapter. Jackson, "Censorship and freedom of expression," 141.
[81] *YL v. Birmingham City Council*, 95 1 AC (2008). Para 128.
[82] Jillian C. York, "Policing Content in the Quasi-Public Sphere," Open Net Initiative, last modified September, 2010, https://opennet.net/policing-content-quasi-public-sphere.

explicitly identifies which companies would be subject to these rules and what would be required of these platforms.

Despite these drawbacks, the merits of the quasi-state approach remain. The arguments made in this area have laid the foundations for future claims against social media companies. These ideas represent a departure from a "a narrow and formalist construction" of the public/private divide which could "could prevent the protection of communications that so vigorously embody First Amendment values."[83] It also acknowledges that the issue of companies and human rights is nuanced and should not be treated as a one-size-fits-all situation. Special care should also be taken when examining self-regulatory or co-regulatory regimes in order to ensure that those who perform public functions do not avoid public responsibilities.[84] This area is particularly adept at characterising the complicated nature of social media platforms as they adopt more state-like features and privileges. It cautions us "not to equate the broader sociopolitical importance of public spaces with a simple (albeit useful) categorisation of places, facilities or environments into boxes of public and private."[85] While these ideas may not be sufficient in and of themselves to inform regulation, they provide a level of complexity and analysis that enriches discussions of corporate human rights obligations.

2.4: Moral Agents

2.4.1: The concept of moral agency:

Another approach is the argument that social media companies are moral agents and that this is an appropriate basis for imposing human rights obligations. The starting point of a moral agency approach is the primacy of the rights-holder. Rights are perceived not merely as a "limitation on the exercise of state power" but rather as attributes that are essential to

[83] Jackson, "Censorship and freedom of expression," 134.
[84] Daithí Mac Síthigh, "Datafin to Virgin Killer: self-regulation and public law," *Norwich Law School Working Papers Series* 09/02 (2009). 21.
[85] Mac Síthigh, "Virtual walls? The law of pseudo-public spaces." 394.

human dignity.[86] These rights must, therefore, be protected regardless of the identity of the violator. Companies, therefore, can be perceived as independent from the individuals who work there and can be treated as moral agents.[87] This approach dovetails with Raz's interest theory of rights, which argues that rights are interests that generate duties.[88] Although this is a moral argument "the very existence of rights is a moral/normative argument external to the law."[89]

Advocates of the moral agency approach view the human rights responsibilities of businesses as independent from the state and stemming from underlying moral grounds.[90] Accordingly, while states may impact the fulfilment of corporate human rights responsibilities (either in positive or negative ways), these responsibilities "will not be conditioned by flowing through the state."[91] Therefore, it becomes irrelevant whether the American government (or any other government for that matter) perceives social media platforms as having human rights obligations as the state interpretation would not affect the existence of moral agency. It is also irrelevant that it would be difficult to impose human rights obligations against social media platforms without cooperation from the American government as proponents of the moral agent argument distinguish between the underlying moral grounding of corporate human rights obligations and the problems that currently exist in holding companies accountable.[92]

These moral rights are also perceived by proponents of moral agency as predating and not being predicated on any specific human rights legislation, thus rendering it irrelevant that this or that social media companies did not ratify any such documents.[93]

---

[86] Thomas terms this the "normative mandate." Thomas, *Public rights, private relations*, 26, 41, 112.
[87] Obara, "What Does This Mean?," 253-54.
[88] Joseph Raz, *The morality of freedom* (Oxford: Clarendon Press, 1988), 166-71.
[89] Thomas, *Public rights, private relations*, 51.
[90] Arnold, "Corporations and Human Rights Obligations." 261-262, Brenkert, "Business Ethics and Human Rights." 288.
[91] Brenkert, "Business Ethics and Human Rights," 290.
[92] That is considered merely an issue of enforcement and divorced from the normative question. Peter Muchlinski, *Multinational enterprises and the law*, 2nd ed., Oxford international law library, (Oxford: Oxford University Press, 2007), 507-36.
[93] See, for example, Sorrell's argument that human rights obligations "are a sub-class of moral obligations" which exist regardless of any enforcement structures. Tom Sorrell, "Business and human rights," in *Human rights and the moral responsibilities of corporate and public sector organisations*, ed. Tom Campbell and Seumas Miller (Dordrecht: Springer, 2005), 134.

Instead, the state-centric view of human rights obligations is seen as a reflection of the historic period in which human rights theories developed, rather than any specific limitation inherent in conceptions of human rights.[94] Raz exemplifies this argument when he explains that "'there is no closed list of duties which correspond to the right. . . . A change of circumstances may lead to the creation of new duties based on the old right."[95] One must therefore set aside specific labels of "state actors" and "non-state actors" and simply consider whether they have a moral obligation to protect human rights.

2.4.2: Why social media companies are moral agents

There are a number of different reasons (which should be considered distinct but complementary) why businesses could be considered moral agents. The first argument is that companies have powerful impacts on employees, customers, the community, and the environment and that agents who have impacts on others usually also have moral responsibilities for those impacts.[96] Sorrell refers to these impacts as the "moral risks associated with one's commercial and other activities" and argues that it is appropriate to justify corporate human rights responsibilities on the basis of these risks.[97] Social media platforms can affect the human rights of their users and the general public in a myriad of ways. Examples include the restriction of content that would ordinarily be considered protected speech, the widescale violations of privacy that have become headline news in the wake of Snowden and Cambridge Analytica, and certain failures to act when users become targeted for campaigns of harassment. Social media companies offer a digital experience that has fundamentally altered human society in a comparatively short period of time. These technologies have become an accepted part of life and have become the norm in many Western societies but also in many developing countries as well. It seems difficult, therefore,

[94] Surya Deva, "Human Rights Obligations of Business: Reimagining the Treaty Business" (presented at the Workshop on Human Rights and Transnational Corporations: Paving the Way for a Legally Binding Instrument, Geneva, March 11-12), 5.
[95] Raz was not explicitly discussing businesses and human rights in this section but it is still relevant to this discussion as he was explaining the flexible nature of these obligations. Raz, *The morality of freedom*, 171.
[96] The exception being when the agent has no decision-making ability and therefore should not bear moral responsibility for its impacts. One could consider how moral responsibility would not apply to a tsunami that devastates a town but would apply to a company that destroys a town because of the negligent handling of toxic chemicals. Brenkert, "Business Ethics and Human Rights," 288.
[97] Sorrell, "Business and human rights," 139.

when considering the scale of impact not to conclude that social media companies are morally responsible for the way people's human rights are affected on their platforms.

A further argument for businesses as moral agents stems from Santoro's "fair share theory" which argues that "human rights are moral rights of such importance" that they give rise to a collective responsibility to help fulfil them, especially in situations where states are failing to meet these requirements.[98] During the 2011 Arab Spring, optimism about social media was at an all-time high as activists used these platforms to spread their message, document state abuse, and organise protests against governments that routinely violated human rights. Indeed, Peter Beaumont, a writer for *The Guardian*, reflected after the Arab Spring, 'The barricades today do not bristle with bayonets and rifles, but with phones'.[99] The Arab Spring was a clear example of how a company can offer the public opportunities to exercise their rights regardless of whether their government is unwilling or unable to fulfil their responsibilities. Whilst social media companies have many negative impacts on human rights, it is important to understand that platforms can also help people to fulfil their rights and they must acknowledge and accept that responsibility. The fair share theory does have an element of controversy as there might be concerns that countries would "pass the buck" and argue that companies should have the primary responsibility for fulfilling human rights. This interpretation distorts the argument for corporate human rights obligations: transforming the contention that businesses must align their corporate activities with human rights principles to the radical assertion that businesses must now become replacement governments and care for all the citizenry of a particular jurisdiction. One should not assume that corporations would somehow supersede states in human rights governance when at most corporations would only complement state protections. Arnold criticises this argument for being resistant to a multi-actor system of human rights and for "arguing for a Westphalian model in a post-Westphalian era." [100] Ultimately, the

[98] Michael A. Santoro, "Post-Westphalia and Its Discontents: Business, Globalization, and Human Rights in Political and Moral Perspective," *Business Ethics Quarterly* 20, no. 2 (2010): 290-91, https://doi.org/10.5840/beq201020221.
[99] Peter Beaumont, "The truth about Twitter, Facebook, and the uprisings in the Arab world," The Guardian, last modified February 25, 2011, http://www.theguardian.com/world/2011/feb/25/twitter-facebook-uprisings-arab-libya.
[100] Arnold, "Corporations and Human Rights Obligations," 274-75.

acknowledgement of additional moral agents does not diminish the moral responsibility of the state. In many situations, states would be directly responsible for human rights violations (such as the arrests and torture of protesters during the Arab Spring) and the existence of a non-state actor that can help the public exercise their rights would be a separate issue.

A third factor in determining moral agency is usually referred to as the "power perspective" but could just as easily be referred to as the globalisation argument. The power perspective "stresses that such is the reach and influence of modern companies, particularly multinationals, state governments are unable or unwilling to control business activity."[101] In other words, "while business has gone global there is not yet a global society able to exercise effective control over transnational enterprises." [102] Wettstein also contends that the capacity of states to govern human rights issues is "shrinking" in the face of globalisation and that consequently, the human rights obligations of businesses have become more important. [103] This is sometimes referred to as the "third agency problem" where the principal (society) can no longer control its agent (a business).[104]

Since this diminishment of states has occurred at the same time as the emerging ascendency of multinational corporations it may be argued that such companies have a moral duty to involve themselves in human rights issues. [105] Indeed, the invocation of globalisation is a frequent theme in any discussion of corporate human rights obligations. For example, the introduction to the UN Protect, Respect, and Remedy Framework, devised by Professor John Ruggie in his capacity as UN Special Representative for Business and Human Rights, states that:

> "The root cause of the business and human rights predicament today lies in the governance gaps created by globalisation—between the scope and impact of economic forces and actors, and the capacity of societies to manage their adverse consequences. These governance gaps provide the permissive environment for

[101] Obara, "What Does This Mean?," 253-54.
[102] Dalia Palombo, "The Future of the Corporation: The Avenues for Legal Change," *Future of the Corporation Working Paper* (2019). 5.
[103] Wettstein, *Multinational corporations and global justice*, 164.
[104] Palombo, "The Future of the Corporation: The Avenues for Legal Change." 4.
[105] Obara, "What Does This Mean?," 253-54.

> wrongful acts by companies of all kinds without adequate sanctioning or reparation."[106]

While globalisation has intensified the issues in business and human rights, this statement seems like an over-simplification. A corporation (or even a small business) can cause serious human rights issues even in a domestic context. Whilst it may be easier to address these problems if they are bounded by national laws, one would still need to marshal the necessary political will and resources to draft and enforce laws on corporate human rights obligations. Social media companies have caused a lot of problems in America, complications that will be explored in this thesis and difficulties that the current laws seem unable to address. Therefore, it seems over-simplistic to argue that globalisation is the "root cause" of business and human rights violations but is likely rather an aggravating factor of problems that have always existed. The social media companies that are headquartered in America have, in fact, caused numerous problems domestically even without the exacerbating influence of borderless activities. That being said, the majority of the social media companies discussed in this thesis are global companies and their unique ability to be ubiquitous while maintaining very few physical headquarters in countries does make them a challenge to regulate.

Finally, one could point to other forms of moral responsibility that companies hold (such as environmental standards, health regulations, or prohibitions on slave labour) and argue that as companies are perfectly capable of bearing those responsibilities then there is no logical reason why they should not be able to manage human rights obligations.[107] Social media companies are already subject to regulations on data protection and play an active role in combatting particularly egregious types of content such as Child Sexual Abuse Material (CSAM). Human rights obligations do not appear fundamentally different from

---

[106] John Gerard Ruggie, *Protect, Respect and Remedy: a Framework for Business and Human Rights*, Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises (A/HRC/8/5), (Geneva: United Nations, 2008), 3.
[107] The examples provided in the brackets are my suggestions. Brenkert, "Business Ethics and Human Rights," 288.

other forms of moral responsibility that social media companies possess and which are reflected in commensurate legal obligations.[108]

In conclusion, this section has examined the moral agency argument and concluded that social media networks are moral agents that have human rights obligations. This contention was based on the impacts of social media companies, their ability to shoulder a "fair share" of the responsibility of facilitating the exercise of rights, and the fact that platforms already have other forms of moral responsibility. Ultimately, the moral agency argument is very compelling but it can only be a point of inspiration as once this moral responsibility has been identified, one would need to translate those responsibilities into legal obligations and enforce them accordingly. The next section focusses more on a theory that is focused on exactly those objectives.

2.5: Business and Human Rights: holding platforms accountable

2.5.1: Introduction to BHR

Even though there are compelling arguments to be made for imposing human rights obligations on social media companies because of their status as moral agents or quasi-public bodies, the simplest approach is offered by way of the new field of Business and Human Rights (BHR). BHR represents a pivot away from more abstract theories of corporate human rights obligations towards a more applied approach. It focusses on the creation and enforcement of legal obligations on businesses with little discussion of the under-pinning ethical or moral justifications for such actions.[109] While it lacks some of the intellectual richness of more established schools of thought, there is an appealing simplicity in a BHR approach, which agitates for the creation of "facts on the ground." If moral agency is the

[108] Admittedly, human rights issue do necessitate rights-balancing, which can be more difficult than more straightforward obligations but platforms often have to juggle a variety of different imperatives and can be incentivised to do so through legal obligations.
[109] Ramasastry, "Bridging the Gap," 240.

product of natural rights theories then BHR represents an approach firmly grounded in legal positivism.

The centrepiece of BHR studies is accountability and, therefore, the quest for binding law and access to remedies.[110] It is different from earlier theories of corporate human rights obligations in a number of ways. First, it emphasises regulation as opposed to the corporate voluntarism that characterises orthodox approaches. After all, "business typically dislikes binding regulations until it sees their necessity or inevitability."[111] BHR also acknowledges that state involvement is essential in creating and enforcing corporate human rights obligations.[112] Palombo, for example, argues that states (at least in Europe) already have duties to protect people from human rights violations by private parties and duties to progressively prevent such abuses.[113] BHR is distinct from the quasi-state approach because it envisions human rights obligations as applying to *all* businesses instead of just a select range of companies that have state-like features. Instead of adopting a case-by-case, specialised approach, the BHR school of thought is a broad church, where human rights obligations are universally applied. Finally, BHR appears to be the product of all the failures of earlier theories, which have been unable to curb serious corporate human rights violations. It argues that "human rights are non-negotiable" for businesses and that "compliance with human rights should be a pre-condition for having the privilege to conduct business in society."[114] It is a more hard-line approach at a time when tolerance for human rights abuses by non-state actors is decreasing. A similarly pragmatic approach is evolving in international law, where "long-standing doctrinal arguments over whether corporations could be 'subjects' of international law… are yielding to new realities. Corporations

[110] Ramasastry, "Bridging the Gap," 238.
[111] John Gerard Ruggie, "Business and Human Rights: The Evolving International Agenda," *American Journal of International Law* 101, no. 4 (2007): 822, https://doi.org/10.1017/S0002930000037738.
[112] Ramasastry, "Bridging the Gap," 237, 47.
[113] Palombo uses these duties (to protect and to fulfil) as the basis for outlining various legal approaches to enforcing these rights through the domestic court system. Palombo, *Business and human rights*. 40.
[114] Deva, "Human Rights Obligations of Business."

increasingly are recognised as 'participants' at the international level, with the capacity to bear some rights and duties under international law."[115]

In some ways, BHR's emphasis on accountability can be seen as a reaction to the UN Guiding Principles, (UNGP's) which focussed more on consensus. As discussed earlier (at 2.2.1), Ruggie perceived his work on the UNGP's as a "principled form of pragmatism" which he saw as combining human rights principles with a pragmatic approach to securing widespread support and adoption of his principles.[116] This pragmatism has been criticised by BHR scholars, who argue that there was always pragmatism in human rights but only in relation to implementation and the UNGP's should not have compromised in identifying human rights norms.[117] The UNGP's may be perceived as the starting point for a new approach but the field of BHR now has radically different expectations from the Protect, Respect, and Remedy framework devised by Ruggie. The language of the UNGP's is a particular issue, with critics pointing out that using words like "responsibility" and "impact" rather than "duty" and "violation" diminishes the development of the "legal constitutionalisation of corporate human rights obligations."[118] BHR academics were highly critical of the convoluted reasoning in the Guiding Principles, arguing that if states are required to ensure businesses comply with human rights obligations (as part of their duty to protect) then this must mean that businesses "are themselves obligated to comply with such requirements. Indeed, if the third parties were not bound by international law to comply

---

[115] *Business and Human Rights: Mapping International Standards of Responsibility and Accountability for Corporate Acts. Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises (A/HRC/4/035)* (Geneva: United Nations, 2007), 7-8.
[116] *Interim Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises (E/CN.4/2006/97)*, at 81.
[117] Bilchitz and Deva give the example of progressive realisation in the ICESCR as an example of being pragmatic about implementation. David Bilchitz and Surya Deva, "Human rights obligations of business: a critical framework for the future," in *Human Rights Obligations of Businesses: Beyond the Corporate Responsibility to Respect?*, ed. Surya Deva and David Bilchitz (Cambridge, UK: Cambridge University Press, 2013), 12.
[118] Surya Deva, "Treating human rights lightly: a critique of the consensus rhetoric and the language employed by the Guiding Principles," in *Human Rights Obligations of Businesses: Beyond the Corporate Responsibility to Respect?*, ed. Surya Deva and David Bilchitz (Cambridge, UK: Cambridge University Press, 2013), 80.

with such requirements, then there would be no reason for the state to ensure that they do so."[119]

2.5.2: Applying BHR to social media

A BHR approach to imposing human rights obligations on platforms would focus on creating binding laws (whether at the national, regional, or international level) and effective enforcement structures. These laws would embody two objectives: the "purpose objective" (businesses should try to create "profitable solutions" to societal and global problems) and the "do no harm" objective (businesses should refrain from creating or exacerbating such problems).[120] These objectives could be seen as mirroring the positive and negative types of human rights. Compliance with these regulations would be a "precondition to doing business"[121] for social media platforms. BHR is a developing field and focusses on combining the philosophical and the applied, "looking back and forth between the two and treating them as mutually supportive, in a kind of reflective equilibrium."[122] This dual approach would fit well with regulating social media as one needs to be able to anticipate new developments in technology (an applied concern) while also articulating normative principles that would be relevant regardless of the specific nature of these platforms (the theoretical element). This is one of the challenges in regulating platforms which will be discussed in Chapters Six and Seven.

BHR will also likely influence discussions of holding social media companies accountable for human rights violations because the field is developing at the same time as discussions about regulating platforms proliferate. Focussing on accountability and enforcement are also essential when effectively regulating human rights on social media platforms as their transnational character can make them difficult to regulate. BHR is eminently practical and focusses on what Deva terms the "three-fold challenge" (the why, what, and how) in corporate human rights obligations.[123] This thesis will attempt to answer

[119] David Bilchitz, "The Necessity for a Business and Human Rights Treaty," *Business and Human Rights Journal* 1, no. 2 (2016): 208, https://doi.org/10.1017/bhj.2016.13.
[120] Palombo, "The Future of the Corporation: The Avenues for Legal Change." 6.
[121] Deva, "Treating human rights lightly," 101.
[122] Thomas, *Public rights, private relations*, 51.
[123] Bilchitz and Deva, "The human rights obligations of business," 1.

all of those questions in relation to social media companies.[124] In particular, it will revisit the work of BHR scholars and methods of accountability in Chapter Seven. BHR, therefore, while new and relatively underdeveloped offers an ambitious but workable set of expectations for companies that has inspired how this thesis explores issues in social media content moderation.

2.6: Conclusion

This chapter has explored different theories of how human rights obligations could be applied to businesses, and social media companies in particular. It should be noted that even though this chapter compartmentalised the theoretical approaches to the human rights obligations of businesses, it is entirely possible to approach this issue holistically. Indeed, the inevitable conclusion must be that "there are multiple, compelling and overlapping justifications of corporate human rights obligations."[125] Therefore, one must conclude that human rights obligations should be applied to social media companies because they derive privileges from their association with human rights, govern their platforms in a manner reminiscent of a state or quasi-state, meet all the criteria of a moral agent, and are unlikely to respect human rights law unless they are held legally accountable.

Imposing human rights obligations on platforms is essential to ensure that we can help regulate their impacts, impacts which will be explored in the next three chapters. This belief is supported by the UN Special Rapporteur on Freedom of Expression, who stated in 2015 that the role of private actors is one of the most pressing human right issues in the digital age.[126] Platforms facilitate an exchange of information, the creation of enterprise, charitable funding campaigns, safety notifications after major incidents, and many other activities that lead to a flourishing of civil society. Conversely, platforms have also led to a lot of problems, whether one is considering electoral manipulation, fake news, extrajudicial

[124] This chapter has discussed the "why", the next three chapters will discuss the "what" and the two solutions chapters will consider the "how."
[125] Arnold, "Corporations and Human Rights Obligations," 255.
[126] David Kaye, "Keynote speech" (presented at the Workshop on human rights and new technologies, University of Connecticut School of Law, Hartford, CT, October 23 2015).

surveillance, terrorist recruitment, hate speech, or harassment. These impacts can be both positive and negative but their scale and potential severity justifies not only an attribution of moral responsibility but the subsequent conclusion that this responsibility should be translated into a set of legal obligations. This chapter, however, has been focussed on the justification of imposing human rights obligations on social media platforms. Chapter Seven will revisit this issue, recall the work done here, and investigate what legislation for corporate human rights responsibilities will entail. First, however, the content moderation process itself must be investigated, to identify what issues exist at the Creation, Enforcement, and Response stages which were earlier described.

Chapter Three: Creation

3.1: Introduction

The first step in the content moderation process is the creation of the rules that govern what content is and is not permissible on the platform. These rules comprise part of the terms and conditions of the platform even though they are often given softer, less formal titles such as "community guidelines" or "community standards."[127] These rules are an example of self-regulatory codes of conduct, which are adopted for a number of reasons: to build trust among users, avoid liability, protect users, raise the public image of an industry, and to prevent command-and-control regulation.[128] Terms and conditions are, therefore, a demonstration of a platform's social responsibility, an attempt to regulate the space for an improved user experience. This is also seen as an aspect of the "success of private ordering in the online environment" as platforms are able to articulate and enforce rules in a way that is not technically feasible for public regulators dealing with that volume of content.[129] It is imperative that we scrutinise how this private ordering occurs, and whether the rules being created embody rule of law principles (as a set of standards reflecting good regulatory practice) and human rights protections.

This chapter will focus more on rule of law ideas than human rights principles (although there will be still be some discussion of rights) as rule of law scholars offer a wealth of useful criteria for judging the substantive content of regulations and identifying the procedural issues in how they are applied. The distinction between human rights and rule of law may be illusory however, as Bingham argues that today, a "thick" definition of the rule of law must necessarily include protecting human rights as one of its criteria.[130] Both rule of law and human rights principle are grounded on the preservation of human dignity (see 3.5)

---

[127] These titles also imply that the community of users assisted in their drafting, when of course most social media platforms operate a command-and-control approach to their terms and conditions.
[128] Damian Tambini, Danilo Leonardi, and Christopher T. Marsden, *Codifying cyberspace: communications self-regulation in the age of internet convergence* (London: Routledge, 2008). 251.
[129] Luca Belli and Jamila Venturini, "Private ordering and the rise of terms of service as cyber-regulation," *Internet Policy Review* 5, no. 4 (2016), https://doi.org/10.14763/2016.4.441. 2.
[130] Tom H. Bingham, *The rule of law* (London: Penguin, 2010).66-67.

and this chapter will detail how the regulations made by platforms can either strengthen or undermine that principle.

These terms and conditions are illuminating as they offer insight into the raison d'êtres of these platforms, how they view their role in society, and the values through which they structure their world. These rules are important both in shaping our online experience (and increasingly our offline behaviour) and articulating the perspective of the platform. When a platform creates content guidelines, they are drafting a constitution to govern their platform, and good rules (rules that respect human rights and embody rule of law principles) can provide a solid foundation for the content moderation process. Conversely, this chapter will show how rules that lack certainty, transparency, and fail to address the values encoded in their guidelines undermine rule of law principles and exacerbate issues in the content moderation process.

3.2: Creating terms and conditions

If one takes an orthodox approach to social media companies (see 2.2), one would likely query whether the rules created by a company to govern expressive activity on their platform should even be a subject of a discussion of this sort. They might argue that this is merely a private body creating a private regulatory model in a private space. The necessary conclusion of that argument is that social media companies are entitled to create any rule they like for their platform so long as they comply with existing legal obligations concerning illegal content. As we have already discussed (at 2.2), the problem with this view is that it disregards the power that social media companies have in society today, with the rules they create having an immediate impact on visibility, attention, and opportunities both online and offline. This echoes Weimer's contention that private standard-setters are often involved in the "authoritative allocation of things of value" [131] In particular, the controversy of misinformation and electoral manipulation that engulfed Facebook demonstrates that these

[131] David L. Weimer, "Puzzle of Private Rulemaking: Expertise, Flexibility, and Blame Avoidance in US Regulation," *Public Administration Review* 66, no. 4 (2006): 575, https://doi.org/10.1111/j.1540-6210.2006.00617.x.

private spaces can interfere with public accountability and decision-making as well as individual human rights. This thesis has already argued that platforms should have responsibilities to the public. Creating terms and conditions that comply with rule of law principles should be one of those obligations.

The first thing to understand about these terms and conditions is that have a discursive power on how users experience these platforms.[132] The system of developing terms and conditions and deciding what content should be prohibited on a platform contributes to larger processes of how we view the world and what we perceive as the norm. Platforms, therefore, are "of particular importance in the production of culture and meaning."[133] It would be erroneous to accept the narrative that social media platforms merely allow us to share our everyday lives or are just a natural evolution of other technologies.[134] Rather, as Van Dijck explains "A platform is a mediator rather than an intermediary: it shapes the performance of social acts instead of merely facilitating them."[135] This section will explore how terms and conditions subtly shape and structure perspectives and value-judgements about what is appropriate behaviour online and what actions or affiliations are worthy of condemnation.

Social media platforms are not value-neutral and any attempt by platforms to appear otherwise is an attempt to shift responsibility for the difficult decisions they must make.[136] The theory of media ecology examines how media technology acts as an environment that

---

[132] I am referring to discursive not in its ordinary meaning (rambling from subject to subject) but rather to the field of discursive sociology, which "focuses on the interpretive systems and practices through which members deal with behaviour." For more on this, see: Jack Blimes, *Discourse and Behaviour* (Boston: Springer, 1986), 187.

[133] Daithí Mac Síthigh, "The mass age of internet law," *Information and Communications Technology Law* 17, no. 2 (2008), https://doi.org/10.1080/13600830802204187. 83.

[134] Indeed, the mere existence of a social media platform may alter a situation (just as the fact that a researcher is watching can affect an experiment). Macdonald and Mair, for example, argue that the increase in terrorist groups beheading hostages can be explained by the fact that the act is dramatic and theatrical, and videos of beheadings are likely to be shared widely online. Therefore, the existence of a technology has actually affected how these groups conduct their activities offline as terrorist acts have always been about communication. See: Stuart Macdonald and David Mair, "Terrorism online: A new strategic environment," in *Terrorism Online: Politics, Law, and Technology*, ed. Lee Jarvis, Stuart Macdonald, and Thomas M. Chen (Padstow, UK: Routledge, 2015).

[135] van Dijck, *The culture of connectivity* 29.

[136] After all, as Gillespie concludes, "what you permit, you promote." Gillespie, *Custodians of the internet*. 153.

structures the world we see, our expectations about society, and the things we value.[137] These processes of norm-building are often "implicit and informal" as individuals believe that they are only interacting with a piece of technology instead of with something that could change their lived experience.[138] Revolutionary technologies (such as the printing press or the computer) "are not therefore simply machines which convey information. They are metaphors through which we conceptualise reality in one way or another."[139] Social media companies, therefore, can structure how we view the world and that process often starts at the terms and conditions stage. Take for example, the fact that YouTube does not consider videos depicting white people in blackface as hate speech in of itself.[140] Whether YouTube decides to explicitly prohibit these videos or not, a decision is being made about their value and that decision can affect the visibility and acceptability of that particular practice, as well as the level of discrimination tolerated on the platform. Gillespie makes this point when he explains that social media platforms are not merely "transmitting what we post," they are "constituting what we see."[141]

Revolutionary technologies such as the printing press and the television often unleash a flood of information which necessitates the creation of new information management systems to act as control mechanisms.[142] These systems delineate between what information is and is not relevant and valued in order to create order from the influx of data and as a result they become entrenched in our society. Postman gives the example of school curriculums that "categorises knowledge" and "systematically excludes" certain information as illegitimate.[143] Social media platforms also have a number of different information management systems such as the algorithms governing newsfeeds, the hashtags

---

[137] Neil Postman, "Reformed English Curriculum," in *High School 1980: The Shape of the Future in American Secondary Education*, ed. Alvin Christian Eurich (London: Pitman, 1970), 160.
[138] Postman, "Reformed English Curriculum," 160.
[139] Neil Postman, *Teaching as a conserving activity* (New York: Delacorte Press, 1979), 39.
[140] The video would have to contain other violations of their hate speech provisions. Catherine Buni and Soraya Chemaly, "Secret Rules of the Internet: The murky history of moderation, and how it's shaping the future of free speech," Verge, last modified April 13, 2016, http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.
[141] Gillespie, *Custodians of the internet*. 21.
[142] Neil Postman, *Technopoly: the surrender of culture to technology*, 1st ed. (New York: Vintage Books, 1992), 61, 65-66.
[143] Postman, *Technopoly*, 61, 63, 74.

that are used to thematically organise content, and the terms and conditions that indicate what is and is not welcome on the platform.

The user's perspective is structured at the earliest point in the content moderation process based on how content is categorised and what value is assigned to various categories. Crawford and Gillespie have identified a number of different approaches to structuring terms and conditions, all of which relate to the mission of the platform and what the platform values.[144] Flickr divides content by its perceived raciness,[145] the *New York Times* according to what a proper and constructive debate entails, while YouTube creates genres of objectionable content such as "impersonation" and "nudity."[146] From the moment that these categories are created, the perceptions of users are affected in such an implicit way that it escapes notice. Langdon Winner argues, that "the same careful attention one would give to the rules, roles, and relationships of politics must also be given to such things as the building of highways, the creation of television networks, and the tailoring of seemingly insignificant features on new machines."[147] The artefacts of everyday life are imbued with hidden politics, an notion that clearly applies to the subtle values and choices encoded into content moderation.

The content that is categorised and prohibited in the platform's terms and conditions will directly reflect the reality that social media creators are attempting to build. Pinterest prioritises a safe and positive environment, the head of the policy team explaining "we help people discover and do what they love by showing them ideas that are relevant, interesting, and personal. For people to feel confident and encouraged to explore new possibilities, or try new things on Pinterest, it's important that the Pinterest platform continues to prioritise

---

[144] Kate Crawford and Tarleton Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint," *New Media and Society* 18, no. 3 (2014): 419, https://doi.org/10.1177/1461444814543163.
[145] "All content on Flickr, public and private, has to be appropriately moderated as "safe", "moderate", or "restricted" using our safety and content filters. If your judgment proves to be poor, we'll moderate your account to match appropriate categorisation for Safe Search and/or content type and send you a warning." "Flickr Community Guidelines," Flickr, accessed February 12, 2018, https://www.flickr.com/help/guidelines.
[146] "YouTube Community Guidelines," YouTube, accessed February 12, 2018, https://www.youtube.com/yt/about/policies/#community-guidelines.
[147] Langdon Winner, *The whale and the reactor: a search for limits in an age of high technology* (Chicago: University of Chicago Press, 1986), 29.

an environment of safety and security."[148] This approach could be contrasted with Twitter's initial mission of being "the free speech wing of the free speech party"[149] where there was a strong presumption against content being removed. This helps illuminate how technology "weighs in on the side of one vested interest over others."[150] Some of the most interesting transformations occur when a social media company begins to shift its objectives, thereby attempting to change the culture on its platform, such as Twitter's attempts to distance itself from harassment complaints and YouTube's prioritisation of citizen journalism as videos of violent but significant events began to appear on their platform. Through the creation and revision of the terms and conditions, a platform is elucidating their mission and influencing the lived experience of their users. The culture on the platform is always changing, because culture in general "is not a fixed collection of texts and practices, but rather an emergent, historically and materially contingent process through which understandings of self and society are formed and reformed."[151]

To conclude, social media is an excellent example of how societies are impacted first by a new media technology and then again by the entrenchment of an information management system that changes perspectives and re-allocates value. Pfaffenberger, for example, argues that the new interpretations and societal restructuring that occurs is not a natural by-product of the technological innovation but rather instigated by the "design constituency" (the people and companies introducing this invention into society) in order to encourage society to embrace the innovation (a process he calls "regularisation").[152] This

---

[148] Adelin Cai, "Putting Pinners First: How Pinterest is Building Partnerships for Compassionate Content Moderation," Tech Dirt, last modified February 5, 2018, https://www.techdirt.com/articles/20180205/10143639158/putting-pinners-first-how-pinterest-is-building-partnerships-compassionate-content-moderation.shtml.

[149] Josh Halliday, "Twitter's Tony Wang: 'We are the free speech wing of the free speech party," The Guardian, last modified March 22, 2012, https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

[150] Helen Nissenbaum, "From Preemption to Circumvention: If Technology Regulates, Why Do We Need Regulation (and Vice Versa)?," *Berkeley Technology Law Journal* 26, no. 3 (2011): 1375.

[151] Julie E. Cohen, *Configuring the networked self: law, code, and the play of everyday practice* (New Haven: Yale University Press, 2012), 17.

[152] Bryan Pfaffenberger, "Technological Dramas," *Science, Technology, and Human Values* 17, no. 3 (1992): 291, https://doi.org/10.1177/016224399201700302. It should be noted this is actually the first stage in Pfaffenberger's theory of technological dramas, which traces how a new technology is introduced, interpreted, establishes a dominant paradigm, and finally becomes a neutral and unquestioned artefact in society.

restructuring dynamic is also captured by Morozov when he argues that "Twitter too is an engine, not a camera; it doesn't just reflect realities—it actively creates them."[153] It is important to remember how much is negotiable when exploring the various issues with terms and conditions discussed throughout this chapter. What seems to be natural or inevitable on social media platforms is only a perception of the current ecology of the platform and can be changed when there are enough interested parties.

3.3: Certainty

3.3.1: Issues with Clarity

From the very beginning, social media platforms have treated terms and conditions as ancillary to the real task of regulating content on their platforms. For example, Twitter was created in 2006 but it wasn't until 2009 that the company created rules about the content permitted on the platform.[154] YouTube spent its first year of existence with nothing more than a one page bullet-pointed list for its moderators and assessors who found themselves frequently asking each other "can I share this video with my family?" as the litmus test for regulation.[155] This de-prioritisation of good regulation means that terms and conditions were (and often continue to be) written very broadly and use language that make the scope of the terms difficult to identify. Users are forced to accept these terms and conditions, as they know that there is no opportunity to renegotiate or modify these terms.[156]

This section will explore how platforms fail to provide users certainty by drafting rules that are vague and difficult to interpret. Rules that are clear to everyone are the foundation of a coherent system of regulation, "one of the essential ingredients of legality."[157] This is Bingham's first principle of the rule of law, that "the law must be accessible, and so

---

[153] Evgeny Morozov, *To save everything, click here: the folly of technological solutionism*, 1st ed. (New York: PublicAffairs, 2013), 150.
[154] Biz Stone, "The Zen of Twitter Support," Twitter Blog, last modified January 15, 2009, https://blog.twitter.com/2009/the-zen-of-twitter-support.
[155] Stone, "The Zen of Twitter Support."
[156] Pasquale, *Black box society*, 144.
[157] Fuller, *The morality of law*, 63.

far as possible intelligible, clear and predictable."[158] Unfortunately, social media platforms are not instituting such laws, choosing instead to create terms and conditions that are so vague that a "Kafkaesque uncertainty" emerges online.[159] If the goal of regulation is to bring about a change in behaviour then the objective of the terms and conditions created by the platforms should be to help users understand what behaviour they should avoid on the platform, not merely providing a basis for content removal. Clarity then, becomes a tool to provide certainty to users. Even if one believes that platforms are entitled to create any substantive rule about content they wish, there must be some expectation that these rules comply with the most basic requirements of the rule of law, and none are so basic as clarity.

Many platforms tend to rely on simple prohibitions that refrain from providing clear definitions to help users make distinctions between what would and would not be acceptable. This is a certainty issue because if there is a penalty for noncompliance (in this case, an account suspension, termination, deletion of content, or demonetisation) then "we ought to be able, without undue difficulty to find out what we must or must not do."[160] For example, YouTube prohibits content it terms as "sexualisation of minors" which includes videos "featuring minors involved in provocative, sexual, or sexually suggestive activities."[161] There is no further explanation of what such a broad statement entails and only a few minutes of searching on YouTube was able to produce three videos that seemed to plausibly entail the sexualisation of minors. The first clip was from a Korean talent TV show and featured a boy and girl who could not be more than eight-years-old doing a sexy dance routine on stage complete with grinding and twerking. The video was even titled "Sexy Kid-Another Troublemaker."[162] The second clip is an infamous scene from the American reality show 'Toddlers and Tiaras' (about child beauty pageants) where a four-year old girl performs a routine dressed as Julia Roberts' sex worker character from 'Pretty Woman'

[158] Bingham, *The rule of law*, 37.
[159] Morozov, *The net delusion*, 102.
[160] Bingham, *The rule of law*. 37.
[161] YouTube Policy Centre, "Child Safety on YouTube," YouTube, accessed April 8, 2020, https://support.google.com/youtube/answer/2801999?hl=en.
[162] Kpop Focus, "Sexy Kid-Another Troublemaker," YouTube, last modified May 5, 2017, https://www.youtube.com/watch?v=vV4r5PV4I2c.

(complete with thigh-high boots).[163] The final clip is from the TV show 'Keeping up with the Kardashians' which features an underage Kylie and Kendall Jenner cavorting on a stripper pole and pretending to be on "Girls Gone Wild."[164] All of these clips could fall into the vague definition that was quoted above, terms and conditions that "are so over-broad and general as to allow almost any kind of regulation" that a social media platform may choose to pursue.[165] In the meantime, YouTube is perceived as condemning child sexualisation in principle without creating practical guidelines that provide certainty for the users who access the platform and protection of a child's right not to be exploited or to view material that is "injurious to his or her mental and physical well-being."[166]

Platforms appear to lack awareness on how difficult it can be to understand and comply with their rules. YouTube, for example, suggests in its Guidelines "Don't try to look for loopholes or try to lawyer your way around the guidelines—just understand them and try to respect the spirit in which they were created."[167] This statement is strange because the YouTube guidelines, like other social media terms and conditions, tend towards the general and therefore, the concept of loopholes seems wholly inapplicable. In addition, the demonization of law is especially frustrating as there is a tendency by social media policy-makers[168] and overly-optimistic academics[169] to compare these terms and conditions to laws (replete with the same procedural assurances) without considering what

---

[163] TLC, "Pretty Woman Toddler: Toddlers & Tiaras," YouTube, last modified September 8, 2011, https://www.youtube.com/watch?v=QAxEt5YL8w4.
[164] April and Fack, "Kendall and Kylie young funny video," Youtube, last modified May 5, 2018, https://www.youtube.com/watch?v=kzFmb5wjM38.
[165] Ben Wagner, "Governing Internet Expression: How Public and Private Regulation Shape Expression Governance," *Journal of Information Technology and Politics* 10, no. 4 (2013): 398, https://doi.org/10.1080/19331681.2013.799051.
[166] Article 17e (well-being), Article 19 (exploitation), and Article 34 (sexual exploitation). *Convention on the Rights of the Child*.
[167] "Policies and Safety," YouTube, accessed April 9 2020, https://www.youtube.com/about/policies/#community-guidelines.
[168] Monika Bickert, for example, calls the meetings where they discuss policy changes "mini-legislative sessions." See: David Talbot and Nikki Bourassa, "How Facebook Tries to Regulate Postings Made by Two Billion People," Medium, last modified October 19, 2017, https://medium.com/berkman-klein-center/how-facebook-tries-to-regulate-postings-made-by-two-billion-people-bca9408b6b4b.
[169] Marvin Ammori claims that "The terms of these policies often take the form of traditional legal rules and standards…they have just as much validity" in M. Ammori, "The 'new' New york times: Free speech lawyering in the age of google and twitter," *Harvard Law Review* 127 (2014): 2263. Kate Klonick claims that "procedurally, platform content-moderation systems have many similarities to a legal system." Klonick, "New Governors," 1664.

responsibilities and values they should also be protecting. This labelling confers a recognition of legitimacy in rule-making that the platforms have not yet achieved in practice. Even rules that try to provide an explanation can be difficult to understand. An example of this is TikTok's prohibition on derogatory slurs, where it outlines the prohibition but then states "However, we are aware of the fact that slurs can be used self-referentially or have been reappropriated, and we may give exceptions when slurs are used in a song or in other instances of a self-referential satirical context and/or reappropriation."[170] Parsing whether these terms are being used in their original context, appropriated, or being "reappropriated" seems like an incredibly complex task (one that would likely entail complicated assessments of whether the speaker of the slur is part of the protected group or not) and it is disappointing that there is no further information or examples provided. While platforms have discretion in what they permit and prohibit (within the confines of the law) they still have an ethical obligation to their users to regulate their space in a way that provides them certainty and accountability. This can also be a pragmatic decision as a failure to provide certainty to users may disincentivise them from using a particular platform, especially if they are content creators. Users are not going to expend time and effort to create content and build an online profile on a platform where they are chronically uncertain about the scope of the rules. This reflects Bingham's explanation that clarity of regulations is important because it facilitates "the successful conduct of trade, investment and business generally."[171]

Accessible language is an important aspect of certainty. Pinterest summarises in laymen's terms every section of their terms and conditions so that users can quickly grasp the meaning of the agreement. For example, after an extended paragraph absolving Pinterest of liability for third-party content, there is a sentence in a different colour that reads: "More simply put: Pins link to content off of Pinterest. Most of that stuff is awesome but we're not responsible when it's not."[172] While Pinterest should be applauded for their proactive and transparent policies, one might cynically ask if moderating content on Pinterest could be compared to playing a video game on "easy mode" as the platform is most famous for boards

[170] "TikTok Community Guidelines," last modified 21 August 2019, https://www.tiktok.com/community-guidelines?lang=en.
[171] Bingham, *The rule of law*. 38.
[172] "Terms of Service," Pinterest, accessed April 8, 2020, https://policy.pinterest.com/en/terms-of-service.

devoted to planning weddings, recipes, craft projects, interior design, and aspirational "dream boards." While the platform is forced to deal with some of the same issues that affect all social media sites,[173] they have largely escaped the high-profile scandals that have embroiled other platforms in controversy, such as fake news, election manipulation, extremist activity, and the harassment of women and minorities.

Another common problem is that platform terms and conditions are replete with language that renders the scope of their limitations uncertain and imprecise. This makes it harder for users to know if they are complying with the rules, a determination that might only become clear when content is flagged and/or removed. Since many platforms have underdeveloped appeals mechanisms (which will be discussed in Chapter Five) it is more pressing that they create clear regulations that provide users with a measure of certainty. This can be compared to traditional principles of administrative law, which have "long presumed that rulemaking procedures help protect against arbitrariness when due process does not apply."[174] The Twitter Rules are a classic example of the vagaries of online terms and conditions. While the Rules are written in accessible language and include rationales for each section, they still include statements such as "Some examples of encouraging or promoting self-harm include (but may not be limited to) encouraging or glorifying…"[175] Not only is there no definition of "encouraging or glorifying" but there is also the qualifier "include (but may not be limited to)" which undermines any certainty that an itemised list of examples could provide. Another example is OnlyFans' (a British platform for monetised profiles) prohibition on content or behaviour that "causes annoyance, inconvenience, or needless anxiety or is likely to upset, embarrass, alarm, or annoy any other person." [176] While this prohibition is incredibly broad, the term that seems particularly mystifying is "needless anxiety" as it

---

[173] Of particular concern on Pinterest is images of self-harm and so-called "thinspiration" images that feature emaciated people, which the site actively removes and re-directs related searches to a message explaining their concerns. Carolyn Gregoire, "Pinterest removes eating disorder-related content, pro-anorexia community continues to thrive," Huffington Post, last modified October 8, 2012, http://www.huffingtonpost.co.uk/entry/pinterest-removes-eating-disorder-content_n_1747279.
[174] Danielle Keats Citron, "Technological Due Process," *Washington University Law Review* 85, no. 6 (2008): 1288.
[175] "Twitter Rules: Glorifying Self-Harm and Suicide," Twitter, accessed April 12, 2020, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.
[176] "Only Fans Terms of Service," last modified August 21, 2019, https://onlyfans.com/terms. Section 9.36

would be very difficult for any user to understand whether their actions were causing an unnecessary amount of discomfort.

3.3.2: Why platforms benefit from a lack of certainty

The vagueness of the terms and the conditions serves a particular purpose: it gives platforms a tremendous amount of flexibility and discretion in how they moderate content. Platforms can create exceptions and alter the threshold of permissibility in a silent response to changing societal responses without publicly amending their rules or making any announcements.[177] Crawford and Gillespie argue this opaqueness allows social media companies to retain the ability to make judgements in high-profile cases "based on ad hoc and often self-interested assessments of the case at hand."[178] This flexibility can be an advantage of private standard-setters as "they adjust standards over time and can respond to issues quicker"[179] but that would be the case even if social media platforms published detailed rules and then engaged in public revisions. As platforms are only legislating for their community, they would still be able to respond much more quickly than traditional lawmakers. These features are further supplemented by unilateral modification clauses in the terms and conditions, which underscore to users that they have no ability to negotiate the current terms and conditions that they are agreeing to but they must also consent to these rules being changed without their consultation in the future, all of which makes "any meaningful user consent largely illusory."[180] Unfortunately, this elasticity is provided at the expense of providing information and certainty to social media users. It also harms the legitimacy of platforms and makes it hard for users to hold platforms accountable as they are not given enough information to understand and challenge the platform's actions. Therefore, the benefits that the current approach offers to platforms is at the expense of users and the perception of these platforms as legitimate regulators.

---

[177] Sarah Roberts, *Behind the screen: content moderation in the shadows of social media* (New Haven: Yale University Press, 2019). 95.
[178] Crawford and Gillespie, "What is a flag for?," 420.
[179] Weimer, "The Puzzle of Private Rulemaking," 575.
[180] Tambini, Leonardi, and Marsden, *Codifying cyberspace*, 120. See also: Pasquale, *Black box society*, 144.

There is a risk that the requirement of clarity can be misinterpreted as a box-ticking exercise and will actually render the laws normatively weaker. Reed argues that "laws drafted in terms of broad and open textured rules have a much stronger normative force, at least in the longer term. Their underlying aims and purposes are more easily understandable by the law subjects, even if it is not necessarily clear precisely what needs to be done to comply with the law."[181] This is a problematic argument because it would only be true in situations where users perceive the rule as being inherently legitimate and actively strive to align their behaviour with the spirit of the law. It is also likely to be most applicable in situations where the law is reflecting social norms and shared interpretations (what Fuller calls "common sense standards lived outside legislative halls"[182]) and will hold no real power when there is societal debate or conceptual uncertainty, such as social media prohibitions against hate speech, graphic violence, or sexualised content. Finally, it runs completely contrary to Bingham's assertion that if a person is to claim rights or perform obligations, "it is important to know what our rights and obligations are."[183]

In conclusion, this discussion of certainty in terms and conditions demonstrates that the very foundation that these regulatory regimes rests upon is flawed because the rules are incoherent to their users. The subjects of these rules are forced to navigate "norms whose meaning is not so obscure or contestable as to leave those who are subject to them at the mercy of official discretion."[184] The result of this structural weakness is that other good governance standards are compromised as clear rules are a precondition to many other rule of law requirements. A number of these factors will be discussed in the next two chapters, but it is important to remember that every problem explored in this thesis is exacerbated by the lack of clear, coherent rules. Social media platforms should revise their terms and conditions, eliminating unclear language and providing guidance notes to explain each provision in enough detail that users can understand what compliance looks like before they

[181] Chris Reed, "How to Make Bad Law: Lessons from Cyberspace," *Modern Law Review* 73, no. 6 (2010): 928, https://doi.org/10.1111/j.1468-2230.2010.00824.x.
[182] Fuller, *The morality of law*, 64.
[183] Bingham, *The rule of law*, 37-38.
[184] Jeremy Waldron, "Rule of law and the importance of procedure," *Nomos* 50 (2011): 3-4.

are penalised by the platform.[185] They should also base those terms and conditions on human rights values, expressing their commitment in policy.[186]

3.4: Transparency

3.4.1: Transparency in Terms and Conditions

Transparency is an important element in any plan for reforming how social media platforms approach content moderation. It facilitates all other objectives because a platform cannot be held accountable if parties outside the platform are unable to understand the actions and motivations of the company. This is a strange predicament for private rule-makers, as Weimer argues that private standard setting "generally involves an open process characterised by evolutionary adjustment"[187] and yet this process is completely closed on social media platforms. The UNGP's also make clear that companies should be transparent, because "in order to account for how they address their human rights impacts, companies should be prepared to communicate this externally."[188] Transparency will be discussed frequently throughout this thesis but this section will provide a brief overview of how companies approach transparency and how these practices specifically apply to social media terms and conditions.

Transparency is a concept that is frequently discussed by social media companies as one of their core principles. For example, in 2009, Sheryl Sandberg said of Facebook "we have really big aspirations around making the world a more open and transparent place."[189] Transparency, however, is not a reciprocal action on social media but rather, as Van Dijck

---

[185] This suggestion is also made in the Santa Clara Principles on content moderation, which states that platforms should include "examples of permissible and impermissible content and the guidelines used by reviewers." "The Santa Clara Principles: On Transparency and Accountability in Content Moderation," last modified 07 May 2018, 2018, https://santaclaraprinciples.org.
[186] Principle 16, *UN Guiding Principles.*
[187] Weimer, "The Puzzle of Private Rulemaking," 575.
[188] Principle 21, *UN Guiding Principles.*
[189] Chris Tryhorn, "Evangelical networker who wants Facebook to open up the world," The Guardian, last modified August 20, 2009, https://www.theguardian.com/business/2009/aug/20/facebook-ceo-sheryl-sandberg-interview.

argues, surprisingly one-sided.[190] Users are increasingly encouraged to share as much as possible on social media platforms, an action that not only populates the platform with original content but also provides valuable data that can be sold to third-party advertisers.[191] Meanwhile, social media companies continue to perform the proverbial dance of the seven veils, obscuring their actions in code and proprietary arguments. Farrand and Carrapico contrast the "secretive negotiation process" of social media platforms with the "overt law making" which public regulatory bodies engage in, arguing that a lack of transparency poses a fundamental legitimacy problem for these companies.[192] This lack of accessibility also forces would-be critics to rely on whatever content is made public or the unsubstantiated claims of inside sources. This challenge goes to heart of the issue of transparency; secrecy impedes reform and accountability, thus undermining good regulation.

Pasquale has written extensively on transparency in Silicon Valley and he argues that tech companies represent only the latest challenge for regulators who are faced with a product that becomes "critical to everyday life," necessitating that they "strike the fairest balance they can between public and private good."[193] But transparency is not a goal in of itself; companies can release huge amounts of information that is so complex that it is "as effective at defeating understanding as real or legal secrecy."[194] Therefore, transparency is a concept that encompasses more than just openness. There must be an element of intelligibility in order to facilitate understanding by ordinary users of how these terms and conditions work.

Transparency is an important principle for anyone seeking to understand and even reform the practices of platforms. This normative value, however, must be translated into concrete objectives that platforms can implement. Platforms should disclose information about how much content violating each specific rule has been flagged and removed. This

---

[190] van Dijck, *The culture of connectivity* 61.
[191] van Dijck, *The culture of connectivity* 61.
[192] Benjamin Farrand and Helena Carrapico, "Networked Governance and the Regulation of Expression on the Internet: The Blurring of the Role of Public and Private Actors as Content Regulators," *Journal of Information Technology and Politics* 10, no. 4 (2013): 362, https://doi.org/10.1080/19331681.2013.843920.
[193] Pasquale, *Black box society*, 91. It should be noted that Pasquale never directly cites Fuller, but their ideas are very complementary.
[194] Pasquale, *Black box society*, 8.

information should be broken down into category of content and by "locations of flaggers and impacted users (where apparent)."[195] Currently, social media platforms generally disclose the amount of content removed after government requests but not the content that violates their terms and conditions, and not broken down by country. One justification for this omission, according to Jørgensen's interviews with Facebook employees, is that they don't perceive their actions as controversial or a human rights issue when they stem from the terms and conditions as opposed to a government request.[196] This explanation is illuminating but troubling as it indicates why social media companies do not prioritise transparency; they do not see themselves as capable of violating human rights values or good governance principles. Another aspect of this omission may be that platforms do not want to be held publicly accountable or required to explain its justifications for certain moderation practices. Platforms, therefore, are able to take advantage of being a private company with a veneer of being a community space, "relying upon their privileged position as private publishers while making public assertions about communication and connecting communities."[197]

3.4.2: Case Study in Transparency: Internal Guides

Thus far, this chapter has argued that platform terms and conditions are often unclear and provide little certainty to users. One might query how the content moderators are able to apply such simple prohibitions when making decisions on content that has been flagged for review. These rules would leave moderators with a lot of discretion but instead, moderators are provided with thick content assessment manuals (internal rules) designed to cover every potential nuance in moderation. At many platforms such as Facebook and YouTube, these internal rules were created before the external, public-facing terms and conditions and are updated much more frequently.[198] This is another problematic example of how the experience of users trying to navigate the rules of a platform is not prioritised as

[195] This is important to understand how flagging and the application of rules differs based on geographical location. "The Santa Clara Principles: On Transparency and Accountability in Content Moderation." "The Santa Clara Principles: On Transparency and Accountability in Content Moderation."
[196] Jørgensen, "Framing Human Rights," 351-52.
[197] Mac Síthigh, "The mass age of internet law," 83.
[198] Klonick, "New Governors," 1648; Buni and Chemaly, "Secret Rules of the Internet."

highly as the internal command-and-control aspects of moderation.

These manuals are not typically disclosed publicly, which is why the manuals that were leaked to The *Guardian* in 2017 were so fascinating. One particularly interesting example was the content assessment manual on terrorism. The manual was highly detailed, splitting terrorist content into a huge number of sub-categories (all of which required different approaches) such as: Contemporary Activity Primary Focus, Contemporary Activity Incidental Focus, Symbols/Leaders: Primary Focus, Symbols/Leaders: Incidental Focus, Historical Activity and Historical Artefacts.[199] This detailed guide on terrorism would complement the moderator's own knowledge as it was disclosed to The *Guardian* that moderators are also provided with a 44-page document featuring the names and faces of terrorist leaders and their groups, and are expected to familiarise themselves with this information (along with the other aspects of content moderation such as the importance of context) within the first two weeks of the job.[200] All of this should be contrasted with the information that is given to Facebook users about what kind of terrorist content is prohibited. Facebook states that they don't allow "organisations or individuals involved in…terrorist activity" and that they also "remove content that expresses support or praise for groups, leaders, or individuals involved in these activities."[201] No list of organisations considered terrorist by Facebook is provided and no further definitions of what expressing "support or praise" entails can be found on the platform. The detailed information provided to moderators in no way reflects the information provided to users. This is especially important as social media becomes more universal as there is a significant amount of debate

[199] "How Facebook guides moderators on terrorist content," The Guardian, last modified May 24, 2017, https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content.
[200] Nick Hopkins, "Facebook struggles with 'mission impossible' to stop online extremism," The Guardian, last modified May 24, 2017, https://www.theguardian.com/news/2017/may/24/facebook-struggles-with-mission-impossible-to-stop-online-extremism; Olivia Solon, "To censor or sanction extreme content? Either way, Facebook can't win," The Guardian last modified May 23, 2017, https://www.theguardian.com/news/2017/may/24/facebook-struggles-with-mission-impossible-to-stop-online-extremismv.
[201] "Facebook Community Standards: Dangerous Individuals and Organisations," Facebook, accessed April 8, 2020, https://www.facebook.com/communitystandards#dangerous-organizations.

around the world on whether various organisations (such as Hamas or the Tamil Tigers for example) should be labelled as terrorists.[202]

Another example of how the external rules and the internal guidelines appear distinct is Facebook's prohibitions on hate speech against women. Facebook's external rules state that hate speech includes "profane terms or phrases with the intent to insult, including but not limited to…bitch" and "violent speech or support in written or visual form."[203] It is therefore surprising that the Facebook internal guidelines explicitly use "to snap a bitch's neck, make sure to apply all your pressure to the middle of her throat" as an example of permitted speech on the platform![204] What should be clear from this example is that it would be very difficult to predict what content is permissible on a platform based solely on the terms and conditions available to users. It also raises questions whether Facebook's rules would be in compliance with the ICCPR's requirement that civil rights be enjoyed equally by people regardless of gender.[205]

Facebook claims they have now published these internal guides[206] but it is obvious that the information they have disclosed is not the full story. Their "internal guidelines" for terrorism, for example, contain no list of terrorist groups or terrorist leaders. Instead, they read more like detailed terms and conditions, a definite improvement over what existed before on the platform but still largely devoid of the examples and contextual information to truly make these rules accessible to users. These guidelines were released in 2018 during the Cambridge Analytica scandal so it is also clear that this bid for transparency was a political decision although it failed to capture the public's attention.

---

[202] Depending on the content, platforms may also be in breach of the ICCPR's rules on war propaganda so care would have to be taken when assessing content. Article 20, *ICCPR*.
[203] "Facebook Community Standards: Hate Speech," Facebook, accessed April 8, 2020, https://www.facebook.com/communitystandards/hate_speech.
[204] Nick Hopkins, "Revealed: Facebook's internal rulebook on sex, terrorism and violence," The Guardian, last modified May 24, 2017, https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence.
[205] Article 3, *ICCPR*.
[206] Monika Bickert, "Publishing our Internal Enforcement Guidelines and Expanding our Appeals Process," Facebook Newsroom, last modified April 24, 2018, https://about.fb.com/news/2018/04/comprehensive-community-standards/.

This discussion should not be perceived as a specific criticism of Facebook as the same problems exist at all platforms. It is difficult to discuss other platform's internal guidelines, however, as they have not been leaked to the public. Therefore, the examples of Facebook's internal guidelines highlight a larger issue in the creation of terms and conditions on these platforms. While these guidelines may have intended to only expand on the publicly available terms and conditions, it is possible that their interpretations of the rules (and the concrete actions taken in response to those rules) might differ markedly from the public's understanding of these regulations. By creating two sets of guidelines (one that users can access and one that is strictly for internal use) the platforms create the risk that there will be a disparity between the rules that are available and the rules that are actually applied (a disparity between regulation and enforcement). While enforcement will be discussed in greater detail in the next chapter, it should be noted that some of the problems platforms face may not lie in the application of the rules but rather in the existence of these two distinct sets of rules. This is highly problematic because when rules are created, a tacit promise is made to the adherents that these are the rules that will be applied to their actions.[207]Applying a different set of rules is a serious breach of the rule of law and is widely condemned by Bingham.[208] This continuity allows people to ascertain prohibitions with certainty and adjust their behaviour to avoid sanction. The creation of these rules is clearly very important and platforms should ensure that they constitute good regulation and reflect the basic principles of the rule of law, such as clarity, stability, and publicity.[209] These principles are focused on the ability of citizens to comply with the law but there are also other considerations such as the procedural protections owed to subjects.[210] Unfortunately, this theme of the publicly-shared content moderation policies and processes being treated as ancillary to the actual task of regulating platforms will be reiterated throughout this thesis.

---

[207] Fuller, *The morality of law*, 40.
[208] See Bingham's first and second principles which requires that the law be predictable and that questions of law should be resolved by applying the law not using discretion. Bingham, *The rule of law*. 37, 48.
[209] These topics come from Lon Fuller's eight principles. See: Fuller, *The morality of law*.
[210] This elucidation of various aspects of the rule of law and how academics have approached them comes from an excellent article by Waldron. See: Waldron, "Rule of law." See also: Reed, "How to Make Bad Law: Lessons from Cyberspace," 917.

It might seem strange for social media platforms to refuse to share their content assessment manuals with the public but this practice offers a number of benefits to the companies. First, this secrecy means that platforms are not forced to justify to the public the strange, often arbitrary distinctions they make when assessing content. In 2012, Facebook hired an outside firm to create a content assessment manual for their content teams. Unfortunately, this document was leaked to Gawker and provoked public ridicule and derision.[211] The manual was lenient on gore and incredibly tough on female nudity, banning "female nipples" and images where the shape of a woman's genitalia was somewhat visible through a pair of pants (which is colloquially known as a "camel-toe.")[212] There has been a similar backlash against the 2017 leak of Facebook's updated content assessment manuals.[213] Of course, media derision and public debate often occur when democratic governments pass laws but social media companies seem determined to avoid this important feature of accountability. Klonick argues that this lack of accountability "lays bare our dependence on these private platforms to exercise our public rights"[214] and it is clear that platforms are not interested in encouraging a culture of justification or even providing explanations for their conduct.

Another reason platforms refuse to share their internal guidelines with users is that the companies are concerned that such information would help disseminators of questionable content game the system.[215] This idea is explained by a moderator who told a researcher "We have very, very specific itemised internal policies . . . the internal policies are not made public because then it becomes very easy to skirt them to essentially the point of breaking them."[216] Platforms may be genuinely worried about users circumventing the rules

---

[211] Jeffrey Rosen, "Delete Squad," New Republic, last modified April 29, 2013, http://www.newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules.
[212] Eva Galperin, "What the Facebook and Tumblr Controversies can teach us about content moderation," Electronic Frontier Foundation, last modified March 2, 2012, https://www.eff.org/deeplinks/2012/03/what-facebook-and-tumbler-controversies-can-teach-us-about-content-moderation.
[213] See, for example, Nick Hopkins, "How Facebook allows users to post footage of children being bullied," The Guardian, last modified May 22, 2017, https://www.theguardian.com/news/2017/may/22/how-facebook-allows-users-to-post-footage-of-children-being-bullied.
[214] Klonick, "New Governors," 1668.
[215] Buni and Chemaly, "Secret Rules of the Internet."
[216] Sarah Roberts, "Commercial Content Moderation: Digital Labourers' Dirty Work," in *Intersectional Internet: race, sex, class and culture online*, ed. Safiya Umoja Noble and Brendesha M. Tynes (New York: Peter Lang, 2015), 152. This explanation was also explored in Buni and Chemaly, "Secret Rules of the Internet."

but this does not seem to be a sufficient justification for keeping content moderation guidelines secret, especially when publicly available rules might actually decrease the amount of impermissible content on the platform. It is arguable that some flagged content is shared by users who are unaware that the content violates the terms and conditions. This approach is contrary to rule of law principles because judging people's behaviour by unpublished laws "is an affront to man's dignity as a responsible agent."[217] Because these internal guidelines would assist users in coordinating their behaviour (what Waldron calls providing "a calculable basis for running their lives or businesses"[218]) and would provide clear assurances that their conduct is permitted, it is a problematic to withhold these and create an atmosphere of uncertainty among users.

The internal moderation guide should be made available to the public. Of course, platforms should redact any graphic imagery in the manuals and merely provide verbal descriptions of any distressing content. This disclosure would help to remedy the disparity of information between the platform and the users, and would provide a much-needed element of certainty on how these rules are interpreted to users. It would allow users to understand the motivations behind these terms and conditions and then challenge the ones that seem to conflict with human rights values. Disclosing the internal guide would also contribute to creating a culture of justification on the platform, whereby those who limit freedom of expression, participation in cultural life, or the right to privacy must justify those limitations to the users.[219] Transparency, therefore, becomes an important tool for the accountability of platforms but it is one that we must demand of platforms because they are unlikely to engage in it willingly. Consequently, Bonnici and De Vey argue that until we acknowledge that transparency is not a priority for these companies (and that we must demand they prioritise it) any plans dependent on these platforms becoming more transparent are doomed to fail.[220]

---

[217] Fuller, *The morality of law*, 162.
[218] Waldron, "Rule of law."
[219] Articles 19 and 17 *ICCPR*. Article 15(a) *ICESCR*.
[220] J. P. Mifsud Bonnici and C. N. J. de vey Mestdagh, "Right Vision, Wrong Expectations: The European Union and Self-regulation of Harmful Internet Content," *Information and Communications Technology Law* 14, no. 2 (2005), https://doi.org/10.1080/13600830500042665.

3.5: Facilitating Participation

Social media platforms should allow more participation from users in the development and revision of their terms and conditions. Baldwin and Cave argue that the underlying rationale of participation is that it constitutes "proper democratic influence over regulation" and has a "legitimating effect" for the regulator.[221] The users of social media sites are not merely clients or customers, they are value creators who produce and share content, contribute to discussions, and possess valuable data that can be sold to third parties.[222] The UNGP's also emphasise the importance of consulting with stakeholders in order to identify and remedy any potential human rights issues. [223] Unfortunately, while platform spokespeople often engage in rhetoric comparing the platforms to communities and democracies,[224] the actual governance style on these platforms is much more authoritarian. In fact, the current approach to the creation and amendment of terms and conditions at most social media platforms resembles a classic command-and-control structure, which can impoverish the legitimacy and efficacy of the moderation process.

Creating spaces where individuals can participate in discussions about the rules that govern their behaviour elevates individuals from a social media *user* (which is quite a diminishing term) to a social media *citizen*. This approach is explained eloquently by Jeremy Waldron, who writes:

> "law is a mode of governing people that treats them with respect, as though they had a view or perspective of their own to present on the application of the norm to their conduct and situation. Applying a norm to a human individual is not like deciding what to do about a rabid animal or a dilapidated house. It involves paying attention to a point of view and respecting the personality of the entity one is dealing with. As such it embodies a crucial dignitarian idea—respecting the dignity of those to whom the norms are applied as *beings capable of explaining themselves*."[225]

---

[221] Robert Baldwin and Martin Cave, *Understanding regulation: theory, strategy, and practice* (Oxford: Oxford University Press, 1999), 79.
[222] van Dijck, *The culture of connectivity* 63.
[223] Principles 18 and 20, *UN Guiding Principles.*
[224] Mark Zuckerberg, "Building Global Community," Facebook, last modified February 18, 2017, https://m.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634.
[225] Waldron, "Rule of law," 17.

Waldron argues that this aspect of procedure is often overlooked by individuals who focus on the substantive content of the law and ignore the opportunities for "argumentation that a free and self-possessed individual is likely to demand."[226] Individuals should be allowed to debate and discuss the rules that govern their behaviour and while social media platforms are not democracies, platforms are dependent on users to generate wealth (content, attention, and personal information) and their input should be valued. One could consider this a softer, digital version of that famous rallying cry 'No Taxation Without Representation.' In particular, as human rights often involve balancing exercises, with regulators trying to determine the appropriate course of action when multiple rights (such as free speech and security of the person) are involved, it becomes even more important that users be allowed to contribute to those discussions on where the line should be drawn.

Participation, therefore, is an important value for perceptions of legitimacy and strengthens rule of law principles but these values need to be focused into workable strategies that could be implemented by social media companies. First, any proposals favouring a "direct democracy" approach to terms and conditions with all users being allowed to vote on what content should be permitted on the platform should be discarded.[227] This approach would likely result in a mass of contradictory rules that may not comply with the legal obligations platforms must comply with (such as laws concerning CSAM) and could privilege powerful groups at the expense of minority interests.[228] What might be possible, however, is to occasionally hold non-binding plebiscites to gauge users' feelings about the content that is and is not allowed on the platform. These polls would be an easy way for

[226] Waldron, "Rule of law," 23-25.

[227] Facebook actually attempted a variation of this called Facebook Site Governance in 2009 where users who liked the site governance page could comment on rule proposals and the ones that generated a high volume of contents were then put to a vote. This experiment will be discussed in greater detail at 5.4.2 but suffice to say it was a failure. For more on this experiment, see: "Mark Zuckerberg: Vote on Facebook Site Governance," Facebook, last modified April 20, 2009, https://web.facebook.com/facebookapp/videos/mark-zuckerberg-vote-on-facebook-site-governance/186119950483/?_rdc=1&_rdr.

[228] It should be noted that Wikipedia and Reddit do have public discussions about what content should be removed. Reddit, however, has a hierarchy of paid administrators (admins) and two levels of voluntary moderators attached to each forum that have the final say in content moderation decisions. In addition, as the goal of Wikipedia is to act as a repository for factual information (which must be cited), moderators will keep discussions on an article's "Talk Page" focused on factual topics and will remove comments that merely express an opinion. Wikipedia, therefore, is not a good comparator to social media platforms which deal with a much more diverse range of content.

platforms to take the community temperature because, as many lawyers who work with social media platforms acknowledge, the norms of social media community are always evolving.[229] It could also be likened to the traditional administrative law procedure of notice-and-comment whereby agencies publish a rule, ask for comment from any interested parties in the public, and issue a final rule that explains their reasoning and responds to any important comments.[230]

Another way for platforms to enhance participation is for them to create a space where users can engage in "vital public negotiations" and debate various aspects of the terms and conditions and the moderation process.[231] These discussions would supplement the formal moderation process and could provide important information to the policy teams at the platform, who would benefit from a forum where they could identify the "evolving expectations from our community."[232] These forums for participation would demonstrate a commitment to users and would improve the perception of platforms as legitimate regulators. This idea will be revisited in Chapter Five.

Finally, individual users should be able to provide feedback on the terms and conditions directly to the platform. Currently, most social media platforms allow users to respond to specific moderation decisions made about them (usually by ticking boxes and choosing preselected answers) but do not provide avenues for concerned users to proactively raise concerns about the content guidelines of a platform.[233] This means that users can only engage directly with the platform when they post prohibited content, which undermines the ability of the majority of rule-abiding platform users to comment on the rules that govern the site. Some critics have contended that users even engage in "frivolous appeals" because the user finds "certain provisions of in the company's Terms of Use

---

[229] Klonick, "New Governors," 1649.
[230] Keats Citron, "Technological Due Process," 1290.
[231] Crawford and Gillespie, "What is a flag for?," 422-23.
[232] Zuckerberg, "Building Global Community."
[233] Flickr, the photo-sharing platform, is one exception as it does have a web form that users can use to provide feedback. Another exception is Tumblr, which sought feedback from its users when it changed its policies in 2013. The General Counsel of Tumblr even personally responded to every e-mail they received containing feedback. See: Ammori, "The 'new' New york times," 2273.

objectionable."[234] While they condemn this action as a waste of resources, the underlying problem is that there is no other feedback mechanism that users can access to communicate directly with the platform. By not providing avenues for feedback, users must rely on the media or create a campaign that is popular enough to get the policy team at social media companies to take action.[235] A report by a number of researchers at the Berkman Klein Centre recommends that a specific feedback form be created that users can fill out when they want to ask questions or share their opinions on the platform terms and conditions.[236] It would not be possible for a social media network to directly respond to every user but these questions and comments should be used as an informal polling method and as inspiration for the platform's FAQ section and for articles written by representatives of the platform.

All of the suggestions that have been discussed above could help create new avenues for participation on social media platforms and would indicate a commitment to enhancing accountability. While these ideas have been focussed on the relationship between users and platforms (as that is the primary focus of this thesis), it should also be noted that platforms should create channels for gathering input from moderators as well. A researcher who interviewed a diverse sample of moderators discovered that many of them were frustrated by the fact that their expertise was not taken into account and their suggestions about policies were neither solicited nor valued.[237] Instead, policies were drafted by senior staff at social media platforms with no input from the individuals moderating content on a day-to-day basis. This command-and-control approach should be reformed at social media companies as the platforms could become more effective and legitimate regulators if they created avenues for participation from all affected stakeholders, users and employees alike.

3.6: Conclusion

[234] Newland et al., *Account Deactivation and Content Removal*, 20.
[235] The various forms of response will be discussed at 5.3.
[236] Newland et al., *Account Deactivation and Content Removal*, 13.
[237] For more on the experience of moderators, see 4.2. Roberts, *Behind the screen*. 97.

Terms and conditions help to structure the world we experience online but are rife with certainty and transparency issues that are contrary to rule of law principles. Concerned parties have been aware of some of the issues explored in this chapter for years. In 2011, the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue, recommended that internet companies take a number of steps to protect these rights including "be transparent about the measures taken" and "establish clear and unambiguous terms of service."[238] Unfortunately, nine years later, the actions of the major social media platforms fall short of these recommendations. This is problematic because these rules matter in a world that is becoming more and more reliant on social media as a mediator for the human experience. The codes of conduct created by platforms have an impact on our lives and these rules exist in an "interdependent relationship with legal codes" because "where one fails, the other is under more pressure to succeed, and where one develops, the other may wither."[239] Platforms have been given a tremendous amount of discretion in how they regulate their spaces because of their technical abilities to handle a large volume of content and the misperception that what happens on these platforms is of no real consequence. The current approach to content moderation, however, has created a lot of issues and the problems that exist at the creation stage are further exacerbated by the enforcement stage, which will be discussed in the next chapter.

[238] Frank La Rue, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue (A/HRC/17/27)* (Geneva: United Nations, 2011).
[239] Tambini, Leonardi, and Marsden, *Codifying cyberspace*. 29.

Chapter Four: Enforcement

4.1: Introduction

This chapter will focus on the enforcement process, which is arguably the most important stage in content moderation because it has the most impact on what content is available and how the platform is governed. The enforcement stage is the only place where users are able to interact with moderators, albeit within a highly structured disciplinary process where moderators control the rules of engagement and request no real input from the users. The enforcement stage is where the concerns over private companies regulating the exercise of free expression online (as well as other right) are at their more pronounced.[240] This period in the moderation process is where decisions are made on a case-by-case basis on whether content should be visible on the platform, a person's account should be suspended, or a group should be removed. The last chapter focused on the creation of terms and conditions and a number of problems were identified in how these standards were developed. The enforcement stage complements the creation stage and can either help or hinder the reform of the creation process. Baldwin and Cave capture this interplay between the various stages of enforcement by explaining that "Astute enforcement can remedy design defects in regulatory mechanisms and ill-enforcement can undermine the most sophisticated designs of regulation."[241]

This chapter will first explain the role of the moderator (4.2) and how the enforcement process occurs (4.3). Section 4.4 will then explore three issues in the enforcement process: bias in decision-making, an over-reliance on efficiency as a solution, and inconsistent enforcement of terms and conditions. Finally, Section 4.5 will suggest that platforms adopt a body of precedents as a tool for accountability and the empowerment of users. Ultimately this chapter will conclude that there are some serious issues in how platforms enforce their rules, and that these directly and clearly engage the human rights of users in a number of

[240] Ben Wagner takes this farther by arguing that "In a more general sense, the focus on non-judicial content regulation rather than the use of established legal procedures to regulate content is one of the hallmarks of expression governance on the Internet." Wagner, "Governing Internet Expression."
[241] Baldwin and Cave, *Understanding regulation*, 96.

ways, but that many of these issues could be ameliorated with some widespread reforms and changes in priorities at social media companies.

4.2: The role of the Moderator

The role of the moderator is to enforce the terms and conditions of the social media platform by making decisions about what content should remain visible on the platform. A moderator can be an algorithm or it can be a human, either employed by the company, outsourced to another company, or acting on a volunteer basis such as on Reddit.[242] It should also be noted of course, that in the report-and-respond systems detailed below, where users flag content as potentially violating the rules, ordinary users act as an informal, first tier of moderators but they will not be discussed in this chapter as they can only flag content and cannot make any decisions on whether the content should remain on the platform.[243] This section will give a brief overview of moderators at social media companies before moving on to the enforcement process in the next section.

4.2.1: The Human Moderators

The first moderators at social media companies typically joined in the start-up stage, and were recent university graduates (often from prestigious universities) who worked in Silicon Valley offices.[244] These in-house moderators still exist, they are typically hired on short-term contracts, kept separate from other employees, and barred from the attractive benefits packages that other employees enjoy.[245] These employees exist in a liminal space,

[242] Technically there are also private companies who will moderate the content posted on a company's social media page for a fee. They check any content the company shares to make sure no mistakes have been made (a service US Airlines could have used when it accidentally tweeted a pornographic airplane-themed image in 2014) and remove any content that users post to the group page which do not fit their brand identity. Roberts terms these moderators "boutique moderators" and they will not be discussed in this thesis as their objective is quite different from other kinds of moderators. See: Roberts, *Behind the screen*, 40.
[243] YouTube, for example, has created the YouTube Heroes programme, which employs users to act as moderators in exchange for small perks such as access to special events and launches.
[244] Buni and Chemaly, "Secret Rules of the Internet."
[245] In interviews with former moderators at the pseudonymised "MegaTech" (which bears a striking resemblance to YouTube), the moderators talked about being barred from the office Christmas party, working in jobs that only lasted two years and offered no hope of a permanent contract, and walking into

bifurcated from the rest of the company despite performing an essential role. As the platforms began to increase in size, however, the volume of content being uploaded on platforms made it difficult to scale. The inevitable result in the late 2000's was that platforms began to open offices around the world, especially in developing countries where individuals could be paid significantly less than what they would make in America and have less labour rights.[246] While contractors were also hired in various locations in America, Ireland, and Italy, a large amount of moderation work is also outsourced to India and the Philippines, traditional destinations for Business Process Outsourcing (BPO).[247] Workers are also hired to do moderation on micro-labour sites such as Amazon Mechanical Turk where they are paid per moderation decision.[248] One job advertisement Roberts includes in her book offered one cent (US) per decision.[249] Companies typically use a hybrid approach, employing in-house moderators, contracted moderators, and micro-labour moderators. [250] This complicated web of outsourcing and secrecy has created what Chen calls a "vast, invisible pool of human labour"[251] and Roberts adds that "this invisibility is by design."[252]

Information about moderators is exceptionally difficult to get from companies. They are invariably incredibly secretive about outsourcing and outsiders are forced to estimate the number of content moderators working for social media platforms as between

work every day past a rock climbing wall in the lobby and being one of only a handful of MegaTech employees not being allowed to use it. Roberts, *Behind the screen*, 82-85.

[246] According to Adrian Chen, as of 2014, Facebook paid the equivalent of $312 US a month to a Filipino moderator although the going rate at other social media companies in the Philippines was $500 a month. This means that "a brand-new American moderator for a large tech company in the US can make more in an hour than a veteran Filipino moderator makes in a day." See: Adrian Chen, "The Labourers Who Keep Dick Pics and Beheadings out of your Facebook Feed," Wired, last modified October 23, 2014, https://www.wired.com/2014/10/content-moderation/.

[247] Sarah Roberts, "Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste," *Wi: Journal of Mobile Media* 10, no. 1 (2016): 3.

[248] For an in-depth investigation into the workers on micro-labour sites, see: Mary L. Gray and Siddharth Suri, *Ghost work: how to stop Silicon Valley from building a new global underclass* (Boston: Houghton Mifflin Harcourt, 2019).

[249] Roberts, *Behind the screen*, 48.

[250] Roberts, *Behind the screen*, 48-49.

[251] Chen, "Labourers."

[252] Roberts, *Behind the screen*, 3.

100,000[253] and 150,000,[254] which would account for roughly half of the total workforce for social media companies.[255] These teams are still kept separate from the rest of the organisation and "siloed" into "isolated corporate enclaves" both in-house at the companies and at special content assessment sites where they must sign strict Non-Disclosure Agreements (NDA's).[256] Most research must therefore be done under the condition of anonymity or with disgruntled employees like Amine Derail, the man who leaked the Facebook guide to Gawker.[257] These NDA's can be very problematic for employees who are struggling with the upsetting content they view at work as it inhibits their ability to get support from friends and family.[258] For example, a recent report on Cognizant, a Florida-based moderation contractor for Facebook, found that many workers were subsequently diagnosed with Post-Traumatic Stress Disorder (PTSD) and were not warned that these jobs would be ill-suited for individuals with a history of anxiety and depression.[259] This raises questions of whether their rights to the "highest attainable standard of physical and mental health" and "just and favourable working conditions" are being respected.[260] These agreements also prevent outsiders from gaining the insight of experienced employees who may have important suggestions for reform. Sarah Roberts, a media academic who pioneered the study of what she terms "commercial content moderation" highlights this issue when she asks how we "effect change on moderation practices if they're treated as industrial secrets?"[261]

---

[253] Chen, "Labourers." This estimate was actually given to the journalist by SSP Blue, an online security consultancy.
[254] Ciaran Cassidy and Adrian Chen, *Moderators* [documentary short], (New York: Field of Vision Films, 2017). In December 2017, Google announced that it was hiring more content moderators for YouTube, bringing the collective number of YouTube moderators to roughly 10,000 people. See: Susan Wojcicki, "Expanding Our Work Against Abuse of Our Platform," YouTube Official Blog, last modified December 4, 2017, https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html.
[255] Chen, "Labourers."
[256] Buni and Chemaly, "Secret Rules of the Internet."
[257] Emma Barnett and Ian Hollinshead, "Dark side of Facebook," Telegraph, last modified March 2, 2012, http://www.telegraph.co.uk/technology/facebook/9118778/The-dark-side-of-Facebook.html.
[258] Roberts, "Digital refuse," 7.
[259] Casey Newton, "Bodies in Seats: At Facebook's worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives," Verge, last modified June 19, 2019, https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa.
[260] Article 12 and Article 7 respectively. *ICESCR*.
[261] This quote came from an interview with Roberts in Chen's piece. Chen, "Labourers."

Despite claiming to disrupt traditional power arrangements, media platforms are replicating the problematic practices of many large corporations. These companies take advantage of the power disparity between the developed and the developing world to concentrate the unsavoury aspects of their businesses in the Global South. These roles are contracted and outsourced so that they can be kept separate from the more attractive aspects of work in Silicon Valley and in order to give platforms a measure of distance, or plausible deniability, from the emotional disturbances and trauma that the moderators they employ will experience.[262] Roberts likens this separation of the roles at a social media company to Western practices of shipping garbage to the Global South, calling the content "techno-trash" and explaining that social media waste products are kept away from the developed world through the out-sourcing of Commercial Content Moderation.[263] It is clear that platforms keep information about these moderators confidential because in many ways, these moderators embody the antithesis of the narrative of social media as an inherently positive force. Despite employing the rhetoric of connectivity and transparency, for example, the moderators are heavily restricted by their NDA's.[264] This level of secrecy seems over-broad and some academics have suggested that we need stricter criteria on when commercial confidentiality is justified and the burden should be on the company to prove the need for non-transparency from the public.[265] By increasing the amount of information shared with the public, users may perceive the platforms as a more legitimate regulator of their online interactions.

The moderator is a blank space upon which the values and debates of social media are projected. They are expected to have an in-depth knowledge of "social norms and mores of the places in the world for which the content is destined" and therefore must situate themselves as an imagined member of that community, regardless of their own beliefs or experiences. Roberts calls this a "phenomenon of cultural and linguistic embodiment"[266] and likens it to out-sourced call-centres but the analogy is debatable as one interacts with call-

[262] Roberts, "Digital refuse," 7.
[263] Roberts, "Digital refuse," 7-9.
[264] Roberts, "Digital refuse," 8.. These NDA's are so strict that moderators cannot even discuss their work with their co-workers.
[265] Baldwin and Cave, *Understanding regulation*, 308.
[266] Roberts, "Commercial Content Moderation," 147.

centre employees whereas moderators work behind the scenes, a faceless embodiment of the platforms' terms and conditions. This experience is rendered more disorienting by the fact that the cultures they inhabit on social media can be distressing, such as when a moderator must become "steeped in the racist, homophobic, and misogynist tropes and language of another culture" in order to make decisions about the context-heavy category of hate speech.[267] These moderators work in Special Economic Zones, places which Saskia Sassen explains are where "an actual piece of land becomes denationalised"[268] and so too the moderators themselves are stripped of any cultural identity and tailored to the identity needs of the corporation. The cognitive dissonance this "denationalisation" process entails is incalculable, but it is likely considered a necessary evil for a job that pays relatively well compared to other jobs for young employees in India and the Philippines and contains few physical dangers. The psychological risks of viewing an endless stream of upsetting content, of course, are harder to quantify.

4.2.2: The Algorithmic Moderator

An increasing number of moderation tasks on social media platforms are conducted by algorithms, which are often perceived as a more efficient and inexpensive way to deal with the volume of content as companies expand. Algorithms perform a number of functions in content moderation. For example, algorithms facilitate ex ante moderation, which prohibits objectionable content from being posted or flags it for consideration by human moderators as soon as it is processed.[269] It is estimated that one-third of all content flagged on Facebook comes from algorithmic identification[270] and the proportion is higher for terrorist content, which is identified by algorithms over half the time.[271] The increasing use of algorithmic moderators, however, is not without its concerns. The nuance that is possible in an assessment by a human moderator is flattened by algorithms. This distillation can be especially problematic when platforms use algorithms to locate content that they are legally

---

[267] Roberts, "Commercial Content Moderation," 148.
[268] Saskia Sassen, *Losing Control? Sovereignty in an Age of Globalization* (New York: Columbia University Press, 1996), 8-9.
[269] Klonick, "New Governors," 1636.
[270] Zuckerberg, "Building Global Community."
[271] Hopkins, "Facebook struggles..."

required to remove (such as CSAM, copyright violations, and certain categories of hate speech) because programmers, who often have no legal background, interpret regulations in ways that are efficient for their system, transforming uncertain norms into programming language that may not reflect the original regulation.[272] This is further compounded when decisions necessitate rights-balancing and competing but valid interests need to be resolved. Algorithms also struggle with tasks that require "computer vision" (computer recognition of images and objects) and humans are often called in to complete what Gray and Suri refer to as "automation's last mile."[273]

Algorithms are also being used to help identify re-posted content that has already been identified as prohibited and removed. One of the most ambitious applications of this idea can be found in the Photo DNA system, which is a joint project between a number of technology companies including Microsoft, Google, Twitter, and Facebook. The PhotoDNA programme was originally designed to combat the dissemination of CSAM but has now been extended to also apply to extremist content. The programme works by attaching unique digital fingerprints (called hashes) to each piece of content that has been deemed CSAM or extremist and then uploading it to a shared database.[274] This means that only one of the partners needs to have identified this content as prohibited and entered into the system so that it can be identified on any other platforms (or on the same platform in the future) and flagged. Hany Farid, the initial developer of PhotoDNA has stated that now that the system is extending beyond relatively clear-cut areas like CSAM, there is a need for an impartial body to monitor the database and also for transparency on how decisions about prohibiting content are made.[275] These assertions are reasonable but unlikely to be fulfilled by a group of companies that has consistently engaged in secretive regulation with no impartial contributions. The PhotoDNA system also raises concerns that after content has been added to the hash database then there will be a presumption in favour of removal communicated

---

[272] Jon Bing, "Code, Access, and Control," in *Human Rights and the Digital Age*, ed. Mathias Klang and Andrew Murray (London: Cavendish Publishing, 2005), 205.
[273] Roberts, "Commercial Content Moderation," 37; Gray and Suri, *Ghost work*, xxii.
[274] Olivia Solon, "Facebook, Twitter, Google and Microsoft team up to tackle extremist content," The Guardian, last modified December 6, 2016, https://www.theguardian.com/technology/2016/dec/05/facebook-twitter-google-microsoft-terrorist-extremist-content.
[275] Solon, "Team up."

to the other platforms, who might delete the content without first considering whether the decision seems correct and also whether the content actually violates their terms and conditions. This issue is further complicated when the PhotoDNA system is used to engage in highly subjective assessments on whether something constitutes extremist content as it is likely the platforms will have a diverse range of views that could be sacrificed in order to achieve consistency and efficiency, and to be seen as being "tough on terrorism."

Clearly, algorithmic moderators raise a number of concerns for lawyers, most of which will be discussed in the subsequent sections of this chapter. One issue that must be raised here, however, is the fact that how algorithms are programmed and how they engage in decision-making is incomprehensible to the majority of people who use social media. This centralisation of technical knowledge in the hands of the people who run platforms creates what Harold Innis refers to as a "knowledge monopoly" whereby the introduction of new technologies allows the destruction of an old knowledge monopoly and the creation of a new one formed of those with the technical insight to guide this new technology.[276] Postman argues that these groups "accumulate power and inevitably form a kind of conspiracy against those who have no access to the specialised knowledge made available by the technology."[277] This has clearly occurred in Silicon Valley, where we interact with algorithms on a daily basis but have no understanding in how they operate, thereby relinquishing any nascent ability we might have to critique and demand changes to algorithmic processing.[278] These algorithmic processes, therefore, lack transparency, which is an important aspect of the procedural protections that underlie any regulatory bid for legitimacy.[279] This unintelligibility is rendered more absolute by the fact that unlike human moderators,

[276] Harold A. Innis, *The Bias of Communication* (Toronto: University of Toronto Press, 1951), 179-80.
[277] Postman, *Technopoly*, 9.
[278] It is hopeful, however, that the introduction of the GDPR may alter this disparity of knowledge about algorithmic processing although the GDPR only covers a fraction of the activity that occurs on social media platforms.
[279] Bert-Jaap Koops, "Criteria for Normative Technology: An Essay on the Acceptability of 'Code as Law' in Light of Democratic and Constitutional Values," in *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes*, ed. Roger Brownsword and Karen Yeung (Oxford: Hart, 2008), 163.

algorithms are incapable of violating a NDA and blowing the whistle on problematic behaviour in Silicon Valley.[280]

It should be noted that while human moderators and algorithmic moderators have issues that are specific to their type, there are also problems with moderators that transcend these categorisations. The clearest example of this is the issue that pervades every aspect of content regulation on social media platforms: transparency. Both the actions of human moderators and algorithmic moderators are kept hidden from ordinary users and are rendered impervious to investigations. This is problematic because these moderators have a significant impact on the world we experience on social media and yet, the policy from senior members of these platforms is to share as little information about how content is regulated with the public as possible. Martin Ammori, who takes an overly optimistic view of these social media platforms, claims that "fifty years from now, though, we will remember these lawyers and their impact on how millions of people experience freedom of expression."[281] This assertion is contradicted by the fact that users will never be able to remember something that they never knew in the first place. Unlike landmark free expression cases that were fought in an open courtroom and publicly reported, the moderators of social media live in a shadow world of NDA's and proprietary knowledge. There is no emphasis on transparency and these freedom of expression decisions that Ammori celebrates lack any form of legitimacy. Ammori also makes the mistake of conflating perceived legitimacy (the fact that social media platforms are regulating speech online with no concerted protest from the public) with the legitimacy a regulator actually deserves.[282] This point is further supported in the next section, which examines the process of moderation on these platforms.

---

[280] Sarah Roberts, "Social Media's Silent Filter," Atlantic, last modified March 8, 2017, https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/.
[281] Ammori, "The 'new' New york times," 2262.
[282] This contrast between actual and deserved legitimacy is made in Baldwin and Cave, *Understanding regulation*, 82.

4.3: The Process of Content Moderation

4.3.1: An overview of flagging and decision-making

Questionable content on the platform is identified by users or algorithms that flag content as potentially violating the content guidelines. A large volume of content is uploaded to services like YouTube and Facebook and the report-and-remove mechanism mobilises the equally vast numbers of users to wade through the deluge. The flag is a symbol of user condemnation, an attempt to codify complaints and translate a sense of distaste into an actionable data-point.[283] This flagging process, however, should not be interpreted as a discourse between the user and the platform as decisions on flagged content are rarely communicated to those who flagged it, leaving them uncertain as to exactly how policies are applied by moderators and depriving them of any opportunity for responding or discussing what should be available on the platform.[284]

These moderation systems can result in very fast response times; YouTube has stated that most videos that violate their terms and conditions are flagged and removed within an hour of dissemination.[285] Morozov calls this process "crowd-sourced censorship"[286] but a more apt description may be crowd-sourced enforcement, as everyday users help to police the site. Users are integral to the flagging system, making them "uncompensated digital labourers" who are unaware of the services they deliver to the platform along with access to their valuable personal data.[287] The majority of social media platforms use this flagging system, reviewing content only when it is brought to their attention. One major exception is the secret-sharing app Whisper, which reviews all content posted on the site, approving each

[283] Crawford and Gillespie, "What is a flag for?," 413.
[284] Crawford and Gillespie, "What is a flag for?," 414. It should also be noted that Facebook now provides notifications of the decision reached and a one-sentence explanation, along with a way to see where your complaint is in the moderation process. But this does not change the fact that users have no idea how this decision was reached, which is arguably the most important part of the process from an academic point of view.
[285] Abraham H. Foxman and Christopher Wolf, *Viral hate: containing its spread on the Internet*, 1st ed. (New York: Palgrave Macmillan, 2013), 107.
[286] Morozov, *The net delusion*, 104.
[287] Buni and Chemaly, "Secret Rules of the Internet."

post before it becomes publicly available.[288] One wonders, however, if this would ever be feasible for larger platforms like Facebook or YouTube.

Once the content is flagged (whether by a human or an algorithm) content moderators decide whether it should be retained, removed, and/or whether the disseminator should have their account suspended. These decisions are usually made by low-level staffers, who use internal content manuals (see 3.4.2), combined with information they have committed to memory in order to speed up the decision-making process. These moderators are expected to evaluate content extremely quickly. One moderator disclosed that staffers must assess at least 2000 pictures an hour, giving them a decision-making window of 33 seconds.[289] This astonishing pace seems generous compared to Facebook's moderation teams, where users are expected to make a decision every *ten seconds*.[290] These strict time-frames makes one wonder whether social media regulators are really moderating on reflex, forced to react instead of being given time to evaluate and consider their response. It is almost as if the expectations for moderators are aligned with algorithmic performance, a daunting prospect for any employee.[291]

Some commentators are overly optimistic of what the content-moderation process entails and might even go so far as comparing it to a fully functioning judicial body. Klonick for example, seems to afford the practices of social media companies too much deference and characterises their processes with too much optimism. For example, she writes "training moderators to overcome cultural biases or emotional reactions in the application of rules to facts can be analogised to training lawyers or judges."[292] She goes on to quote research by Kahan that examined how judges, lawyers, and law students analyse situations designed to trigger political bias.[293] The results indicated that judges, and to a lesser degree lawyers, came to consistent conclusions free of bias whereas law students (like the general public)

---

[288] Chen, "Labourers."
[289] Cassidy and Chen, *Moderators* [documentary short]
[290] Solon, "Facebook can't win."
[291] Especially as platforms are increasingly relying on algorithms (for decisions as well as flagging-see above) and will likely use them as a benchmark for performance in the content moderation process.
[292] Klonick, "New Governors," 1463.
[293] Dan M. Kahan et al., "Ideology' or 'Situation sense'? an experimental investigation of motivated reasoning and professional judgment," *University of Pennsylvania Law Review* 164 (2016): 354-55.

made decisions that conformed to their individual political biases. Klonick argues that "the experiments by Kahan and his co-authors demonstrate empirically what Facebook learned through experience: people can be trained in domain-specific areas to overcome their cultural biases and to apply rules neutrally. Just as this truth is an essential part of the legal system, it is an essential part of Facebook's moderation system."[294] Klonick's reasoning is flawed because most moderators are only given a couple of weeks of training (presumably the law students in the study would have had more education than that) so it is unrealistic to assume that these moderators will behave as objectively as judges and lawyers. Moderators are also expected to make decisions within seconds, a parameter that would be inconceivable in the legal world. Similar legitimising arguments are made by Ammori, who argues that "The terms of these policies often take the form of traditional legal rules and standards…they have just as much validity."[295] This is an overestimation of the content guidelines at platforms, which Chapter Three argued are seriously flawed. It also seems unlikely that a valid legal system would be so lacking in procedural safeguards.

Unfortunately, effective platform regulation has become conflated with mere deletion by social media companies.[296] This is evident in articles like "The Delete Squad"[297] and in interviews with content regulators who indicated that they did not perceive the enforcement of content rules as a freedom of expression issue (unlike governmental requests for removal) so they were unconcerned about human rights issues.[298] Deletion is a weak remedy because it does not induce a change in human behaviour (in this case, the posting of less objectionable content) and has little value as a preventative factor. Baldwin and Cave explain that one of the key questions in regulatory enforcement is how much the group of potential offenders will be deterred by the regulator's current or prospective approach to enforcement."[299] Removing content is only a superficial remedy and offers no real promise of a reduction of antisocial behaviour on the platform in the future. The current content moderation regime,

---

[294] Klonick, "New Governors," 1464.
[295] Ammori, "The 'new' New york times," 2263..
[296] Mac Sithigh characterises it as the "delete now, ask questions later" approach. Mac Síthigh, "The mass age of internet law," 82.
[297] Rosen, "Delete Squad."
[298] Jørgensen, "Framing Human Rights," 351.
[299] Baldwin and Cave, *Understanding regulation*, 110-11.

therefore, does not represent an effective approach because, as Julia Black argues, good regulation focuses on achieving outcomes rather than technical compliance.[300]

The majority of platforms, therefore, fail to create practices that can initiate change in the behaviour of regulatees, thus failing to institute Black's conception of outcome-led regulation. Platforms instead focus on removing content, a practice that they also insist remains shrouded in secrecy. Most distressingly, deletion becomes a benchmark for success on social media platforms and censorship becomes a moot point, an irrelevant concept for a plugged-in world. Nowhere is this more evident than in Monika Bickert's comment that "I've been in this role just over four years and I would say that in general [Facebook's policies] have gotten more and more restrictive and that's true not just at Facebook but for all the large social media companies."[301] Bickert is saying that this is a *positive* change, that restrictiveness shows that platforms will not tolerate antisocial content, but it seems hard to imagine a traditional media company using rates of censorship as a proxy for success just as a police force should not conflate arrest rates with larger crime prevention objectives.

The removal of content carries a particularly powerful condemnatory weight in a society where the allocation of attention is considered of central importance, where "visibility is perceived to be a proxy for legitimacy."[302] A new form of communications technology creates another "arena in which thoughts develop"[303] so the removal of certain contributions from this public conversation should not be treated lightly. To delete something from social media, to deem a category of content as too objectionable to remain, is to alter the public conversations that occur on social media and the opinions that are formed. Neil Postman, the media theorist, once wrote "technological change is not additive; it is ecological."[304] If social media is an environment, an ecology in which we find ourselves then to render a category of content prohibited is to earmark a species for extinction. This is not to argue that social media platforms should not prohibit and remove content, but rather that they should

---

[300] Julia Black, "Constructing and contesting legitimacy and accountability in polycentric regulatory regimes," *Regulation and Governance* 2, no. 2 (2008): 157, https://doi.org/10.1111/j.1748-5991.2008.00034.x.
[301] Talbot and Bourassa, "How Facebook Tries..."
[302] Crawford and Gillespie, "What is a flag for?," 422.
[303] Postman, *Technopoly*, 6.
[304] Neil Postman, "Five Things We Need to Know about Technological Change," University of California, Davis, last modified March 28, 1998, https://web.cs.ucdavis.edu/~rogaway/classes/188/materials/postman.pdf.

consider the implications of these actions and how the decisions they make affect the digital ecosystem, before certain categories of expression become endangered species on social media. This is an important feature of any speech-based regulation because the act of expressing oneself is devoid of any real power if it cannot be heard by others.[305] This idea certainly predates the Internet[306] but private control of essential platforms for expression (social media) has rendered this point more salient.

The content moderation process must also evolve to stay abreast of new trends in social media behaviour. An early challenge for platforms was when ISIS began to establish a presence on social media, creating a patchwork of violent videos, domestic pictures, and groups designed to facilitate the recruitment and transportation of both men and women to Iraq and Syria. This content was interspersed with seemingly innocuous content designed to normalise ISIS such as 'The Cats of Mujahedeen' which featured fighters in combat gear cuddling kittens.[307] Most of this content has now been removed, with Monika Bickert of Facebook announcing "if it's the leader of Boko Haram and he wants to post pictures of his two-year-old and some kittens, that would not be allowed."[308]

4.3.2: The Challenges of Live-Streaming

A recent and disturbing problem for some social media platforms is the live-streaming of criminal acts as was tragically demonstrated in the Christchurch attacks. These videos are inherently unpredictable, difficult to interrupt, and are not subject to algorithmic moderation because the content is simultaneously shared and uploaded to the platform.[309] They weaken the power of the moderator and allow some of the most disturbing acts

---

305 See: Brian W. Esler, "Filtering, Blocking and Ratings: Chaperones or Censorship?," in *Human Rights and the Digital Age*, ed. Mathias Klang and Andrew Murray (London: Cavendish Publishing, 2005), 99.

306 One famous statement of this concept was by Justice William Brennan in *Lamont v. Postmaster General*, 381 U.S. 301(1965).: "the dissemination of ideas can accomplish nothing if otherwise willing addressees are not free to receive and consider them. It would be a barren marketplace of ideas that had only sellers, and no buyers."

307 James Vincent, "I can haz Islamic State Plz: ISIS Propaganda on Twitter Turns to Kittens and LOLspeak," Independent, last modified August 21, 2014, https://www.independent.co.uk/life-style/gadgets-and-tech/isis-propaganda-on-twitter-turns-to-kittens-and-lolspeak-i-can-haz-islamic-state-plz-9683736.html.

308 Klonick, "New Governors," 1652.

309 Natasha Lomas, "Facebook's content moderation rules dubbed 'alarming' by child safety charity," TechCrunch, last modified May 22, 2017, https://techcrunch.com/2017/05/22/facebooks-content-moderation-rules-dubbed-alarming-by-child-safety-charity/.

captured on film to be shared with no oversight. This content can also be used to incite violence against others, is a violation of the privacy rights of the victims, and if shared without restrictions would be accessible to children which could be "injurious to his or her well-being."[310] While some of these acts fall into the newsworthiness exception many platforms have as they depict police brutality such as the Philando Castile case, other violent films have no redeeming relevance for society-at-large. This was certainly the case with the live-streamed murder of an 11-month-old baby by her father, a video that was viewed over 350,000 times (and shared widely) in the twenty-four hours that it remained up on the platform before it was finally removed.[311] Other platforms with a live-streaming feature (such as Periscope (later acquired by Twitter), YouTube Live, Instagram Live, and Twitch's Lifestreaming feature) have similarly struggled with the murders, sexual assaults, physical attacks, and suicides that people choose to film. Surette argues that there is a history of performative crime that predates social media (terrorist and anarchist groups for example often engaged in such behaviour) but there was also a place for performative justice (most notably public executions) that no longer exists today.[312] This hearkens back to the discussion about the discursive power of deletion (at 4.3.1), an act that is diametrically opposed to a portrayal of performative justice. Platforms need to develop new tools to detect live-streaming crimes but one wonders if perhaps these mechanisms should also contribute to a performative justice agenda, even if the justice meted out is not punitive as in the past but rather of a restorative nature.

One may ask whether there should be any difference in expectations in how platforms handle live consent as opposed to uploaded content. Perhaps platforms assumed that they had less responsibility in relation to live content than they did to uploaded content, which they have designed algorithms and trained human moderators to handle. They may have assumed that live-streaming was the equivalent of wandering around a public space, aware

---

[310] If that incitement was on the grounds of national, racial, or religious hatred then it would be a violation of Article 20(2) *ICCPR*. Article 17 (privacy) *ICCPR*. If children could view it, it would be a violation of Article 17(e), *Convention on the Rights of the Child*.

[311] Samuel Gibbs, "Facebook under pressure after man livestreams killing of his daughter," The Guardian, last modified April 25, 2017, https://www.theguardian.com/technology/2017/apr/25/facebook-thailand-man-livestreams-killing-daughter.

[312] Raymond Surette, "Performance Crime and Justice," *Current Issues in Criminal Justice* 27, no. 2 (2015): 195, https://doi.org/10.1080/10345329.2015.12036041.

that there is always a risk that you may see something illegal or upsetting. The more apt comparison, however, would be to liken live-streaming to live-broadcasting a television programme.[313] Over the years, many horrific acts have been broadcasted live on television, content that would have never made it to the screen if it had been pre-recorded.[314] Producers of live-broadcast shows might know that it is easier to prevent harmful content on pre-taped shows but they are still aware that they have responsibilities to their viewers and have put in place certain measures to respond to these risks. These measures include broadcasting the programme with a slight delay in order to pre-empt any harmful content and watching the programme at all times so that the transmission can be ceased the moment anything goes wrong.[315] Another possible distinction is that live-broadcast was the original way television programmes were created as video recording had not been invented yet.[316] At the time, it was not technologically feasible to prevent harmful content being broadcast on television. Social media companies, however, should have been able to identify and address many of subsequent issues in live-streaming before they ever introduced the product to the public. This failure to anticipate issues and develop new methods of moderation resulted in a serious governance gap on social media. It is also a violation of the UNGP's, which state that business should "avoid causing or contributing to adverse human rights impacts" by identifying and preventing impacts before they occur.[317] Finally, regardless of their

---

[313] While live broadcasting was more popular in the first few decades of television, it is still frequently used for sporting events, awards shows, special episodes of TV shows, and the 'live-on-location' segments of news programmes. Of course there are important distinctions between live broadcasts and platforms that rely on user-generated-content but it should have been clear to social media companies that some of the same challenges would be present in all live-streaming services.

[314] Some notable examples include the assassination of Lee Harvey Oswald, the on-air suicide of news presenter Christine Chubbuck, and the murders of news reporter Alison Parker and cameraman Adam Ward during a live-on-location report. One of the most disturbing examples was a 1998 Los Angeles police standoff that ended in the on-camera suicide of Daniel V. Jones (and the killing of his dog). This event occurred on a Thursday afternoon and broadcasters interrupted their ordinary programming to show the standoff. Unfortunately, many of these programmes were children's cartoons and so many children saw the horrific acts of violence. For more information on these cases, see: James Sterngold, "After a Suicide, Questions on Lurid TV News," New York Times, last modified May 2, 1998, https://www.nytimes.com/1998/05/02/us/after-a-suicide-questions-on-lurid-tv-news.html; Erin Kelly, "Eight Shocking Deaths that Happened as TV Cameras were Rolling," All That's Interesting, last modified October 17, 2017, https://allthatsinteresting.com/live-deaths-tv.

[315] Contrast this approach with the practice of social media companies to only have moderators view content once it has been flagged, which results in a significantly longer lag-time.

[316] Mitchell Stephens, "History of Television," Grolier Encyclopedia, 2000, https://www.nyu.edu/classes/stephens/History%20of%20Television%20page.htm.

[317] Principles 13 and 17, *UN Guiding Principles.*

difficulties in handling live-streamed content, platforms should have been mindful of the fact that by allowing this functionality, they were permitting a large volume of content (some of which depicted illegal acts) to be associated with their brand and stream to their users. Therefore, there was a responsibility to prevent dangerous content from achieving this access and legitimacy by predicting these issues and creating solutions, such as a delay in broadcast to disrupt the performativity of crimes.

In conclusion, the content-moderation process is a complicated and nuanced interplay of various values and concerns and platforms are faced with unanticipated consequences from the new products they introduce and the increasing penetration of social media around the world. This, however, does not mean that platforms should not be held accountable for the problems that their platforms create or the pre-existing issues that manifest in their sphere of control. The next section will consider a number of these problems in greater depth and identify some solutions for reform.

4.4: Issues in the Enforcement Stage of Content-Moderation

4.4.1: Bias/cultural issues for human moderators

This chapter argues that enforcement is likely to be the stage where the most human rights violations occur on social media platforms, particularly those related to freedom of expression and the right not to be discriminated against.[318] One of the causes of this situation is that limitations on expression are applied inconsistently and may replicate the biases experienced by the predominantly white and male staffers at social media platforms.[319] These staff members devise content assessment strategies to be employed by content assessment teams and algorithms. Bias is a problematic quality in regulation because it calls into question the fairness of the regulations. It also reduces certainty as users will be unable

[318] Article 19 (expression) and Article 3 (gender-based discrimination) and Article 26 (discrimination). *ICCPR*.
[319] A CNN Money Study of 20 tech companies found that a significant majority of leadership positions were held by white males. For an interactive breakdown of how different companies compare on diversity, see: Julianne Pepitone, "How Diverse is Silicon Valley?," CNN Money, accessed June 8, 2018, https://money.cnn.com/interactive/technology/tech-diversity-data/.

to predict how hidden factors will contribute to an assessment of their content. While this section can only provide an overview of bias in technology it will attempt to examine a number of ways that differential treatment infiltrates the content-moderation process.

Both human and algorithmic moderators hold biases because the prejudices and assumptions that organically occur in humans are held by both moderators and the programmers who create algorithms. The choice of training data and the definition of parameters for algorithmic regulation are not neutral activities and can result in biased processes that might only grow more prominent through machine-learning processes.[320] There is a perception that algorithms are inherently more objective than human moderators but this is a myth. As Brown and Marsden write "code is no more neutral than regulation, with each subject to monopoly and capture by commercial interests."[321] No technology is neutral, it is embedded with values and politics that differ only from human-centric processes in their comprehensibility by lay-people. Pasquale contends that all that algorithmic processes have done, therefore, is to "drive discrimination upstream."[322]

Human moderators can express bias in a number of ways. The first, shared with algorithms, is that they can express the value-choices and beliefs of the executive policy teams at social media platforms. These activities often reify the status quo embraced by the Western white males who dominate leadership positions in Silicon Valley. A clear example of this is the topic of blood. There are a number of examples of platforms distinguishing between depictions of blood in the context of accidents and violence and depictions of menstrual blood.[323] It might be reasonable to assume that menstrual blood is a normal

---

[320] Claire Cain Miller, "When Algorithms Discriminate," New York Times, last modified July 9, 2015, https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html; Martin Degeling and Bettina Berendt, "What is wrong about Robocops as consultants? A technology-centric critique of predictive policing," *AI and society* 33, no. 3 (2018), https://doi.org/10.1007/s00146-017-0730-7.

[321] Ian Brown and Christopher T. Marsden, *Regulating code: good governance and better regulation in the information age*, Information revolution and global politics, (Cambridge, MA: MIT Press, 2013), xix.

[322] Pasquale, *Black box society*, 35.

[323] Clementine de Pressigny, "Instagram Deleted Harley Weir's Account over Period Blood," I-d Magazine, last modified September 7, 2016, https://i-d.vice.com/en_us/article/a3gxx4/instagram-deleted-harley-weirs-account-over-period-blood.. See also: Radhika Sanghani, "Instagram deletes woman's period photos – but her response is amazing," Telegraph, last modified March 30, 2015, https://www.telegraph.co.uk/women/life/instagram-deletes-womans-period-photos-but-her-response-is-amazing/. This should be contrasted with social media's approach to graphic violence or injuries. One needs

feature of half of the world's lived experience and is related to important topics such as women's health and should not be sanctioned on social media platforms.[324] It also might be reasonable to assume that gratuitous pictures of injuries and violent attacks are more upsetting than normal health processes and should not be widely accepted on social media platforms.

The reality, however, is that male discomfort with menstrual blood and endorsement of violent images[325] results in a bias against allowing women to share on social media the aspects of their lives that differ from the male experience. This means that it is perfectly permissible in Facebook's moderation guidelines to post a picture of a man shot in the head, lying in a pool of his own blood as long as the caption is "condemning rather than celebratory" [326] but not an image of a woman with a blood stain on the back of her sweatpants.[327] This discomfort with women has also led to prohibitions on female nipples, breast-feeding pictures,[328] and pictures depicting the outline of female genitalia and nipples while fully clothed.[329] In addition, while revealing pictures of plus-size women or women with body hair were removed, similar photos of slimmer, hairless women remained

---

only to search on Instagram for example using hashtags like "injury", "accident", or "blood" to view many graphic images.

[324] Although of course male anxieties around menstrual blood pre-date social media. One ancient example dates from AD 78 when Pliny the Elder wrote in his *Natural History* that when a woman is on her period, "Her very look, even, will dim the brightness of mirrors, blunt the edge of steel, and take away the polish from ivory. A swarm of bees, if looked upon by her, will die immediately." Pliny also thought menstruating women could wither trees and plants, rust metals, and turn dogs insane. See: Chapter 13 (15) Remarkable Circumstances connected with the Menstrual Discharge. Pliny the Elder, *Natural History*, ed. John Bostock and H. T. Riley (Medford, MA: Trustees of Tufts University, 2004).

[325] See, for example: Lauren Schulte, "Facebook & Google Block Period Language, OK Video of Man Shooting Himself in the Face," Medium, last modified July 27, 2016, https://medium.com/the-fixx/facebook-google-block-period-language-ok-video-of-man-shooting-himself-in-the-face-ac9c8c2e50d8; Natasha Hinde, "Musician slams Facebook for Removing Post about Period Pain Labelling it a 'Disgusting Hit of Oppression," Huffington Post, last modified July 19, 2017, https://www.huffingtonpost.co.uk/entry/melody-pool-facebook-status-about-period-pain-removed-from-facebook_uk_578dec2fe4b0885619b11978.

[326] The Guardian, "How Facebook guides moderators on terrorist content."

[327] Sanghani, "Instagram deletes woman's period photos."

[328] Alex Hern, "Facebook's changing standards from beheading to breastfeeding images," The Guardian, last modified October 22, 2013, https://www.theguardian.com/technology/2013/oct/22/facebook-standards-beheading-breastfeeding-social-networkin.

[329] Adrian Chen, "Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where 'Camel Toes' are more Offensive than 'Crushed Heads," Gawker, last modified February 16, 2016, http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads.

online.[330] Feminist writer Jessica Valenti argues that these images represent the reality of womanhood but that "it's men that social media giants are "protecting" - men who have grown up on sanitised and sexualised images of female bodies. Men who have been taught to believe by pop culture, advertising and beyond that women's bodies are there for them."[331] Women have often struggled with the fact that the values of social media platforms have frequently privileged the interests of males under the guise of impartiality, a state of affairs that was most evident in Twitter's protracted debate over whether free speech on the platform could include the protracted harassment and threats of sexual violence levied against female users. It should be noted that many social media platforms have changed their policies over time, making them more responsive to the concerns of women and persons of colour, but it is important to understand that these changes were often the product of concerted campaigning and media attention and may not represent the platform's original intentions. For example, after concerted campaigning, in 2015, Facebook clarified its rules on nudity to make an exception for breastfeeding and images depicting post-mastectomy scarring.[332] The fact remains that this activism was required to change a status quo that should not have been replicated in the online environment and that the default gaze of a content policy-developer is a male perspective and remains unchanged.

It is important to remember that the status quo envisioned by executive policy teams at social media platforms comes from a particular lived experience that differs widely from large numbers of social media users. It is difficult to identify how these beliefs about the world and what should be prohibited in the public discourse affect the social media experience but these assumptions must be identified and challenged. It is also important to agitate for alternative perspectives on the world to be included, because otherwise, as

---

[330] Elizabeth Plank, "This photo was banned by Instagram – Thanks to Society's Sexist Double-Standards," Mic, last modified January 20, 2015, https://mic.com/articles/108624/this-banned-instagram-photo-exposes-the-latest-double-standard-in-censorship#.cOQiZv3Dw; Hern, "Facebook's changing standards from beheading to breastfeeding images.".

[331] Jessica Valenti, "Social Media is protecting men from periods, breast milk and body hair," The Guardian, last modified March 30, 2015, https://www.theguardian.com/commentisfree/2015/mar/30/social-media-protecting-men-periods-breast-milk-body-hair.

[332] Facebook actually claimed in 2015 that it had 'always' allowed breastfeeding photos (hence the use of the term 'clarifying') but the facts do not support this claim as there were many examples of breastfeeding photos being removed. "Facebook Clarifies Breastfeeding Pics OK, updates Rules," CBC, last modified March 16, 2015, https://www.cbc.ca/news/world/facebook-clarifies-breastfeeding-pics-ok-updates-rules-1.2997124.

Crawford writes "we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with its old, familiar biases and stereotypes."[333]

Another way that human moderators can express bias is in making decisions in relation to flagged content. Because moderators must make decisions so quickly, it is likely that they resort to heuristics and schemas to come to a conclusion, using what Kahneman would refer to as 'System 1 thinking.'[334] This can be compared to the problem of juries bringing their misconceptions about sexual assault (so-called "rape myths") into trials and then comparing the facts they are presented with to their schema of what a rape "should" look like to reach a conclusion about whether a sexual assault occurred.[335] These schemas will be particularly powerful in situations where it is less clear whether content falls into a particular category, such as the prohibitions against hate speech, terrorist content, and bullying. This can result in differential outcomes for similar content, such as the research finding that Islamist accounts faced much more suspension pressure than white supremacist ones on Twitter.[336]

Algorithms have the capacity to be more problematic than human moderators because while they are also likely to display embedded biases, they are perceived by the general public as inherently objective, resulting in users placing more faith in them than they might endow a human moderator. One of Kranzberg's six laws of technology states that "technology is neither good nor bad; nor is it neutral."[337] Unfortunately, too often the partiality of technology is obscured by layers of programming and faux-objectivity that make it harder to uncover. Algorithms are programmed by people and "allow prejudices to become

---

[333] Kate Crawford, "Artificial Intelligence's White Guy Problem," New York Times, last modified June 25, 2016, http://mobile.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?0p19G=c.

[334] System 1 thinking is fast, intuitive thought. System 2 thinking is more calculated, conscious, and measured. Daniel Kahneman, *Thinking, fast and slow*, 1st ed. (New York: Farrar, Straus and Giroux, 2011).

[335] Sokratis Dinos et al., "A systematic review of juries' assessment of rape victims: Do rape myths impact on juror decision-making?," *International Journal of Law, Crime and Justice* 43, no. 1 (2015), https://doi.org/10.1016/j.ijlcj.2014.07.001.

[336] J. M. Berger, *Nazis vs. ISIS on Twitter: A Comparative Study of White Nationalist and ISIS Online Social Media Networks* (Washington, DC: George Washington University, 2016).

[337] Melvin Kranzberg, "Technology and History: 'Kranzberg's Laws'," *Technology and Culture* 27, no. 3 (1986): 545, https://doi.org/10.2307/3105385.

embedded in the technology, making their effect biased and less transparent."[338] Facial recognition software designed for white faces may struggle to identify people of other races and algorithms modelled on Islamic jihadism may not catch material from other extremist groups such white supremacist organisations. Crawford writes "Histories of discrimination can live on in digital platforms, and if they go unquestioned, they become part of the logic of everyday algorithmic systems."[339] In addition, it is very difficult for outsiders to identify these biases as the algorithms are designated by companies as proprietary knowledge.[340]

In conclusion, it must be presumed that the content moderation process is suffused with biases and it would naive to assume that the system could be rendered entirely impartial and objective. There is room for improvement, however, and the first step must always be transparency. Details of the enforcement process must be made public so that biases can be identified and platforms can be held accountable. Transparency is frequently discussed in this project because secrecy is one of the main weaknesses in the content moderation process developed by platforms and very few improvements can be initiated without a commensurate increase in transparency. A more ambitious solution that will help minimise bias (as well as other issues examined in this chapter) will also be explored at 4.5.

4.4.2: The Efficiency Narrative

One problem with the enforcement of terms and conditions on platforms is that whenever they are faced with a new issue or controversy, social media companies focus on providing narrow solutions centred on efficiency rather than considering the underlying problems that they routinely face. For example, platforms often fixate on how fast content can be removed after it is posted, a narrative that this section will argue has also influenced political discussions around social media. Arguably, the reason for this obsession with removal is largely due to the fact that it is a simpler variable for social media companies to address and it is a parameter that can be measured and adjusted using technological methods. Morozov borrows a term from architecture to describe this behaviour as

[338] Esler, "Filtering, Blocking and Ratings," 99.
[339] Crawford, "Artificial Intelligence's White Guy Problem."
[340] Crawford, "Artificial Intelligence's White Guy Problem."

"solutionism." It is also a feature of coding, where the goal is to eliminate problems and less attention is paid to how these solutions are achieved. This term connotes a preoccupation with sweeping technological solutions without any concerted attempt to examine larger issues at play.[341] Solving any problem (no matter how minor) is celebrated, despite "completing neglecting more burning, but less obvious, issues."[342] Consequently, larger questions about how much power a private company should have in regulating speech and how prohibited content is defined are displaced by a narrow focus on efficiency.

Efficiency as a goal in of itself was a by-product of the Industrial Revolution, where using machinery to increase productivity was treated as an achievement in of itself.[343] Alfred North Whitehead once wrote that the greatest invention of the nineteenth century was the idea of invention itself.[344] A similar trend has emerged today, spurred on by the rise of Big Data and the notion that patterns that were unmeasurable and goals that were unattainable ten years ago are now possible today. Postman concludes that after the Industrial Revolution, "we had learned how to invent things, and the question of why we invent things receded in importance. The idea that if something could be done it should be done was born in the nineteenth century."[345] This is a strikingly apt description of the prevailing attitude in social media development, which makes it all the more surprising to learn that Postman articulated this idea in 1992.

This narrative of efficiency can be identified in the institutional culture that exists at social media companies. Jon Bing explains that programmers take principles that are capable of having more than one valid interpretation and convert them into the "certain norms" that represent what is "efficient or appropriate within the framework of the system being developed."[346] Efficiency, therefore, becomes a clearly achievable goal that is amenable to all

---

[341] Morozov, *To save everything, click here*, 6.
[342] Morozov, *To save everything, click here*, 149.
[343] Of course, the Industrial Revolution also produced a number of prominent resistance groups to this notion such as the Luddites. One wonders if the Digital Revolution will produce its own version of Neo-Luddites, concerned about employment prospects and the rights of the individual in the digital age.
[344] Alfred North Whitehead, *Science and the modern world*, Lowell lectures, (New York: Simon and Schuster, 1970 [1925]), 96.
[345] Postman, *Technopoly*, 42.
[346] Bing, "Code, Access, and Control," 205.

kinds of metrics which loftier objectives frustrate. A common maxim at these companies is "code wins arguments" and this belief leads to "the primacy of technical solutions, where technology has the capacity and the performativity to solve all problems within the platform."[347] Clearly, larger debates about human rights, good governance, and societal goods are de-prioritised by this code-centric approach. This echoes Thoreau's famous pronouncement that our inventions are but improved means to an unimproved end.[348] Platforms often respond to scandals by increasing efficiency especially in their moderation procedures. One of the first examples of this was in 2011, when Google and Facebook responded to concerns raised in a presidential summit about teenage cyber-bullying by introducing systems that would make it easier for bullying content to be flagged and removed.[349] This response side-steps larger questions of whether teenagers should be allowed on social media and the effects that using social media as a young adult will have on their development and experiences in life. It also fails to consider that regulators should not aim for "perfect compliance or the complete elimination of hazard" as it is unrealistic and will reach a point where "the costs of further enforcement are not justified by the gains."[350]This pattern of addressing symptoms of larger societal conflicts with simplistic responses that rely on increasing productivity still exists today. When the controversies broke over murders and sexual assaults being live-streamed, Facebook's response was to announce the hiring of 3000 new moderators to better police content online.[351] There were no public discussions about whether live-streaming was an essential feature of social media or whether it should merely be considered a failed experiment that was incompatible with Facebook's current content moderation capabilities.

The legal embedding of the narrative of efficiency is evident in Germany's passing of the Network Enforcement Act in Summer 2017. The law requires large social media platforms to remove illegal content (violations of 22 provisions in the German Criminal

---

[347] Wagner, "Governing Internet Expression," 396. See also: Jørgensen, "Framing Human Rights," 345.
[348] Henry David Thoreau, *Walden* (London: Penguin Classics, 2016 [1854]), 31.
[349] Brown and Marsden, *Regulating code*, 128-29.
[350] Baldwin and Cave, *Understanding regulation*, 110.
[351] Samuel Gibbs, "Facebook Live: Zuckerberg adds 3,000 moderators in wake of murders," The Guardian, last modified May 3, 2017, https://www.theguardian.com/technology/2017/may/03/facebook-live-zuckerberg-adds-3000-moderators-murders.

Code) within 24 hours of it being reported or face a fine of up to 50 million Euros.[352] This law emphasises the necessity of rapid responses to hate speech without considering whether deletion is always the best course of action or acknowledging the risk that platforms will delete more content than is strictly necessary, a risk exacerbated by the lack of judicial oversight or the right to appeal in the legislation.[353] Human Rights Watch has cautioned that this law encourages the creation of "no accountability zones" where the state is able to exert pressure on private companies to censor content free from judicial scrutiny.[354] Three countries (Russia, Singapore, and the Philippines) have already announced that they intend to draft similar laws and this is particularly concerning as all three countries have a chequered history of protecting human rights. This particular strain of the efficiency narrative is especially troublesome because "forcing companies to act as censors for government is problematic in a democratic state and nefarious in countries with weak rule of law."[355] Another example of this narrative being adopted within the legal, and law enforcement, framework is the 2016 announcement by the London Metropolitan Police that the National Counter Terrorism Internet Referral Unit, a specialised unit of the Met Police dedicated to addressing terrorist and extremist material online, remove around 2000 items of harmful material every week.[356] Once again there is a clear equation of deletion with success, one of the more common elements of the efficiency narrative perpetuated in the age of social media.

Social media's approach to moderation is rife with efficiency narratives, including the very limited time-frames moderators are given to make a decision. It is as if moderators are

---

[352] *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [Act to Improve Enforcement of the Law in Social Networks]*, BGBl. I S. 3352.

[353] This means that even though YouTube, Facebook, and Twitter have an appeals system, in place for users who have content removed or their accounts suspended because of a violation of the terms and condition, users who run afoul of the German law cannot appeal the decision.

[354] Similar laws are being considered in Kenya and Venezuela as well. See: "Germany: Flawed Social Media Law," Human Rights Watch, last modified February 14, 2018, https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

[355] Human Rights Watch, "Germany: Flawed Social Media Law."

[356] "250,000th piece of online extremist/terrorist material to be removed," Metropolitan Police News, last modified January 9, 2017, http://news.met.police.uk/news/250000th-piece-of-online-extremist-slash-terrorist-material-to-be-removed-208698.

expected to behave like algorithms, a theme reminiscent of Postman's work when he writes that the direct effects of a technology-fixated society are;

> the beliefs that the primary, if not the only, goal of human labour and thought is efficiency; that technical calculation is in all respects superior to human judgment; that in fact human judgment cannot be trusted, because it is plagued by laxity, ambiguity, and unnecessary complexity; that subjectivity is an obstacle to clear thinking; that what cannot be measured either does not exist or is of no value…[357]

Moderators are perceived as mere conduits for the policy decisions made by the leadership teams at the company and are expected to exercise as little discretion as possible in decision-making. This is apparent in an interview that Dave Willner, a senior lawyer at Facebook's policy team, gave when he told a researcher "Effectively, we ask whether something is blue or red, not beautiful or ugly."[358] It might be defensible to reduce complicated rules down to binary criteria that can be applied mechanistically if the policy-making teams at the top had already confronted the larger issues (to use Willner's words, whether something is beautiful or ugly) but the narrative of efficiency is all-encompassing and ensures that every level of the hierarchy is equally concerned with getting results fast.

The efficiency narrative challenges important human rights protections but it also represents an important challenge for the rule of law. This is because it undermines the legal arrangements that have been developed and articulated for centuries between the state and its citizens. It de-prioritises non-economic ends and social obligations which are "legislative goals of at least equal importance."[359] It is also problematic when the argument is made that the private regulation of speech by social media platforms is beneficial because they are able to remove unwanted content that would otherwise be protected by free expression laws if the state were the regulator.[360] This could also be seen as an articulation of the efficiency narrative, arguing that the advantage of social media platforms moderating speech is that they are not beholden to human rights laws on free expression. This trend in outsourcing

---

[357] Postman, *Technopoly*.
[358] Ammori, "The 'new' New york times," 2278.
[359] Baldwin and Cave, *Understanding regulation*, 82.
[360] Klonick argues that if social media platforms couldn't remove the content then we would have an "internet nobody wants" without explaining why social media necessitates a different arrangement for free speech than other areas of American society. See: Klonick, "New Governors," 1659.

censorship (in a manner similar to the outsourced moderation that platforms employ)[361] represents a serious challenge to the continued existence of a robust right to free expression at the very same moment that social media is being lauded as an essential forum for free speech.[362] It also misses the point that platforms also permit content that could be seen as breaching rights to privacy, security of the person, or hate speech. This avoidance of judicial or legislative channels undermines the legitimacy of technological solutions and the accountability of their creators.[363]

In conclusion, an excessive focus on efficiency pervades discussions of social media controversies but this myopic view of regulation need not be so prevalent. Postman writes that in a technological world, "Time, in fact, became an adversary over which technology could triumph. And this meant that there was no time to look back or to contemplate what was being lost."[364] It is time for those would-be critics of the practices of social media, whether they be governments, journalists, or academics to extricate themselves from this set of artificial rhetorical parameters and investigate the normative and social issues that are not so easily solved with a clever algorithm or a hiring spree. It is only by setting aside the terms of engagement that these platforms have created that critics will be able to develop solutions that rely on more than just clever code, solutions that have democratic legitimacy and are based on a respect for human rights and the rule of law.

4.4.3: Inconsistent enforcement

Inconsistent enforcement is a serious problem in the content moderation process at many social media platforms. The previous chapter discussed how vague terms and conditions could be, and this problem is exacerbated by enforcement applied in a piecemeal

---

[361] Because, as Gray and Suri remind us, "outsourcing was never simply about cost cutting. It was also about the growing resistance to unionisation and evading long-standing labour regulations." Gray and Suri, *Ghost work*, 55.
[362] See, for example, *Packingham v. North Carolina*, 137 S. Ct.
[363] Winner, *The whale and the reactor: a search for limits in an age of high technology*, 19. See also Robin Mansell's assertion that "many of the judgments and social values that appear to have achieved a consensus are subject to misapplication as we come to rely on technology to implement the law" Robin Mansell, "Introduction and Equity in Cyberspace," in *Human Rights and the Digital Age*, ed. Mathias Klang and Andrew Murray (London: Cavendish Publishing, 2005), 10.
[364] Postman, *Technopoly*, 45.

fashion. Inconsistent enforcement reduces user certainty and makes the unclear rules even harder to contextualise to their behaviour. This situation, however, allows platforms to remain flexible and able to react to situations quickly. Platforms can "retain the ability to make judgments on content removal, based on ad hoc and often self-interested assessments of the case at hand."[365] This flexibility was useful to YouTube when the radical preacher Anwar al-Awlaki increasingly became notorious for his connections to violent attacks like the Fort Hood shooting and the attempted murder of British MP Stephen Timms. YouTube began to remove more and more of his sermons from the platform, even though some of the sermons had been on the site for long periods of time and did not appear to violate the terms and conditions.[366] Awlaki's newfound infamy as a preacher who inspired terrorists made YouTube make an exception for content that appeared legal. Unfortunately, this exercise of discretion causes uncertainty for all users, and violates Bingham's second principle on the rule of law: "questions of legal right and liability should ordinarily be resolved by application of the law and not the exercise of discretion."[367] It is clear that what is most expedient and useful for regulators may not be beneficial for users, or in compliance with the principles underpinning the rule of law. This practice of quietly removing content that was previously deemed acceptable without announcing any changes in the terms and conditions exemplifies inconsistent enforcement and is caused by three main factors; popularity, accepted narratives, and newsworthiness.

The first of these factors that we turn to now is the relative popularity of flagged content. Interviews with content moderators indicate that the popularity of a piece of content is a factor that is considered when assessing flagged content. The more a piece of content is viewed or shared, the more likely moderators are to decide that it is permissible because it is in the interests of the company to keep up content that will generate attention and income.[368] On the other hand, the amount of flags a piece of content receives (a more notorious form of popularity) can lead to content "queue-jumping" in the moderation process. Twitter, for example, states in the Twitter Rules that "the number of reports we

[365] Crawford and Gillespie, "What is a flag for?," 420.
[366] Jessica Stern and J. M. Berger, *ISIS: The state of terror* (Glasgow: William Collins Publishers, 2015), 132.
[367] Bingham, *The rule of law*, 48.
[368] Roberts, "Commercial Content Moderation," 150.

receive does not impact whether or not something will be removed. However, it may help us prioritise the order in which it gets reviewed."[369] Factoring in popularity is problematic because platforms are treating coherent enforcement as incidental to the real task of regulating content on their platforms.

Inconsistent enforcement can also often stem, secondly, from some narratives being privileged over others (the problem of accepted narratives). These narratives can lead to some content being excepted from the terms and conditions but it can also result in other content being singled out for strict treatment. Examples like the Anwar al-Awlaki situation and Monika Bickert's comments about terrorists not being permitted to share videos of their cats demonstrate that Islamist content has been consistently singled out for wide-spread deletion. The problem with accepted narratives is also present when social media companies decide what constitutes terrorist content (the definition of which lacks global consensus). It is clear that there is a serious difficulty in defining terrorism, a conceptual uncertainty that has forestalled many efforts to enact multi-lateral terrorism treaties.[370] While acknowledging this lacunae in the law, however, it must be noted that social media platforms have ignored the previous debates over what constitutes terrorism and failed to apply their rules on terrorism to groups that would likely not be considered 'edge-cases'. For example, the Facebook Files leak found that Facebook featured multiple slides prohibiting any positive statements about the Irish Republican Army (the IRA) but no accompanying slides indicating that the same treatment should be applied to other para-military groups from Northern Ireland like the Ulster Defence Association (UDA) and its sub-group the Ulster Freedom Fighters (UFF), groups that have been considered terrorist for decades.[371]

A final example of how accepted narratives shape how enforcement occurs on social media is the disparity of treatment in YouTube videos depicting graphic violence from Syria and Mexico. Technically, the videos all violated the terms and conditions as YouTube's rules

---

[369] "The Twitter Rules: hateful conduct policy," Twitter, accessed February 1, 2017, https://support.twitter.com/articles/18311.
[370] Conor Gearty, *Terror* (London: Faber and Faber, 1991), 10.
[371] Indeed, the British named the UFF as a proscribed terrorist organisation in 1973 and the UDA in 1992. See: *Proscribed Terrorist Organisations* (London: Government of the United Kingdom, 2019); The Guardian, "How Facebook guides moderators on terrorist content."

focus on the graphic nature of the content, not on the particular importance of one region over another.[372] YouTube's policy team, however, informed the moderators that the Syrian videos should stay up to raise awareness of the situation there while other graphic videos (such as content depicting the violence in Mexico's narco-wars) would continue to be removed.[373] These decisions are important because they convey attention and legitimacy to some global issues and deny that publicity to others; it is the virtual equivalent of pulling someone's chin to direct their gaze towards something and away from something else. This prioritisation is not often made on a reasoned basis and is more likely to stem from the personal beliefs of the policy-makers about what is legitimate, although in the Syria/Mexico case, Roberts also makes the point that YouTube's decisions aligned with American Foreign Policy: to support certain groups and narratives in Syria while denying any responsibility for the narco-wars in Northern Mexico.[374] These discursive decisions have a substantial impact in a world where attention is the most valuable resource[375] and people mobilise around certain issues on social media. Unfortunately, no matter how uninterested in making value-judgements platforms profess to be [376] inconsistent enforcement because of accepted narratives demonstrates that these judgements do frequently occur.

We see inconsistent enforcement, thirdly, on the grounds of newsworthiness. Cases like police brutality and global events (such as democratic protests around the world) have led to platforms making exceptions to their rules on graphic violence if the content is considered newsworthy (a factor that is never defined and appears to occur on a largely case-by-case basis). If the first phase of social media usage can be characterised by largely apolitical content, the Arab Spring in 2011 signalled a major paradigm shift which forced social media companies to consider what role they wanted to play in global politics.[377] The

[372] Buni and Chemaly, "Secret Rules of the Internet."
[373] Roberts, "Social Media's Silent Filter."
[374] Roberts, "Social Media's Silent Filter."
[375] James Williams, *Stand out of our light: freedom and resistance in the attention economy* (Cambridge, UK: Cambridge University Press, 2018), xi-xiii.
[376] Ammori's interviews with lawyers at social media companies found that "the lawyers generally attempt to avoid making judgment calls about the value of particular speech." Ammori, "The 'new' New york times," 2276.
[377] An earlier version of this did occur in the 2009 Iranian protests. Morozov, however, believes that the utility of social media during those protests was largely exaggerated. See: Morozov, *The net delusion*, 1-33.

Arab Spring marked a serious enhancement of social media's legitimacy as it demonstrated how the same platforms that could be used to share cat videos and family pictures could also be used to organise protests and document state brutality. This forced many platforms to leave up content that they would ordinarily remove because it depicted graphic violence. A more controversial case, however, occurred five years later, when a social media company was faced with the question of whether it is ever appropriate to allow images of child nudity to remain on the platform.

In 2016, the Norwegian newspaper *Aftenposten* announced that Facebook had removed the famous "Terror of War" picture that it had shared and called Mark Zuckerberg, the CEO of Facebook, "the world's most powerful editor." [378] The picture depicted a Vietnamese girl (Kim Phuc) running naked and crying after she was burned by Napalm. This statement caught the attention of the Norwegian Prime Minister Erna Solberg, who then posted the photo as an act of solidarity. It transpired that not only had Facebook intentionally removed the photo but also that the picture was used in training sessions of content assessors as an example of a post that *should* be removed since it featured a distressed, naked child.[379] Once the controversy went public, with major media sources reporting on the issue, Facebook reversed its decision and announced that it would now weigh newsworthiness more heavily in its decisions in the future. In an interesting aside, it should be noted that this photo is so compelling that that it has a history of causing media companies to make exceptions to their regulatory regimes. [380] While this case might seem like a positive development, it cannot be denied that by introducing an element of newsworthiness, platforms have reduced certainty and rendered their rules more inconsistent.

Another example of inconsistent enforcement on the grounds of newsworthiness occurred in 2018 when Hungary's chief of staff to Prime Minister Viktor Orban put up a racist video complaining about migrants. Facebook removed it on the grounds that it was hate

---

[378] Time photo, "The story behind the 'Napalm Girl' photo censored on Facebook," Time, last modified September 9, 2016, http://time.com/4485344/napalm-girl-war-photo-facebook/.
[379] Reuters, "Facebook and YouTube use automation to remove extremist videos, sources say," The Guardian, last modified June 25, 2016, https://www.theguardian.com/technology/2016/jun/25/extremist-videos-isis-youtube-facebook-automated-removal.
[380] When the photo was first published, major media outlets like The New York Times chose to ignore their anti-nudity rules on the grounds that the picture was so newsworthy. Time photo, "Napalm Girl."

speech and then put it back up a couple of hours later on the grounds that it was newsworthy.[381] This incident shows how confusing the rules become when the newsworthy element is applied to content moderation decisions. There are also serious concerns that the result of this decision is that hate speech is available on the platform purely because it was shared in a political capacity, thus legitimising its presence and setting aside the concerns of immigrants who might be concerned that the Hungarian government is permitted to use social media to disseminate hate speech.

The creation of the newsworthiness exception may have been done with the best of intentions but it is questionable whether another vague parameter being added into the content assessment process is a positive outcome. This notion of what is worthy of public attention is also an example of perpetuating accepted narratives as social media platforms are privately deciding what is so important that the public must be able to access it regardless of a general prohibition against that type of content. This lack of certainty is further complicated by the fact that users have little understanding of what content has been treated as unworthy of the newsworthiness exception and removed from the platform. The likely result will then be a self-reinforcing cycle where content deemed newsworthy is allowed to stay up on the platform, where it generates more discussion and appears even more vital to public conversations as a result. Meanwhile, content that is removed will not be reported and shared as widely and nor will it capture the public's attention to the same degree.

In conclusion, inconsistent enforcement seriously undermines the principle of certainty, one of the key pillars of the rule of law and of good regulation. It reduces the substantive dimension of fairness, an aspect of regulation that Baldwin and Cave define as "the quality of outcomes of regulatory procedures and whether the actual policies, rules, and decisions that regulators arrive at are coherent, intelligible, and fair between different parties."[382] In order to increase the fairness of their regulatory regime (and thereby enhance

---

[381] David Gilbert, "Why Facebook censored a "racist" video from Hungary's Government—then put it back," Vice News, last modified March 9, 2018, https://news.vice.com/en_us/article/gy87m4/why-facebook-censored-a-racist-video-from-hungarys-government-then-put-it-back.
[382] Baldwin and Cave, *Understanding regulation*, 314.

the legitimacy of their authority), platforms must make more explicit rules with more detailed explanations of what they entail (and any exceptions or mitigating factors that could be applicable) and then enforce these rules consistently. This should increase the source, process, and outcome legitimacy of the content moderation these platforms employ.[383] It cannot be denied that this approach will be less convenient for companies, who have enjoyed a significant amount of discretion in enforcement, but this sacrifice would be justified on the grounds that the users would experience much more certainty on the platforms and perceive these companies as much more legitimate regulators, organisations that respect both the principles behind the rule of law and the rights they as users enjoy.[384]

4.5: Enhancing Accountability and Transparency

Currently there is a lack of accountability and transparency on social media platforms and this must be addressed before any other reforms can be undertaken. Brin argues that both transparency and accountability are essential for compliance with the rule of law because "without the accountability that derives from openness, enforceable upon even the mightiest individuals and institutions, how can freedom survive?"[385] Every stage in the content moderation process, therefore, must be transparent, accountable, and mindful of the companies' human rights obligations. This is echoed by one of the reports written by David Kaye on social media and free expression, where he called for 'radical transparency, meaningful accountability and a commitment to remedy in order to protect the ability of individuals to use online platforms as forums for free expression, access to information and engagement in public life."[386]

[383] Brown and Marsden, *Regulating code*, 19.
[384] They would also perceive platforms as less likely to engage in discriminatory practices as there would be a greater degree of accountability.
[385] David Brin, *The transparent society: will technology force us to choose between privacy and freedom?* (Reading, MA: Perseus Books, 1998), 13.
[386] David Kaye, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (A/HRC/38/35)* (Geneva: United Nations, 2018), 19.

Transparency and accountability are interrelated because neither objective holds much value without the corresponding one. Without methods of holding platforms accountable, openness is no different from impunity. Without transparency, it is impossible to gather the relevant information to induce any changes in how platforms operate. At times, content moderation appears shrouded in excessive secrecy, with platforms refusing to share even the most basic information such as the standards they use to assess content or even how many content moderators work for the company in order to pre-empt any demands for change.[387] Farrand and Carrapico contrast the "secretive negotiation process" of social media platforms with the "overt law making" which public regulatory bodies engage in, arguing that a lack of transparency poses a fundamental legitimacy problem for these companies.[388] This lack of accessibility also forces would-be critics to rely on whatever content is made public or the unsubstantiated claims of inside sources. This challenge goes to heart of the issue of transparency and accountability; secrecy impedes reform, thus undermining good regulation. Accountability and transparency, therefore, are important goals for anyone seeking to reform the practices of platforms. These normative values, however, must be translated into concrete objectives that platforms can implement.

A publicly-available body of precedents needs to be created to detail how enforcement occurs on social media platforms. To make this set of precedents user-friendly, they would be organised according to which provision in the terms and conditions they concerned. The case-studies would be anonymised and any content that raised data protection problems or was too graphic would be replaced with a description of the content. There would be a short explanation of why this content was flagged and what the decision of the moderator entailed. These precedents would give shape to the terms and conditions, expanding on the various categories of prohibited content and indicating how borderline cases are decided. Not every decision would be publicly available but there should be at least

[387] Twitter executive Nick Pickles once refused to disclose how many staff Twitter employed to work in content moderation on the grounds that if he gave an exact number, he would invariably be told: 'it is not enough.' This severe aversion to criticism seems bizarre and excessive. Alan Travis, "Face-off between MPs and Social Media Giants over Online Hate Speech," The Guardian, last modified March 14, 2017, https://www.theguardian.com/media/2017/mar/14/face-off-mps-and-social-media-giants-online-hate-speech-facebook-twitter.

[388] Farrand and Carrapico, "Networked Governance and the Regulation of Expression on the Internet."

a few examples for any potential violation of any rule and this resource should be updated when changes in the content policy for the platform are introduced. This tool would give users the knowledge and resources to participate effectively in appeals of moderation decisions and in larger campaigns to change how a platform regulates certain categories of content. By enhancing the democratic aspects of the content-moderation process, the platform can increase their perceived legitimacy with users.[389] This solution would require an initial outlay of effort and resources by the platforms but then it would quickly become an established part of the social media experience, one which represents an important bid for enhanced transparency and accountability.

There is a database that shares some similarities with this body of precedents proposal. The Lumen database is a joint project between the Berkman Klein Centre and a number of independent organisations, some of which are law clinics at universities.[390] It collates cease-and-desist letters that relate to social media content. There are, however, some important differences between Lumen and the proposed body of precedents. The first is that the subjects covered by Lumen are more narrow in focus than my proposal. Lumen generally focusses on intellectual property issues, although it does include other subjects like the right to be forgotten, defamation, and a general category for law enforcement requests. My proposed body of precedents would include all subjects in the platforms' terms and conditions including substantive topics like terrorism, nudity, and hate speech. Second, Lumen does not include an explanation of what decision was made and they are unable, of course, to explain the reasoning behind the decision. Because my proposed body of precedents would be run by the policy teams at the platforms, clearer explanation would be possible. Third, the appearance and functionality of Lumen is clearly designed for academics. It is not particularly user-friendly, can be hard to navigate, and is unlikely to be of interest to the everyday user. The body of precedents, however, would be organised according to specific terms and conditions and would try to illuminate these decisions for users. Finally, Lumen is a voluntary body that collects cease-and-desist letters people send them, it therefore may not be reflecting the most important or controversial decisions. While Lumen

[389] Baldwin and Cave, *Understanding regulation*, 79.
[390] For more information, see: https://lumendatabase.org

is a very interesting project, the proposed body of precedents would be very different in its objectives and approach.[391]

This database would be individual to the company (so Twitter and Facebook would each have a separate body of precedents) however it would be possible in the future to create a joint body of precedents in areas of content moderation that are relatively harmonised such as terrorist content. The body of precedents would be created by the policy team at a social media platform, as they are responsible for outlining the content moderation rules and determining how cases that hover on the border of permissibility will be decided. The policy team will also be best-suited to keep the precedents updated as minor policy changes can occur frequently. The policy teams also have access to the relevant data to create and maintain a body of precedents whereas would-be reformers outside the company have to rely on anecdotal evidence and documents that are voluntarily shared either by the platform or by other activists.[392]

A publicly available database would help hold platforms accountable by providing users with knowledge to challenge specific decisions or identify problematic themes in the content moderation process. It is therefore similar to the databases of decisions maintained by Press Councils (self-regulatory bodies for print journalism) across Europe.[393] Such an approach would enhance procedural fairness on the platform by making the regulatory process more open, transparent, and accessible to the public.[394]In fact, "publication of basic regulatory data" is considered "a generally accepted standard for transparency of regulation" particularly when human rights are involved.[395] Issues of bias and inconsistent enforcement could be identified much more easily and it would provide a benchmark for

[391] The other limited comparison is Facebook's promise to publish all of the decisions of the new Facebook Oversight board in a public ledger. This does seem connected to the idea of a body of precedents but the board will be unable to generate enough decisions (as they are only forty people- working in a part-time capacity) to create the detailed tapestry of examples that my proposal envisions. This oversight board will be discussed in greater detail at 5.4.2 and 7.4.6 but for an overview, see: Brent Harris, "Preparing the Way Forward for Facebook's Oversight Board," Facebook Newsroom, last modified January 28, 2020, https://about.fb.com/news/2020/01/facebooks-oversight-board/.
[392] The Lumen database for example relies on parties voluntarily sharing their cease-and-desist letters with them.
[393] Tambini, Leonardi, and Marsden, *Codifying cyberspace*, 69.
[394] Baldwin and Cave, *Understanding regulation*, 314.
[395] Tambini, Leonardi, and Marsden, *Codifying cyberspace*, 24.

users in understanding how decisions on content are made, which would increase certainty. Social media users would have a clearer understanding about what is and is not prohibited on the platform and they would not have to waste time posting content that is immediately removed, or flagging content that is objectionable to them but complies with the terms and conditions.

This would not be a judicial body, it would a policy-based scheme, but it would strengthen good governance and rule of law principles on the platforms as well as signalling the company's commitment to transparency.[396] The body of precedents, however, would have some enforceability in court. While the specific decisions in the body of precedents would not be justiciable, there would be a legal requirement that platforms enact and maintain a set of procedural protections, and this would include the body of precedents (this will be discussed in greater detail in Chapter Seven). Creating a body of precedents, therefore, would be a simple but powerful way for social media platforms to indicate that they value users and are willing to move beyond a rhetoric of democracy into the adoption of feasible measures designed to enhance the democratic potential of these platforms. It would also contribute to creating a culture of justification on the platform, whereby those who limit human rights such as expression or privacy must justify those limitations to their users.

4.6: Conclusion

One of the most fervent critics of laissez-faire governance in the twentieth century was economist and lawyer Robert Lee Hale. Hale spent decades studying how corporations and government interact with citizens and concluded that personal freedom required good corporate governance.[397] This assertion is even more true online where private companies

[396] Thus fulfilling Principle 21's requirement of communication to stakeholders. *UN Guiding Principles.*

[397] Robert Lee Hale, *Freedom through law: public control of private governing power* (New York: Columbia University Press, 1952).

now exert a significant level of control over the exercise of free expression and other rights on their platforms.

This chapter has explored how this power is exercised in the enforcement stage of the moderation process, arguably the most important phase of content moderation. A solution was then proposed that represents the first step (but certainly not the last) in rectifying some of the larger issues in content moderation: a body of precedents that could function as a sort of 'case-law' to illuminate and empower user inter- actions with social media platforms. These User-Empowerment Tools (UET's) are an important aspect of the central question for social media companies today: should platforms be a reflection of reality (with all the ugliness and disturbing content that entails) or an idealised, utopian place where negative content is eradicated? The answer, of course, will lie on a spectrum between these two extremes and platforms have tended to vacillate over time in a search for equilibrium but UETs, such as a body of precedents, would allow users more opportunities to participate in this essential discussion on the future of social media. This chapter and the preceding one both focused on the actions of the architects of social media but it is also important to examine how users have responded to the actions of these companies.

Chapter Five: Response

5.1: Introduction

This chapter will discuss appeals, both the internal processes that exist within social networks to handle appeals and the external channels that activist groups access when they cannot appeal through the company to initiate change. The central theme running throughout this chapter is how social media companies respond to challenges to their regulatory process, whether coming from users or from public outcry over scandals. This chapter highlights some of the unique tensions and pressures at play in the discourse that content moderation creates. This discourse could be construed as a dialogue about power as "power is relational, and as a result, the power to influence how the Internet is regulated is also relational."[398] The appeals process is an important procedural mechanism that can be used to challenge substantive rules but its current incarnation at social media companies is weak and anaemic. Activism outside of the platform offers some measure of correction for these deficits but this avenue is not a suitable substitution for a robust appeals process.

The previous two chapters have shown that content moderation is inconsistent and unpredictable. The line between what is permitted and what is prohibited advances and recedes constantly in relation to public concern, geopolitical events, media coverage, and governmental pressure. Content moderation can be construed as a dialogue between a platform and its users (with other voices, such as government leaders, chiming in) about expectations, both what the platform expects from users but also what users expect from the platform. This dialogue is reflected in Wagner's claims that social networks have become important sites of "regulatory contestation."[399] Even at the granular level, this negotiation between the platform and a concerned party exists through an appeal over a piece of content being removed.

This chapter will explore both the internal appeals processes at these platforms and the external actions that groups employ to demand a change in the content-moderation process.

---

[398] Farrand and Carrapico, "Networked Governance and the Regulation of Expression on the Internet," 358.
[399] Wagner, "Governing Internet Expression," 399.

This will be followed by an inquiry into the larger issues present in the response stage and some proposals on how the current appeals processes employed by social networks should be restructured and bolstered through increased accountability measures. Ultimately, this chapter will explore the discourse that occurs between users and platforms (both through formal and informal channels) and argue that this discourse is an important instrument for improving the procedural tools of the platform and demonstrating a commitment to rule of law principles and therefore (as the right to due process is one of its central guarantees) to human rights protection as well.

5.2: Internal Appeals System

5.2.1: The importance of an appeals system

An appeals system is a highly important feature in any system that purports to be fair and accountable. Waldron characterises the right to appeal to a higher tribunal and the right to hear reasons from the tribunal when it comes to a decision as important procedural aspects of the rule of law.[400] An appeals system serves a number of important functions for an institution and the people it serves. First, appeals offer the opportunity for a decision to be checked for any obvious errors.[401] Error correction is one of the key advantages of an appeals system, with error being defined by Langbein as either "good faith differences of opinion about finding the facts or about formulating or applying rules of law."[402] It can prevent miscarriages of justice and ensure the affected parties receive a fair hearing.[403] These systems also empower individuals by dignifying the participants and making "meaningful the interaction between individuals and the state."[404]

---

[400] Waldron, "Rule of law," 4.
[401] Rossman, therefore, characterises appeals as the "quality control mechanism" of a system. See: David Rossman, "'Were There No Appeal': The History of Review in American Criminal Courts," *Journal of Criminal Law and Criminology* 81, no. 3 (1990): 519, https://doi.org/10.2307/1143847.
[402] John H. Langbein, Renée Lettow Lerner, and Bruce P. Smith, *History of the common law: the development of Anglo-American legal institutions* (New York: Aspen Publishers, 2009), 416.
[403] Peter Marshall, "A Comparative Analysis of the Right to Appeal," *Duke Journal of Comparative and International Law* 22, no. 1 (2011): 2-3.
[404] Judith Resnik, "Precluding Appeals," *Cornell Law Review* 70, no. 1 (1985): 619.

An appeals system strengthens a regulatory institution because it encourages consistency in hearings and uniformity in decisions. [405] Uniformity in decision-making contributes to a sense that the institution is fair and that there is an element of accountability, which increases the perceived legitimacy of the decision-making body.[406] Appellate review has the added benefit of helping to clarify and interpret the relevant regulations, "encouraging the development and refinement of legal principles." [407] While appeals emerged in the continental systems earlier than they did in England, appeals have become an established part of the majority of legal systems today owing partially to the growth of human rights law, in particular the procedural protections found in the right to a fair trial.[408] Appeals systems (and the remedies they offer) are referred to as "secondary rules" in Hart's conceptualisation and these procedural protections are essential in creating an effective regulatory system.[409] These rules are even more important in systems that are not directly governed by the courts because "procedural due process and the promise of fairness, transparency, and accountability has often served as replacements for individual adjudications."[410] Appeals systems can also function as a form of grievance mechanism, which is an important aspect of the corporate duty to respect human rights. Grievance mechanisms act as an early warning system for human rights issues and can help prevent or resolve serious human rights issues.[411]

5.2.2: Appeals Systems at Social Media Platforms

While a social media network serves a different function than a court, the benefits that appeals offer remain largely the same regardless of whether an institution deals with criminal law, civil law, tribunal issues, or one of the many other grievances that can surface

---

[405] Marshall, "A Comparative Analysis of the Right to Appeal," 3.
[406] Marshall, "A Comparative Analysis of the Right to Appeal," 3-4.
[407] Cassandra Robertson, "The Right to Appeal," *North Carolina Law Review* 91 (2013): 1225.
[408] Marshall, "A Comparative Analysis of the Right to Appeal," 2. See also: Article 14(5) *ICCPR:* "Everyone convicted of a crime shall have the right to his conviction and sentence being reviewed by a higher tribunal according to law." The European Convention on Human Rights does not require states to provide a system of appeal (although if they do, the right to a fair trial standards apply) but the right to appeal is found in Article 2 of Protocol No.7 to the European Convention. While it is optional, almost every member of the Council of Europe has ratified the protocol.
[409] H.L.A. Hart, *The Concept of Law*, 3rd ed. (Oxford: Oxford University Press, 2012). 94.
[410] Keats Citron, "Technological Due Process." 1251.
[411] Principle 29, *UN Guiding Principles.*

in society. Appeals systems, therefore, are important even outside of a formal judicial process and should be a feature of any system that allocates a "scarce social good"[412] in a way that is perceived as legitimate by the populace.

The appeals processes that exist at social media companies are, however, varied in their scope and effectiveness. This stage of the moderation process is under-developed and has received significantly less academic analysis and media attention than the first two stages. This stage is arguably not being respected by companies as one study of 150 social networking platforms found that 88% of platforms "explicitly foresee that platforms providers may terminate a specific user account without previous notice or the possibility to challenge the decision." [413] It can also be difficult for even a diligent user to find information about appealing social network decisions either on the platforms or by searching the web.[414]

Appeals systems may be perceived as optional accoutrements by certain social media platforms[415] but it is becoming increasingly clear that they are an essential aspect of a user-oriented moderation process. An appeals system provides an assurance to all users that content that has been erroneously removed will be restored and that the rules they are expected to obey are applied consistently. From a practical perspective, appeals would also be of interest to the multitudes of users who use platforms to store content (especially images) that they routinely delete from their phones to save space. An erroneous moderation decision may, therefore, remove valued content that users have no other way of accessing. A report by the Berkman Centre argues that in light of this user tendency, regardless of the outcome of any appeals, users should be provided the opportunity to export their content

[412] Florian Waldow, "Conceptions of Justice in the Examination Systems of England, Germany, and Sweden: A Look at Safeguards of Fair Procedure and Possibilities of Appeal," *Comparative Education Review* 58, no. 2 (2014): 323, https://doi.org/10.1086/674781.
[413] Belli and Venturini, "Private ordering and the rise of terms of service as cyber-regulation." 9.
[414] Search results are complicated by the existence of so many articles providing advice on how to create "appealing content" on social media.
[415] Flickr, for example, appears to have no appeals process while Instagram has a limited one that does not permit individual content appeals in most cases. This will be explored in greater depth at 5.2.3.

from the platform.[416]

A number of platforms have developed relatively strong appeals systems: in particular Twitter and Facebook. Facebook's appeal system is more developed than other networks. The platform allows users to appeal individual content decisions and it provides a dashboard where they can track their reports and appeals and the decisions made about them.[417] Facebook has recently unveiled a new appeals system with a number of benefits for users. It now allows users to appeal individual pieces of content that were removed and states that reviews will occur by a person "typically within 24 hours."[418] Most intriguingly, Facebook indicates that it is currently developing a separate appeal process whereby people who have reported content can appeal a decision to allow the content to remain on the platform.[419]

The appeals system at Twitter is especially interesting, a result of the highly diversified enforcement process that exists on the platform. Objectionable content at Twitter may trigger a range of different actions: hiding the tweet until the user agrees to delete it (instead of the platform removing the tweet), limiting the visibility of the tweet so it is still on the platform but less prominently placed, blocking a user from tweeting but still allowing them to maintain a Twitter account and read other people's tweets, and hiding tweets in specific countries (known as country-withheld content). [420] These actions can all be appealed

---

[416] Newland et al., *Account Deactivation and Content Removal*, 15. It is presumed that this opportunity would not be provided to users who have uploaded illegal content such as Child Sexual Abuse Material.
[417] "How to Appeal," Online censorship, accessed October 24, 2018, https://onlinecensorship.org/resources/how-to-appeal.
[418] Bickert, "Publishing our Internal Enforcement Guidelines and Expanding our Appeals Process."
[419] Bickert, "Publishing our Internal Enforcement Guidelines and Expanding our Appeals Process." Facebook does not indicate the timeline envisioned for this project. The company hinted in 2016 that it was "reviewing its appeals process in response to public feedback" (See: "Facebook Execs feel the heat of the platform's biggest content controversies," Fortune, last modified October 26, 2016, http://fortune.com/2016/10/28/facebook-media-content-controversy/.) After the controversy of the Terror of War photo when many critics complained that the lack of individual content appeals on Facebook meant that users had no recourse if they shared the photo and then it was removed. Many people might have thought that individual appeals would soon be possible on the system but this development took two years to come to fruition. The new system whereby you can appeal a decision to leave content up on the platform may, therefore, not occur in the near future.
[420] Twitter Help Centre, "Our range of enforcement options," Twitter, accessed October 24, 2018, https://help.twitter.com/en/rules-and-policies/enforcement-options.

and Twitter provides a support form for account-level actions[421] and an e-mail notification providing information about the action and opportunities for appeal in other situations.[422] Even if an appeal is denied, "In most cases users will be able to fully access their account and data even if it has been suspended—Twitter will fully disable accounts only in egregious cases of abuse."(as opposed to Instagram which will not allow a user to export their data if their account is deactivated).[423]

A robust appeals system will also be of central importance to two particular categories of users: social media entrepreneurs and activists. An increasing number of users have found ways to earn money from their activities on social media. The most common method is creating content that meets a particular threshold of popularity and is then "monetised" so that users earn income from the advertisements displayed with their content.[424] Other methods are creating content for subscription services (such as Spotify Premium or YouTube Red), operating a voluntary patronage option through platforms like Patreon, displaying merchandise that could be purchased, or diverting users to a website where the entrepreneur offers other services or products and advertises events. The deletion of content (whether that is a full profile or just individual posts) or the decision to de-monetise a profile can therefore have a significant impact on an entrepreneur.[425] It begins to resemble an employment issue and one that is certainly deserving of an oversight mechanism to ensure a decision was made correctly. An unreliable appeals system can cause quite a lot of uncertainty for social media entrepreneurs and when uncertainty of regulatory

---

[421] This form is available at: Twitter Help Centre, "Appeal an account suspension or locked account," Twitter, accessed October 24, 2018, https://help.twitter.com/forms/general?subtopic=suspended.
[422] Twitter Help Centre, "Help with locked or limited account," Twitter, accessed October 24, 2018, https://help.twitter.com/en/managing-your-account/locked-and-limited-accounts.
[423] Online censorship, "How to Appeal."
[424] Monetised YouTube videos, for example, can be very lucrative. Jenna Marbles, a popular YouTube personality, earns roughly $350,000 (US) a year. Jim Edwards, "Yes, you can make six figures as a YouTube star and still end up poor," Business Insider, last modified February 10, 2014, https://www.businessinsider.com/how-much-money-youtube-stars-actually-make-2014-2/?IR=T.
[425] The effects of de-monetisation became headline news in 2018 when Nasim Aghdam, a vegan activist who posted workout videos and content about being Persian-American, opened fire at YouTube headquarters, injuring three people before committing suicide. Aghdam was furious that her videos had been de-monetised and a popular ab-workout video she created was tagged age-restricted. See: Harriet Alexander and Nick Allen, "YouTube HQ shooting: Father of dead female suspect warned police on day of attack she 'hated' company," Telegraph, last modified April 4, 2018, https://www.telegraph.co.uk/news/2018/04/03/gunshots-heard-outside-youtube-office-california/.

outcomes becomes endemic, "not only are people's expectations disappointed, but increasingly they will find themselves unable to *form* expectations on which to rely, and the horizons of their planning and their economic activity will shrink accordingly."[426]

Activists would also specifically benefit from a robust appeals system because the content they post may be in the grey area between what is permissible and what is prohibited (a wide area due to the vagaries of content regulations and the complicated application of factors like the newsworthiness exception) so they need to appeal frequently. Those seeking to raise awareness of human rights issues around the world have also started using social media platforms as "privately owned evidence lockers."[427] It can be dangerous for activists to keep controversial content on their computers and phones in case they are arrested by hostile forces[428] so pseudonymous accounts on social media act as a storage facility that is stable and less risky as it becomes possible to quickly disseminate the content around the world. Evidence on social media has already been used in prosecutions for war crimes in Germany and Sweden, and has played a significant part in the arrest warrant of a Libyan commander issued by the ICC. It is also the primary source of evidence for the new UN International, Impartial and Independent Mechanism which is collecting material on war crimes in Syria.[429] When platforms remove this content and do not have robust appeals systems, the effect on future evidence-gathering abilities is disastrous.[430] In 2013, over 80% of the content showing the Syrian regime using chemical weapons on the civilians of Damascus was deleted from Facebook, destroying content that might have been essential to future legal operations.[431] There is an interesting tension at play here

---

[426] Waldron, "Rule of law." 23.
[427] Per Christoph Koettl, a senior analyst at Amnesty International, commenting in the article: Avi Asher-Schapiro, "YouTube and Facebook are Removing Evidence of Atrocities, Jeopardizing Cases against War Criminals," Intercept, last modified November 2, 2017, https://theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya/.
[428] According to Youmans and York, many activist groups even create contingency plans to delete an individual's social media accounts if they are arrested as it has become common practice to demand an individual's social media log-in information when they arrested in countries such as Syria and Iran. See: William Lafi Youmans and Jillian C. York, "Social Media and the Activist Toolkit: User Agreements, Corporate Interests, and the Information Infrastructure of Modern Social Movements," *Journal of Communication* 62, no. 2 (2012), https://doi.org/10.1111/j.1460-2466.2012.01636.x.
[429] Asher-Schapiro, "YouTube and Facebook."
[430] Asher-Schapiro, "YouTube and Facebook."
[431] Asher-Schapiro, "YouTube and Facebook."

where social media platforms often benefit from their associations with human rights activists (it gives them legitimacy and adds credence to the notion that they are worthy of a certain protected status) but the infrastructure of the platforms do not accommodate the specific needs of these activists. In addition to the problems of the uncertain evidence locker, pseudonymous accounts may be deleted because they run afoul of "real-name policies," (which could be a privacy rights issue) and users may be suspended from platforms as suspected spammers because they sent too many messages or friend requests.[432]

While the conflict between a platform's interest in serving humanitarian activists and other considerations will likely continue, one important safeguard for the people risking their lives to gather evidence and raise awareness is an appeals process that will assure them that if their profiles are erroneously deleted because they triggered an automated spammer alert or their content was flagged as "terrorist content" by pro-government flaggers then they will have some recourse to a procedure open to them. This same promise must be extended to the people who earn their livelihoods from social media and to the scores of ordinary users who trust that so long as they abide by the rules then they will always have access to the decades of images, videos, and notes they've created online. These assurances of certainty, which users often struggle with throughout the content moderation process, are what makes a strong appeals system so important on social media. It signals to users that their efforts are respected, which exemplifies the connection between procedural concepts of the rule of law and preserving human dignity.[433]

As platforms move towards automating more and more of their flagging and moderation processes, errors will be made that an appeals system can help rectify. In the last two years, YouTube has automated many of its decisions about which channels qualify for monetisation and when a channel might lose its monetised status. Algorithms also flag videos on monetised channels that are considered "unsuitable" to advertisers even though there is little information provided of what factors are considered when making these decisions.[434]

---

[432] Youmans and York, "Social Media and the Activist Toolkit."
[433] Waldron, "Rule of law." 16.
[434] Sam Levin, "YouTube's small creators pay price of policy changes after Logan Paul scandal," The Guardian, last modified September 18, 2017, https://www.theguardian.com/technology/2018/jan/18/youtube-creators-vloggers-ads-logan-paul.

YouTube has publicly encouraged users to appeal demonetisation decisions to help the algorithms improve but the platform also announces that it will prioritise videos that had more than 1000 views in the last seven days and that it will not consider appeals of demonetisation if a channel has less than 10,000 subscribers.[435] This is problematic because YouTube seems to be relying on its appeal process to compensate for a sub-standard algorithm and putting the burden on users to correct these errors. This issue is compounded by YouTube's indication that it will not consider the appeals of users with less subscribers even though they seem just as likely to be affected by algorithmic errors as the larger channels. A weak or non-existent appeals system sends a clear message to users that their contributions to the platform, the value they generate for other users,[436] and the personal and career development they have generated for themselves is not valued and can be arbitrarily diminished. This deficit exacerbates other issues in content-moderation, "escalating the situation from an instance of censorship to exile from the platform altogether."[437] Offering users certainty and the chance to participate in regulatory decisions as they pertain to their profile is, therefore, not a "prioritised part of platform moderation systems."[438] These issues are symptoms of larger problems in the platform appeals systems, which will now be discussed in greater detail.

5.2.3: Problems with Appeals Systems

There are a number of serious overarching problems with the appeals systems that are currently in place at social networks. We have seen that the appeals systems at platforms lack certainty but they also lack transparency and accountability. All of these principles, which are fundamental to a regulatory system that embodies rule of law principles leads to a deficit of legitimacy at these platforms. Legitimacy is best construed as "the collective

---

[435] Erik Kain, "YouTube Wants Content Creators To Appeal Demonetisation, But It's Not Always That Easy," Forbes, last modified September 17, 2017, https://www.forbes.com/sites/erikkain/2017/09/18/adpocalypse-2017-heres-what-you-need-to-know-about-youtubes-demonetisation-troubles/#1614ffd6c267n.

[436] For example, as of 2013, 4 percent of YouTube users provided almost three-quarters of the site's content. See: van Dijck, *The culture of connectivity* 116.

[437] Ilana Ullman, Laura Reed, and Rebecca MacKinnon, *Submission to UN Special Rapporteur for Freedom of Expression and Opinion David Kaye: Content Regulation in the Digital Age* (Amsterdam: Ranking Digital Rights, 2017), 16.

[438] Klonick, "New Governors," 1665.

acceptance of an authority claim by the overwhelming majority of those to whom the claim is addressed."[439] Private regulators need to earn legitimacy and this is best achieved by implementing strong procedural tools that enhance these principles.[440] Ensuring procedural rule of law principles operate effectively in the quasi-public and private spheres matters all the more as the influence of non-judicial regulators becomes more prominent.[441] Appeals systems are often overlooked when discussing content moderation but this has allowed the system to stagnate and suffer from a lack of principled reform. This section will discuss some of the larger issues that exist at these platforms.

The first problem has to do with the scope of a prospective appeal. Not all platforms allow appeals on the removal of individual pieces of content and will only allow appeals for the deletion of entire profiles or pages.[442] Instagram and Google+ (when it still existed) do/did not permit appeals on individual removals and neither did Facebook until April 2018.[443] While the motives for this limitation are obvious (lowering the number of prospective appeals that the platform has to consider), the impact on users appears disproportionate. A single video may have taken between 10 and 60 hours of editing[444] while a single post on Instagram may have taken up to an hour-and-a-half to stage.[445] Of course, the amount of time spent on content must not outweigh any clear violations of the rules but users should be afforded the chance to appeal a decision if they believe that the

---

[439] Chris Reed and Andrew Murray, *Rethinking the jurisprudence of cyberspace*, Rethinking law, (Cheltenham, UK: Edward Elgar, 2018).
[440] Maurizia De Bellis, "Public law and private regulators in the global legal space," *International Journal of Constitutional Law* 9, no. 2 (2011): 429.
[441] Waldron, "Rule of law." 18.
[442] Of course, platforms appeals policies change over time. As discussed at 5.2.2, Facebook has recently overhauled its appeals system and made it more user-focused. See: Bickert, "Publishing our Internal Enforcement Guidelines and Expanding our Appeals Process."
[443] Instagram does, however, permit individual appeals in the case of copyright and trademark removals. Onlinecensorship.org has collated the appeals processes for a number of social networks, a compendium that they strive to keep up-to-date. See: Online censorship, "How to Appeal."
[444] This range was based on a number of discussions between YouTube creators found at: "How long do YouTubers take to edit their videos?," Quora, last modified July 19, 2018, https://www.quora.com/How-long-do-YouTubers-take-to-edit-their-videos. And: "How long does it take to make a YouTube video?," YT Talk, last modified August 10, 2012, http://yttalk.com/threads/how-long-does-it-take-you-to-make-a-youtube-video.11340/.
[445] Jenn Herman, "How much time should it take to create an Instagram post?," Jenn's Trends in Social Media Management, last modified January 27, 2016, https://www.jennstrends.com/how-much-time-should-it-take-to-create-an-instagram-post/.

decision was a mistake. Being denied the opportunity to appeal is disempowering especially as even when parties lose their appeal, research shows that they feel a sense of fairness and are more positive about the outcome because they were given the opportunity to be heard.[446] An appeal, therefore, legitimises the user's concerns and reaffirms their status as a valued participant on a platform even if the appeal is unsuccessful.[447] It also signals that platforms are not placing a high value on accountability as they have designated a whole category of decisions (which may have been made by an algorithm) as beyond the scope of appeal.

A second problem has to do with the limited scope of remedies available through the appeals process on social networks. Currently a successful appeal will result in the reinstatement of content (or a profile) on the platform. In some cases, this will be an adequate remedy and will rectify any damage caused by the original decision. Some content, however, is particularly time-sensitive and a subsequent decision to reinstate the content will be too late as the "window" in which this content was relevant and could inspire collective action will have closed. Youmans and York give the example of videos depicting current events being removed and then reinstated on appeal. They argue "even when videos are restored, however, the impact on behalf of activists may be diminished by the loss of viewers and because the video may be overtaken by more recent events."[448] Another group that may be unduly affected by a decision even if it is successfully appealed are the users affected by demonetisation as users will not be compensated for the period in the video's "life-cycle" where the video was available but no longer generating income as "by the time the appeal goes through, the lion's share of the views that that video is ever going to get have probably been and gone, so you've lost out on the majority of your income from that video."[449] The UN Special Rapporteur on free expression has also raised these concerns, calling the scope of remedies available on platforms as "limited or untimely to the point of

---

[446] Scott Barclay, *An Appealing Act: Why People Appeal in Civil Cases* (Evanston: North-Western University Press, 1999), 101-12.

[447] This should also apply to digital contractors who do the moderation piecework that was referenced at 4.2.1. Gray and Suri discovered that these contractors often often lack any ability to appeal the terminations of their account (which freezes any money they have earned on the site and haven't collected yet), the poor ratings given by employers who may be avoiding payment, or random glitches that lock them out of their account for days on end, depriving them of their livelihood. Gray and Suri, *Ghost work*. 80-90.

[448] Youmans and York, "Social Media and the Activist Toolkit," 320-21.

[449] Kain, "YouTube Wants."

non-existence" and arguing that reinstatement was an insufficient remedy if removal resulted in a specific harm (whether physical, reputational, or financial) to the person posting the content or if the suspension occurred during a time of political protest and could have influenced the debate.[450] He therefore recommends that remediation programs be created that include as options "reinstatement and acknowledgment to settlements related to reputational or other harms."[451] The UNGP's outline a number of different types of remedy: satisfaction (confirmation and apology), restitution, guarantee of non-repetition, rehabilitation (providing resources to restore the victim), and compensation.[452] These types of remedy could serve as inspiration for platforms to expand their remedial options. Platforms have a serious impact on the enjoyment of human rights and the very least that we could expect of them is that they create remedial structures that respect the value of those rights.

A third problem is that the appeals systems that exist at social media platforms are extremely narrow and do not provide any opportunity for users to make representations on why the content in question should represent an exception to a platform's current rule or why the rule itself should be changed. This deficit might not be so troubling if there were other formal avenues where users could make these arguments, but they are noticeably lacking[453] (with the exception of an obscure option at Facebook, which will be discussed at 5.4.2). Some commentators argue that users should not appeal decisions on content that clearly violate the terms of use "even if she may find certain provisions in the company's ToU objectionable" because "frivolous appeals divert resources from legitimate appeals."[454] This statement disregards the fact that users may be initiating these policy-based appeals because they have been denied any other avenue through which to advocate for a change in the rules and that debates about the governance of platforms are often far from frivolous. Waldron

---

[450] Kaye, *A/HRC/38/35*, 13, para. 38. The importance of effective remedies was also raised by an earlier: La Rue, *A/HRC/17/27*.
[451] Kaye, *A/HRC/38/35*, para 59.
[452] Principle 22, *UN Guiding Principles.*
[453] YouTube in particular has been criticised for having "no centralised location to communicate with their content creator partners, and they have made no effort to really establish one either." This is especially problematic when users run monetised channels and rely on a good relationship with YouTube for their livelihoods. See: Kain, "YouTube Wants."
[454] Newland et al., *Account Deactivation and Content Removal*, 20.

would characterise this discouragement of policy appeals as a "command and control" approach to regulation, which ignores the essential role that argumentation over norms plays in the rule of law. He writes that "we don't just obey them or apply the sanctions that they ordain; we argue over them adversarially, we use our sense of what is at stake in their application to license a continual process of argument back and forth, and we engage in elaborate interpretive exercises about what it means to apply them faithfully as a system to the cases that come before us."[455] Without a formal process that allows argumentation to occur (even through an individualised appeals system that does not set precedents) users are denied any opportunity to participate in the discourse that Waldron views as integral to a procedural rule of law. This is why Chapter Three suggested the creation of forums of participation on platforms, an idea which will be revisited later in this chapter.

Finally, there is an issue with how little information is provided to users throughout the moderation process, a deficit that continues through to the appeals system. The process of appeals is so opaque that the NGO Onlinecensorship.org has even begun collating information on how users can appeal to each social network,[456] a clear indication that platforms are failing to keep users appraised of such processes. When information is provided, it is often only available to users that have been the subject of a removal decision and it is not otherwise accessible to users. For example, Instagram offers a FAQ section that states that if a user thinks their account was disabled by mistake then they can open the app and follow the on-screen instructions to lodge an appeal.[457] This is not particularly illuminating for users who have not been blocked and just want more information.

This informational asymmetry puts users at a substantial disadvantage when trying to engage with these platforms and Pasquale contends that this lack of transparency is entirely by design. He argues that "the challenge of the "knowledge problem" is just one example of a general truth: What we do and don't know about the social (as opposed to the natural) world is not inherent in its nature, but is itself a function of social constructs."[458]

---

[455] Waldron, "Rule of law," 20.
[456] Online censorship, "How to Appeal."
[457] Instagram Help Centre, "What can I do if my account has been disabled?," Instagram, accessed October 24, 2018, https://help.instagram.com/366993040048856?helpref=search&sr=2&query=appeal.
[458] Pasquale, *Black box society*.

This issue leads to the "knowledge monopoly" that was discussed in Chapter Four, whereby an elite is able to control the dissemination of knowledge and the power that knowledge entails.[459] It is therefore in the best interests of the platform to provide very little information about the appeals systems. Even Facebook, which has a relatively strong appeals system, has not unified all of the information about the appeals process in one place. This is further complicated by the existence of a Facebook Help Community which purports to assist people but which consist mostly of questions that are only relevant to that user and often go unanswered by anyone but other equally confused users.[460]

Other platforms also offer limited information about appeals with one platform in particular being a notable violator. When Onlinecensorship.org began to collate information about the appeals processes at social networks, Flickr not only refused to explain their appeals process but also would not confirm that an appeals system even existed.[461] This is precisely why a 2011 report from the Berkman centre stressed that platforms needed to provide information to the affected user concerning why a particular action was taken and what they could expect from the platform appeals process.[462] The lack of transparency that exists throughout the content moderation process is therefore equally present at the appeals stage. This deficit precludes the refinement of any concrete adjudication principles that are identifiable by users and results in them being unable to cite any sort of precedent or previous experience. While social networks may emphasise the consistency of their appeals systems, users have no way of evaluating the veracity of this claim. This lack of transparency seems, therefore, to be designed as a tool to minimise accountability on the platform. This is a problem because accountability is at the core of an appeals system that respects a procedural rule of law.[463] If there is an imbalance of resources or access to information

---

[459] Harold Innis, *Empire and Communications* (Oxford: Clarendon Press, 1950). 161.
[460] See, for example, "I want to Appeal Support Decision' Facebook Help Community," Facebook, accessed October 24, 2018, https://www.facebook.com/help/community/question/?id=10202808946613203.
[461] Online censorship, "How to Appeal."
[462] Newland et al., *Account Deactivation and Content Removal*, 3.
[463] Accountability underscores a number of Waldron's criteria for procedural rule of law including a right of appeal and "a right to hear reasons from the tribunal when it reaches its decision, which are responsive to the evidence and arguments presented before it." Waldron, "Rule of law." 7.

between the stakeholders and the company, "it can reduce both the achievement and perception of a fair process and make it harder to arrive at a durable solution."[464]

The first step platforms must take to address these issues is to prioritise developing a strong internal appeals system designed to meet the users' needs. This process must be transparent and provide detailed information to the user about why the original action was taken, their opportunity to appeal, and how to navigate the appeals process. This system would help businesses to comply with the principles enshrined in the UN Guiding Principles on Business and Human Rights, which states that when a company identifies a situation where it has caused or contributed to adverse impacts on human rights then it has a responsibility to engage in remediation of these impacts.[465] These principles and their connection to remedial systems will be discussed in greater detail in Chapter Seven. The NGO "Ranking Digital Rights" has also suggested that as governments are putting increasing pressure on social media platforms to respond quickly to violent content (and negative content more generally) they should also "not only support but participate in the development of effective grievance and remedy mechanisms."[466] The necessary changes to platform appeals processes will be discussed in greater detail (at 5.4). An appeals system, however, is primarily focused on individual cases while users may want to initiate a larger change in how content is regulated on social media. The next section will discuss the activism that has emerged in response to the platforms' content moderation processes and how these platforms have reacted.

5.3: External Pressures

5.3.1: An Overview of Collective Action

An individual appeals system is incapable of addressing all the grievances that users may have about content moderation. Parties may have general concerns about the availability of certain categories of content (something that an individual appeal cannot

[464] Principle 31, *UN Guiding Principles.*
[465] *UN Guiding Principles.*
[466] Ullman, Reed, and MacKinnon, *Content Regulation in the Digital Age*, 3. It should also be noted that this obligation complies with the Remedy Principles of the UN Guiding Principles on Business and Human Rights, see p. 25 as an overview.

capture) especially as visibility has become a proxy for legitimacy in contemporary society. The previous chapter discussed how some platforms have prohibited images of menstrual blood as an example of this denial of legitimacy. Gillespie offers the example of a Tumblr decision to provide no results when you searched for certain hashtags, including #gay.[467] The action was primarily aimed at hashtags associated with pornography but by including #gay many users searching for content on gay rights, support communities, and queer bloggers were unable to easily find new content. Many users might have had a problem with rendering gay content harder to locate but, as was discussed in Chapter Three, there is often no way to directly contact a platform to share your opinions. Practically as well, social media platforms, some of which have a user population in the millions (and even billions), are unlikely to consider an individual user's policy concerns, even if those concerns are about a discriminatory police that reduces a marginalised group's visibility.

Users, therefore, lack the ability to register their complaints through the social network's formal processes and without participation, users are disempowered, power being understood as "an actor's capability to enact favoured decisions" in an "often asymmetrical relationship among social actors."[468] Pfaffenberger's theory of technological drama states that the final stage of the process of technology adoption is "designification" (or normalisation) where the politics of the technology is no longer visible and the technology is perceived as neutral with no competing discourses.[469] Collective actions seek to bring these discourses back to the forefront for discussion and possible revision. When they are successful, these groups have a significant impact as social media companies are so few in number that they "operate as convenient choke points under pressure."[470] This power can therefore be important although it does pose some concerns. This section will discuss

---

[467] It should be noted the underlying content was not removed but you could no longer look for all content categorised under the prohibited hashtags (the primary purpose of the hashtag system). It is worth mentioning that #straight was not banned. Gillespie, *Custodians of the internet*, 182-84.
[468] See: Laura Stein, "Policy and Participation on Social Media: The Cases of YouTube, Facebook, and Wikipedia," *Communication, Culture and Critique* 6, no. 3 (2013): 356, https://doi.org/10.1111/cccr.12026.. See also: Anthony Giddens, *The constitution of society: outline of the theory of structuration* (Berkeley: University of California Press, 1986); Manuel Castells, *Communication power* (New York: Oxford University Press, 2009).
[469] Pfaffenberger, "Technological Dramas," 308-09.
[470] Kyle Langvardt, "Regulating Online Content Moderation," *Georgetown Law Journal* 106, no. 5 (2018): 1386.

how groups engage in collective action and the possible issues that it represents.

As a result of the lack of channels to allow users to provide input and influence how platforms moderate content, an increasing number of activist campaigns have found other methods of achieving their goals. Some collective campaigns are seeking to permit a previously prohibited category of content on the platform such as the campaigns demanding that images of breastfeeding and female nipples be allowed. Other campaigns argue that a certain category of content is problematic and must be banned, such as eating disorder content and cyber-bullying. These campaigns might be protesting about social media practices but they also benefit from the advantages social media provides activists. A study of how social media was used by pro-democracy protesters in a number of countries found that it could bolster collective action by making it easier for disaffected citizens to act publicly in coordination and dramatically increasing publicity through diffusion of information to regional and global publics.[471]

Feminist groups have been particularly successful at recruiting members, securing media attention, and pressuring advertisers, often coordinating their efforts on the very platforms they are trying to change. These campaigns are often laudable, an example of what Laidlaw argues is the internet's main contribution to democracy: a facilitator of participation.[472] The first major success by one of these campaigns was when three female activists contacted advertisers on Facebook and showed them screenshots of their advertisements next to graphic content featuring rape and domestic violence.[473] Advertisers were also bombarded by over 60,000 tweets as more and more users became involved in the campaign. A number of important advertisers such as Nissan and Nationwide responded by pulling their advertisements off the platform. Facebook responded by committing itself to identifying "gender-based hate" on its platform, removing it, and improving the training of its moderators on the issue. [474] The activists were successful by raising awareness,

---

[471] Marc Lynch, "After Egypt: The Limits and Promise of Online Challenges to the Authoritarian Arab State," *Perspectives on Politics* 9, no. 2 (2011): 304-05, https://doi.org/10.1017/S1537592711000910.
[472] Laidlaw, *Regulating speech in cyberspace*, 22.
[473] Ellen Wauters, Eva Lievens, and Peggy Valcke, "Towards a better protection of social media users: a legal perspective on the terms of use of social networking sites," *International Journal of Law and Information Technology* 22, no. 3 (2014): 292, https://doi.org/10.1093/ijlit/eau002.
[474] Wauters, Lievens, and Valcke, "Towards a better protection of social media users," 292.

participating in collective action, and targeting Facebook's business interests, a very effective strategy. Klonick argues that how platforms respond to collective action can also affect their perceived legitimacy among users as these platforms, while not democratic, "arise out of a democratic culture" and borrow the language of democracy.[475]

Media attention can result in rapid changes from social media companies. For example, in 2012, a number of large outlets (such as the Huffington Post and The Atlantic) ran investigation pieces on how Pinterest, Tumblr, and Instagram all allowed large pro-eating disorder groups (often called "thinspiration" groups) to flourish. All three platforms responded by publicly announcing that they would now block searches for thinspiration and try to minimise the spread of these groups.[476] Collective groups understand the power of the media and the normative power that "the ability to speak in media systems and to influence the structure and regulation of communication resources" holds in democratic societies.[477]

Another major example of collective pressure was the "Terror of War" controversy, which was discussed in Chapter Four. After the controversy, Sheryl Sandberg publicly apologised to the Norwegian Prime Minister Erna Solberg for removing her post of the photo. The apology is particularly strange in light of the fact that Facebook had a sound justification for removing the photo and any subsequent decision to change their policies did not necessitate a public apology. It should also be noted that apologising only to the most high-profile individual in the campaign instead of all the activists who demanded the photo be available was not the highpoint of Facebook's democratic pedigree.[478] This case study provides insight into how high-profile controversies can force platforms to contort their practices, making exceptions for issues that are sufficiently popular. This may seem laudable, but the next section will discuss some of the issues with this situation.

---

[475] Klonick, "New Governors," 1653.
[476] Ysabel Gerrard, "Beyond the hashtag: Circumventing content moderation on social media," *New Media and Society* 20, no. 12 (2018), https://doi.org/10.1177/1461444818776611.
[477] Stein, "Policy and Participation on Social Media," 354.
[478] Klonick, however, characterises this tendency to react more to the opinions of famous people (rather than that of the average user) as an issue that traditional media companies also share. See: Klonick, "New Governors," 1654.

5.3.2: Issues with Collective Action

While collective actions may seem wholly positive and able to offer a powerful check on social media companies' power, these campaigns should not be treated as a reliable solution. First, these campaigns may have a democratic element (in the sense that they are majoritarian) but they cannot be construed as a complete solution for protecting human rights on any given platform. This avenue for reform is only accessible for causes that have popular support and is not a suitable substitution for human rights law, which is often employed to protect unpopular speech and minority groups.[479] Offline inequalities are accordingly replicated online as the same groups that are unable to secure resources and align social interests in their favour are equally unable to initiate change online.[480] Langvardt makes this point when he argues that "the most likely reason that Facebook's content moderation policies are so broadly accepted is that most of the burden falls on marginal or unpopular speakers—exactly the speakers whom the law of free speech is traditionally concerned with protecting."[481] Minority causes will be unable to secure the necessary resources, public support, and media attention to force a response from a social media company. As platforms grow, it can also be harder to reach that critical mass of users organising on a subject, leading to a situation where a robust activist culture is "modified or softened through the influence of new users."[482]

Second, these campaigns may achieve their stated goal but the result represents only a small victory (a proverbial Band-Aid for the problem) when what is needed is wholescale reform of the content-moderation system. A clear example of piecemeal reform was

[479] According to Alexander Bickel, appellate review in general has strong counter-majoritarian leanings. He argues that some of the most important decisions of the American Supreme Court have centred on protecting politically marginalised against majority rule. See: Alexander M. Bickel, *The least dangerous branch: the Supreme Court at the bar of politics*, 2nd ed. (New Haven: Yale University Press, 1962), 16-17.
[480] The UN Special Rapporteur also discusses the disproportionate impact of censorship on minority groups using social media. See: Association for Progressive Communications, *Content Regulation in the Digital Age: Submission to the United Nations Special Rapporteur on the Right to Freedom of Opinion and Expression* (Geneva: Office of the United Nations High Commissioner for Human Rights, 2018), 2.
[481] Langvardt, "Regulating Online Content Moderation," 1385. See also the statement made the by the Online Censorship Organisation that "Often . . . the people that are censored are also those that are least likely to be heard." "What We Do," Online censorship, accessed October 18, 2018, https://onlinecensorship.org/about/what-we-do.
[482] Mac Síthigh, "The mass age of internet law." 84.

Facebook's response to the "Terror of War" controversy, which was to introduce a general exception to their rules governing permissibility if the content in question was deemed "newsworthy."[483] Chapter Three, which focused on the creation of terms and conditions, documented how many of these content policies developed over time in a haphazard manner that was often a reaction to political events. After creation, these policies remain changeable and sensitive to public pressures, which only results in more piecemeal reform that lacks a coherent set of overarching values. It should be acknowledged that private standard-setting often entails "an open process characterised by evolutionary adjustment"[484] but these adjustments must occur in a reasoned and rational way rather than just as reactionary solutions. A consistent set of content policies that reflects established human rights standards would offer a principled approach that ensures certainty, justification, and could offer guiding principles when a social network is faced with a new challenge or unforeseen issue which might necessitate a response. This would ensure a sense of coherency to the guidelines as well, since the proliferation of activist campaigns may actually pressure platforms into enacting contradictory changes.

A cynical interpretation of how networks respond to these campaigns would also argue that acquiescing to activists on single issues that are usually low-priority for the platform (such as images depicting breastfeeding as opposed to high-priority content such as terrorist material) allow platforms to avoid larger questions about their approaches to regulation, their business models, and the relationships they build with problematic governments. Collective actions can therefore act as a "pressure valve" that stops frustration with content moderation policies building up to a critical point where users may start leaving the platform or demanding greater regulatory intervention from governments. Therefore, Youmans and York argue that "social media provide the tools for organised dissent yet can constrain

---

[483] Joel Kaplan and Justin Osofsky, "Input from Community and Partners on Our Community Standards," Facebook Newsroom, last modified October 21, 2016, https://newsroom.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards/.

[484] Weimer, "The Puzzle of Private Rulemaking," 575. Indeed, these words are echoed by Nicola Wong, a prominent lawyer who has worked at a number of social media companies, who acknowledges that online speech is going through a "norm-setting process" that changes so rapidly that that it's hard to "figure out these norms, let alone create policy to reflect them." What is notable, however, is that platforms are able to reflect rapidly-changing norms much more effectively than public bodies. Klonick, "New Governors," 1628.

collective action."[485]It might, therefore, be in the interests of platforms to occasionally be seen to concede to popular demands about low-stakes issues so that users are assuaged.

This section has clearly demonstrated that collective action is an insufficient method for initiating widespread change on social media. Activism in of itself cannot remedy the human rights issues in social media because as a solution, it lacks source, outcome, and process legitimacy.[486] What is required is solutions that offer greater accountability. Morozov sums up the deficit that online activism embodies when he argues that the internet (particularly social media) "may have made the revolutions of the Arab Spring possible, but "the Internet"—at least the blind, unquestioning faith in the superiority of decentralised and horizontal networks—is making those revolutions very difficult to complete."[487] A similar point is made by Mac Síthigh when he writes "the particular issue with these spaces is that they create the *impression* of democratic participation while offering no guarantee as to the *realisation* of this goal."[488] The next section will discuss how we can move beyond spotty appeals systems and haphazard collective action towards a viable set of procedural protections and remedies for social media users that improves certainty, accountability, and legitimacy.

5.4: Solutions

There are a number of changes that platforms must introduce to remedy the wide array of problems that current exist in the response stage of the content-moderation process. This section will make a number of proposals to ensure improvements in what the UN Guiding Principles terms "access, procedures and outcomes" of the "remedial process."[489] The proposals will include a stronger appeals process, a forum for participation, and an industry-

[485] Youmans and York, "Social Media and the Activist Toolkit," 316.
[486] Brown and Marsden, *Regulating code*, 373.
[487] Morozov, *To save everything, click here*, 128.
[488] Mac Síthigh, "Virtual walls? The law of pseudo-public spaces." 400.
[489] *UN Guiding Principles.*

wide appeals mechanism. These reforms will complement the larger, more radical suggestions made in Chapter Seven.

5.4.1: Reforming the Internal Appeals System

Social networks must develop robust internal appeals processes that rectify some of the current deficits of platform appeals. A good starting point would be the Santa Clara principles, which were created by academics, industry representatives, and NGO's as general guidelines for content moderation. These principles are quite brief (and do not explicitly discuss human rights) but they do recommend that appeals include "human review by a person or panel of persons that was not involved in the initial decision, an opportunity to present additional information that will be considered in the review, notification of the results of the review, and a statement of the reasoning sufficient to allow the user to understand the decision."[490] Users need to be provided with information about the appeals process in a centralised location that they can access even when they're not currently disputing an enforcement decision. When content they have posted is removed, users need to be provided with information about the removal decision (including reasons) and how to appeal that decision. If a user has flagged content, a platform should also contact them when a decision is made to inform them about the outcome and the reason for that decision. Users who have flagged content should also be given the option to appeal a decision not to remove this content as it is possible that an error was made in the original decision and problematic content is being retained on the platform. By providing all of this information, platforms are not only making a bid for legitimacy and accountability but they will also be incentivised to improve their policies and processes. Integrating due process principles into a private regulatory scheme can consequently lead to stronger and more efficient standards.[491]

The possible remedies available should also be expanded beyond reinstatement (or removal) of the content or profile. The UN Guiding Principles identifies a litany of potential remedies including: "apologies, restitution, rehabilitation, financial or non-financial

[490] "The Santa Clara Principles: On Transparency and Accountability in Content Moderation." "The Santa Clara Principles: On Transparency and Accountability in Content Moderation."
[491] Bellis, "Public law and private regulators in the global legal space." 435.

compensation and punitive sanctions (whether criminal or administrative, such as fines), as well as the prevention of harm through, for example, injunctions or guarantees of non-repetition."[492] While not all of these remedies may be feasible (or desirable) in the case of social networks, the current system should be broadened to include a diverse array of options. Compensation should be available for monetised accounts which have had content removed (especially if it was during the peak earning period). The platform could publicly acknowledge an error by maintaining a correction register or affixing a symbol to the content to indicate that it was wrongfully removed. This is similar to the corrections that are mandated by press regulators in the UK such as IMPRESS and IPSO.[493] IMPRESS even points out in its guidance on corrections that in high-profile cases, corrections should be "pinned" to the top of online editions for a reasonable amount of time so that newer stories do not bury these corrections.[494] These public acknowledgements would also provide important data for academics and journalists trying to understand social media content moderation. It could also be possible to tweak newsfeed algorithms to ensure that wrongfully removed content, once reinstated, has a prominent place so that it will be highly visible to others and that it might recoup some of the traffic it was denied earlier.

When considering what principles matter in an internal appeals system, the UN Guiding Principles provides a set of factors that determine the effectiveness of what they term "Non-judicial grievance mechanisms." Article 31 of the Principles states that these systems should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, a source of continuous learning, and based on engagement and dialogue.[495] Platforms should assess their appeals systems against these principles and provide transparency reports on how their processes measure up and how they will rectify any deficits. These principles provide

---

[492] *UN Guiding Principles.*

[493] IMPRESS focuses more on corrections as their primary remedy whereas IPSO also includes "a private letter of apology from the editor, an undertaking as to further conduct by the newspaper" or, in cases that involve legal claims, the option of a "low-cost arbitration scheme" where complainants can be awarded up to £60,000. See: Impress, "Guidance on the Impress Standards Code," 2020, https://www.impress.press/downloads/file/impress-code-guidance-2020.pdf. IPSO, "Complaints-Frequently Asked Questions," 2018, https://www.ipso.co.uk/faqs/complaints/#what-are-some-of-the-ways-that-my-complaint-might-be-resolved.

[494] Impress, "Guidance on the Impress Standards Code."

[495] *UN Guiding Principles.*

a common language that all platforms can use to communicate their findings as well as a set of values that critics of social networks (such as Onlinecensorship.org) can use to evaluate a platform's progress. Any appeals system that embodies these eight factors is likely to be a robust process that forms a strong foundation for user/platform interaction.

5.4.2: Forums for Participation

As was discussed in Chapter Three, social media companies need to create a space on the platform where users can discuss broader issues they have with the network's policies and processes of content moderation. This forum would signal to users that platforms value their feedback and want them to participate in platform governance. It would reaffirm the fact that regulatory systems that embody rule of law principles have an element of discourse or argumentation and treat "ordinary citizens with respect as active centres of intelligence."[496] This would also be beneficial to the companies because they would be able to identify weak spots in their current approach and address issues before they begin to generate negative publicity. In addition, the more platforms are able to shift the public campaigning element of collective activism onto a designated space on the platform, the more legitimacy the platform will gain, specifically input legitimacy, which "concentrates on the participatory nature and inclusiveness of the norm-creation process."[497] Participation and legitimacy are inherently connected, as attention should be paid to "the levels of participation that regulatory decisions and policy processes allow to the public, to consumers, and to other affected parties" because of its "legitimating effect."[498]

Facebook is the only major platform which has created a forum where users can comment and vote on proposed policy changes.[499] This page, named 'Facebook Site Governance' was launched on the 7th of April 2009. The 2009 video where Mark Zuckerberg

---

[496] Waldron is expanding on MacCormick's idea that law is an argumentative discipline and that this is an important aspect of the rule of law. Waldron, "Rule of law." 20, Neil MacCormick, *Rhetoric and the Rule of Law: A Theory of Legal Reasoning* (Oxford: Oxford University Press, 2005), 14-17.
[497] Reed and Murray, *Rethinking the jurisprudence of cyberspace*. 174, Phillip Paiement, "Paradox and Legitimacy in Transnational Legal Pluralism," *Transnational Legal Theory* 4, no. 2 (2013), https://doi.org/10.5235/20414005.4.2.197. 213-215.
[498] Baldwin and Cave, *Understanding regulation*, 79.
[499] See: "Facebook Site Governance," Facebook, accessed October 12, 2018, https://www.facebook.com/fbsitegovernance.

introduces the Site Governance model references a number of important values that this thesis has argued are lacking in social media content moderation. These values include transparency, participation, and responsible governance.[500] The process would start with a user liking (or in the terminology of Facebook back in 2009, 'fan-ing') the relevant Facebook page and would then be kept appraised of new policies introduced by Facebook.[501] The novel aspect of this proposal was that if a new policy received more than 3000 comments then it would go to a vote on whether the policy should be vetoed but it required 30% of all Facebook users to participate in the vote to make it binding.[502] This experiment ended in 2012 after a proposed policy received just 0.038 percent participation and Facebook eliminated the voting system.[503] Facebook also seemed to have lost interest in maintaining the Site Governance page. Policies were posted quite frequently at the beginning but in the last three years only one set of changes has been posted (on 4 April 2018) and as voting has been discontinued, this policy only invited users to comment.[504] Even this policy, a Data Privacy update introduced only eighteen days after The Guardian broke the explosive Cambridge Analytica story, garnered only 331 comments and 236 shares, a miniscule percentage of the users on Facebook in 2018.

There are a number of reasons why the Site Governance experiment failed. First, and arguably most importantly, was a lack of publicity. Most users were and remain completely unaware of the Site Governance page (which still exists on Facebook). Of the 2.3 billion of users who hold Facebook accounts, only three million users have liked and followed the page.[505] This feature is so niche that it is rarely mentioned even in scholarship focusing on Facebook. It is surprising that Mark Zuckerberg does not mention this forum in more of his

---

[500] Facebook, "Mark Zuckerberg: Vote on Facebook Site Governance."
[501] Facebook, "Mark Zuckerberg: Vote on Facebook Site Governance."
[502] Gillespie, *Custodians of the internet*. 209-210.
[503] Casey Johnston, "Whopping 0.038% of Facebook Users Vote on Data Use Policy Change," ArsTechnica, last modified October 18, 2018, https://arstechnica.com/information-technology/2012/06/whopping-00038-of-facebook-users-vote-on-data-use-policy-change/.
[504] The post also stated "Once finalised, we'll notify you and ask you to review our updated Terms and Data Policy" which seems to indicate that Facebook's grand democratic experiment was now reduced to another "click-through and accept" exercise. See: "Facebook Site Governance post on Data Policy," Facebook, last modified April 4, 2018, https://www.facebook.com/fbsitegovernance/.
[505] The exact number, as of 12 October 2018, is 2,997,836 people.

Facebook posts and that so few users have heard of it.[506] The Site Governance experiment would be the equivalent of a previously autocratic state introducing municipal elections but failing to invest any resources into making citizens aware of their new democratic rights and the existence of an electoral process. One must query whether this feature was an experiment in participation that was designed to fail, or at the very least was so under-resourced that its lack of impact was inevitable. Second, the Site Governance model introduced by Facebook may have simply expected too much from its users. The policies that Facebook posted on the site were lengthy and overly technical[507] and would have proved discouraging for many users to understand even if they were willing to make the time commitment. Finally, the required numbers of user participation were too high when considered in combination with the poor publicity and difficult commitment that the Site Governance model represented. Today, Facebook has over two billion users so it would be difficult to presume that users would be able to marshal campaigns and attain the participation required to trigger a response from Facebook but it seems an impossible task even in 2009 when Facebook had 200 million users.[508]

Platforms should use the Site Governance experiment not as a cautionary tale against participation but rather as a case-study replete with valuable lessons they could employ when developing their own participatory forums. Indeed, Gillespie argues that the Site Governance model could have worked if Facebook had engaged in improvements by "expanding participation, earning the necessary legitimacy, developing more sophisticated forms of voting, and making a more open process."[509] Perhaps instead of a voting threshold, platforms could just solicit votes and opinions on summaries of proposed changes. Social networks should also create mechanisms for consultations on the forum and encourage the creation of working groups and dedicated pages for discussing specific policy and process changes. The creation of a participatory forum is one of the simplest changes that this thesis proposes and this would help enhance the legitimacy of the platforms in question.

---

[506] Johnston argues that "While Facebook didn't make its policy changes a secret, it scarcely tried to bring it to users' attention." See: Johnston, "Whopping 0.038% of Facebook Users Vote on Data Use Policy Change."
[507] Gillespie, *Custodians of the internet*, 250-51.
[508] Facebook, "Mark Zuckerberg: Vote on Facebook Site Governance."
[509] Gillespie, *Custodians of the internet*, 209-10.

While the existence of a forum for participation would be a positive addition to the online environment, this would not be a full remedy to the issues explored in this chapter, even if it was coupled with an internal appeals process. Stein would characterise participation forums on social networks as a "consultation mechanism" which (as opposed to stronger forms of participation) "allow users minimal influence or control, but only at the discretion of the platform owner."[510] The ability of users to effect change is thus, highly conditional on platform acquiescence and cannot act as a complete safeguard for important human rights considerations. A greater assurance of accountability and oversight is required as the human rights issues posed by social networks become more pressing every year. The next suggestion will explore how these safeguards can be integrated into the current approach to content appeals.

5.4.3: Industry-wide Appeals Mechanisms

Currently, any appeals decision offered by a platform is final and users cannot appeal further.[511] Their only other option is to engage in campaigning and attempt to get media attention, but many users will be unwilling or unsuccessful at employing this strategy. The lack of a higher body to adjudicate moderation decisions is problematic because users have no real assurance that the platform is behaving impartially or that an appeal constitutes anything more than a rubber-stamping of the original decision made by the platform. This concern was echoed by the Guiding Principles, which concludes by stating "Since a business enterprise cannot, with legitimacy, both be the subject of complaints and unilaterally determine their outcome, these mechanisms should focus on reaching agreed solutions through dialogue. Where adjudication is needed, this should be provided by a legitimate, independent third-party mechanism."[512] An independent appeals body could make assessments using a modified form of judicial review and explore how a decision was reached in an impartial manner.[513] Ideally, the decisions it would make would also be

[510] Stein also characterises appeals processes as consultation mechanisms as there is still a substantial power asymmetry between the one seeking an appeal and the decision-maker. Stein, "Policy and Participation on Social Media," 360.
[511] With the exception of the Facebook Oversight Board which will be discussed later in this section.
[512] *UN Guiding Principles.*
[513] A. W. Bradley and K. D. Eving, *Constitutional and administrative law*, 15th ed. (New York: Pearson Longman, 2011), 613.

applicable to systemic issues on the platform and could help create a measure of consistency in how platforms employ content moderation.

There should be an industry-wide appeals mechanism for social media platforms, a tribunal funded by the companies to hear complaints once users have exhausted internal appeals. This suggestion has also been raised by David Kaye, whose report on free expression online raised the possibility of "company-specific or industry-wide ombudsman programmes" such as "an independent 'social media council,' modelled on the press councils that enable industry-wide complaint mechanisms and the promotion of remedies for violations" which could hear individual complaints as well as investigate more systemic issues.[514] Press Councils are a good model as they give "legalistic treatment" to rules "that belong in the realm of ethical/moral and/or professional conduct and not in that of ordinary law."[515] Another source of inspiration could be the ICANN (Internet Corporation for Assigned Names and Numbers) dispute resolution policy, which outlines a number of procedural steps that claimants should take when initiating a dispute.[516] This appeals mechanism would only be available for users who have already used the internal appeals methods at platforms. They would then make an application to the appeals mechanism, filling in a short online form with the relevant information and their representations. Of course, it is important to "build disincentives into the framework to dissuade the casual complainer." [517] The appeals mechanism would therefore include an initial assessment stage to consider the substantiality of the claim to exclude claims where there is no prospect of a successful appeal (such as pornographic content). After meeting this substantiality threshold, there would be an option for mediation with the platform or it would pass on to the adjudication stage.[518]

An industry-wide appeals board would have a number of advantages: they are capable of efficiently handling a large volume of cases (as compared to a normal court) and they can provide adjudicators that have specialised knowledge in the relevant subject.

---

[514] Kaye, *A/HRC/38/35*, 18.
[515] Tambini, Leonardi, and Marsden, *Codifying cyberspace*. 74.
[516] ICANN, "Uniform Domain Name Dispute Resolution Policy," last modified October 24, 1999.
[517] Laidlaw, *Regulating speech in cyberspace*. 262.
[518] A similar model is proposed by Laidlaw for considering human rights issues in the digital world. See: Laidlaw, *Regulating speech in cyberspace*. 261-262.

Creating such a board would also signal to the public that platforms valued their users and were committed to creating transparent systems that would benefit them. Platforms would also be able to use the board to share best practices and to assist emerging social networks in developing their appeals systems. By focussing on a judicial review-type approach (where decisions need only fall within a range of reasonable results based on the policies and processes at that particular platform), platforms would also be able to maintain a measure of discretion on what values they encode into their platforms. An appeals board would also strengthen a platform's commitment to the rule of law as Waldron has written that one of the key protections for a procedural rule of law is the right to appeal to a higher tribunal.[519] He claims that "law comes to life in institutions," and this is arguably similar in relation to other types of regulators.[520]

Users would also benefit from an appeals board. First, this board could become a location for contestation, where would-be reformers could focus their efforts and make arguments about why platforms should change their rules. Second, this appeals board could help redress the balance of power between users and platforms by allowing users to make representations to a board that is separate from the moderation structure of a particular platform. This is also a more accessible solution than encouraging users to use the court system as the appeals board could accept digital applications and users would not be barred from making appeals because of location or financial resources. An independent appeals board could also help limit the amount of inappropriate government pressure placed on platforms as some states may be concerned that their attempts to co-opt these platforms into assisting them with human rights violations (such as handing over information about anti-government activists) would emerge at the appeals board. Finally, a second level of appeals would also provide an extra level of assurance to groups who are particularly affected by poor appeals decisions, such as activists and entrepreneurs. The creation of a separate tribunal would be a positive step for social networks and would offer users a higher degree of certainty, increased transparency, and heightened accountability.

---

[519] Waldron, "Rule of law," 4.
[520] Waldron, "Rule of law," 12.

This social media appeals board would, therefore, transform the current single-appeals systems at platforms into an interlocking set of review mechanisms. It would offer assurances to users of impartiality and accountability, and would be a positive publicity move for platforms as well as an opportunity to share best practices. This expanded appeals system will sit within a larger set of regulatory reforms that will be outlined in Chapter Seven.

5.4.4: Case Study: Facebook Oversight Board

Facebook has already debuted a second level of appeals, an idea that was first suggested by Mark Zuckerberg in 2018 as a "supreme court" for Facebook.[521] The Facebook Oversight Board (which will be composed of forty experts in relevant fields from around the world) considers individual appeals from users who have had content removed and exhausted earlier appeals and also "significant and difficult" cases referred by Facebook itself.[522] Facebook can also request a policy advisory statement from the board clarifying a previous decision or providing guidance on possible changes to Facebook's policies.[523] The decisions made by the Oversight Board would be binding except for the policy advisory statements.[524] The Board will also make all decisions publicly available in a database of case decisions (which was an idea touted by Chapter Four of this thesis) and will release annual reports that detail number and type of cases, case summaries, the region and source of referral, the international human rights issues in the cases, and how Facebook has implemented their decisions.[525] The Oversight Board is a very interesting development in the world of social media regulation. It is also highly unusual, with BSR explaining that "to our knowledge, no company in any industry has ever established an oversight mechanism

---

[521] Alex Hern, "Facebook among 30 organisations in UK political data inquiry," The Guardian, last modified April 5, 2018, https://www.theguardian.com/technology/2018/apr/05/facebook-mark-zuckerberg-refuses-to-step-down-or-fire-staff-over-mistakes.
[522] Brent Harris, 'Preparing the Way Forward for Facebook's Oversight Board' Facebook Newsroom. 28 January 2020. https://about.fb.com/news/2020/01/facebooks-oversight-board/ Accessed 25 March 2020.
[523] Facebook, "Oversight Board Charter," 2020, https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf. 7.
[524] Facebook, "Oversight Board Charter." 3,7.
[525] Section Four: Transparency and Communications. Facebook Oversight Board, "Bylaws," 2020, https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf. 15.

with binding decision-making power."[526] Facebook has outlined a mechanism that promises to boost transparency, accountability, and legitimacy of its services, offering users an impartial second opinion from a diverse group of reputable experts.

There are, however, some concerns about the Oversight Board that can be identified at this stage, although some of these issues may be ironed out when the Board is fully operational. First, there is a scale problem. At its capacity, Facebook's board will only be comprised of forty individuals working in a part-time capacity and it is hard to predict how many cases they will be able to handle. In the first quarter of 2019, there were 25 million appeals against content removal on Facebook.[527] The Oversight Board, therefore, will have to ensure that the decisions they make will be applicable across large content sets or else their decisions will be symbolic gestures. Facebook has committed to "undertake a review to determine if there is identical content with parallel context" that remains on Facebook after the Board has recommended removal and "take action on that content" if its technically feasible.[528] It is difficult at this juncture, however, to predict whether this will be an adequate solution. Another issue is that at the Board will only make decisions about whether content should remain or be removed, which reaffirms social media's persistent fixation on deletion as the only appropriate recourse. The Bylaws do state that in the future "subject to Facebook's technical and procedural improvements" the Board will be able to assess other kinds of enforcement actions and new services, but there is no definite timeline for this change.[529]

The most significant issue is that it is hard to tell what role human rights will play in the Board's decisions. While the Charter and Bylaws do reference human rights (such as in the transparency provision quoted above) it is hard to tell the effect in practice. The Charter clearly states that "Facebook has a set of values that guide its content policies and decisions. The board will review content enforcement decisions and determine whether they were consistent with Facebook's content policies and values."[530] This thesis has consistently

[526] BSR, *Oversight Board Review*. 4.
[527] BSR, *Oversight Board Review*. 10.
[528] Board, "Bylaws." 21.
[529] Board, "Bylaws." 16.
[530] Facebook, "Oversight Board Charter." 5.

argued that Facebook's values occasionally align with human rights principles but are just as likely to come into conflict. By privileging Facebook's values as the primary standard of review, the Charter and Bylaws fail to explain the appropriate course of action when these values come into conflict with human rights. This is also apparent when the Bylaws state what kinds of cases will not be eligible for the Board to review. These cases include ones "where the underlying content is criminally unlawful in a jurisdiction with a connection to the content" and a Board decision could result in either criminal liability or "adverse governmental action against Facebook."[531] This means that the Board will not review cases where content was posted in countries that have criminalised certain forms of speech (such as blasphemy or lèse-majesté), regardless of whether these laws would comply with international human rights standards. The predicted consequences that preclude eligibility are also broad, with "adverse governmental action against Facebook" indicating that any case that might result in Facebook suffering any harm to its business will be precluded from oversight. This means that the Oversight Board risks becoming a mechanism that can only review cases from countries that already have broad free speech protections and where people would have other access to other forums for sharing their opinions. This would just reaffirm the status quo that already exists offline, negating the democratising effect of these technologies. It also makes one wonder if the much-touted diversity of Oversight Board members will be largely symbolic as the Board will primarily be considering cases from Western liberal democracies.

While Facebook has signalled a commitment to human rights in the creation of the Oversight Board, some of its decisions run contrary to this aim. Facebook hired BSR (Business for Social Responsibility) to conduct a Human Rights Impact Assessment (HRIA) on the Oversight Board. This shows an admirable commitment to transparency and human rights but ultimately Facebook chose to release a final Charter without waiting for the completion of the HRIA or consulting with BSR.[532] One wonders, therefore, if the HRIA exercise was more of a public relations exercise than a genuine bid to improve their processes. The final human rights issue is that while occasionally referencing "international

[531] Board, "Bylaws." 17.
[532] Facebook, "Oversight Board Charter." 9.

human rights," most of the references to rights in the Charter and Bylaw are confined to free expression. In fact, the Bylaws start by saying "the purpose of the Oversight Board is to protect freedom of expression" and the Charter states that the Board will "pay particular attention to the impact of removing content in light of human rights norms protecting free expression."[533] It seems that the decisions the Board makes will balance Facebook's values and free speech against each other as opposed to the more nuanced rights-balancing exercises that would be required.

The Oversight Board is an interesting idea but it might even be more effective if the oversight board covered more than just Facebook and Instagram or if this board existed *in addition* to the industry-wide appeals mechanism for a number of reasons. First, an industry-wide appeals board would make clear that the importance of appeals decisions transcends any particular platform and that the decisions made by platforms should not be construed as entirely within their discretion to make. Second, an appeals board could help identify problematic trends or best practices across the different platforms, and would therefore have more of an impact on improving content moderation practices. Third, by creating an industry-wide board, platforms would be able to pool resources and create a larger body that is capable of handling a high volume of cases. The progress of the Oversight Board, however, will be monitored with great interest and is sure to have an impact on how Facebook governs its platforms.

5.5: Conclusion

One of the biggest misconceptions that people can make when thinking of social media is to consider the activities that occur on these platforms as trivial and undeserving of legal protection. This view does not reflect the rich array of activities that occur on social media. Users engage in activism and citizen journalism, create and promote businesses, share thoughts and opinions about subjects that are important to them, and connect with other groups and people around the world. Social media comprises a rich landscape of

[533] Board, "Bylaws." 5. Facebook, "Oversight Board Charter." 5.

human activity and the interference with these activities (regardless of whether the platform is right to do so) can be distressing to users. They deserve the benefit of a set of interlocking review mechanisms that they can appeal to when they feel their rights have been violated. Users can then connect their concerns with a broader discourse about human rights and procedural due process, what Citron calls "a common structure for debating and addressing concerns about the propriety of administrative actions."[534] A regulatory body may still decide that the enforcement decision was valid but users will have an assurance that they can seek a second (and even third) opinion to ensure that the decision was justified. These reforms would have a significant impact on how social networks engage in content moderation and would empower users in their dialogues with the platforms.

This chapter has discussed the response stage of the content moderation process: how users respond to the terms and conditions that are created and the enforcement decisions of a platform. These responses can be categorised into two general categories: responses that occur within the structured processes of the platform (primarily the appeals system when an enforcement action is taken) and responses that come from outside the platform (most notably collective actions by campaigners). The external campaigning may be perceived as remedying the lack of participation in the platform (just a robust appeals system is perceived as a mitigating factor for a content moderation system that has procedural flaws)[535] but this perspective is overly simplistic. Every stage in the content moderation process must be transparent, accountable, and mindful of the companies' human rights obligations. This is echoed by one of the reports written by David Kaye on social media and free expression, where he called for "radical transparency, meaningful accountability and a commitment to remedy in order to protect the ability of individuals to use online platforms as forums for free expression, access to information and engagement in public life."[536] There will be errors of course, but the policies and processes in place should be structurally sound. A strong appeals system is not enough if users must grapple with uncertain rules, inconsistent enforcement, and a general feeling that the platforms are acting

[534] Kaye, *A/HRC/38/35*, 1252.
[535] Think, for example, of YouTube's encouragement that users who felt that that the flawed algorithms had demonetised their accounts should appeal in order to help make these systems better.
[536] Kaye, *A/HRC/38/35*, 19.

in an arbitrary way. Collective actions can allow users to participate in important discussions about digital citizenship and corporate responsibility but it is not a panacea. It should never be more than a complement to a robust internal appeals process and a strong commitment by platforms to align their practices with human rights and the rule of law. The solutions have encompassed improving internal appeals, creating forums for participation, and creating a new appeals body. These solutions not only benefit users but also the platforms themselves by enhancing their legitimacy.

The previous three chapters, including this one, have explored the stages of the content moderation process: creation, enforcement, and response. Each chapter has identified a number of serious problems with each of the stages and offered some specific solutions on how these issues can be remedied. The proposals however are narrow in scope and the issues that they seek to address could rightly be construed as symptoms of a more general condition, a structural defect, of social media content moderation. This is the issue of accountability. Whatever changes are suggested and whatever new measures are introduced by platforms will not remedy the rule of law and human rights issues that have been identified here unless there is an assurance of accountability. The next chapter will evaluate a number of proposals made by other academics and regulators in the field. The final chapter of this thesis will then go on to outline a new proposal for how these platforms can be held accountable for the consequences of their content moderation practices.

Chapter Six: Other Proposals

6.1: Introduction

In 2006, after existing for a year, YouTube had 20 million unique users, 100 million videos a day were being watched, and the platform accounted for 60% of all videos viewed online.[537] Despite these high numbers, YouTube had a content moderation team of only ten people, who had to work in shifts around the clock.[538] Their internal moderation guidelines amounted to a single piece of paper, folded in half, with a bullet-pointed list of things to remove. When moderating they asked themselves "can I share this video with my family?" as a benchmark for moderation. Buni and Chemaly observed that "this small team of improvisers had yet to grasp that they were helping to develop new global standards for free speech."[539] This anecdote illustrates a number of issues in how content is moderated by social media companies. First, content moderation is frequently treated as a secondary concern (even though it is fundamental to a site's continued popularity and ability to attract investment) with serious issues in staffing and labour conditions.[540] Second, moderators are referred to as a "team of improvisers" and that same pattern of behaviour still pervades content moderation: rules are developed or services are introduced with little consideration of the risks they pose or how negative consequences could be mitigated.[541] Finally, moderation is equated simply with removal, with no acknowledgement of the complex rights-balancing exercises that content regulation should reflect.

Social media content moderation is a relatively new phenomenon and academic interest in how companies can better regulate content is growing.[542] Postman writes that "it is inescapable that every culture must negotiate with technology, whether it does so

---

[537] Reuters, "YouTube serves up 100 million videos a day online," USA Today, last modified July 16, 2006, https://usatoday30.usatoday.com/tech/news/2006-07-16-youtube-views_x.htm.
[538] This anecdote is taken from Catherine Buni and Soraya Chemaly. Buni and Chemaly, "Secret Rules of the Internet."
[539] Buni and Chemaly, "Secret Rules of the Internet."
[540] See, for example, Newton, "Bodies in Seats: At Facebook's worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives."
[541] See, for example, the live-streaming discussion at 4.3.2 or this chapter's discussion of Myanmar at 6.2.3.
[542] With works by Tarleton Gillespie, Kate Klonick, and Sarah T. Roberts are of particular significance. See: Gillespie, *Custodians of the internet*; Klonick, "New Governors."; Roberts, *Behind the screen*.

intelligently or not."[543] The same could be said of content moderation: there are very few people today who believe that social media content should not be moderated. The true challenge is determining how these processes should occur in a way that maintains the strengths of private regulation while still respecting the rights that moderation processes can imperil. The objective of this chapter, and the next, is to identify solutions to this problem.

This chapter will discuss a number of legal solutions that have been proposed to address the issues that exist in the social media content moderation process and which we have discussed in earlier chapters.[544] Due to the variety of proposals that exist, the chapter is by no means an exhaustive survey of all options.[545] Instead, these three sections represent three kinds of regulatory arrangements that could be imposed on social media companies: self-regulation, direct regulation (substantive regulation) and co-regulation (the proposed duty of care). The various proposals will be organised by the level of state interference they require, starting with the least amount of interference (self-regulation) and then moving on to consider substantive regulation, and ending with the newest and most radical proposal: a

---

[543] Postman, *Technopoly*, 5.

[544] These issues are content moderation is not being treated as seriously as it should by platforms and the policies and processes that are being put in place are arbitrary, uncertain, both under- and over-regulate certain areas, and do not reflect human rights principles of the spirit of the rule of law.

[545] For example, this chapter will not consider addressing social media content moderation issues through consumer protection or contract law. These suggestions were researched but were ultimately discarded so that the chapter would be able to explore a number of proposals in greater depth. Both consumer protection and contract enforcement place a disproportionate level of responsibility on individual users to understand and enforce their rights through the legal system. In consumer protection situations, users will often be disinclined to bring claims because it will be time-consuming, costly, and the financial penalties initiated against powerful companies will be relatively small. Users, may therefore, feel that these claims represent a proverbial 'drop in the bucket.' There are academics, however, who do not argue for a direct application of consumer protection laws to the harms arising from using the internet but do suggest there should be more "thoughtful conversation and translation between two bodies of law that have a common history and more in common than scholars and lawyers sometimes realise." Grimmelmann suggests that since consumer protection is an area of law focussed on protecting people from unnecessary risks then it could be used as a tool to address social media harms such as privacy issues. While this is an interesting exercise, it is at best a nascent solution, one that would require further exploration and interpretation before it could be practically implemented. Contract law is also problematic because it presumes the principle of freedom to contract, which is based on the idea that parties are free to negotiate the terms of the contract. Social media platforms, however, offer terms and conditions which are drafted unilaterally and by the stronger party in the relationship. This situation weakens the claim that contract law can effectively protect users and act as a check in the activities of platforms. For further reading on the role of consumer protection and contract rights in social networking, please consult: James Grimmelmann, "Privacy as Product Safety," *Widener Law Journal* 19 (2010); Wauters, Lievens, and Valcke, "Towards a better protection of social media users."

social media duty of care. Ultimately, these three solutions will all be found to be incomplete or otherwise unsatisfactory and so the next chapter will offer an entirely new approach to resolution of the issues that are at the core of this thesis.

6.2: Self-Regulation

6.2.1: The Advantages in a Self-Regulatory Scheme to protect human rights

Self-regulation was, until recently, seen as the obvious choice to encourage companies to respect human rights. It is "intrinsically linked" to the concept of Corporate Social Responsibility (CSR) because it is typically deployed to achieve CSR objectives.[546] The concept of self-regulation encompasses a number of different regulatory arrangements that are aligned along a spectrum based on the degree of legislative constraint, outsider participation in rule creation and enforcement, and accountability.[547] At one end of the spectrum is industry regulation by a collective body made of representatives of that industry (collective self-regulation) and at the other end of the spectrum is the situation where a single company governs itself (individualised self-regulation).[548] This section will focus on individualised self-regulation (and will refer to it as self-regulation) because of its applicability to social media companies. Whilst collective self-regulation is typical in many industries, social media platforms largely govern themselves internally, bounded only by direct regulation on certain categories of content (such as Child Sexual Abuse Material). This is typical of internet regulation more generally, where self-regulation often means "private actors' CoC's [codes of conduct] with very little or no sanction for non-compliance."[549]

---

[546] Laidlaw, *Regulating speech in cyberspace*. 67.
[547] Anthony Ogus, "Rethinking Self-Regulation," *Oxford Journal of Legal Studies* 15, no. 1 (1995): 99-100, https://doi.org/10.1093/ojls/15.1.97.
[548] Julia Black, "Constitutionalising Self-Regulation," *Modern Law Review* 59, no. 1 (1996): 26-27, https://doi.org/10.1111/j.1468-2230.1996.tb02064.x.
[549] Tambini, Leonardi, and Marsden, *Codifying cyberspace*, 29.

Self-regulatory initiatives offer some important benefits for states, companies, and other actors who may be affected.[550] States benefit from companies bearing the compliance costs of human rights initiatives and the probability that companies will adhere to the self-regulatory schemes they create for themselves.[551] There is also a certain logic in placing the burden of regulating these activities on the parties that benefit from it. From a practical perspective for states, self-regulation also recognises that the major social media platforms have significant resources (including a pool of intellectual talent) and the ability to deploy these resources in an efficient, scalable way. Generally, content moderation processes at social media result in very quick responses even though they are handling a large volume of content, an efficiency that is unmatched by other regulators.[552] While currently these processes are not entirely human rights-compliant, the same efficiency could be advantageous in deploying a content moderation system that respects human rights. It is clear that some element of self-regulation will always be necessary to ensure that these moderation processes remain responsive and it is entirely plausible that this system could be aligned along stronger principles that address the issues identified by this thesis.

Activists and the wider public also derive benefits from self-regulatory schemes. These codes act as a set of promises from companies that stakeholders can use to make their demands. Advocates and rightsholders can structure their demands from companies using the promises made in these codes.[553] The most obvious advantage to concerned parties is that companies will perceive the policies they develop as a legitimate expression of their objectives and might be more inclined to adhere to these policies. Another benefit that self-regulation offers is that these policies may offer more protection and higher standards in jurisdictions with weak legal controls, thus providing a tangible benefit to the affected

[550] Of course many of the same advantages would be offered by a mandatory human rights due diligence scheme as well. This will be outlined in the next chapter.
[551] Whether this increased compliance is actually true is hard to determine but this perception seems like a reasonable proposition to states who are engaged in solving a multitude of governance problems and rely on companies to design their own human rights programmes.
[552] Of course, there is a danger in attributing an excessive amount of value to efficiency as that can contribute to the efficiency narrative discussed at 4.4.2. This paragraph, however, is merely identifying the advantages social media platforms have as regulators and one of those advantages is the ability to handle a high volume of content and respond very quickly.
[553] Thomas, *Public rights, private relations*, 6.

parties.[554] In theory, if companies were to prioritise strong human rights protections through their codes of practice, then they could have an impact on the lived experience of whole societies even where the relevant national government was unwilling or unable to offer similar levels of protection. It should be acknowledged that corporate cooperation is an invaluable asset in any attempt to address the human rights implications of private companies. Dan Danielson argues that corporate governance is often misunderstood and underestimated, and that corporations can affect the creation, interpretation, and application of legal regimes to such an extent that they "produce effects on social welfare similar to the effects resulting from rule- making and enforcement by governments."[555] The power of platforms to dictate rules to other bodies is already evident. Media organisations, for example, attracted to the volume of activity and high degree of user attention on platforms, have aligned their rules and activities with social media terms and conditions.[556]

This is one of the primary advantages of individualised self-regulation: a company can enact policies that drastically, and positively, change the conditions people live in, regardless of what the laws of that country dictate.[557] An example of this ability to create human rights "facts on the ground" is the actions of PepsiCo. In America, Pepsi responded to the Civil Rights movement by becoming the first major American corporation to hire an African-American as a company executive and altering its hiring practices at every level of the business to encourage more minority candidates and reduce discrimination.[558] In South Africa, Pepsi went further, attempting to navigate the inherent injustice of the apartheid regime by paying black employees above-average pay and creating opportunities for career advancement.[559] When Pepsi divested from South Africa in 1984 (one of the first American companies to do

---

[554] Dan Danielson, "How Corporations Govern: Taking Corporate Power Seriously in Transnational Regulation and Governance," *Harvard International Law Journal* 46, no. 2 (2005): 424.
[555] Danielson, "How Corporations Govern," 412.
[556] Priyanjana Bengani, Mike Ananny, and Emily J. Bell, *Controlling the Conversation: The Ethics of Social Platforms and Content Moderation* (New York: Columbia University - Tow Centre for Digital Journalism, 2018), 10.
[557] Of course, this is also an accurate description of how companies can negatively impact people when they engage in practices that constitute human rights abuses but this section is considering the positive side of this possibility.
[558] John Kirby Spivey, "Coke vs. Pepsi: The Cola Wars in South Africa during The Anti-Apartheid Era" (Master of Arts (MA) Master Thesis, Georgia State University, 2009), 19-20.
[559] Spivey, "Coke vs. Pepsi," 26-27.

so) it withdrew completely, ceasing to sell Pepsi products.[560] Nelson Mandela even supported Pepsi in the early nineties, drinking the beverage in public on his tour of America (including in Coke's hometown of Atlanta) and requiring that the hotels he stayed at remove all Coke products before he arrived.[561] When Pepsi re-entered South Africa in 1994 after the election of Nelson Mandela and calls for renewed investment, it took pains to ensure that the new Pepsi bottling plant would be managed and owned by black South Africans.[562] Throughout its history in South Africa, Pepsi has managed to offer employees a working environment premised on equality and black economic empowerment even when the prevailing government of the time refused to do so.[563] This is an excellent example of how individualised self-regulation can offer the experience of having one's rights respected in the workplace even when the local government is not involved in that process.

Self-regulatory policies to protect human rights can also offer serious advantages to companies. The companies may benefit from an enhanced public image, thus reducing the possibilities of negative coverage and boycotting.[564] Adherence to human rights codes (whether drafted internally at the company or by members of an organisation such as the GNI) may also prevent activities that pose a commercial risk to the company (aka risks to operational continuity) such as labour unrest or friction with the local community resulting in a withdrawal of their so-called "social license to operate."[565] There is also a reduction in

---

[560] "South African cola wars III," Brand South Africa, last modified May 30, 2006, https://www.brandsouthafrica.com/south-africa-fast-facts/media-facts/pepsi.
[561] Spivey, "Coke vs. Pepsi," 22.
[562] The capitol raised for this venture also contained investments from prominent African-Americans including Danny Glover, Shaquille O'Neal, and Johnny Cochrane. Tracy Connor, "Pepsi re-entering South Africa," United Press International, last modified October 3, 1994, https://www.upi.com/Archives/1994/10/03/Pepsi-re-entering-South-Africa/3971781156800/.
[563] This section, however, should not be taken as a wholehearted endorsement of the Pepsi company but only an expression of approval of Pepsi's actions in the segregation era of America and South Africa. Pepsi's actions in Cuba after the Cuban revolution were less laudable, when Pepsi's Vice-President Robert Geddes Morton attempted to organise a coup to overthrow Castro (this occurred before the Bay of Pigs) and was planning to use Pepsi plants in Cuba and delivery trucks to house and transport commando units. This bizarre decision was motivated by the fact that Pepsi relied on Cuban sugar plantations for its production and was concerned about the company's ability to operate in Cuba after the revolution. Spivey, "Coke vs. Pepsi," 22.
[564] Sean D. Murphy, "Taking Multinational Corporate Codes of Conduct to the Next Level," *Columbia Journal of Transnational Law* 43, no. 2 (2005): 404.
[565] Secretary of State for Foreign and Commonwealth Affairs, *Good Business: Implementing the UN Guiding Principles on Business and Human Rights* (London: HM Government, 2013), 6; Murphy, "Taking Multinational Corporate Codes of Conduct to the Next Level," 404.

the risk of litigation for human rights abuses and it can help forestall more stringent mandatory regulation being imposed.[566] Another practical advantage that self-regulation (in service of human rights) offers is that it is the method generally endorsed by the American government (with some exceptions such as CSAM) as it complies with the First Amendment.[567] As the majority of social media companies began in America, self-regulation offers the promise that the rules created by these companies will not be struck down by the courts.[568] Companies that respect human rights may increase their customer base, find it easier to attract and retain good staff and appeal to institutional investors and potential investment partners (both business and government) who are all increasingly concerned with ethical investments.[569] It can appease would-be critics and provide assurances that governments do not need to allocate valuable resources to regulate industries. Content moderation is often perceived as an act of corporate responsibility by platforms and also by the wider public.[570] A moderation approach that is perceived as effective by users (aligning with how they would like to experience the platform) can also help to attract more users. This is an example of how companies can use self-regulation to create standards (which then affect the quality of their offerings) and compete more effectively against other companies.[571]

We can see that a self-regulatory approach could offer some important advantages in protecting human rights. These benefits are accrued by states, companies, and affected parties. It must be acknowledged, however, that self-regulation has some serious disadvantages as an approach to regulating compliance with human rights in the fields of activity with which this thesis is concerned.

---

[566] A cynic may say that this is one of the most attractive reasons for companies to be "seen doing something" although such an assessment disregards all of the other benefits self-regulatory human rights schemes confers. Secretary of State for Foreign and Commonwealth Affairs, *Good Business*, 6.
[567] Pamela G. Smith, "Free Speech on the World Wide Web: A comparison between French and United States Policy with a focus on UEJF v. Yahoo," *Penn State International Law Review* 21, no. 2 (2003): 324.
[568] This, of course, is the orthodox view and it may change over time. Cases like *Packingham v. North Carolina*, 137 S. Ct. may signal the beginning of greater judicial scrutiny in the social media sphere, a trend that in the future may lead to cases that assess how social media companies create and enforce rules on their platforms.
[569] Secretary of State for Foreign and Commonwealth Affairs, *Good Business*, 6.
[570] Klonick, "New Governors," 1626.
[571] Ogus, "Rethinking Self-Regulation," 103.

6.2.2: Why self-regulation is insufficient

The primary issue with self-regulation is it suffers from a low degree of compliance. Even when platforms make their own rules or voluntarily accede to codes, they do not necessarily adhere to them. This self-regulatory regimes, therefore, can often "amount to little more than declarations of goodwill."[572] This reflects the concern that self-regulation by social media companies will never be sufficient to protect fundamental rights, that "an over-emphasis on non-interventionist techniques" could embody platforms with "unintentionally significant power in violation of the communicative rights of individual users."[573] This is a problem that does not exist in just the realm of human rights but in all voluntary (including self-regulatory) approaches to regulation. In the UK, for example, businesses participate in only 6% of Consumer Ombudsman cases where their involvement is voluntary.[574] Even the most famous voluntary human rights scheme, the UN Guiding Principles, suffers from a lack of support. A report by the UN Working Group on business and human rights found that seven years after the introduction of the UNGPs, the majority of the companies they had assessed did not demonstrate compliance.[575] The scheme was designed to be pragmatic and reasonable for businesses to adopt, but one must remember that it is still easier for companies to choose *not* to comply with a voluntary scheme. This section will attempt to explain why companies may choose not to embrace self-regulatory schemes, but it is important to remember that all of these issues only serve to illuminate the fact that the evidence demonstrates that the vast majority of companies do not voluntarily adhere to human rights schemes.[576]

An example of the practical failure of self-regulation is the prelude to the introduction of the German Network Enforcement Act. In 2011, German government officials and stakeholders in the tech sector announced that that they had developed a self-regulatory code for social networks with German users. The code was specifically focussed on data

[572] Tambini, Leonardi, and Marsden, *Codifying cyberspace*, 4.
[573] Mac Síthigh, "The mass age of internet law." 86.
[574] BEIS, *Consumer green paper: modernising consumer markets* (London: HM Government, 2019), 48.
[575] *Report of the Working Group on the issue of human rights and transnational corporations and other business enterprises (A/73/163)* (Geneva: United Nations, 2018), 8.
[576] Or if they do, compliance is often piecemeal and unsatisfactory to achieve real change.

protection, consumer protection, and the protection of children. In 2013, however, word leaked out that Facebook, Google, and LinkedIn had refused to sign the code.[577] The companies stated that they preferred supporting international self-regulation attempts instead of national regimes.[578] Google released a statement explaining that the international nature of their services meant they could not participate in self-regulatory initiatives in each country. The question remains, however, whether Google could not or simply would not participate in the scheme. This anecdote might perhaps explain why Germany subsequently introduced the Network Enforcement Act. After trying to entice platforms into a self-regulatory regime that met German values and receiving an uninterested response, Germany instead chose to simply transform it into a legal obligation and draft direct regulation. The fact that platforms obeyed this law after rejecting a more specific, voluntary scheme provides a powerful example that self-regulatory approaches to human rights issues in content moderation are not enough.

Companies may show an initial commitment to respecting human rights but translating that commitment into sustained action does not always occur. Therefore, even when companies do indicate some adherence to a human rights framework, this compliance is often patchy, and this inevitably decreases in effectiveness.[579] For example, in a study conducted by Shift, 88% of companies in their research sample had a policy commitment to respect human rights.[580] However, when it came to activities that represent the implementation of these policies, the numbers fell dramatically. Only 16% of companies evidenced a robust due diligence process and only 12% had a strong system in place for assessing the human rights impacts their activities present.[581] While one study may not be a

[577] Wauters, Lievens, and Valcke, "Towards a better protection of social media users," 293.
[578] A cynical interpretation of this preference might be that international self-regulatory codes tend to lack clear accountability structures and what social networks were actually supporting was the approach that afforded them the widest discretion.
[579] Bilchitz explains that "purely voluntary instruments will be dependent upon corporations accepting any standards that emerge; they will also depend on corporate goodwill to give effect to them." Bilchitz, "The Necessity for a Business and Human Rights Treaty," 212.
[580] *Evidence of Corporate Disclosure relevant to the UN Guiding Principles on Business and Human Rights* (New York: Shift project, 2014), 8.
[581] 56% disclosed *some* evidence of a human rights due diligence procedure but Shift did not consider these procedures to be robust while 47% disclosed *some* processes for assessing impact. *Evidence of Corporate Disclosure*, 9-10.

perfect representation, the Shift study included major corporations such as Coca Cola and BP and fielded results from eight major industry sectors.[582] If this study is representative then many companies are failing to move beyond an initial commitment to respect human rights and towards a fully integrated process. This is another weakness in self-regulation: by allowing companies to determine what compliance looks like, many platforms will fall short of what is actually required to protect human rights in their content moderation processes.

A further weakness of self-regulation in service of human rights is that companies may believe there is a business case for noncompliance. Voluntary codes become less attractive if companies perceive adherence with these codes as putting them at a competitive disadvantage.[583] An example of this is the relative outcomes for Coke and Pepsi in South Africa. As discussed previously, Pepsi implemented policies in South Africa first to diminish the effects of apartheid on its employees and then to help encourage black economic empowerment in the post-apartheid landscape. Pepsi hoped that its ethical position would help it to supplant Coca-Cola, the soft drink of choice in South Africa during apartheid.

Coca-Cola has had bottling plants in South Africa since 1928 and the beverage's ubiquity led one commentator to state that Coke had "painted South Africa red."[584] By the 1980's Coke held 90% of the soft drinks market in South Africa.[585] Unlike Pepsi, Coca-Cola became a target for anti-apartheid activists because of a number of controversial practices at their South African manufacturing facilities. Not only did Coca-Cola pay black workers less but there were even allegations that one of their South African bottling plants had a government contract to use black prison labour in the factory, a group who, legally, could be paid even less than other black workers.[586] In 1978 Coke refused to cooperate with a US

---

[582] The sectors were oil & gas/extractives; fast moving consumer goods; apparel; food, beverage & agriculture; information and communications technology; banking & finance; automotive; and pharmaceuticals. For more information on methodology, *Evidence of Corporate Disclosure*, 5.
[583] Mandatory regimes, of course, actually result in consistent results as all competitors are forced to adhere to the same requirements.
[584] "Coca-Cola moves Africa HQ to Jozi," Brand South Africa, last modified August 21, 2006, https://www.brandsouthafrica.com/investments-immigration/africa-gateway/coke-210806.
[585] "Coke sweetens apartheid," Chicago Anti-Apartheid Movement Collection, accessed July 6, 2019, https://caamcollection.omeka.net/items/show/13.
[586] Pepsi also initially paid its black workers less but changed its practices earlier than Coke and offered more progressive policies afterwards. Myron P. Curzan and Mark L. Pelesh, "Revitalizing Corporate Democracy:

Senate Foreign Relations Subcommittee that was investigating American business practices in South Africa, provoking condemnation from the committee.[587] One famous protest poster advocating for a boycott of Coca-Cola stated: "Coke Sweetens Apartheid."[588] In 1986, Coca-Cola divested from South African by selling its holdings but crucially, Coca-Cola products continued to be sold in South Africa throughout the intervening years with no change in the availability of the beverage or the amount of advertising.[589] At the time, an anti-apartheid activist criticised these actions, stating "it is not divestment, it is warehousing. Foreign companies have maintained their operations on a franchise basis."[590] Coke's market share even increased in South Africa at the time because of the absence of competitors.[591]

Post-apartheid, Pepsi's optimism about being rewarded for its ethical stance quickly faded. Nelson Mandela, realising that Coke's dominance in South Africa could provide a host of jobs and economic benefits, stopped publicly snubbing the company.[592] Pepsi's bottling company in South Africa folded only three years after it opened; Pepsi's belief that consumers would choose the more ethical company was shattered.[593] Pepsi had failed to

---

Control of Investment Managers' Voting on Social Responsibility Proxy Issues," *Harvard Law Review* 93, no. 4 (1980): 672, https://doi.org/10.2307/1340521.

[587] Coke ignored Committee requests for information about their pay-scales, promotion, hiring practices, and other activities in South Africa, claiming that the information was confidential. This is strongly reminiscent of the many times social media companies have refused to disclose information on the grounds that it was proprietary information or could be used to game the system. In both situations, one suspects that the primary motivation was a desire to avoid public criticism if further details about questionable practices were revealed to the public. J. C. Louis and Harvey Z. Yazijian, *The cola wars*, 1st ed. (New York: Everest House, 1980), 164.

[588] Chicago Anti-Apartheid Movement Collection, "Coke sweetens apartheid."

[589] Coke simply moved its production plants to Swaziland and shipped products to South Africa. During this time, Tandi Gcabashe, the daughter of a former ANC leader and currently resident in Atlanta (the city where Coke is headquartered) continued to agitate for a total boycott of South African business by Coke. She argued that "for every 80-cent bottle of Coke sold in South Africa, the apartheid government collected 10 cents in taxes to support their regime." Spivey, "Coke vs. Pepsi," 34, 35, 47.

[590] Anthony Robinson, "Inside, Doubt Takes Root; Disinvestment From South Africa," *Financial Times*, June 16 1987.

[591] When Coke sold off its holdings in South Africa, it did so through a franchise system. White businessmen then purchased these franchises (along with other major American company franchises such as General Motors, General Electric, and I.B.M.) The new businessmen often then eliminated the limited social programmes American companies had in place to benefit black workers. Divestment, therefore, likely had a negative impact on black South Africans who had worked at American companies in the short-term. Assessing the long-term impacts of divestment is a complex undertaking and beyond the scope of this thesis. Anatole Kaletsky, "Coca-Cola 38 Per Cent Up to $934M," *Financial Times*, February 20 1987.

[592] Mark Pendergast, *For God, Country, & Coca-Cola: The Definitive History of the Great American Soft Drink and the Company That Makes It* (New York: Basic Books, 2000), 394.

[593] Brand South Africa, "South African cola wars III."

understand that their complete divestment in the last days of apartheid, laudable as it was, had allowed Coke the opportunity to further consolidate their dominance in South Africa, a business reality that seemed to trump Pepsi's ethical claims. In 1997, an independent survey on consumer brand loyalties found that Coca-Cola was both the "Most Popular" and "Most Admired" brand in South Africa.[594]

The Pepsi and Coke saga has some important lessons for would-be reformers. The first is that companies are not always rewarded for ethical stances (and consumers have short memories) and any scheme that assumes that companies will voluntarily change their practices must be mindful of this reality, especially if compliance is expected to be an on-going process. The second lesson is that when companies are given broad discretion in how they implement an obligation (such as divestment) some companies will behave like Coke and attempt to do only the bare minimum of what is required of them while disregarding the underlying objective of the law, while others will take a more principled approach even if it harms their interests. In South Africa, the decision to keep selling Coke products and use a franchise system to get around divestment requirements is a clear example of this issue. The situation Pepsi and Coke faced in South Africa is also interesting because by the end, neither company was presented with any real incentives to prioritise human rights, which is a central weakness in a voluntary scheme; at some points the disincentives of compliance will outweigh the incentives and there are no legal consequences for companies that stop abiding by human rights principles.

Another flaw in self-regulation, as previously observed, is a lack of effective accountability mechanisms. Instead, platforms are often expected to assess their own compliance and change their practices accordingly, effectively "marking their own homework." [595] This approach seems inherently flawed and unlikely to produce significant changes. Danielson contends that as private companies engage in important governance activities "accountability for the social welfare effects of regulatory outcomes should not fall exclusively on "public" regulators and the actions and decisions of "private" corporate actors

[594] Spivey, "Coke vs. Pepsi," 54.
[595] Select Committee on Communications, *Regulating in a digital world (Second Report of Session 2017–2019. HL Paper 299)* (London: House of Lords, 2019), 16.

should not be exempt from public participation, review, and political contestation."[596] Self-regulatory schemes, however, do purport to keep the actions of private regulators exempt from public input, even when in doing so, they undermine the effectiveness of the regime.

Occasionally, social media platforms voluntarily join organisations that have drafted codes of practice (a move towards collective self-regulation), codes that may purport to be more principled than completely individualised self-regulation.[597] While it is entirely possible to voluntarily accede to legal schemes which then mandate compliance,[598] the organisations that platform join tend to lack strong accountability measures. One need only think of all the platforms that are members of the Global Network Initiative (such as Facebook and Google) who have voluntarily agreed to abide by the GNI Guidelines, which requires (among other things) that companies "avoid or minimise the impact of government restrictions on freedom of expression."[599] However, there are ample examples of GNI members entering into secretive agreements with states to voluntarily remove content that would likely be protected by the right to free expression.[600] For example, Facebook and Twitter have both been criticised for treating Zionists as a protected group and accordingly removing most of the negative content referred to them by the cyber unit of the Israeli government while allowing incitements to violence and hateful content about Palestinians to remain available.[601] Collective systems of self-regulation, therefore, may represent a

---

[596] Danielson, "How Corporations Govern," 423-24.

[597] Black argues that the "essence of self-regulation is a process of collective government" and argues that this situation is more typical, with self-regulation usually describing "the situation of a group of persons or bodies, acting together, performing a regulatory function in respect of themselves and others who accept their authority." I would argue, however, that social media companies do not typically engage in collective self-regulation. The only real examples, such as the Global Network Initiative, appear more symbolic than anything else, with the real work of regulation occurring in-house at the specific social media companies. Black, "Constitutionalising Self-Regulation," 27.

[598] The Rome Statute, for example, which opens countries up to the jurisdiction of the International Criminal Courts should they be the victims or perpetrators of international crimes.

[599] "The GNI Principles: Freedom of Expression," Global Network Initiative, accessed September 22, 2019, https://globalnetworkinitiative.org/gni-principles/.

[600] Association for Progressive Communications, *Content Regulation in the Digital Age*, 10.

[601] In July-December 2016, 85% of the content referred by the Israeli government was removed by Twitter. This is controversial because the Israeli Basic Law states "Nothing in the law allows state authorities to censor content based solely on an administrative determination." To put those numbers into perspective, 29% of Canadian government requests and 21% of UK requests resulted in content removal. Meanwhile, 7amhel (a Palestinian human rights charity) released a report claiming that a hateful post about Palestinians is uploaded every 46 seconds but is not being removed. "Israel's 'Cyber Unit' operating illegally to censor social media content," Adalah, last modified September 14, 2017,

harmonisation and codification of principles but they are essentially identical to self-regulation in practice (but with an added sheen of legitimacy), especially when compliance is entirely discretionary and no strong accountability measures exist.

Finally, it is important to understand that companies approach regulation differently from public bodies and this can make them unsatisfactory regulators when one applies a human rights perspective. Bonnici and de vey Mestdagh argue that instead of procedural formalities, private regulators are concerned with whether the regulation is efficient and meets the needs of both their customers and the company. Accordingly, principles like accountability and transparency are not given a high priority.[602] While efficient regulation is important, human rights considerations may not always be compatible with efficiency or the interests of the company. In fact, human rights often entails protecting minority interests so these decisions may not even be supported by a majority of users. When creating regulations, however, economic efficiency should not be "an overriding legislative goal" and "social obligations are legislative goals of at least equal importance."[603] It is clear that one must import the perspective of these corporations into any considerations of how a plan to protect human rights can be implemented on the platforms. However, being sensitive to practical realities and allowing companies broad discretion in how they define their human rights responsibilities are very different propositions. Self-regulation allows platforms to create a hierarchy of their priorities, a hierarchy that might differ quite markedly from the approach legislators envisioned. Bilchitz contends that self-regulation relies "on an ability of corporations to think in a manner that considers their wider social impact where the incentives for their decision-makers are often focused on shorter-term profit maximisation."[604] This tendency would be manageable in a more heavily regulated system,

---

https://www.adalah.org/en/content/view/9228; "7amleh releases new racism index exposing heightened Israeli online incitement against Palestinians," 7amleh, last modified March 5, 2018, https://7amleh.org/2018/03/05/7amleh-releases-new-racism-index-exposing-heightened-israeli-online-incitement-against-palestinians/.

[602] Mifsud Bonnici and de vey Mestdagh, "Right Vision, Wrong Expectations: The European Union and Self-regulation of Harmful Internet Content," 145.

[603] Baldwin and Cave, *Understanding regulation*, 82.

[604] Bilchitz, "The Necessity for a Business and Human Rights Treaty," 213.

where the private and public regulators may engage in discourse, but could lead to ineffective results otherwise.

6.2.3: Case Study: Myanmar

A grim example of the futility of recognising and protecting human rights through self-regulation may be seen in Facebook's response to the genocide in Myanmar.[605] In 2017, tensions were high in Myanmar, with a Rohingya (a Muslim minority) militant group attacking military forces and harsh government reprisals that forced hundreds of thousands of Rohingya civilians to flee Myanmar.[606] These events coincided with a spike in hate speech against the Rohingya people on Facebook and incitements to violence. It must be noted that Facebook holds a special position in Myanmar because, as a recent UN report explains, Myanmar only came online in 2010 and "for most users, Facebook is the Internet."[607] This pre-eminence of Facebook in Myanmar is confirmed by a further report that found that users in Myanmar considered Facebook the only internet entry point for information and a majority perceived Facebook posts as news.[608] Researchers have also found that Myanmar suffers from low rates of digital literacy (where users struggle to identify digital misinformation) and that many of the "legal, political and cultural assumptions (such as freedom of speech and rule of law)" that Facebook's approach to content moderation is predicated on do not apply in Myanmar.[609] Unfortunately, as a result of all of this, the platform became a useful tool for military figures, government officials, and radical

---

[605] A UN Fact-Finding Mission has confirmed that "Myanmar incurs State responsibility for committing genocide and is failing in its obligations under the Genocide Convention to investigate and, where appropriate, prosecute genocide. It is also failing to enact effective legislation criminalising and punishing genocide. The State of Myanmar continues to harbour genocidal intent and the Rohingya remain under serious risk of genocide." Human Rights Council, *Detailed findings of the Independent International Fact-Finding Mission on Myanmar (A/HRC/42/CRP.5)* (Geneva: United Nations, 2019), para 213.

[606] Libby Hogan and Michael Safi, "Revealed: Facebook hate speech exploded in Myanmar during Rohingya crisis," The Guardian, last modified April 3, 2018, https://www.theguardian.com/world/2018/apr/03/revealed-facebook-hate-speech-exploded-in-myanmar-during-rohingya-crisis.

[607] Human Rights Council, *Report of the independent international fact-finding mission on Myanmar (A/HRC/39/64)* (Geneva: United Nations, 2018), 14.

[608] It should be noted that Myanmar's press is strictly controlled by the government so it is natural for citizens to assume that they would be more likely to find the truth on Facebook. GSMA and LIRNEasia, *Mobile phones, internet, and gender in Myanmar* (London: GSMA, 2016).

[609] BSR, *Human Rights Impact Assessment: Facebook in Myanmar* (New York: Business of a Better World, 2018), 24, 13.

supporters to call for violence against the Rohingya as well as to spread false information about the activities of Rohingya militants. Facebook had become a modern-day Radio Rwanda and the consequences for the Rohingya people were serious. Initially, Facebook was slow to react and the UN has publicly criticised Facebook's response to the crisis as "slow and ineffective."[610]

In 2018, Facebook hired BSR (Business for Social Responsibility), an independent NGO, to conduct an in-depth Human Rights Impact Assessment (HRIA) on Facebook's activities in Myanmar. Facebook publicly shared the report, a positive move towards transparency as the report had found serious failings and human rights violations by Facebook in Myanmar. The report contained a variety of recommendations on how the platform could improve their conduct in Myanmar including employing Burmese moderators, creating a secure storage space for the preservation of digital evidence, and encouraging initiatives on digital literacy. [611] There have been some important recommendations, however, that the platform has inexplicably failed to implement.[612] These oversights, which will be detailed below, highlight the weakness of self-regulation in recognising and protecting human rights: no matter how important the issue may be, in a self-regulatory environment the platforms can always choose to do nothing.

First, BSR recommended that Facebook create a stand-alone human rights policy. This was to be a place where Facebook would publicly commit to abiding by the UDHR, ICCPR, and ICESCR as well as other relevant human rights treaties. It would also provide information on the human rights risks and opportunities that Facebook posed and information about their processes and policies.[613] A stand-alone policy is an excellent idea because it would demonstrate that the role that social media platforms play in human rights is nuanced and complex, and that platforms should commit to transporting human rights values into their activities, including content moderation. Having a stand-alone human rights

---

[610] Human Rights Council, *Report (A/HRC/39/64)*, 14.
[611] BSR, *Facebook in Myanmar*, 5.
[612] BSR made many recommendations but these three recommendations seem very important and it is surprising that Facebook did not implement them. Perhaps it would be easier to understand if Facebook had provided an explanation for why they chose not to adopt certain recommendations.
[613] BSR, *Facebook in Myanmar*, 42.

policy would also clarify that social media community guidelines are a different species from human rights policies, at times overlapping but at other times being very different.[614] Despite this suggestion being relatively easy to implement, Facebook chose not to create a standalone human rights policy.[615] Perhaps Facebook preferred the flexibility of a set of guidelines they interpret themselves (except for national legal interventions on specific categories of content) rather than established legal principles that might require they make decisions that are contrary to their business interests.

BSR also suggested that Facebook publish a country-specific enforcement report on Myanmar.[616] Facebook usually publishes a community standards enforcement report that provides information about how much content Facebook removes globally in each category (such as hate speech or nudity). BSR argued that it was important to have a report solely for Myanmar which also "describes elements of local process, such as the nature of government relationships, how certain standards are interpreted locally, and whether the government submits content-removal requests through the Community Standards process, rather than via law enforcement relationship channels."[617] The clear benefit of a country-specific report is that it would then be possible to compare reports from other quarters to identify patterns. Of course, these reports would also be useful in assessing the effectiveness of Facebook's enforcement in Myanmar (although only as it applied to removals), which would provide a cynical explanation for why Facebook has chosen *not* to create country-specific reports, even in its most challenging countries.

BSR then recommended that Facebook undertake HRIA's in other high-risk countries and create a forum where these HRIA's could be shared with the public.[618] While Facebook intimated to BSR that other HRIA's *were* being conducted at the same time as Myanmar, none of these HRIA's have ever been made public. Facebook does share news releases on human

---

[614] Of course, over time one would expect to see less and less divergence between the two policies. Ideally, community guidelines would function as a sector-specific interpretation of international human rights principles.
[615] As of 10 October 2019, anyways, Facebook has not enacted a stand-alone human rights policy.
[616] BSR, *Facebook in Myanmar*, 28.
[617] BSR, *Facebook in Myanmar*, 28.
[618] Other human rights updates should also be shared on this forum, including perhaps an annual human rights report. BSR, *Facebook in Myanmar*, 42-43.

rights with its users (through the Facebook newsroom) but there is no dedicated forum for this material and finding it necessitates sifting through the myriad of press releases Facebook produces on a variety of different subjects. Creating a human rights-specific area with a regularly updated collection of HRIA's would be an incredibly useful tool for academics, activists, and concerned users, although perhaps it might be *too* useful from Facebook's perspective. Just as displaying the calorie count of baked goods at a café might make customers less likely to purchase them, becoming aware of how many human rights issues really exist at Facebook could leave users disenchanted and willing to delete their accounts.

Facebook has not only ignored some of the recommendations in the BSR report, they have also acted in contradiction to some of its suggestions. In February 2019, Facebook announced that it had banned four armed organisations in Myanmar (all of whom were ethnic independence movements) and would be removing any content that was in support of the groups.[619] This was a controversial decision and ran counter to the recommendations in the HRIA, which advised Facebook to "narrow its existing definition of terrorist organisations…to exclude organisations considered to be legitimate combatants in conflict, such as officially recognised ethnic armed organisations (EAOs)."[620] The report emphasised that this was particularly important in Myanmar, "where there is a history of toxic nationalism and state-mandated violent oppression of ethnic groups, as well as the presence of multiple legitimate secession movements."[621] It should be noted that the report wasn't saying that content by these organisations could not be removed if it violated other prohibitions (such as hate speech or graphic violence) but simply that these organisations should not be banned entirely from the platform.[622] The situation in Myanmar is still ongoing, with the International Court of Justice (ICJ) ruling in January 2020 that the Rohingya people were at serious risk of genocide and that Myanmar must "take all measures within its

[619] Facebook Newsroom, "Banning More Dangerous Organisations from Facebook in Myanmar," Facebook, last modified February 5, 2019, https://newsroom.fb.com/news/2019/02/dangerous-organizations-in-myanmar/.
[620] BSR, *Facebook in Myanmar*, 47.
[621] BSR, *Facebook in Myanmar*, 47.
[622] Terrorist organisations are banned on Facebook, no matter what content they post. BSR was recommending that these organisations not be banned in a similar fashion.

power" to prevent it.[623] UN investigator who participated in a fact-finding mission to Myanmar have recently criticised Facebook for insufficient action, stating "Facebook's actions can only be described as minimal."[624] The posts that explicitly call for violence are being removed but the posts denigrating the Rohingya people remain available, a concerning development in a country where genocide has occurred and again seems imminent.[625] These issues will likely become more pressing in the months to come as the country gears up for a general election in 2020 and the coronavirus pandemic allows Myanmar to keep foreign journalists out.

The Myanmar situation illustrates why self-regulation is an unsatisfactory way to safeguard human rights. A serious problem emerged in a country with a history of human rights abuses, a situation that the platform had likely not foreseen.[626] The company was therefore slow to respond and eventually, sensing that outside expertise was needed, the platform hired BSR to conduct a HRIA. It was then up to the platform's discretion how many of the recommendations from the report they would choose to implement. It should be noted again that Facebook has followed some of the report's suggestions, such as banning key Myanmar military figures in the genocide and employing a hundred moderators who speak Burmese.[627] These concessions, however, do not diminish the central issue in this section. Self-regulation is rife with discretionary decisions and selective adherence. A voluntary commitment to respect human rights loses its legitimacy unless it is complemented by accountability measures, remedies, and contains some level of independent monitoring.[628]

---

[623] *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (The Gambia v. Myanmar) - Provisional measures. Order of January 23, 2020* (The Hague: ICJ, 2020), 25.
[624] Patrick O'Neill, "Facebook's Efforts 'Not Nearly Sufficient in Genocide-Torn Myanmar,' UN Investigator says," Gizmodo, last modified March 4, 2019, https://gizmodo.com/facebooks-efforts-not-nearly-sufficient-in-genocide-tor-1833719999.
[625] O'Neill, "Facebook's Efforts 'Not Nearly Sufficient in Genocide-Torn Myanmar,' UN Investigator says."
[626] The next chapter will focus on the proposal of mandatory human rights due diligence and one of the requirements of that process is that companies assess the context of a country and the human rights issues of operating there. Had Facebook done that when it began to offer Facebook Basics in Myanmar then it might have foreseen and prevented it becoming a platform to incite genocide, thus fulfilling Principle 13's requirement that companies avoid "causing or contributing to adverse human rights impacts" and preventing impacts caused by their business relationships. Principle 13, *UN Guiding Principles.*
[627] Facebook Newsroom, "An independent assessment of the human rights impact of Facebook in Myanmar," Facebook, last modified November 5, 2018, https://newsroom.fb.com/news/2018/11/myanmar-hria/.
[628] Murphy, "Taking Multinational Corporate Codes of Conduct to the Next Level," 420-32.

It is clear that there is much to be gained from a self-regulatory scheme: it offers the most efficient use of resources and experience and retains the capacity to handle a volume of content that would be inconceivable to traditional public regulators.[629] Any successful scheme to align content moderation practices with human rights must, therefore, involve the cooperation of social media companies. This does not, however, necessarily require the *voluntary* cooperation of these companies and it certainly should not entail an entirely voluntary scheme where companies determine every aspect of the regime, as this has proven to be an unsatisfactory approach. Any proposed plan of action must, therefore, preserve the strengths of a self-regulation regime but account for its weaknesses by designing a co-regulatory regime. The state must outline a set of expectations on how human rights should be protected by these companies and the platforms should be afforded discretion in how they achieve those goals. This approach must contain accountability measures to ensure that the discretion exercised by these platforms is bounded by expectations from the state and affected parties.

6.3: Substantive Regulation

6.3.1: Overview and Advantages of Substantive Regulation

Substantive regulation is the most common trend in regulating the harms caused by social media platforms. It is predominantly used to alter the content moderation process and typically it mandates the removal of certain categories of content because of public concerns.[630] This kind of directed regulation (where the government passes a law containing specific requirements for regulatees) tends to focus on a single issue and is very specific in its application, the polar opposite of the broad discretion afforded social media platforms in

[629] This is the stumbling block in Pamela Wu's argument that content moderation (at least as it relates to extremist content) should be handed over to a UN body. The resources required to make this scheme work would not be allocated and the backlog of content decisions would become unmanageable. See: Paulina Wu, "Impossible to Regulate: Social Media, Terrorists, and the Role for the UN," *Chicago Journal of International Law* 16, no. 1 (2015): 309.

[630] Categories of content that have already sparked public campaigns include cyber-bullying, self-harm, pro-eating disorders, fake news, white supremacist, and terrorist content.

a self-regulatory approach. While a discussion of substantive regulation sometimes feels less like a proposal than it does a description of the current situation, there are still strengths in this approach that can be investigated. It is also clear that this approach has human rights applications, as the key driver in this area of regulation is to protect the public from perceived harms on social media platforms such as childhood exposure to inappropriate content, hate speech, extremist content, and non-consensual sexual imagery.[631] Could the human rights concerns that have been addressed in this thesis be addressed in targeted legislation addressing specific issues in the social media content moderation process?

Substantive regulation offers a number of advantages as an approach to resolving the human rights issues on social media platforms. First, because these laws tend to be narrower in their focus and are results oriented, they can be applied with almost surgical precision to the problem. If one is concerned about the existence of self-harm imagery on social media (and the effect this would have on children's rights)[632] for example, one could pass a law that directly addresses this issue without trespassing into other areas of regulatory intervention. Second, the regulatory regime that is designed can be tailored exactly to the technology in use with little need for the convoluted interpretations that occur when old laws are interpreted for new technologies. A law, therefore, can be directly applicable to social media instead of retrofitting an old law that concerned telephones, print communications, or traditional media. Finally, these laws offer the practical advantage that their specificity makes it easier to foster political support and achieve consensus as opposed to broad, holistic regulatory schemes for entire industries. It is easier, for example, to pass a law criminalising up-skirting photos then it is to pass a law criminalising the many coercive forms of abuse (both of a sexual and non-sexual nature) that women suffer online.

---

[631] Although so far there does not seem to be any regulations demanding that content remain on the platform on free expression grounds. Those demands tend to be dealt with through the collective action approach explored at 5.3.

[632] Specifically Article 17(e) which states that children should be protected from "information and material injurious to his or her well-being." *Convention on the Rights of the Child*.

6.3.2: Disadvantages of Substantive Regulation

Despite these advantages, there are some serious weaknesses to relying on substantive regulation in this sphere. First, it must be acknowledged that relying primarily on substantive regulation to reform content moderation results in governance gaps. Specific regulatory regimes are reactive and siloed without acknowledging that many issues are interconnected (to both present and emerging situations).[633] Patchwork regimes are difficult for the public to understand and can be contradictory, thus reducing their certainty on what is permissible and what is not. It will be recalled that Fuller identifies clarity and the avoidance of contradiction as two of his criteria in assessing the quality of a law whilst Bingham emphasised that the law should be "intelligible, clear, and practicable."[634] One could only imagine the chaotic results if every crime was separately legislated against instead of coherent criminal codes being created.[635]

An example of this complexity can be found in the laws that have been precipitated by a growing concern over new technologies being used to create and disseminate sexual images of individuals who have not consented to these actions. It is difficult to encompass all the related issues in a single sentence (which is why the previous sentence is deliberately broad) but it is clear that there is a wide range of content and troubling behaviour that would be deemed harmful to individuals or society in general. Instead of considering a broader, more holistic, approach, the UK originally singled out one specific aspect of this problem, the disclosure of private sexual images or what is often referred to as "revenge porn."[636] This

[633] Damian Tambini made a similar point at the House of Lords inquiry when he stated that before now regulatory measures had been implemented in "a fragmented way across different areas. The solution to the current impasse is not going to be a tweak here or there, but a policy response that is coordinated across multiple policy areas." Select Committee on Communications, *Regulating in a digital world*, 61.
[634] Fuller, *The morality of law*, 63-70. Bingham, *The rule of law*. 37.
[635] Of course, new crimes are created and these are often introduced through substantive regulation, but it is far more common to amend existing criminal codes to retain an element of coherence in these regulations.
[636] Some academics such as McGlynn, Rackley, and Houghton find the term "revenge porn" inappropriate because it focusses on the motives of the perpetrator instead of the experience of the victim, it also disregards the many motives a person may have in sharing these images (such as to make money, gain notoriety, or sexual gratification) and separates these actions from other forms of abuse including domestic violence. The term "porn" may also be inappropriate as it affords these images a measure of legitimacy, a concern that also led to the term "child pornography" being replaced with "Child Sexual Abuse Material." Clare McGlynn, Erika Rackley, and Ruth Houghton, "Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse," *Feminist Legal Studies* 25, no. 1 (2017): 38-40, https://doi.org/10.1007/s10691-017-9343-2; Clare McGlynn and Erika Rackley, "Not 'revenge porn', but abuse: let's call it image-based sexual abuse," Inherently Human,

was criminalised in England & Wales in Section 33 of the Criminal Justice and Courts Act 2015. This provision specifically addressed "disclosing private sexual photographs and films with intent to cause distress." It soon became clear, however, that the narrow scope of this provision meant that there were equally troubling actions that weren't captured by the provision such as disclosure for any motivation other than causing distress and disclosure to anyone who features in the images (even if it is done to cause distress).[637] The emphasis on private sexual acts also limited the offence, which led to the introduction of another law targeting "upskirting" images in The Voyeurism (Offences) Act 2019.[638] Upskirting is colloquially defined as "taking a picture under a person's clothing without them knowing, with the intention of viewing their genitals or buttocks."[639] While laws like the up-skirting law were widely celebrated,[640] the fact remains that by drafting such specific provisions, much of what McGlynn, Rackley, and Houghton term "the continuum of image-based sexual abuse" still falls outside the remit of the laws on revenge porn and upskirting.[641] These

last modified February 15, 2016, https://inherentlyhuman.wordpress.com/2016/02/15/not-revenge-porn-but-abuse-lets-call-it-image-based-sexual-abuse/.

[637] Alisdair Gillespie, "'Trust me, it's only for me':'revenge porn' and the criminal law," *Criminal Law Review* 11 (2015): 868.

[638] The Voyeurism (Offences) Act creates two new offences under section 67 of the Sexual Offences Act 2003. Some of the actions described in this act were previously prosecuted under the common law offence of Outraging Public Decency or the existing voyeurism offence under section 67. There were some gaps in those laws, such as the OPD offence requiring that two or more people be present in the immediate area (so a train carriage with only the complainant and perpetrator might not be covered) and that it happen in a public place (the definition of public is unclear). Meanwhile, the old version of section 67 was confined to places where one would reasonably expect privacy so public places did not usually qualify. Criminal Courts and Criminal Law Policy Unit, *Voyeurism (Offences) Act 2019: Implementation of the Voyeurism (Offences) Act 2019* (London: Ministry of Justice, 2019).

[639] The full legal definition is "Someone who operates equipment or records an image under another person's clothing (without that person's consent or a reasonable belief in their consent) with the intention of observing or looking at, or enabling another person to observe or look at, their genitals or buttocks (whether exposed or covered with underwear), or the underwear covering the genitals or buttocks, where the purpose is to obtain sexual gratification or to cause humiliation, distress or alarm." Criminal Courts and Criminal Law Policy Unit, *Voyeurism (Offences) Act 2019: Implementation of the Voyeurism (Offences) Act 2019*; "Press Release: 'Upskirting' law comes into force," Ministry of Justice, last modified April 12, 2019, https://www.gov.uk/government/news/upskirting-law-comes-into-force.

[640] See, for example: Sonia Elks, "The UK has just introduced a new law to protect women," World Economic Forum, last modified April 15, 2019, https://www.weforum.org/agenda/2019/04/the-uk-has-just-introduced-a-new-law-to-protect-women; "Upskirting now a crime after woman's campaign," BBC News, last modified April 12, 2019, https://www.bbc.com/news/uk-47902522.

[641] They write "Understood as a continuum, the concept of image-based sexual abuse is sufficiently broad and flexible both to embrace new ways of perpetrating, and experiencing, these forms of abuse. Thus far, law, policy and public discourse has tended to concentrate on specific categories of activity, harm, or particular motives, often focussing on one particular example, only to find other forms of abuse excluded and ignored." McGlynn, Rackley, and Houghton, "Beyond 'Revenge Porn'," 28.

actions include sexualised photoshopping (AKA deepfakes),[642] sexual extortion,[643] and down-blousing[644] with the possibility that other image-based abusive acts may emerge in the future. McGlynn, Rackley, and Houghton argue that the problem with highly specific statutory interventions of this sort is that "such provisions are a clear demonstration of how legal categories based on supposedly isolated forms of conduct and for specific purposes leave many victim-survivors unprotected. Laws, as currently interpreted and enacted, therefore largely fail to cover the range of experiences of abuse."[645] While it takes longer to draft a comprehensive code on a particular issue (such as image-based sexual abuse) the results will likely be more coherent and will better address the spectrum of harmful behaviours regulators are attempting to address. Unfortunately, "governments are frequent failures in learning lessons from regulatory history"[646] and the most common approach to online problems is the demand for specific substantive regulations.

The image-based sexual abuse example not only demonstrates the complexity of a patchwork regime, but it also illustrates another issue with substantive regulation: certain situations will be unintentionally excluded from the overlapping regulatory regimes. For example, by focussing on explicit self-harm imagery, we fail to consider how these technologies may contribute to mental health issues on a broader level.[647] After all, some

---

[642] Sexualised photoshopping is defined as when "without consent, a pornographic image is superimposed onto an individual's head/body part, such that it looks as if that individual is engaged in the pornographic activity." McGlynn, Rackley, and Houghton, "Beyond 'Revenge Porn'," 33. A study found that 96% of Deepfakes were pornographic in nature. Ivan Mehta, "A new study says nearly 96% of deepfake videos are porn," The Next Web, last modified October 7, 2019, https://thenextweb.com/apps/2019/10/07/a-new-study-says-nearly-96-of-deepfake-videos-are-porn/.

[643] "Sexual extortion (often colloquially known as 'sextortion') generally describes practices whereby perpetrators coerce individuals, often young people, into creating and/or sharing private sexual images, as well as deploying threats to force further image-creation. Alternatively, webcams, phones or data storage areas such as the iCloud are hacked to obtain consensually taken images or videos without the person's consent, with perpetrators using threats and blackmail to solicit further images and/or sexual practices and, in some cases, money." McGlynn, Rackley, and Houghton, "Beyond 'Revenge Porn'," 34.

[644] Down-blousing is the act of taking images down women's shirts. It was frequently combined with up-skirting in decisions which has made the decision to specifically criminalise up-skirting but not down-blousing baffling. This article discusses these two actions and predates the up-skirting law. Clare McGlynn, "We Need A New Law to Combat 'Upskirting' and 'Downblousing' " Inherently Human, last modified April 15, 2015, https://inherentlyhuman.wordpress.com/2015/04/15/we-need-a-new-law-to-combat-upskirting-and-downblousing/.

[645] McGlynn, Rackley, and Houghton, "Beyond 'Revenge Porn'," 32.

[646] Brown and Marsden, *Regulating code*, xvii.

[647] See, for example: Andrew K. Przybylski et al., "Motivational, emotional, and behavioral correlates of fear of missing out," *Computers in Human Behavior* 29, no. 4 (2013): 1841,

impacts of technology "are not yet understood or have yet to galvanise a set of activists to demand change."[648] Platforms themselves also engage in this piecemeal reform, which can result in uneven results. For example, Islamist terrorist groups were initially targeted for removal on Twitter with more intensity than other terrorist groups.[649] After the Charlottesville rally and attack (which resulted in the death of Heather Heyer) many tech companies publicly pledged to ban white supremacist groups or people espousing that ideology.[650] Once again, platforms chose to reactively ban a category of harmful people without considering whether there might be other kind of groups with similar qualities who could pose a problem in the future. This approach is sometimes termed "regulating by outrage" and occurs when "in the absence of an effective regulatory framework "outrage, campaigning and lobbying" intensified by media coverage have stimulated *ad hoc* responses to online harms."[651] By adopting a tunnel-vision approach, some issues, often more nuanced ones which are less capable of an easy political win are ignored. This approach might, therefore, simultaneously overregulate and underregulate a field and result in unpredictable outcomes.

These problems exist even if one considers the use of substantive regulation to mandate human rights protections on the platforms. Take, for example, concerns about censorship in content moderation. It would be possible to pass a specific law obliging platforms to respect the right to free expression when moderating content. This approach, however, is problematic for a number of reasons. First, it singles out a particular kind of business as having human rights obligations without considering whether other companies should have similar responsibilities. If future legislation is then passed addressing the human rights

---

https://doi.org/10.1016/j.chb.2013.02.014; Rhys Edmonds, "Anxiety, loneliness and Fear of Missing Out: The impact of social media on young people's mental health," Centre for Mental Health, accessed October 8, 2019, https://www.centreformentalhealth.org.uk/blog/centre-mental-health-blog/anxiety-loneliness-fear-missing-out-social-media.

[648] Catherine Miller, Jacob Ohrvik-Stott, and Rachel Coldicutt, *Regulating for Responsible Technology: Capacity, Evidence and Redress: a new system for a fairer future* (London: Doteveryone, 2018), 71.

[649] Berger, *Nazis vs. ISIS on Twitter: A Comparative Study of White Nationalist and ISIS Online Social Media Networks*.

[650] Keith Collin, "A running list of websites and apps that have banned, blocked, deleted, and otherwise dropped white supremacists," Quartz, last modified August 16, 2017, https://qz.com/1055141/what-websites-and-apps-have-banned-neo-nazis-and-white-supremacists/.

[651] Per the Communications Chambers. Select Committee on Communications, *Regulating in a digital world*, 22.

responsibilities of related sectors such as news media or interactive media then a plethora of regulatory regimes for different sectors may spring up, an outcome that will only grow more complicated as more companies develop technological capabilities that may result in them being covered by multiple, potentially contradictory regimes.[652] Second, instead of considering human rights responsibilities more broadly, substantive regulation may focus on a specific right (such as free expression) which can further complicate understandings of what legal obligations these companies may have. This is especially problematic because human rights often entails rights-balancing (when rights come into conflict) and if companies only have legal obligations in relations to some rights, they will naturally prioritise those rights over other rights.[653] Any subsequent amendments to include more rights will only render the situation more complex. Instead, human rights legislation must be comprehensive and any disparities in obligations on different business sectors or in relation to specific categories of rights must be reasonable, clearly articulated and justifiable.

Further, as previously mentioned, substantive regulation can both underregulate and overregulate an area. In terms of overregulation, there is a troubling tendency to neglect valid human rights concerns surrounding censorship and impose stricter expectations on platforms than on similar offline offerings. The government White Paper on online harms, for example, identifies twenty-three discrete categories of harms that exist on the

---

[652] This point was also made in the House of Lords hearings when discussing competition law as a number of representatives (the British Computing Society, Pinar Akman (an academic), and Hugh Milward, Director of Corporate, Legal and External Affairs at Microsoft) stated that it was it is hard to predict what companies could be considered "tech" companies as new areas of digitisation emerge. Pinar Akman gave the example of Amazon acquiring Whole Foods and asking whether they were in the same industry or not, an important consideration in competition law and one that the rate of digitisation across industries has made complicated to answer. Select Committee on Communications, *Regulating in a digital world*, 40; Select Committee on Communications, "The internet: to regulate or not to regulate. Uncorrected transcript of evidence given by Hugh Milward, Director of Corporate, Legal and External Affairs, Microsoft; Katie O'Donovan, Public Policy Manager, UK, Google; Rebecca Stimson, Head of Public Policy, UK, Facebook.," Houses of Lords, last modified October 30, 2018, http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/oral/92263.html.

[653] This same critique is made of the German Network Enforcement Act as as it provides obligations for social media companies on the removal of hate speech and other proscribed categories of content but does not implement similar obligations on the protection of free expression. Human Rights Watch strongly criticised the Act, explaining that "the law fails to provide either judicial oversight or a judicial remedy should a cautious corporate decision violate a person's right to speak or access information. In this way, the largest platforms for online expression become "no accountability" zones, where government pressure to censor evades judicial scrutiny." Human Rights Watch, "Germany: Flawed Social Media Law."

Internet.[654] Many of these harms, such as "the glamorisation of gang life" would be perfectly legal in other contexts and it is concerning that this White Paper may prompt legislation rendering this content impermissible online.[655] This outcome has already happened in some countries after the 2016 fake news scandal, which produced a flurry of national legislation aimed at the problem of fake news without any consideration of how this regulation would address issues such as censorship.[656] Even if substantive regulation is subsequently used to protect human rights considerations, it is unlikely that these two categories of legislation would converge easily and it is possible that the result would be significantly watered-down human rights protections in the online world.[657]

It should also be emphasised that this form of regulation does not place enough emphasis on procedural protections. Its objective is either the specific prohibition of something or introducing a new obligation on social media companies. Little consideration is afforded to how these regulatory schemes should be implemented or what principles should be integrated into these processes. Conversely, as this thesis has argued, procedural regulation is essential to ensure that content moderation processes are transparent, accountable, and respect both human rights and the rule of law.

At the heart of these regulations is the assumption that the old laws simply cannot cope with the unprecedented harms that have arisen from new technologies.[658] This is an example of technological determinism (also sometimes called technological exceptionalism) where the introduction of a technology "into the mainstream requires a systematic change to the

---

[654] DCMS, *Online Harms White Paper* (London: HM Government, 2019).
[655] For example in films, TV shows, and music (especially music that reflects urban realities such as rap music). It should also be noted that one person's interpretation of "glamorisation" might match another person's attempt to depict a way of life in a nuanced, sensitive way. The famous rap group N.W.A. was frequently criticised for creating music that promoted violence and criminality. In the N.W.A. biopic *'Straight Outta Compton'* a news reporter challenges the band about their glamorisation of gang life, drugs, and violence. The rapper Ice Cube responds "our art is a reflection of our reality. What do you see when you go out your door? I know what I see, and it ain't glamorous…freedom of speech includes rap music." Unfortunately, while freedom of speech might include rap music, proposals such as the Online Harms White Paper would ensure that it doesn't include social media. DCMS, *Online Harms White Paper*, 13, 67.
[656] Germany's Network Enforcement Act was used as inspiration for similar laws in the Philippines, Russia, and Singapore, prompting concerns that the issue fake news would be used by governments to crack down on content they perceived as anti-government. Human Rights Watch, "Germany: Flawed Social Media Law."
[657] Which is not that different from the current status quo in the online world.
[658] In many ways, this is the polar opposite of the approach that argues that social media companies should be regulated like any other media company and the same regulations and principles should apply.

law or legal institutions in order to reproduce, or if necessary displace, an existing balance of values."[659] While it is clear that the Internet has radically altered society, this narrative must be scrutinised carefully to ensure that substantive regulation is not used as a method of circumventing established legal protections.[660] One example of this was the 2014 proliferation of content by ISIS, which led to an intense interest in regulating terrorist content on social media.[661] France, for example, introduced a law that empowered government authorities to order Internet Service Providers (ISP's) to block websites that promote terrorism without the need for a court order.[662] There may be times when new laws are needed for the online environment but it essential to scrutinise legislative attempts to ensure that important legal principles are not being discarded. There may be situations where new legislation is not even necessary, and where the "pacing problem" (the notion that law must keep up with technology) is nothing but a straw man argument that means that lawmakers must "unnecessarily accept a degree of irrelevance."[663]

Finally, a serious weakness of substantive regulation is that its specificity means that it can become irrelevant quickly as technology develops and new services are introduced.[664] This is the inherent flaw in laws driven by techno-determinism because by arguing that these

---

[659] Ryan Calo, "Robotics and the Lessons of Cyberlaw," *California Law Review* 103, no. 3 (2015): 552.
[660] A similar point is made by Mac Síthigh, who explains that "An account of media regulation that is unduly influenced by technological determinism would not be useful. There is little for the legal scholar to say if broad assumptions about the consequences of technological change are left unchallenged." Daithí Mac Síthigh, *Medium law*, Routledge studies in law, society and popular culture, (Abingdon, Oxon: Routledge, 2017). 10.
[661] For an overview of how ISIS used social media in 2014, see: Faisal Irshaid, "How ISIS is spreading its message online," BBC News, last modified June 19, 2014, https://www.bbc.co.uk/news/world-middle-east-27912569.
[662] Of course, this law was also a response to the Charli Hebdo shootings but it is clear that there was a lot of concern about online terrorist content at that time in France. Décret du 6 Février 2015 relatif au blocage des sites provoquant à des actes de terrorisme ou en faisant l'apologie et des sites diffusant des images et représentations de mineurs à caractère pornographique. (Decree of February 6, 2015 related to the blocking of sites that provoke acts of terrorism or act as an apology for terrorism and the websites that disseminate pornographic images or representations of minors). *Décret n° 2015-125 du 5 février 2015 relatif au blocage des sites provoquant à des actes de terrorisme ou en faisant l'apologie et des sites diffusant des images et représentations de mineurs à caractère pornographique*.
[663] After all, as Meg Leta Jones makes clear: "If technology is the driving force of law, law will always follow technology." Meg Leta Jones, "Does Technology Drive Law: The Dilemma of Technological Exceptionalism in Cyberlaw," *University of Illinois Journal of Law, Technology and Policy* 2 (2018): 256, https://doi.org/10.2139/ssrn.2981855; Gary Elvin Marchant, Braden R. Allenby, and Joseph R. Herkert, *Growing gap between emerging technologies and legal-ethical oversight: the pacing problem*, International library of ethics, law and technology,, (Dordrecht: Springer, 2011), 22-23.
[664] Reed, "How to Make Bad Law: Lessons from Cyberspace." 905.

technologies are exceptional and require specific laws they are building in a measure of obsolescence that will be quickly reached. Regulators are then faced with the challenge of either interpreting this old legislation for technologies that were not envisioned by the original drafters or by passing new substantive legislation. This form of regulation is frequently referred to as "rules-based regulation" whereby the regulator specifically addresses how companies must comply with these rules with little discretion afforded to them.[665]

Instead, there are increasing calls for laws to be technologically neutral, which is defined as "legislation which targets specific types of behaviour regardless of the medium."[666] These regulations outline a set of principles that can be applied to a myriad of different situations whether they have been foreseen or not and act as an overarching framework through which specific laws can be articulated and against which compliance can be measured. Technology-neutral laws offer the promise of more sustainability, that the law will not require frequent amendments to remain relevant.[667] They are particularly beneficial in industries that are rapidly developing such as the technology sector and they allow for more detail to be developed through guidelines, codes of practice and certification that flow from the principles." [668] Of course, making laws technologically neutral does not guarantee their permanent relevance and laws regulating the digital sector will still require more frequent updates than other, more settled areas of law.[669] It should also be noted that principles-based regulation provides slightly less certainty to companies (who may be unsure whether they are in compliance) but this uncertainty can be undercut by affording companies a measure of discretion in how they meet these objectives and focussing on due diligence

---

[665] Select Committee on Communications, *Regulating in a digital world*, 14.
[666] Select Committee on Communications, *Regulating in a digital world*, 14.
[667] This principle of sustainability was outlined in Fuller's Morality of Law, where he cautions that "a law that changes every day is worse than no law at all." Fuller, *The morality of law*, 37; Bert-Jaap Koops, "Should ICT Regulation Be Technology-Neutral," in *Starting Points for ICT Regulation, Deconstructing Prevalent Policy One-Liners*, ed. Bert-Jaap Koops et al. (The Hague: TMC Asser Press, 2006), 87.
[668] Q115: Select Committee on Communications, "The internet: to regulate or not to regulate. Corrected transcript of evidence given by Elizabeth Denham, Information Commissioner, Information Commissioner's Office (ICO)," Houses of Lords, last modified September 11, 2018, http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/oral/89766.html.
[669] Koops, "Should ICT Regulation Be Technology-Neutral," 88.

procedures.[670] It is also important that the law does not abstract so much from the technology that it lacks specificity, as this will also cause uncertainty.[671] Laws that focus on principles and are technologically neutral, however, are still more promising than specific, directed regulation that solve today's problem with no thought of tomorrow's challenge.

In conclusion, it must be acknowledged that some substantive regulation is necessary in any regulatory scheme and that it is within a country's discretion to identify issues on social media about which they are particularly concerned, such as the posting of content promoting Holocaust denial in those countries that explicitly ban it.[672] Substantive regulation can be used to rectify gaps in current regimes and can be targeted in its application and responsive to emerging issues. It can also complement other forms of regulation that are more comprehensive in their application. The issue, however, is when this checkerboard regime of substantive regulation is treated as a panacea for the problems that exist on social media platforms and when societal issues result in reflexive, piecemeal regulation from governments. The objective instead should be to introduce comprehensive principles-based legislation that primarily focusses on procedural issues in social media content moderation. The substantive regulations that would subsequently be introduced would exist in a fundamentally different digital landscape and would be applied in a way that is consistent with the overarching principles and within a framework that reflects a procedural rule of law.

---

[670] These due diligence requirements will be explored in detail in the next chapter.
[671] Koops, "Should ICT Regulation Be Technology-Neutral," 98.
[672] Facebook has been particularly resistant to removing such content. Elizabeth Schumacher, "Facebook refuses to censor Holocaust Denial as social media sites struggle with German laws," Deutsche Welle, last modified July 27, 2018, https://www.dw.com/en/facebook-refuses-to-censor-holocaust-denial-as-social-media-sites-struggle-with-german-laws/a-44855519.

6.4: Duty of Care

6.4.1: What does the duty of care entail?

The newest and most radical proposal for addressing issues in social media content moderation is the notion of a social media duty of care, introduced in 2018 by Lorna Woods and William Perrin for the Carnegie Trust.[673] The concept of a duty of care for social media platforms is beginning to pick up speed in the UK. It was supported by the NSPCC in their report on online grooming and CSAM[674] and was discussed in the House of Lords report on regulating the internet.[675] The Perrin and Woods report has also clearly been a source of inspiration for the Online Harms White Paper, which explicitly called for a social media duty of care and the government response to the initial consultation, which reaffirmed this stance.[676]

The Perrin and Woods report argues that there are a significant number of harms on social media platforms that the companies are not addressing in an adequate manner. These harms are diverse but many of them are related to the content that is available on the platform.[677] Perrin and Woods advocate for a statutory duty of care whereby social media companies would be obligated to identify the harms that could arise on their platforms and take steps to address these harms to ensure their service was safe for users.[678] The report advocates a harm reduction cycle where harms are measured, changes are implemented, then harms are measured again and the process continues.[679] This is not an obligation to eliminate all harms but rather to take "sufficient care" to avoid systemic failures in their systems. [680] They emphasise safety in design and ensuring the processes they use to

---

[673] William Perrin and Lorna Woods, *Online harm reduction: a statutory duty of care and regulator* (Dunfermline: Carnegie UK Trust, 2019).
[674] *Taming the Wild West Web: How to regulate social networks and keep children safe from abuse* (London: NSPCC, 2019), 8-11.
[675] Select Committee on Communications, *Regulating in a digital world*, 53.
[676] DCMS, *Online Harms White Paper*, 65; DCMS, *Online Harms White Paper: Initial Consultation Response* (London: HM Government, 2020).
[677] See, for example, the categories of harms detailed in Perrin and Woods, *Online harm Reduction*, 35-42.
[678] Perrin and Woods, *Online harm Reduction*, 40.
[679] Perrin and Woods, *Online harm Reduction*, 43, 48.
[680] Perrin and Woods, *Online harm Reduction*, 29; Select Committee on Communications, *Regulating in a digital world*, 53.

moderate content do not have an unacceptable level of risk. The report uses analogies from other industries that manage harms and other situations where a duty of care may arise to outline the contours of their regulatory scheme. Ofcom would be given the responsibility of enforcing this duty of care and the companies would be fined if they did not have the requisite processes in place for addressing these harms.[681]

A similar approach is adopted by the government White Paper, which enumerated a variety of different harms before concluding that a duty of care for social media companies needed to be introduced. Platforms would have a general duty to "to take reasonable steps to keep their users safe and tackle illegal and harmful activity on their services."[682] In addition, platforms would also have to comply with specific codes of conduct in relation to certain illegal harms such as Child Sexual Abuse Material (CSAM) and terrorist content.[683] These codes will outline the policies and processes platforms need to adopt to meet their duty of care, including technology, training, investment, and staffing. This duty of care will be overseen and enforced by a regulator, which will likely be Ofcom.[684]

6.4.2: The Advantages of a Duty of Care

There are some advantages in using a duty of care model for addressing issues in social media content moderation. First, it represents an attempt to shift the balance of power between businesses and the state, "by giving parliament and a regulator responsibility for setting the terms for what the UK, as a society, considers harmful and wants to eradicate."[685] The current emphasis on self-regulation (except in limited circumstances such as CSAM) affords social media platforms a significant amount of power in how they regulate content and this approach could help rebalance the relationship between public and private bodies. Designating a regulator for social media users also benefits users who have an inequality of

---

[681] Perrin and Woods, *Online harm Reduction*, 56; DCMS, *Online Harms White Paper: Initial Consultation Response*.
[682] DCMS, *Online Harms White Paper*. 42.
[683] DCMS, *Online Harms White Paper*. 8.
[684] DCMS, *Online Harms White Paper*. 42-43. DCMS, *Online Harms White Paper: Initial Consultation Response*.
[685] Jacob Ohrvik-Stott and Catherine Miller, *Digital duty of care: Doteveryone's perspective* (London: Doteveryone, 2019).

arms against large platforms.[686] The government White Paper even envisions a role for "super-complaints," which are complaints made by designated organisations to the regulator when they are concerned that users are not receiving redress in serious situations.[687] The White Paper also envisions a range of very strong enforcement and sanctioning methods such as issuing fines, publishing public notices, disrupting business activities through the removal of the company from search results, app stores, or links in posts, ISP blocking, and senior management liability (both civil and criminal).[688]

Second, it obliges companies to consider the risks of the services and products they are developing and holds them responsible for a lack of foresight. This approach is the opposite of the Silicon Valley maxim of "move fast and break things" and could help to prevent the controversies that can erupt from hastily introduced products (such as the live-streaming of crimes at 4.3.2). These consequences are often treated as externalities by platforms, who might consider the social impact of their technologies as being beyond the scope of their responsibility especially in the early stages of the company. The duty of care approach forces companies to "internalise these costs" which seems fitting as they are the direct beneficiaries when these services become successful.[689]

Finally, a duty of care approach represents an attempt to create an objective legal framework that one could compare the actions of social media companies against and respond accordingly.[690] This proposal to create a technology-neutral framework could offer more certainty to both users and companies and would help avoid the reactive approach where new legislation is introduced only after a scandal erupts. The strongest aspect of the report is the architecture Perrin and Woods envision and how detailed it is in its explanation of funding, implementation, and sanctions. Their framework emphasises due diligence (although that term is not explicitly used) and the creation of a regulator to hold platforms

---

[686] Select Committee on Communications, *Regulating in a digital world*, 53.
[687] DCMS, *Online Harms White Paper*. 46.
[688] DCMS, *Online Harms White Paper*. 59-60.
[689] Perrin and Woods, *Online harm Reduction*, 8; Ohrvik-Stott and Miller, *Digital duty of care*.
[690] Of course, as the next section will argue, it is difficult to have an objective debate while using the highly subjective framework they employ in relation to harms. "Perrin and Woods Duty of Care," Open Rights Group Wiki, last modified January 22, 2019, https://wiki.openrightsgroup.org/wiki/Perrin_and_Woods_Duty_of_Care.

accountable and ensure they do have the necessary procedures in place to prevent and manage issues.[691] The same advantage exists in the government White Paper, which emphasises the role of codes of conduct and the implementation of a "transparency, trust, and accountability framework" by the regulator. The White Paper also states that the regulator should ensure that platforms comply with their own terms and conditions in a consistent manner and have effective "user redress mechanisms,"[692] which are issues that have been discussed throughout this thesis and which are important aspects of a strong regulatory regime.

6.4.3: Problems with the Duty of Care Model:

6.4.3.i: The conceptualisation of harm

The central issue with Perrin and Woods' proposal, however, is the organising principle they have built their regulatory framework around: the concept of harm. Harm is poorly defined in the duty of care report and instead the report provides many examples of harms such as cyberbullying, sexual harassment, and fraud.[693] Examples, however, fail to properly delineate the boundary of what would constitute harm in their view. The same criticism can be levied against the Online Harms White Paper, which never defines harms, choosing instead to provide examples of 23 different kinds of harm and a list of harms that will be excluded including harms to organisations, harms resulting from data protection issues, harms from cybersecurity and harms stemming from the dark web.[694] This is unfair as platforms need a measure of certainty as to what they would be required to address to avoid

---

[691] The proposal I outline in Chapter Seven has these features as well but organised along a much different set of principles.
[692] DCMS, *Online Harms White Paper*. 54.
[693] Perrin and Woods do explain why they decided not to offer a definition of harm. First, they felt it would be more appropriate for the regulator to define harms. This is problematic because it is the equivalent of stating "first let's implement this whole scheme and then we can actually discuss what we would be regulating." Second, they explain that harms can be identified from broad descriptions and that bodies like Ofcom have expertise in defining and identifying harmful content. These answers seem incomplete and fail at convincing the reader that a definition of harm, the central objective of this regulatory scheme, is unnecessary, especially as the regime they envision is significantly broader and more onerous than the regime imposed on broadcasters. Perrin and Woods, *Online harm Reduction*, 41.
[694] DCMS, *Online Harms White Paper*. 31-32.

sanctions, and users deserve certainty on laws that affect their speech.[695] The free speech NGO Index on Censorship has accordingly raised concerns that demanding that social media companies regulate "harmful content" creates confusion as there are no agreed definitions of what is harmful beyond what has already been made illegal.[696] It is also extremely hard for academics or policy-makers to assess the validity of a scheme that has such a vaguely defined subject because the scope of the definition would have a bigger impact on this proposal's efficacy than any other aspect. It would be the equivalent of a report claiming that new laws are needed to address "problems" in social media. The same problems existed in earlier attempts to articulate the human rights responsibilities of businesses, with the UN Norms using the poorly defined "corporate spheres of influence" as a guiding principle. Perrin and Woods do go on to explain that the appointed regulator would define the parameters of harm and that it would be better for a regulator to delineate these harms in detail because they would not be a political actor like Parliament.[697] This definitional plan is relatively vague and fails to address the fact that a regulator is less directly accountable to the public than politicians so there may be some concerns in allowing them to essentially define their own mandate.[698]

Failing to supply a definition of harm might be excusable if an established legal regime was truly being imported wholescale into a new context. The report's conceptualisation of harms, however, is much broader than any current legal definition. For example, in relation to the category of "emotional harms" the report states "we suggest that emotional harm is reasonably foreseeable on some social media and that services should have systems in place

---

[695] Graham Smith, "Rule of Law and the Online Harms White Paper," Cyberleagle, last modified May 5, 2019, https://www.cyberleagle.com/2019/05/the-rule-of-law-and-online-harms-white.html; Graham Smith, "Users Behaving Badly – the Online Harms White Paper," Cyberleagle, last modified April 18, 2019, https://www.cyberleagle.com/2019/04/users-behaving-badly-online-harms-white.html.

[696] "Wider definition of harm can be manipulated to restrict media freedom," Index on Censorship, last modified February 18, 2019, https://www.indexoncensorship.org/2019/02/wider-definition-of-harm-can-be-manipulated-to-restrict-media-freedom/.

[697] Perrin and Woods, *Online harm Reduction*, 40.

[698] This concern was echoed by the House of Lords select committee on communications .It was best articulated by Laurie Laybourn-Langton, Senior Research Fellow, Institute for Public Policy Research, who said that addressing regulatory challenges should be "undertaken according to democratic principles, in the same way that we have provided regulation in other key areas of society and the economy through a democratic mechanism—Parliament and the people who represent us." Select Committee on Communications, *Regulating in a digital world*, 3.

to prevent emotional harm suffered by users such that it does not build up to the current threshold of a recognised psychiatric injury."[699] Another controversial inclusion was the category "harms to justice and democracy" and that which examined problems that "impact society as a whole."[700] Both of these kinds of harm could be extremely difficult for platforms to predict and neither of which appear actionable offline. There are similar examples in the White Paper such as "coercive behaviour" and "intimidation."[701] This broad conception of harm is particularly challenging for social media companies as a duty of care is usually applied in situations where the harm is an "objectively ascertainable injury" and can be managed "by the responsible party through their own prior actions"[702] which may not be the case in a social media duty of care.

The use of the word "harm" seems to obscure the fact that in many situations, both the government and Perrin and Woods are actually referring to content, and therefore speech. This might be intentional as the human rights implications of the report and White Paper are particularly troubling (this will be discussed at 6.4.3.iv) and substituting words like speech and content with harm can allow for more regulatory eclecticism and comparison to industries that do not regulate speech, such as the environmental agencies. Perrin and Woods also claim that focussing on harm-reduction means that their policy is content-neutral as "no particular content is targeted by the regulation; indeed, no types of speech are so targeted."[703] This is circular reasoning, however, as many of the harms they list are clearly a manifestation of particular kinds of content such as terrorism, hate speech, and threats to kill.[704] It is perfectly reasonable to regulate such content on social media platforms so it seems strange to claim to be content-neutral as a matter of semantics rather than reality.

---

[699] The same problem exists with the Online Harms White Paper. Perrin and Woods, *Online harm Reduction*, 38.
[700] Perrin and Woods, *Online harm Reduction*, 39.
[701] DCMS, *Online Harms White Paper*. 31.
[702] Open Rights Group Wiki, "Perrin and Woods Duty of Care."; Graham Smith, "A Ten Point Rule of Law Test for a Social Media Duty of Care," Cyberleagle, last modified March 16, 2019, https://www.cyberleagle.com/2019/03/a-ten-point-rule-of-law-test-for-social.html.
[703] Perrin and Woods, *Online harm Reduction*, 39.
[704] Perrin and Woods, *Online harm Reduction*, 36, 38.

Harm is a loaded word, a call for a legal response to protect people, but Perrin and Woods fail to properly demonstrate that these harms exist at all or that they are any different from the issues that exist in offline society.[705] This is essential in a report advocating the imposition of a wide range of obligations on social media platforms as this set of reforms will be complex and likely to raise human rights concerns, so one should be assured that these harms exist and that this approach could ameliorate them. A similar problem exists in the government White Paper, which presumes that everyone is in agreement on what constitutes an actionable harm and makes over-general statements like "illegal and unacceptable content and activity is widespread online."[706] It is also generally expected that duties of care are applied to situations where the harm can be clearly identified when it occurs (such as physical injury or financial loss) or, in cases where a broader set of harms are considered, or where a special relationship exists between the two parties such as a relationship of employment.[707] Instead, Perrin and Woods briefly explain the precautionary principle, which states "on the basis of the best scientific advice available in the time-frame for decision-making: there is good reason to believe that harmful effects may occur to human, animal or plant health, or to the environment; and the level of scientific uncertainty about the consequences or likelihoods is such that risk cannot be assessed with sufficient confidence to inform decision-making."[708] While it is already controversial to apply the precautionary principle to speech (as many of the categories in the report constitute), this principle is still predicated on the existence of scientific advice and a "good reason" to believe that harm is occurring. Perrin and Woods do not, however, explicitly cite the diverse array of research that would be necessary to demonstrate that each of the issues they have

[705] This is particularly important as many of the problems they mention pre-date the Internet such as misogyny, threats, and terrorism so there would be a responsibility to demonstrate that the current legal approaches are insufficient.
[706] DCMS, *Online Harms White Paper*.5.
[707] Open Rights Group Wiki, "Perrin and Woods Duty of Care."; Graham Smith, "Take care with that social media duty of care," Cyberleagle, last modified October 19, 2018, https://www.cyberleagle.com/2018/10/take-care-with-that-social-media-duty.html.
[708] United Kingdom Interdepartmental Liaison Group on Risk Assessment, "The Precautionary Principle: Policy and Application," Health and Safety Executive, last modified July 17, 2018, http://www.hse.gov.uk/aboutus/meetings/committees/ilgra/pppa.htm.

identified is having a harmful effect on humans. Instead, they simply conclude, without further references:

> "We believe that – by looking at the evidence in relation to screen use, internet use generally and social media use in particular – there is in relation to social media "good reason to believe that harmful effects may occur to human[s]" despite the uncertainties surrounding causation and risk. On this basis we propose that it is appropriate if not necessary to regulate…"[709]

By failing to properly justify the existence of social media harm, Perrin and Woods's proposals seem more tenuous and less urgent. The precautionary principle argument is just a claim that one should be "better safe than sorry" and while that might be perfectly reasonable when considering, for example, the safety of consuming alcohol while pregnant or the effects of vaping tobacco, it seems less reasonable when discussing speech-based issues. It is hard to assess the validity of a scheme that purports to be objective but renders even the definition of harm or its existence in any given scenario as subjective. Even though harm is the central feature of the duty of care plan for social media platforms, it also represents the primary weakness of the scheme. A similar flaw exists in the White Paper, which presumes all of the issues on social media should be converted into legal obligations for platforms to address, failing to heed Lord Rodger's famous warning that "the world is full of harm for which the law furnishes no remedy."[710]

6.4.3.ii: The problem of analogies

The Perrin and Woods duty of care proposal (but not the government White Paper) is replete with analogies, comparing social media to an office, bar, theme park, waste disposal unit, and polluting company. [711] The one analogy they explicitly reject is the publisher/distributer analogy citing "the need to deploy regimes and enforcement at scale."[712] While there is nothing inherently objectionable in analogies, there are a number of

[709] Perrin and Woods, *Online harm Reduction*, 11.
[710] *JD v East Berkshire Community Health NHS Trust and others* 2 WLR 993(2005). At [100].
[711] Perrin and Woods, *Online harm Reduction*, 12.
[712] Perrin and Woods, *Online harm Reduction*, 12, 21.

issues in how Perrin and Woods employ them to explain and justify their duty of care proposal.

First, the reasoning employed in choosing these analogies is inductive instead of deductive, with Perrin and Woods selectively employing comparisons to satisfy the requirements of their theory. They never really justify their analogies, simply claiming that "social media networks should be seen as a public or (given that they are privately owned) a quasi-public place – like an office, bar, or theme park."[713] This statement fails to explain why a platform should be treated like a public place (or a "corporate owned public space"[714]) but instead jumps straight to concluding that it would be expedient to do so, stating that "an appropriate analogy for social media platforms is that of a public space. The law has proven very good at this type of protection in the physical realm."[715] The analogies employed in the report (such as a safety issue at a theme park or a company dumping hazardous waste) naturally lead one to conclude that everyone would be better off if the harm had never occurred and this is concerning when it is applied to content issues where there may be valid arguments on both sides. Indeed, duties of care typically assume that it is "uncontroversial that the harm event should be avoided, even if there is controversy about who should manage that risk."[716] This distinction between the social media situation and the orthodox duty of care approach is clearly demonstrated when Perrin and Woods equate social media harms with pollution.[717] If harms are often a proxy for speech then it is strange to compare them to pollution, which is inherently negative and should be limited as much as possible. This is reminiscent of Orin Kerr's assessment of the tendency to regulate the internet by analogy, and how one must be careful because the analogy we choose will likely affect the

[713] Perrin and Woods, *Online harm Reduction*, 12.
[714] Perrin and Woods, *Online harm Reduction*, 28.
[715] Perrin and Woods, *Online harm Reduction*, 28.
[716] Open Rights Group Wiki, "Perrin and Woods Duty of Care."
[717] "Harm emanating from a company's activities has, from a micro-economic external costs perspective, similarity to pollution and we also discuss environmental protection." Perrin and Woods, *Online harm Reduction*, 21.

conclusions we draw.[718] In the Perrin and Woods report, it appears as if the conclusions were drawn and then the analogies were chosen accordingly.

Analogising social media to a public space like a theme park is questionable in of itself but Perrin and Woods also do not outline a duty of care that corresponds to that analogy. First, the majority of duties owed to visitors of a public space are safety-related such as personal injury and damage to the property and "relate to what is done, not said, on their premises."[719] Conversely, the Perrin and Woods duty of care is much broader, encompassing all manner of activities that relate to how users interact with each other in the online world. Admittedly, there are situations where a duty of care is applied to protect visitors from harming each other (such as in a bar or at a football game) but that duty of care only applies to physical injury.[720] The report insists that "a mass membership, general purpose service open to children and adults should manage risk by setting a very low tolerance for harmful behaviour, in the same way that some public spaces, such as say a family theme park take into account that they should be a reasonably safe space for all"[721] but fail to explain why speech on social media networks should be equated with safety in the offline world in their duty of care; as theme parks do not monitor what visitors say to each other and theme parks are not liable for what a third party says to a parkgoer.

Perrin and Woods claim that general laws requiring harm reduction work particularly well in multifunctional places like houses, parks, and pubs as "duties of care work in circumstances where so many different things happen that you would be unable to write rules for each one." [722] This is a misconception as the harms in these spaces are relatively circumscribed (physical injury and damage to property) but rather it is the *methods* of sustaining this harm that can be diverse (so the law must consider all kinds of

---

[718] Orin Kerr, "The problem of perspective in internet law," *Georgetown Law Journal* 91 (2003): 87. Orin Kerr is also referencing Froomkin's work on analogies. See: A. Michael Froomkin, "The Metaphor is the Key: Cryptography, the Clipper Chip, and the Constitution," *University of Pennsylvania Law Review* 143 (1995): 718.
[719] See, for example, the *Occupiers Liability Act, 1957*, c. 31 (Eng. and Wales).. For more on this analogy: Smith, "Take care."
[720] See, for example, *Cunningham v. Reading Football Club Ltd*, 153 Times LR(1991); *Everett v Comojo (UK) Ltd (t/a Metropolitan)*, EWCA Civ 13(2011).
[721] Perrin and Woods, *Online harm Reduction*, 43.
[722] Perrin and Woods, *Online harm Reduction*, 28.

physical injury). That is quite a different proposition from the myriad of harms that Perrin and Woods identify in their report.

The entire report hinges on analogies. It is arguing that one must regulate social media companies *as if* they were a public space and one must handle online speech *as if* it were a safety issue in the offline world. Unfortunately, as this section has made clear, the analogies the report uses seem particularly ill-chosen for the challenges that these platforms pose. This is regulation by slogan, a quotable idea that policymakers and journalists can disseminate with ease. Perrin and Woods could have chosen different analogies, situations where a duty of care was applied in relation to the things visitors said to each other. The problem with this approach is that no such duty exists and so Perrin and Woods are forced to argue that the harms that exist on social media platforms are different from offline spaces while simultaneously insisting that these spaces are still analogous, making what Graham Smith terms "an argument from difference, not similarity."[723] It is clear that a more coherent approach to addressing issues in social media content regulation is to set aside the analogies, scrutinise the issues that exist online, and suggest an approach that reflects the reality of the online world.[724]

6.4.3.iii: Too Onerous for Social Media Platforms

While the self-regulatory approach is overly lenient on social media companies, a social media duty of care imposes a heavy burden on platforms. This is problematic for a number of reasons. First, platforms will expend more energy and resources contesting a set of regulations that they perceive as excessively harsh. This is not an absolute barrier to regulation but it will make it more difficult to achieve compliance if the regulatory scheme is not perceived as legitimate by the regulatees as the best regulatory schemes are neither "solely deterrent" nor "solely cooperative."[725] Second, there are surely valid concerns as to whether it is even possible for platforms to comply with such a broad, heavy-handed

[723] Smith, "Take care."
[724] The next chapter will attempt to do this and does not employ analogies but rather argues that social media companies, as businesses, have a responsibility to respect human rights and that states can obligate platforms to engage in human rights due diligence to meet this responsibility.
[725] Vibeke Lehmann Nielsen and Christine Parker, "Testing responsive regulation in regulatory enforcement," *Regulation and Governance* 3, no. 4 (2009): 376, https://doi.org/10.1111/j.1748-5991.2009.01064.x.

approach to regulation. This raises questions of whether the proposed law has sufficient output legitimacy, which focusses on the nature of the rules and whether they are likely to achieve their desired results.[726] Reed and Murray contend that "even if its aims are fair and just, it is neither fair not just to demand that the law's addressee should engage in behaviour which is unlikely to result in those aims being achieved."[727] Finally, these strict requirements increase the risk that the only way to achieve compliance is to become censorial and overly risk-averse, thus diminishing the positive aspects of social media. This is even more likely when one considers the stringent sanctions considered by the Online Harms White Paper.

Overall, the expectations of what platforms can achieve are extremely high. Perrin and Woods state "we list here some areas that are already a criminal offence –the duty of care aims to prevent an offence happening and so requires social media service providers to take action before activity reaches the level at which it would become an offence." [728] This emphasis on intervening before conduct reaches the level of an offence is related to the report's assertion that "given the lax enforcement of the criminal law, it is unlikely that the existence of the criminal offence has much deterrent effect."[729] This seems like an overly critical assessment of criminal law and an effort to shift some of the responsibilities (and compliance costs) of the criminal law system onto social media platforms, which raises the same accountability, legitimacy, and transparency issues as allowing platforms to make human rights decisions. The high expectations in the duty of care model may also unjustly penalise platforms if harms increase for reasons that are beyond their control.[730]

The duties Perrin and Woods envision for social media platforms also have an overly broad scope. For example, the report does not confine their duty of care framework to users

---

[726] Paiement, "Paradox and Legitimacy in Transnational Legal Pluralism," 213-15; Reed and Murray, *Rethinking the jurisprudence of cyberspace*, 174.
[727] Reed and Murray, *Rethinking the jurisprudence of cyberspace*, 194.
[728] William Perrin and Lorna Woods, "Reducing Harm In Social Media Through A Duty Of Care," Carnegie UK Trust, last modified May 8, 2018, https://www.carnegieuktrust.org.uk/blog/reducing-harm-social-media-duty-care/.
[729] Perrin and Woods, *Online harm Reduction*, 31.
[730] Ohrvik-Stott and Miller give the example of hate speech increasing as a political climate grows more polarised. Graham Smith argues that for the proposed duty of care to have "sufficient certainty and precision" then "the risk of any particular harm has to be causally connected (and if so how closely) to the presence of some particular feature of the platform." Ohrvik-Stott and Miller, *Digital duty of care*; Smith, "A Ten Point Rule of Law Test for a Social Media Duty of Care."

of the social media service, which would already be difficult enough as they would be managing the interactions between thousands if not millions of users. Instead, the report explains that "people are harmed by content on social media and messaging services when they themselves are not customers of those services."[731] Perrin and Woods give the example of revenge porn posted on a service where the complainant is not a customer and state that "extending the statutory duty to individuals who are not users of the service is important as it is far from certain that, under the common law duty of care, a duty would arise to such an individual."[732] While this is a very sympathetic example (which is repeated in the White Paper)[733] these writers once again fail to articulate the scope of the obligations imposed on platforms, seeming to articulate an unlimited duty to an unlimited number of people simply because it would be expedient for such a duty to exist. These standards appear unacceptably high and will be likely to produce a high degree of uncertainty for platforms as to what is required of them.

Perrin and Woods also criticise the moderation processes at platforms as "unacceptably opaque and slow."[734] While this thesis has identified serious issues with transparency at every stage in the moderation process, it is not accurate to call these moderation processes slow. A third of all content that is flagged on Facebook is reviewed by algorithms at the moment it is posted (algorithms even prohibit some content from being posted at all)[735] and YouTube has stated that most videos that violate their terms and conditions are flagged and removed within an hour of dissemination.[736] This thesis has consistently shown that while platforms have many issues in their content moderation process, speed has never been one of them, and indeed this thesis has actually criticised platforms for a *preoccupation* with speed at the expense of more nuanced concerns (see 4.4.2). It is unclear whether this criticism was made from a lack of knowledge or as an attempt to justify the imposition of

[731] Perrin and Woods, *Online harm Reduction*, 31.
[732] Perrin and Woods, *Online harm Reduction*, 31.
[733] "This broader application of the duty, beyond simply users of a particular service, recognises that in some cases the victims of harmful activity – victims of the sharing of non- consensual images, for example – may not themselves be users of the service where the harmful activity took place." DCMS, *Online Harms White Paper*. 42.
[734] Select Committee on Communications, *Regulating in a digital world*, 53.
[735] Zuckerberg, "Building Global Community."
[736] Foxman and Wolf, *Viral hate*, 107.

stricter measures on the platforms, but either way it is inaccurate and overly harsh on platforms.

6.4.3.iv:. Free expression issues are poorly addressed

The primary human rights concern in both the Perrin and Woods report and the government White Paper is the threat to users' Article 19 right to free expression.[737] Perrin and Woods frequently address free expression in their proposals but their answers are often rushed or unsatisfactory.[738] For example, after a number of critics raised concerns that their proposed social media regulator (as well as the regulator in the Online Harms White Paper) would be making decisions that could cause human rights concerns,[739] Perrin and Woods responded by explaining that as the regulator would be a public body, section six of the Human Rights Act 1998 (which requires public bodies to carry out their duties in accordance with Convention rights) would apply.[740] This is a weak answer because their scheme envisions platforms having a wide discretion in how they produce the requisite outcomes so the real free expression issue is not in how the regulator applies the law (although that would still be a concern) but rather how platforms will be legally incentivised to censor an increasing amount of content, thus exacerbating human rights issues that already exist on these platforms. Accordingly, Graham Smith argues that any conceptualisation of a duty of care should be able to address situations where the exercise of a duty of care would cause collateral damage to lawful speech and there should be clear tests for when such a duty of care would be negated by those risks.[741]

---

[737] *ICCPR*.
[738] This criticism was also raised by the Open Rights Group who said that "the proposal claims to sidestep free expression concern but doesn't justify that or explain how." Open Rights Group Wiki, "Perrin and Woods Duty of Care."
[739] See, for example: Graham Smith, "A Lord Chamberlain for the internet? Thanks but no thanks," Inforrm's Blog: International Forum for Responsible Media Blog, last modified October 21, 2018, https://inforrm.org/2018/10/10/a-lord-chamberlain-for-the-internet-thanks-but-no-thanks-graham-smith/.
[740] William Perrin and Lorna Woods, "Whose duty is it anyway? Answering some common questions about a duty of care," Carnegie UK Trust, last modified August 2, 2019, https://www.carnegieuktrust.org.uk/blog/duty-of-care-faq/.
[741] Smith, "A Ten Point Rule of Law Test for a Social Media Duty of Care."

Free expression is also barely discussed in the government White Paper and response to consultation. The White Paper states that "The regulator will also have an obligation to protect users' rights online, particularly rights to privacy and freedom of expression. It will ensure that the new regulatory requirements do not lead to a disproportionately risk averse response from companies that unduly limits freedom of expression, including by limiting participation in public debate."[742] No detail, however, was initially provided on how platforms should address the competing requirements of protecting free speech and targeting a variety of content that does not currently violate any laws. The response to the consultation then explained that in order to protect free expression, there would be "differentiated expectations" for content that is illegal and content that is legal but harmful.[743] This response still fails to acknowledge that categorising content as harmful will still affect how companies treat it, especially when the White Paper envisions a wide range of severe sanctions for companies who get it wrong.

Incentives are an important issue in the duty of care conceptualised by both Perrin and Woods and the Online Harms White Paper. The Open Rights Group has stated that "there remains a worrying presumption inherent within a 'duty of care' that action to limit risks should be the paramount policy driver" which makes sense in a model focused on harm.[744] This means there is no concept of achieving any positive results through regulation, only avoiding negative consequences. Conversely, there is no incentive against over-reaction and the censorship of content that should properly be protected by human rights.[745] The duty of care report does emphasise that the proper exercise of a duty of care can include a wide range of regulatory measures (not just removal) such as filtering or age verification, which could be proportionate responses to some content.[746] However, it is likely that many platforms, fearing a hefty fine (or some of the other measures mooted by the government) will adopt a risk-averse approach and choose to remove content that could have been

[742] DCMS, *Online Harms White Paper*. 56.
[743] DCMS, *Online Harms White Paper: Initial Consultation Response*.
[744] Open Rights Group Wiki, "Perrin and Woods Duty of Care."
[745] Of course this issue is not unique as platforms currently remove plenty of content that is lawful and would comply with human rights principles, particularly because as private companies they have traditionally been treated as not having human rights obligations. The issue with the duty of care is that this tendency could be exacerbated by a system that incentivises even more censorship.
[746] Perrin and Woods, *Online harm Reduction*, 17.

subjected to a less invasive intervention.[747] There should also be procedures to ensure that the content that is targeted is identified accurately and the Open Rights Group has stated that an appeals process would not be sufficient here as users may simply elect not to appeal a mistaken decision. [748] It is clear that there are many incentives to remove content in the duty of care scheme but very few incentives to allow content to remain or to encourage other positive values on the platform such as citizen journalism or robust debate. A similar criticism was made about the German Network Enforcement Act (see 4.4.2) but the duty of care framework would have even more of a chilling effect on speech as the Network Enforcement Act at least confined itself to illegal content whereas the duty of care encompasses a variety of legal content.

A final unsatisfactory response is that Perrin and Woods argue that even if the duty of care resulted in content being removed from one platform, this result is mitigated by the existence of other platforms where this content would be available so human rights are still being respected.[749] This is a problematic argument for a number of reasons. First, human rights are not assessed as an aggregate across an industry, where if Platforms A and B are initiating policies that comply with human rights principles and Platform C is not then the industry will be seen as compliant. Therefore, "any restriction imposed through a duty of care must be assessed against its own impacts, and not dismissed on the basis that the expression can go elsewhere" especially as the audience of each platform has a right to receive information.[750] Second, as Perrin and Woods also believe that harms may necessitate a response from multiple platforms then it would behove all implicated platforms to remove that content. Third, platforms are increasingly converging in their moderation approaches due to collective programmes such as the PhotoDNA programme, where a single platform's decision that something is extremist or depicts child abuse will result in it being technologically impossible to post on any participating platform. Accordingly, it is unlikely

[747] It should also be noted that the Online Harms White Paper focusses almost exclusively on removal, a fact that Perrin and Woods criticise in their response to the White Paper. What this indicates, however, is that even the less problematic and more nuanced aspects of Perrin and Wood's idea will likely be abandoned if implemented by the government. Perrin and Woods, "Whose duty is it anyway?"
[748] Open Rights Group Wiki, "Perrin and Woods Duty of Care."
[749] Perrin and Woods, *Online harm Reduction*, 17.
[750] Open Rights Group Wiki, "Perrin and Woods Duty of Care."

that in the future, there will be much difference in which content is available on which platform. Perrin and Woods even seem to acknowledge and encourage this outcome, writing "the process of regulation could bring service providers of all types together to share knowledge about harms within and between platforms, putting commercial interests to one side."[751] It therefore seems odd to encourage the diversity of platforms and harmonisation in the same breath, all in an attempt to circumvent serious human rights questions. Finally, while it is easy to refer to the multitude of platforms that exist, the truth remains that a number of platforms have a high number of users and even own other popular platforms (such as Facebook with Instagram and Google with YouTube) so if content is removed from those platforms, its continued existence on more niche platforms with smaller user-bases may not be a satisfactory compromise.

In conclusion, the duty of care envisioned by Perrin and Woods (and expanded on by the government White Paper) offers an attractive approach for holding platforms accountable and provides a workable framework for how these ideas could be implemented. It makes a convincing case for affording platforms a measure of discretion in how they meet their obligations and explains how this new regulatory scheme would be administered. The duty of care proposal, however, is replete with issues related to how such an obligation is conceptualised and the effect this inevitably has on the protection of human rights on social media platforms. The proposal outlined in the next chapter will attempt to preserve some of the strongest elements of the government's framework and Perrin and Wood's architecture but will also abandon the duty of care framework in favour of an idea rooted in an approach requiring human rights due diligence.

6.5: Conclusion

The aim of this chapter was not to argue that the solutions presented to the problems social media content moderation pose are misguided or harmful. Rather, the conclusion must be that these solutions are *incomplete.* There are many promising elements that can be

[751] Perrin and Woods, *Online harm Reduction*, 31.

garnered from the implementation of self-regulatory regimes, substantive regulation, and a social media duty of care. Substantive regulation delivers a precision that offers certainty to companies and regulators as to what is expected. In a self-regulatory regime, companies bear the compliance costs and direct resources to developing scalable and efficient processes. Finally, the Perrin and Woods social media duty of care offers a risk-oriented strategy that compels platforms to consider the processes they put in place. Each of these proposals, however, can only form one part of a grander plan to redesign how these companies interact with society and the legal institutions that protect it.

It has to be acknowledged that the activities of social media companies have disrupted so many fields of law that creating a solution based on only one approach or a single legal perspective will be inherently unsatisfactory. Regulators, however, should start with human rights law and administrative law because these fields interact with so many areas of the law and human activity. If solutions can be found in these areas, then reform of other areas will be easier to achieve. The aim of the final substantive chapter is to propose widescale, procedural solutions that will form a bedrock on which any subsequent targeted regulations can be developed.

Chapter Seven: Mandatory Human Rights Due Diligence

7.1 Introduction

This thesis has made suggestions in almost every chapter on how social media content moderation could be reformed. Chapter Three argued that there needed to be more clarity and detail provided in social media terms and conditions and more opportunities for users to provide feedback on the rules (at 3.5). In Chapter Four, the idea of a body of precedents (which would provide more certainty for users and would act as a User Empowerment Tool) was introduced (see 4.5). Chapter Five stated that social media companies should introduce forums of participation and improve their appeal mechanisms so that users could properly contest moderation decisions (at 5.4). All of these suggestions would benefit users and represent an improvement in how content moderation occurs. But larger, more systemic changes need to occur and this chapter will offer a model for a broader, overarching set of reforms that these earlier, more specific suggestions could complement.

This thesis has also argued that there must be a significant increase in transparency across every stage of the content moderation process. Transparency can be beneficial in "mitigating threats to freedom of expression" as well as other rights.[752] But of course, without broader reforms, transparency and disclosure requirements represent, at best, a "mild astringent" for social media companies.[753] To achieve true change in how corporations operate, would-be reformers need to "find ways to affect the decision-making of these corporate institutions, to shape their incentives, bargaining power, and business strategies."[754]

The final theme that has been explored throughout this thesis and is necessary in any proposal for regulating social media platforms is an emphasis on procedural protections. As discussed in the previous chapter (at 6.3.2), many critics and would-be regulators of social

[752] Kaye, *A/HRC/38/35*, 3.
[753] Langvardt, "Regulating Online Content Moderation," 1357.
[754] Danielson, "How Corporations Govern," 424.

media are focussed on the substantive content of the rules employed by these platforms.[755] While this approach is valid, it excludes many important procedural considerations.[756] The current approach often pays lip-service to human rights but by deeming the removal of content as the ultimate priority, concerns about censorship as well as inconsistent applications of other rights are largely ignored. A new approach is needed, one where regulators create powerful legal incentives to protect the rights of their users.[757] This chapter will focus on the procedural nature of content moderation and how reform in these areas is urgently required.

This thesis owes a debt to Jeremy Waldron's scholarship on how important procedural protections are in upholding the rule of law and how often this area is ignored. Waldron defines procedural to mean that we are concerned not with what the law says (the substantive aspects) but rather "with the ways in which a system of rules for governing human conduct must be constructed and administered."[758] If would-be reformers focus only on regulating the content available on social media platforms and fail to address the actual processes of content moderation, then many of the underlying problems in moderation will persist and human rights violations will continue. This chapter embraces a concept that Fuller developed and Waldron extended: the notion that the rule of law acknowledges "man's dignity as a responsible agent."[759] Fuller argues that good law acknowledges human agency and Waldron extends this concept by contending that human dignity is even more closely aligned to procedural protections.[760] Waldron explains that it is often through legal procedures (such as court hearings or appeals) where people are given the opportunity to

---

[755] An exception is a recent blog piece for the official Mozilla blog on their reaction to the UK Online Harms White Paper. The piece consistently emphasised the importance of focussing on "practices over outcomes" and that getting the right "regulatory architecture" was the crucial first step in address online harms. Owen Bennett, "Building on the UK white paper: How to better protect internet openness and individuals' rights in the fight against online harms," Open Policy and Advocacy Blog (Mozilla), last modified July 2, 2019, https://blog.mozilla.org/netpolicy/2019/07/02/building-on-the-uk-online-harms-white-paper/.
[756] A similar point is made by Langvardt, who argues that "any reasonable statutory framework would therefore try to focus judicial, regulatory, and corporate attention on sound content moderation *policies* rather than fussing over individual cases." Langvardt, "Regulating Online Content Moderation," 1376.
[757] Kaye, *A/HRC/38/35*, 3.
[758] Waldron, "Rule of law," 9.
[759] This is why Fuller called his eight principles of good law a "morality." Fuller, *The morality of law*, 162.
[760] Waldron, "Rule of law," 16.

provide their views and participate in the rules that govern their lives.[761] This is an important objective and this thesis has proposed many ideas to empower users through both smaller and larger reforms. The heart of the problem, however, is that the current content moderation processes are structurally flawed in that they fail to reflect the spirit of the rule of law and cause serious human rights issues, all of which undermines the transformative benefits these platforms offer to society.

The last chapter considered a number of solutions to reform the social media content moderation process to ensure better protection of human rights and align the moderation process with rule of law principles. It explored potential solutions from a number of areas of law and the advantages and disadvantages they offered. This chapter will offer a fresh solution that preserves some of the strengths the other solutions offered: certainty for companies, efficient use of resources, strong enforcement methods, and an emphasis on a procedural framework. This solution is the creation of a mandatory human rights due diligence [HRDD] scheme for companies. This chapter will attempt to set out the justifications, substantive content, and procedural aspects of implementing a legal requirement that social media companies engage in HRDD.[762] Baldwin and Cave write that "regulatory processes can be thought of as comprising three stages: the enactment of enabling legislation; the creation of regulatory administrations and rules; and the bringing to bear of those rules on persons or institutions sought to be influenced or controlled."[763] This chapter will address all three of these stages in outlining the proposed HRDD scheme. It will discuss the justifications of HRDD, examine real attempts to introduce HRDD laws, explore the components of a due diligence process and consider how such a law could be implemented in the UK and perhaps, in due course, further afield.

---

[761] Waldron, "Rule of law," 17.

[762] It should be noted that many of the issues and ideas discussed in this section would also be applicable to other sectors and that mandatory HRDD could be implemented for all companies operating in the UK, not just social media companies. This thesis, however, is focussed on social media companies so it will only comment on the utility of a scheme regulating such platforms.

[763] Baldwin and Cave, *Understanding regulation*, 96.

7.2: The foundations of human rights due diligence

7.2.1: Introduction to Due Diligence

Due diligence is defined in the Protect, Respect, and Remedy Framework as "a process whereby companies not only ensure compliance with national laws but also manage the risk of human rights harm with a view to avoiding it."[764] Therefore, due diligence contains both a preventative and remedial element and is a cyclical process that should be regularly updated instead of carried out only once. Indeed, the "prophylactic element" helps to distinguish important human rights due diligence projects like the UN Guiding Principles [UNGP's] from traditional enterprise liability approaches which focus on attributing "ex-post responsibility."[765] Risk, therefore, is a term describing "knowledge deficiencies" about a business's potential negative impact on human rights, which a due diligence process can help to rectify.[766] This risk assessment and management strategy is a similar approach to Perrin and Wood's duty of care framework (at 6.4) although the central objective (addressing human rights issues as opposed to reducing social media harms) is quite different. The specific processes of HRDD will be explained at 7.3.

Many of the requirements in the UNGP's focus on the procedural aspects of human rights compliance, from the creation of a statement of principles, the practice of HRIA's, and the offering of remedies. These procedural guarantees are particularly important in the field of corporate human rights obligations as it strikes a fair balance between companies that offer a good to society and the public who are dependent on that good.[767] It would be feasible for a state to mandate these actions while still allowing a measure of discretion in how companies implement these practices and the substantive results. The objective, therefore, is to catalyse corporate reform by "constitutionalising" a commitment to human rights in the

[764] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 9.
[765] Björn Fasterling, "Human Rights Due Diligence as Risk Management: Social Risk Versus Human Rights Risk," *Business and Human Rights Journal* 2, no. 2 (2017): 228, https://doi.org/10.1017/bhj.2016.26.
[766] Fasterling, "Human Rights Due Diligence as Risk Management: Social Risk Versus Human Rights Risk," 236.
[767] Thomas, *Public rights, private relations*, 223, 25, 32.

"corporate psyche and culture."[768] Legislators may also want to consider the unique cultural context of a country and its corporate environment when designing these laws.[769] It is likely that after some initial implementation challenges, business practices in the states that have mandated procedures based on the UNGP's will result in less human rights issues, offer more certainty to the public, and sustain a positive reputation for the companies that are in compliance.[770]

The Protect, Respect, Remedy Framework and the UNGP's can be a starting point for an actionable plan on human rights due diligence.[771] There have already been important initiatives in a number of countries mandating human rights due diligence. These include both general due diligence requirements (in France and soon in Germany) and due diligence requirements on certain topics (modern slavery in the UK and child labour in the Netherlands). These national laws will be discussed later in this section as well as developments at the regional and international level. There have also been some important academic contributions from scholars who support mandatory HRDD. These academics (whose work was also discussed at 2.5) include David Bilchitz, Surya Deva, and Florian Wettstein.[772]

This chapter is proposing that many of the human rights issues in social media content moderation could be addressed with the introduction of mandatory HRDD. This proposal would take the framework created by the UNGP's and use them to inform a set of legal

---

[768] Peter Muchlinski, "Implementing the New UN Corporate Human Rights Framework: Implications for Corporate Law, Governance, and Regulation," *Business Ethics Quarterly* 22, no. 1 (2012): 157, https://doi.org/10.5840/beq20122218.

[769] For example, a country that is dependent on business from the extractive sector may frame its corporate obligations differently from a country that is dependent on textile manufacturing or the service industry. This is not an argument for laxer regulations but merely stating that national due diligence laws should consider the local corporate context for maximum efficacy.

[770] Of course, these are many of the same advantages that the original Guiding Principles offered but if businesses are not complying with these rules then the positive effects cannot be evidenced.

[771] Ruggie himself characterised the two documents in this way: "the Protect, Respect and Remedy Framework addresses *what* should be done to move in this direction; the Guiding Principles show *how*." Ruggie, "Global Governance and 'New Governance Theory'," 9.

[772] See, for example: Bilchitz, "The Necessity for a Business and Human Rights Treaty."; Deva, "Human Rights Obligations of Business."; Surya Deva and David Bilchitz, eds., *Building a treaty on business and human rights: context and contours* (Cambridge, UK: Cambridge University Press, 2018; Wettstein, *Multinational corporations and global justice*.

obligations for companies operating in the UK jurisdiction.[773] Voluntary HRDD is self-regulation by any other name so the problems with self-regulation that were identified (at 6.2.2) are just as pressing when they are interpreted through the lens of the UN Guiding Principles. In fact, a recent UN report has found that many businesses still have not implemented the UNGP's.[774] The UNGP's can therefore act as both a "precursor to hard law" and as a "supplement to a hard-law instrument"[775] if mandatory HRDD is introduced.

7.2.2: The state of mandatory HRDD across the world:

For years, academic and jurisprudential debate has focused on the question of how to justify imposing human rights on private companies. While this debate has been occurring,[776] transnational corporations and the effects of globalisation have widened the governance gap. The human rights issues caused by companies were more manageable when companies were based in the same jurisdiction as their would-be regulators. It is becoming apparent that voluntary human rights schemes have been insufficient (see 6.2.2) and enforcement tools are necessary to catalyse true reform. Consequently, a pragmatic approach is emerging as states are beginning to directly impose human rights obligations on companies as older regulatory approaches seem increasingly out-of-step with the necessities of contemporary life.[777] Nowhere is this more true than in the flurry of discussions on how social media companies should be regulated. Social media poses a set of unique challenges to settled law across a number of different fields (including human rights) and has become embedded into the fabric of society while simultaneously unravelling it. This chapter, therefore, represents an attempt to understand how we can regulate social media

[773] This approach could also be implemented in any other jurisdiction concerned about human rights, social media, and the rule of law. Indeed, as will be shown throughout this chapter, mandatory HRDD is a principle first implemented in mainland Europe but which is now slowly being introduced into British law and could expand to more jurisdictions in the future.
[774] "UN experts report: Business 'dragging its feet' on human rights worldwide," United Nations News, last modified October 16, 2018, https://news.un.org/en/story/2018/10/1023312.
[775] Dinah Shelton, "Normative Hierarchy in International Law," *American Journal of International Law* 100, no. 2 (2006): 320-21, https://doi.org/10.1017/S0002930000016675.
[776] For a fuller account of the various justifications, see Chapter Two.
[777] See the next section for examples of these changes.

companies as part of a wider effort to impose human rights obligations on companies and what this would entail for platforms.

After being bogged down in theoretical arguments for decades, states and multilateral institutions are trying to move beyond these concepts and address emerging issues in contemporary life. Obara characterises this transition as shifting "the focus in public policy and the media from questions of 'why' to 'how'. That is, from *why* companies should observe human rights, to *how* they can contribute towards the protection and realisation of human rights."[778] Ramasastry takes this argument further, by claiming that the business and human rights debate has already pivoted towards binding law, compliance, and state enforcement. [779] The "how" question, might therefore, have already been answered and the next challenge will be to refine its legal implementation. Of course, theoretical justifications are important, but new academic theories will likely follow from an examination of how these responsibilities have been applied by countries and how these principles could be refined.[780]

Mandatory HRDD can help states to fulfil their duty to protect human rights. Pillar One of the Protect, Respect, Remedy framework (state duty to protect) and Pillar Two (corporate duty to respect) are often conceived as "distinct rather than integrated and complementary." [781] This is myopic as states can protect rights by mandating that corporations engage in activities like due diligence (which would also fulfil the corporate duty to respect). The UNGP's make clear that states should "enforce laws that are aimed at, or have the effect of, requiring business enterprises to respect human rights."[782] This will likely entail a "smart mix of measures – national and international, mandatory and voluntary."[783] Because there was no requirement that the UNGP's be rendered mandatory, they are often perceived as soft law or as essentially voluntary. This is a misconception as

[778] Obara, "What Does This Mean?," 251.
[779] Ramasastry, "Bridging the Gap," 237.
[780] For a great overview on how these theories are being applied in a European context, see: Palombo, *Business and human rights*.
[781] Karin Buhmann, "Neglecting the Proactive Aspect of Human Rights Due Diligence? A Critical Appraisal of the EU's Non-Financial Reporting Directive as a Pillar One Avenue for Promoting Pillar Two Action," *Business and Human Rights Journal* 3, no. 1 (2018): 25, https://doi.org/10.1017/bhj.2017.24.
[782] *UN Guiding Principles.*
[783] *UN Guiding Principles.*v

Ruggie instead left it up to the discretion of states how they would fulfil their duty to protect human rights, but he always envisioned that these laws would be a combination of voluntary and mandatory.[784] This echoes Thomas's notion of a normative mandate, where she argues that states are committed to protecting rights against all potential violators but it is the method of doing so that is "vague and under-theorised."[785] A mandatory approach to the UNGP's is, therefore, perfectly permissible and sends a strong signal that states are serious in their efforts to protect human rights, regardless of who might violate them. It is important, however, that the obligations of businesses be construed differently than the obligations on states, who still bear the primary duty for protecting human rights. These laws must focus on procedural protections and afford platforms discretion in how they achieve their objectives. They must strike a fair balance between what is feasible for a company and what might be required to remedy or prevent human rights issues. It is necessary that these laws comply with rule of law principles and that they be carefully scrutinised.

Specialised laws are increasingly being imposed on social media companies and these laws are incentivising a wide range of behaviours that may pose challenges for human rights. Consequently, it is important that human rights obligations also be imposed so as to act as a counterbalance (and potentially a foil) for laws that reward censorial or otherwise problematic behaviour. Failing to move human rights online and regulating the conduct of private platforms will result in the concept of human rights being divorced from the lived experience of people and human rights losing its currency as technology develops. This is a pragmatic approach that attempts to preserve some of the discretion platforms have enjoyed in the past (which has helped them innovate and scale up) but also provide a new incentive structure for changing how they moderate content and what factors they prioritise in platform governance. This chapter offers an actionable plan on how human rights

---

[784] This has recently been confirmed by John Ruggie and Rachel Davis, who acted as senior legal advisor to Ruggie during his mandate. See: Rachel Davis, "Beyond Voluntary: What it Means for States to Play an Active Role in Fostering Business Respect for Human Rights," Shift, last modified February, 2019, https://www.shiftproject.org/resources/viewpoints/beyond-voluntary-states-active-role-business-respect-human-rights/; John Gerard Ruggie, "Letter to Ms. Saskia Wilks and Mr. Johannes Blankenbach (Business and Human Rights Resource Centre)," Business and Human Rights Resource Centre, last modified September 19, 2019, https://www.business-humanrights.org/sites/default/files/documents/19092019_Letter_John_Ruggie.pdf.

[785] Thomas, *Public rights, private relations*, 33-34.

obligations can be imposed on social media platforms.[786] It attempts to strike a fair balance between the stakeholders and offer a proactive approach to protecting rights in the digital world. The proposal addresses the problems in the models previously explored in Chapter Six while building on many of the advantages they were seen to offer.

7.2.3: Evidence of a Paradigm Shift: Countries that have implemented mandatory HRDD

Despite the continued debates over whether businesses should have human rights responsibilities, there are indications that countries are recognising that voluntary due diligence by corporations is an imperfect solution and have legislated accordingly.[787] This has been termed "the beginning of a paradigm shift" as more countries introduce legislation mandating HRDD.[788] These examples may not seem conclusive in of themselves but when assessed in aggregate, they indicate a growing acceptance of legislating for human rights responsibilities for companies and a "global diffusion of human rights due diligence as a norm of conduct."[789]

7.2.3.i: Regional and International developments

There have been developments at the regional and international level on the issue of mandatory HRDD. These initiatives are intriguing because they could indicate that mandatory HRDD is beginning to generate widespread support, and this in turn could facilitate cooperative efforts and/or attempts at harmonisation. Many of these developments have occurred in a European context although there are some interesting discussions occurring at the international level as well.

The European Parliament and the Council of Europe have both stressed the need for mandatory HRDD in reports in the last few years. These reports are particularly instructive

[786] And all businesses more generally.
[787] These issues were discussed at 6.2.2. The crux of the issue, however, is that voluntary due diligence means that companies may choose not to comply and the evidence indicates that, in practice, they do not comply with voluntary schemes.
[788] Amnesty International and Business and Human Rights Resource Centre, *Creating a Paradigm Shift: Legal Solutions to Improve Access to Remedy for Corporate Human Rights Abuse* (London: Amnesty International, 2017).
[789] *Report (A/73/163)*, 8.

when considering whether a consensus is emerging on the necessity of mandatory HRDD, at least within a European context. In 2018, the Report on Sustainable Finance argued that the EU should use the French duty of vigilance law (which will be discussed at 7.2.3.i) as inspiration for an overarching mandatory due diligence law in the EU.[790] The 2016 European Parliament Report on corporate liability for serious human rights abuses in third countries stated that there is an urgent need for "binding and enforceable rules and related sanctions and independent monitoring mechanisms."[791] These finding were echoed by the Working Group on Responsible Business Conduct, which called for the adoption of mandatory HRDD.[792] A final example comes from 2016, when members of parliament in eight EU States announced a "Green Card" asking the European Commission to introduce mandatory HRDD and an environmental duty of care. It was particularly telling that the eight countries supporting the Green Card were spread throughout the European Union, from the Baltic region, the Southern Mediterranean, Eastern Europe and Western Europe.[793] There appears to be a growing level of support at the European level for the view that mandatory HRDD is the next step in protecting human rights. This flourishing of reports citing mandatory HRDD seems to reflect a concern that voluntary regimes have failed to have the same impact as enforceable measures, such as decisions handed down by the European Court of Human Rights [ECtHR], thus reflecting an interest in establishing mandatory human rights requirements in Europe.[794] There is a difference, however, in simply recommending

---

[790] It should be noted that there is some general due diligence obligations in place in the EU already. The 2014 EU Non-Financial Reporting Directive requires that companies annually disclose their principal risks including environmental and human rights impacts and the due diligence policies they have in place. This disclosure requires that companies also explain the outcomes of these due diligence policies. These reporting requirements include the company, their supply chain, and business operations. This directive, however, is limited to large and listed companies. *Directive 2014/95/EU*, O.J. (L 330).

[791] Committee on Foreign Affairs, *Report on corporate liability for serious human rights abuses in third countries: 2015/2315 (INI)* (Brussels: European Parliament, 2016), para. 28.

[792] "Shadow EU Action Plan on Business and Human Rights," Responsible Business, last modified March 19, 2019, https://responsiblebusinessconduct.eu/wp/2019/03/19/shadow-eu-action-plan-on-business-and-human-rights/.

[793] The eight countries were France, UK, Italy, Estonia, Lithuania, Slovakia, Portugal, and The Netherlands. "Members of 8 European Parliaments support duty of care legislation for EU corporations," European Coalition for Corporate Justice, last modified May 31, 2016, http://corporatejustice.org/news/132-members-of-8-european-parliaments-support-duty-of-care-legislation-for-eu-corporations.

[794] Of course, mandatory HRDD also has opponents. The Swiss debates on imposing a mandatory HRDD have been protracted and polarising. This is not surprising in a country with a strong emphasis on business as well as a positive human rights record. See: Jessica Davis Plüss and Andrea Tognina, "Responsible business initiative heads closer to a national vote," Swissinfo, last modified March 12, 2019,

mandatory HRDD and in actually legislating it and it is likely that there would be a significant amount of debate at the European level before a law requiring due diligence was introduced.[795]

Mandatory HRDD is also being explored at the international level. In 2017, the UN Committee on Economic, Social and Cultural Rights affirmed that states have the duty to establish HRDD obligations for companies and to improve access to remedies (including through corporate liability). [796] This duty, however, seems unenforceable but it does represent an opportunity to exercise soft-law influence. There have been other developments in mandatory HRDD at the international level as well. The UN Working group on Transnational Corporations and other Business Enterprises with respect to Human Rights published a draft document in 2018 for a legally binding treaty requiring states to legislate in the area of mandatory HRDD.[797] There are a number of interesting aspects of this draft but Article 9(1) serves as an excellent introduction to this area:

> "State Parties shall ensure in their domestic legislation that all persons with business activities of transnational character within such State Parties' territory or otherwise under their jurisdiction or control shall undertake due diligence obligations throughout such business activities, taking into consideration the potential impact on human rights resulting from the size, nature, context of and risk associated with the business activities."[798]

The UN draft consolidates some of the best features of the national systems which will be explored later (at 7.2.3.i). It adopts a general duty of due diligence, thus negating the

---

https://www.swissinfo.ch/eng/corporate-responsibility_responsible-business-initiative-heads-closer-to-a-national-vote/44818824.

[795] Some commentators wonder if the 2020 German presidency of the Council of the European Union will increase momentum on a harmonised approach to due diligence. See: Saskia Wilks and Johannes Blankenbach, "Will Germany become a leader in the drive for corporate due diligence on human rights?," Business and Human Rights Resource Centre, last modified February 20, 2019, https://www.business-humanrights.org/en/will-germany-become-a-leader-in-the-drive-for-corporate-due-diligence-on-human-rights.

[796] Economic and Social Council, *General comment No. 24 (2017) on State obligations under the International Covenant on Economic, Social and Cultural Rights in the context of business activities (E/C.12/GC24)* (United Nations: Geneva, 2017).

[797] UN Human Rights Council, "Legally binding instrument to regulate, in international human rights law, the activities of transnational corporations and other business enterprises. Zero Draft Bill," Office of the United Nations High Commissioner for Human Rights, last modified July 16, 2018, https://www.ohchr.org/Documents/HRBodies/HRCouncil/WGTransCorp/Session3/DraftLBI.pdf.

[798] UN Human Rights Council, "Zero Draft Bill."

risk of a patchwork regime and provides a harmonised approach that still affords states some discretion in how they interpret rights, impose liability, and enforce the law. It has two categories of actions: preventative and restorative. The preventative aspect (Article Nine) outlines the requirements of a due diligence process including the establishment of a fund for future actions that necessitate compensation.[799] The restorative section (Article Ten) applies to actions that arise once the human rights abuse has occurred. This section encompasses both civil and criminal liability.[800] The draft treaty has an emphasis on victims bringing these companies to court but the draft does provide a number of articles designed to improve access to justice such as financial aid for victims, waiving costs, and an absolute ban on requiring victims to reimburse the legal expenses of the other party to such claims.[801]

The draft treaty does not rule out the possibility of a regulator. The text of the document states "states shall take all necessary legislative, administrative or other action including the establishment of adequate monitoring mechanisms to ensure effective implementation of this Convention."[802] It is, however, more focussed on enforcing human rights obligations in court. The draft treaty does require states to set up a compensation fund, which is an important aspect of addressing business and human rights issues.[803] This working group document offers a framework that could be a useful starting point for national policymakers.

The draft is, however, not without its critics. Palambo rightfully points out that the treaty imposes obligations on states to pass due diligence laws, instead of directly imposing obligations on corporations, which would have been much more progressive.[804] The most prominent critic is John Ruggie, who has argued that a single treaty cannot capture the

---

[799] Article 9(2)h, UN Human Rights Council, "Zero Draft Bill."
[800] Criminal liability would only apply to "human rights violations that amount to a criminal offence, including crimes recognised under international law, international human rights instruments, or domestic legislation." Criminal liability is beyond the scope of this thesis as this thesis has focussed on rule of law and human rights issues. Legally binding instrument to regulate, in international human rights law, the activities of transnational corporations and other business enterprises. Article 10(8), UN Human Rights Council, "Zero Draft Bill."
[801] Article 8(5)-(6), UN Human Rights Council, "Zero Draft Bill."
[802] Article 15(1), UN Human Rights Council, "Zero Draft Bill."
[803] Article 8(7), UN Human Rights Council, "Zero Draft Bill."
[804] Palombo, *Business and human rights*. 23.

diverse range of concerns in business and human rights.[805] He claims that any attempt to address all the relevant issues in one treaty "would have to be pitched at such a high level of abstraction that it would be largely devoid of substance, of little practical use to real people in real places, and with high potential for generating serious backlash against any form of further international legalisation in this domain."[806] Ruggie's criticisms of the draft treaty seem hollow, however, as Deva points out that the same concerns about comprehensively addressing all issues in a single document could be levied against the UNGP's.[807] In any case, the UN draft treaty is primarily focussed on creating a procedural framework for mandatory due diligence as opposed to exhaustively dictating the substantive content of the rules. Accordingly, it represents a useful resource that can help inform national due diligence laws regardless if the treaty is ever successful.

7.2.3.ii: National laws

In 2017, France introduced a law requiring French companies (above a certain size)[808] to create and publish plans that identify and seek to prevent potential risks to human rights, safety, and the environment.[809] This "duty of vigilance" law was the first generalised obligation of due diligence to human rights and has inspired much of the legislation that has subsequently been enacted or considered.[810] It is particularly ground-breaking because it imposes a general duty to not only consider the company's own actions but also monitor the

[805] John Gerard Ruggie, "Quo Vadis? Unsolicited Advice to Business and Human Rights Treaty Sponsors," Institute for Human Rights and Business, last modified September 9, 2014, https://www.ihrb.org/other/treaty-on-business-human-rights/quo-vadis-unsolicited-advice-to-business-and-human-rights-treaty-sponsors.
[806] Ruggie, "Quo Vadis?"
[807] Surya Deva, "Corporate Human Rights Abuses and International Law: Brief Comments," James G. Stewart Blog, last modified January 28, 2015, http://jamesgstewart.com/corporate-human-rights-abuses-and-international-law-brief-comments/.
[808] With either 5000 employees in France or 10,000 employees worldwide.
[809] Law number 2017-399 of March 27th, 2017 relating to the duty of vigilance of parent and instructing companies**.** English translation available at: "Chronology on the Law on the duty of vigilance," Business and Human Rights in Law, accessed April 26, 2019, http://www.bhrinlaw.org/law-duty-of-vigilance-2-versions-en-october-2018.pdf. This law also allows civil liability if companies fail in their due diligence obligations and any harm results from that failure.
[810] "Evidence of Mandatory Human Rights Due Diligence: Policy Note," European Coalition for Corporate Justice, last modified May, 2019, http://corporatejustice.org/news/9189-evidence-for-mandatory-human-rights-due-diligence-legislation-in-europe.

human rights issues in their supply chain.[811] The law requires that companies create, publish, and implement vigilance plans for addressing environmental and human rights risks posed by their business operations.[812] Companies can be penalised for not creating such plans and they can also be found liable and ordered to pay compensation for harms that would not have occurred if they had conducted due diligence. Of course, this duty of vigilance law does not apply to any of the major social media companies as they are not established in France and nor do they have a subsidiary in France.[813] In 2018, the UN Working Group on Business and Human Rights welcomed the French legislation as "a development that other governments should learn from" and recommended that countries use "legislation to create incentives to exercise due diligence, including through mandatory requirements."[814] There have already been some interesting cases in France under the law. Lawsuits have been brought against Samsung France (over child labour and labour conditions) and against the Cement company LafargeHolcim (for allegedly funding armed groups in Syria to keep its plant open).[815]

As we shall see, France has been a catalyst for other countries adopting similar laws and the lessons learned from these early experiments could be very useful when targeting social media companies for regulation. Specifically, there are some issues in the French duty of vigilance law that could be avoided by policymakers in the UK. First, the government did not designate a regulator to monitor compliance with this law. Instead, concerned parties have to bring the company to court, which shifts the onus of ensuring compliance onto the civil society and ensures that many companies will not be held accountable because of a lack of resources or attention. If one were to consider social media companies in particular, civil

---

[811] Dalia Palombo, "The Duty of Care of the Parent Company: A Comparison between French Law, UK Precedents and the Swiss Proposals," *Business and Human Rights Journal* 4, no. 2 (2019): 275, https://doi.org/10.1017/bhj.2019.15.

[812] Article One, Law number 2017-399 of March 27th, 2017 relating to the duty of vigilance of parent and instructing companies**.** English translation available at: Business and Human Rights in Law, "Chronology on the Law on the duty of vigilance."

[813] If Ireland were to implement a similar law, however, many European subsidiaries of social media companies would fall within the scope of that law.

[814] Paragraphs 67 and 93 (a), *Report (A/73/163)*.

[815] Ben Chapman, "Samsung faces charges over 'misleading' ethics claims after alleged labour abuses in factories," Independent, last modified July 4, 2019, https://www.independent.co.uk/news/business/news/samsung-france-legal-case-child-labour-factories-a8988446.html.

society groups may be unable to engage in protracted litigation with wealthy platforms that are based in other countries. It therefore becomes difficult to truly hold companies to account. Indeed, In the two years since the law's introduction, accountability has emerged as a serious problem, with a quarter of companies having failed to publish a vigilance plan despite this being a legal requirement under Article 1 of the law.[816] Since it is estimated that the law applies to between 100-150 companies in France[817] this means that between 25-37 companies have failed to comply with the law. As filing legal complaints against all of them would be too onerous for civil society groups, many companies might be tempted to play the numbers and ignore the law.[818] The statistic on noncompliance also illustrates a second issue with this law: non-transparency. There is no comprehensive list of companies subject to the law published by the government, despite repeated civil society requests.[819] It is very difficult for outsiders to ascertain which companies this law applies to because of the need for specific information on how many employees a subsidiary has worldwide or whether a company employs contractors. Therefore, one could expend time and resources taking a company to court only to find that the law did not actually apply to them. A number of NGO's have tried to resolve this issue by creating a database of companies that they believe are subject to the law[820] but the burden of ensuring compliance with the law should properly fall on government. There is, therefore, a serious concern in France that many companies can escape compliance because of a lack of accountability and transparency measures, thus diminishing the efficacy of this ground-breaking law. These issues could be avoided in the UK by designating a regulator to hold companies accountable for their due diligence

---

[816] Sherpa and CCFD-Terre Solidaire, "NGOs launch a new tool to track companies subject to the French duty of vigilance law," European Coalition for Corporate Justice, last modified July 1, 2019, http://corporatejustice.org/news/16294-ngos-launch-a-new-tool-to-track-companies-subject-to-the-french-duty-of-vigilance-law.

[817] "French Corporate Duty of Vigilance Law: Frequently Asked Questions," European Coalition for Corporate Justice, last modified March 24, 2017, http://www.respect.international/french-corporate-duty-of-vigilance-law-english-translation/.

[818] In fact, it actually took two years before the first formal complaint was made. This was against the oil company Total over environmental and human rights concerns. Sandra Cossart and Lucie Chatelain, "What lessons does France's Duty of Vigilance law have for other national initiatives?," Business and Human Rights Resource Centre, last modified June 27, 2019, https://www.business-humanrights.org/en/what-lessons-does-frances-duty-of-vigilance-law-have-for-other-national-initiatives.

[819] Juliette Renaud et al., *The Law on Duty of Vigilance of Parent and Outsourcing Companies: Year 1: Companies Must Do Better* (Montreuil: ActionAid, 2019).

[820] See: Sherpa, CCFD-Terre Solidaire, and Business and Human Rights Resource Centre, "Duty of Vigilance Radar," Vigilance Plan, accessed October 22, 2019, https://vigilance-plan.org.

practices and ensuring greater transparency in the entire process. This thesis has shown that social media companies have serious issues with transparency so any attempt to regulate these platforms must rest on more openness and disclosure.

The German approach is different from France as the country attempted to enact a voluntary regime but, dismayed at the low uptake, began to threaten a mandatory approach. In 2016, Germany published a National Action Plan to implement the UN Guiding Principles and stated that it would introduce legislation requiring HRDD if less than half of the large German companies do not adopt HRDD by 2020.[821] This was already a strange approach, with the German government insisting that HRDD was important but also tacitly permitting half of German companies to not engage with these processes.[822] In 2019, a draft of the mandatory legislation was leaked to the media. The draft bill envisions strong penalties including "fines of up to five million Euros, imprisonment and exclusion from public procurement procedures in Germany."[823] This German example shows yet again how voluntary schemes, even ones made under the threat of further regulation with sanctions,[824] fail as companies are unwilling to change or engage in resource-intensive reforms if not all companies (including their competitors) are required to do so. This draft bill also reflects Germany's interest in closely regulating companies that operate in Germany and should be

[821] Interministerial Committee on Business and Human Rights, *National Action Plan: Implementation of the UN Guiding Principles on Business and Human Rights (2016–2020)* (Berlin: Federal Foreign Office, 2017).
[822] It is likely that Germany will continue to raise its expectations of the proportion of companies who employ due diligence but it is still a strange approach, allowing human rights compliance to be assessed as an aggregate just as 6.4.3.iv criticised Perrin and Woods for doing.
[823] "German Development Ministry drafts law on mandatory human rights due diligence for German companies," Business and Human Rights Resource Centre, accessed July 1, 2019, https://www.business-humanrights.org/en/german-development-ministry-drafts-law-on-mandatory-human-rights-due-diligence-for-german-companies.
[824] The mandatory bill also included more companies. The Original Action Plan was aimed at German companies with more than 500 employees but the draft bill included companies with more than 250 employees. This was also threatened in the plan, with the Plan stating "In this context, the Federal Government will also examine, in consultation with the National Regulatory Control Council, the necessity of the corporate compliance costs arising from this plan and will consider a widening of the number of enterprises to be reviewed, in order to potentially include enterprises with fewer employees in future assessments and subsequent additional measures." Interministerial Committee on Business and Human Rights, *National Action Plan: Implementation of the UN Guiding Principles on Business and Human Rights (2016–2020)*, 10; Business and Human Rights Resource Centre, "German Development Ministry drafts law on mandatory human rights due diligence for German companies."

read in conjunction with other recent reforms such as the Network Enforcement Act.[825] Once again, however, as the law only applies to German companies, the major social media companies would not be caught by its terms. This is unfortunate as mandating human rights responsibilities could help to restore the balance after the Network Enforcement Act tipped the scales too strongly in favour of censorship. That being said, the draft bill offers the potential for strong accountability measures and rectifies the enforcement issues in the French law. It would be better in the future, however, if mandatory HRDD was not used as a threat to achieve partial compliance with a voluntary initiative as it results in a more adversarial approach that inevitably makes consultation with stakeholders more difficult.

Beyond Europe, the United States of America has also produced new laws mandating HRDD, albeit in more limited circumstances. For example, Section 1502 of the Dodd-Frank Wall Street Reform Act obliges publicly traded companies to ensure that the minerals they use in production are not related to the conflict in the Democratic Republic of the Congo (DRC).[826] The law requires that companies analyse their supply chain and exercise "due diligence on the conflict minerals' source and chain of custody" in order to "help end human rights abuses in the DRC caused by the conflict."[827] It is interesting that laws like Dodd-Frank, passed in the wake of the 2008 recession, acknowledge that businesses left unregulated can pose a danger to society. In some ways, the last four years have seen a similar reckoning for social media companies, with concerns about privacy, extremism, and election manipulation catalysing regulatory efforts in many countries.

The US State Department also mandates that America companies (and individuals) investing in Myanmar must comply with new reporting requirements on human rights issues (which specifically reference the UN Guiding Principles) as well as the Burmese Sanctions Regulations more generally. These requirements apply to a number of activities such as purchasing a share of ownership in the economic development of resources (which includes

---

[825] On which see 4.4.2. This Act is controversial but does not display a strong interest in regulating business activities.

[826] A similar, but broader, regulation now exists in the EU. The EU Conflict Minerals Regulation requires that EU importers exercise due diligence in investigating their supply chain of tin, tantalum and tungsten, their ores, and gold originating from conflict-affected and high- risk areas. *Regulation (EU) 2016/679*, O.J. (L 119).

[827] *Conflict Minerals: Final Rule (Release No. 34-67716; File No. S7-40-10)* (Washington, DC: SEC, 2012), 8, 10.

natural, agricultural, commercial, financial, industrial and human resources) or entering into a contract providing for the participation in royalties, earning, or profits in the economic development of resources located in Myanmar.[828] Despite the recent controversy over Facebook's actions in Myanmar, it does not appear that Facebook would be covered by these requirements as they only apply to American companies investing over five million dollars.[829] It can be very difficult to quantify Facebook's "investment" in an area, as opposed to a more straightforward consideration of profits because of the company's business model, where the services are provided for free and then Facebook earns profits through advertising and data acquisition. The American examples demonstrate how it can often be easier to pass due diligence and reporting requirements on specific issues (even in pro-business countries like America) rather than draft general obligations. The main issue, however, is the patchwork regime that is created, a regime that may overlook social media companies. Myanmar and the DRC both have serious human rights issues but unfortunately that does not make them unique in the world. Singling out specific countries or issues will inevitably create governance gaps and unnecessary complexity.[830] This chapter, therefore, advocates an overarching approach to HRDD.

It should be noted that these case-studies are just a few examples and other countries have also passed due diligence laws or are currently debating them.[831] There are, therefore,

---

[828] Rae Lindsay et al., "Briefing Note on US State Department Guidance on Reporting Requirements for Responsible Investment in Myanmar," Clifford Chance, last modified October, 2013, https://www.cliffordchance.com/content/dam/cliffordchance/briefings/2013/10/us-state-department-guidance-on-reporting-requirements-for-responsible-investment-in-myanmar-october-2013.pdf.

[829] It also applies to all companies investing in the oil and gas sector regardless of the level of investment. It should also be noted that the original requirements applied to companies investing over 500,000 dollars but the threshold was raised in 2016.

[830] The same critique could be levied against *Modern Slavery Act, 2015*, c. 30 (UK). (which will be discussed at 7.2.3.iii).

[831] These laws include both general due diligence obligations (like the French model) and more specific due diligence requirements (such as in *Modern Slavery Act, 2015*.). The Netherlands has passed a Child Labour Due Diligence Law. Nigeria requires all licensed petroleum operators in Nigeria to provide plans to ensure that environmental and human rights objectives are being met. Switzerland is considering two different due diligence proposals at this time and the state of California obliges large Californian companies to conduct due diligence in relation to human trafficking and slavery in their supply chains. For a full list of all countries that have some form of due diligence or reporting requirements (or are currently considering them) see: "Examples of government regulations on human rights reporting & due diligence for companies," Business and Human Rights Resource Centre, last modified September 30, 2017, https://www.business-humanrights.org/en/examples-of-government-regulations-on-human-rights-reporting-due-diligence-for-

an increasing number of jurisdictions that are moving away from traditional voluntary approaches and embracing a more hard-line approach to businesses and human rights. The laws that are being drafted, however, represent a variety of approaches. Due diligence laws are interesting because while the spirit of the law may remain the same, there is a high degree of diversity in how these objectives are implemented. It is important that while examining the evidence that a paradigm shift is beginning, we also identify best practices and pitfalls to avoid if the UK were to implement a mandatory due diligence law of its own.

7.2.3.iii: The United Kingdom

In the UK, a limited form of HRDD is required in the Modern Slavery Act 2015 which requires companies with an annual turnover of £36 million or over, based in the UK or conducting business there, to disclose the activities they have undertaken to ensure their supply chains do not feature slavery or human trafficking in an annual statement.[832] Section 54 explicitly mentions "due diligence processes in relation to slavery and human trafficking" and also demands that companies assess their efforts in combatting these problems and provide educational training about slavery and human trafficking to their staff. [833] Facebook, Twitter, and Google have all submitted statements pledging their commitment to abide by the Modern Slavery Act and provide details about how they are addressing slavery and trafficking.[834] This law was seen as ground-breaking when passed as it obliged companies to engage in assessments of how their actions might contribute to slavery and human trafficking.[835]

companies; European Coalition for Corporate Justice, "Evidence of Mandatory Human Rights Due Diligence: Policy Note."

[832] Section 54, *Modern Slavery Act, 2015*.

[833] Section 54(5)c-d, f, *Modern Slavery Act, 2015*.

[834] The Modern Slavery Registry has copies of all these statements. For the Facebook statement, see: Patricia Carrier, "Facebook's Anti- Slavery and Human Trafficking Statement," Modern Slavery Registry, last modified April 26, 2018 https://www.modernslaveryregistry.org/companies/18576-facebook-inc/statements/27127; Patricia Carrier, "2017 Modern Slavery Statement," Modern Slavery Registry, last modified April 30, 2019, https://www.modernslaveryregistry.org/companies/19013-google-inc; Patricia Carrier, "Twitter UK anti-slavery statement for the 2017 financial year," Modern Slavery Registry, last modified March 07, 2019, https://www.modernslaveryregistry.org/companies/26502-twitter-uk-limited.

[835] *Independent Review of the Modern Slavery Act 2015: Final Report* (London: Government of the United Kingdom, 2019), 15.

The Modern Slavery Act, however, should not be viewed as a truly mandatory HRDD obligation for a number of reasons. First, companies are not actually required to do anything beyond making a report and there is no obligation that companies attempt to ameliorate the problems they might happen or be obliged to identify. This means that companies can disclose in their report that they are doing nothing to address modern slavery issues and intend to keep doing nothing (although there may be reputational consequences for such a brazen statement).[836] A similar law in California, the California Transparency in Supply Chains Act, has resulted in many companies disclosing that they do not intend to take any of the voluntary steps the government has suggested in reducing slavery.[837] The Home Office has subsequently recommended that companies should no longer be allowed to state that they intend to do nothing and that sanctions should be applied in response to a refusal to act.[838]

Regardless of these new laws, it is likely that a disclosure requirement is not enough to catalyse real reform in an industry.[839] Disclosure generally focusses on accounting for actions that have already occurred whilst the prevention of human rights abuses is arguably more essential and requires more sustained corporate reform.[840] Second, despite the lenient requirements in the 2015 measure, the Home Office found that approximately 40% of the companies that the Modern Slavery Act applied to were not complying with the reporting requirements and that "there have been no penalties to date for non-compliant organisations."[841] Another study found that of the reports that were submitted, 35% of them did not discuss their risk assessment processes ,which is surprising for statements that are

---

[836] Annie-Marie Barry, "The UK Modern Slavery Act and corporate responsibility: progress and challenges," Centre for the Study of Modern Slavery, accessed October 26, 2019, https://www.stmarys.ac.uk/research/centres/modern-slavery/articles/corporate-responsibility.aspx.
[837] Barry, "The UK Modern Slavery Act and corporate responsibility: progress and challenges."
[838] *Independent Review of the Modern Slavery Act 2015*, 15.
[839] A study by Babbington et al found that reporting on human rights and environmental issues had little effect on organisational decision-making to reduce negative impacts on society. Jan Bebbington, Elizabeth A. Kirk, and Carlos Larrinaga, "The production of normativity: A comparison of reporting regimes in Spain and the UK," *Accounting, Organizations and Society* 37, no. 2 (2012/02/01/ 2012): 78, https://doi.org/10.1016/j.aos.2012.01.001.
[840] Buhmann, "Neglecting the Proactive Aspect of Human Rights Due Diligence? A Critical Appraisal of the EU's Non-Financial Reporting Directive as a Pillar One Avenue for Promoting Pillar Two Action," 26.
[841] *Independent Review of the Modern Slavery Act 2015*, 14.

intended to be based around a due diligence approach."[842] There was also a "failure to connect compliance with the law with the awarding of public procurement contracting."[843] Once again, the primary weakness of this scheme appears to be enforcement, with companies choosing not to comply (or complying but not taking any action beyond reporting) because there is no real threat of punishment. The Home Office accordingly recommended a regulatory body to enforce compliance (and sanction non-compliance) and also an amendment to the Act so that a failure to act when slavery or trafficking is identified would result in penalties.[844] The report even acknowledged that the French duty of vigilance law signalled a growing interest in mandating corporate compliance and the Modern Slavery Act should reflect this by requiring more from companies.[845] The final issue is that in a similar fashion to the French duty of vigilance law, there is no publicly accessible registry which lists which companies are covered by the Modern Slavery Act. This will inevitably hinder attempts by civil society groups to hold companies accountable and it is unlikely an NGO would be able to identify which companies fall within the parameters of the Act as it is estimated that there are over 17,000 companies that may qualify.[846] The same themes emerge from the Modern Slavery Act as from other due diligence laws: lack of transparency and enforcement that will surely lead to a lack of compliance.

The preceding paragraphs have demonstrated that the UK has due diligence initiatives, although they are limited to reporting requirements or voluntary schemes.[847] There is, however, interest in introducing a broader duty and these various initiatives could help inform how that obligation could be conceptualised in the British context, so they are worth noting here. A general duty on all companies to prevent human rights abuses was

---

[842] "Reporting on Modern Slavery: The current state of disclosure," Ergon Associates, last modified May, 2016, https://ergonassociates.net/wp-content/uploads/2017/06/Reporting-on-Modern-Slavery2-May-2016.pdf.
[843] *Report (A/73/163)*, 18.
[844] *Independent Review of the Modern Slavery Act 2015*, 15.
[845] *Independent Review of the Modern Slavery Act 2015*, 15.
[846] Barry, "The UK Modern Slavery Act and corporate responsibility: progress and challenges."
[847] Another example of human rights reporting obligations on UK companies can be found in section 414(c) of the UK Companies Act, which states that directors must provide strategic reports to members of their company on a number of issues including information about environmental matters and social, community and human rights issues and "information about any policies of the company in relation to those matters and the effectiveness of those policies." Section 414c: contents of a strategic report. *Companies Act, 2006*, c. 46 (UK).

recommended by the UK Parliament's joint Committee on Human Rights in 2017. This proposal would require that companies perform HRDD and would allow civil remedies against parent companies if human rights violations happened.[848] It should also be reiterated that the UK supported the Green Card campaign in the European Union. It is possible that recent concerns about social media companies and the introduction of the GDPR (which imposes numerous obligations on companies) could also contribute to a climate where mandatory due diligence is seriously considered by UK policymakers. It is likely that actions by other countries, such as France and possibly Germany, could further encourage the UK to introduce mandatory HRDD.

The UK can become a high-water mark for businesses, ensuring that companies meet a certain standard of human rights protection and provide evidence of due diligence.[849] A similar argument was once made by Schrage, who argued that the US should use "its prestige and credibility to serve as an honest broker to endorse or 'qualify' serious [private voluntary initiatives] that address labour standards violations."[850] The UK could play a similar role in relation to all human rights (and not on a voluntary basis) by holding companies to account and giving civil society a common language (or what Thomas refers to as a "normative vocabulary"[851]) in which to structure their expectations and demands. As long as the regulations developed in the UK are clear, attainable, and offer the possibility of avoiding the bad publicity and shareholder concern that serious human rights scandals can bring then it is possible that companies will comply with human rights obligations. This, in turn, could help to improve outcomes in other countries that are in a weaker bargaining position vis-à-vis social media platforms than is the United Kingdom. Danielson argues, therefore, that sometimes it is more effective to focus on agitating for higher corporate standards in "key

---

[848] House of Lords, House of Commons, and Joint Committee on Human Rights, *Human Rights and Business 2017: Promoting responsibility and ensuring accountability (Sixth Report of Session 2016–17)* (London: House of Lords, 2017), 6.

[849] This is still possible in a post-Brexit Britain although whether there currently exists the political will to do so is another question. That being said, the current interest in regulating social media companies could be used as an impetus for a larger discussion about regulating the social impacts of businesses.

[850] Elliot Schrage, *Promoting International Worker Rights through private voluntary initiatives: Public Relations or Public Policy* (Iowa City: University of Iowa Centre for Human Rights 2004), 5-6. A copy can be found at: https://www.business-humanrights.org/en/codes-of-conduct-new-report-examines-their-effectiveness-points-to-need-for-greater-us-government-support Accessed 10 June 2019.

[851] Thomas, *Public rights, private relations*, 3.

markets" and then using "corporate ordering to internationalise those standard rather than seeking to alter the rules in countries with lower standards on a jurisdiction by jurisdiction basis or through a supranational harmonisation process."[852] In some ways, Germany's Network Enforcement Act is playing a similar role right now, inspiring other countries around the world to enact laws targeting social media platforms with heavy penalties for non-compliance. The Network Enforcement Act, however, is primarily focused on the removal of content (see 4.4.2) and should not be treated as a positive model for regulating social media. The UK can instead offer a different, rights-based model for regulating social media platforms, one that could have positive effects in other jurisdictions with less influence.

This section has considered laws and reports from a range of different national, regional, and international resources and attempted to identify patterns that may inform future regulations in the UK.[853] It has also argued that this wave of legislation seems to reflect the demand for greater responsibilisation of businesses as societal expectations on private companies increase and the scale of their impact becomes clear. Mandatory HRDD may be a controversial topic as it entails the constraining of companies that have previously enjoyed a high degree of discretion in relation to their human rights practices. However, just like the Western labour regulations of the early twentieth century (which resulted in greater safety, better workplace standards, and limits on working hours), mandatory HRDD may seem like an impossibility or an unnecessary interference now but over time (if implemented correctly) it could become another uncontroversial and essential aspect of business regulation.

While mandatory HRDD should be implemented for all businesses operating within the UK (and the suggestions made in the next two sections are applicable to other industries), this thesis has specifically examined the risks social media platforms pose for human rights. The string of recent controversies in social media (such as the Cambridge Analytica scandal) have also led to calls for stronger regulations targeting these platforms. While these

[852] Danielson, "How Corporations Govern," 419.
[853] The need for transparency and accountability, common themes in this thesis, is particularly evident.

regulations can take many forms, this thesis contends that mandatory HRDD offers the best framework to effectively regulate social media and protect human rights online. It is arguable that mandatory HRDD could be integrated into social media companies with more ease than in others because their practices are less dependent on global inequality and less intimately intertwined with development issues than other sectors such as resource extraction or clothing manufacturing.[854]

Of course, by mandating HRDD, the UK must also refrain from passing laws that require platforms to deviate from human rights responsibilities except in the most extreme circumstances.[855] After all, states do not relinquish their obligations when they "privatise the delivery of services that may impact upon the enjoyment of human rights."[856] The UK must assess whether any current laws on internet content (and any proposed laws like the Online Harms White Paper) violate human rights and they must ensure that any future laws be measured against the same yardstick.[857] It has become too commonplace for countries, both tyrannies and democracies, to use social media companies as proxy censors or to hand over information about dissidents just as during the 'war on terror'; extraordinary rendition was used to torture detainees in a country where the West could claim plausible deniability.[858] An example of this proxy role is the willingness of tech companies to comply with strict blasphemy laws, which pose serious threats to freedom of expression and are "the most

[854] With the notable exception of the human moderators in developing countries that social media companies employ. That being said, the working standards of these moderators could be improved without much sacrifice from the companies.

[855] With permissible exceptions being determined by the relevant human rights instruments and judicial decision-making. Currently, states may not realise how often they violate free expression. Frank La Rue, the former UN Special Rapporteur on Free Expression once informed the UN Human Rights Council that "states' use of blocking or filtering technologies is frequently in violation of their obligation to guarantee the right to freedom of expression," since these techniques were not clearly established in law, were for purposes not listed in the International Covenant on Civil and Political Rights, used secret blocking lists, were unnecessary or disproportionate, and lacked review by a judicial or independent body. Although it should be noted that LaRue did make an exception for Child Sexual Abuse Material (CSAM) provided that the national law was sufficiently precise and there were effective safeguards against abuse or misuse, including oversight and review by an independent and impartial tribunal or regulatory body." Even in relation to CSAM, LaRue argued that states were relying too heavily on blocking when they should be attempting to prosecute creators and distributors. The merits of that point might be debatable due to the transnational nature of CSAM but the more general point stands. See: La Rue, *A/HRC/17/27*, 10, para. 31-32.

[856] Commentary to UNGP 5. *UN Guiding Principles.*

[857] A report by the UN Special Rapporteur made clear that "States should repeal any law that criminalises or unduly restricts expression, online or offline." Kaye, *A/HRC/38/35*, 19.

[858] See: Chapter 4 of Peter Gibson, "The Report of the Detainee Inquiry," (2013). 30-43.

commonly cited instrument for platforms to "pro-actively" take down content." [859] Mandatory HRDD would give social media platforms a way to push back against oppressive governments that make inappropriate requests for information, violate privacy, or attempt to use the platform as a tool for widescale censorship. It would address the governance gap that results from the lack of legal guidance on how social media platforms should interact with repressive regimes whereas there are laws that regulate "the export of a wide variety of materials and products, including certain technologies such as cryptographic software."[860] After all, as Deibert and Villeneuve argue, the enhanced filtering and blocking opportunities that technology provides the state means that "we can't assume technology naturally leads to liberalisation."[861] The UK would therefore be required to abide by the same high principles that it imposed on businesses in order to achieve consistency. This compliance would also maintain its legitimacy as a human rights-oriented regulator as the UK would be adhering to the set of standards "against which the addressees of that law assess the propriety of making that claim to authority."[862]

7.3: The Process of Mandatory Human Rights Due Diligence: The Application to Social Media

7.3.1: Introduction

The goal of HRDD is for companies to "to identify, prevent, mitigate and account for how they address their adverse human rights impacts" (which can be both actual and

[859] Google, YouTube's parent company, for example, has agreed to cooperate with the Pakistani government to make all blasphemous content inaccessible to Pakistani youtubers. This agreement ended the three-year ban on accessing YouTube in Pakistan. Other common categories of content that are targeted include extremist content and content that violates copyright. Association for Progressive Communications, *Content Regulation in the Digital Age*, 10, 14.

[860] It should be noted that Youmans and York did not discuss mandatory HRDD in this paper. They focused their analysis on how more laws were needed regulating how social media platforms interacted with regimes that regularly violate human rights. Their concerns, however, are relevant to a number of different situations. Youmans and York, "Social Media and the Activist Toolkit," 324.

[861] Ronald J. Deibert and Nart Villeneuve, "Firewalls and Power: An overview of global state censorship of the Internet," in *Human Rights and the Digital Age*, ed. Mathias Klang and Andrew Murray (London: Cavendish Publishing, 2005), 11.

[862] Reed and Murray, *Rethinking the jurisprudence of cyberspace*, 169.

potential).[863] The process of HRDD can be split into a number of discrete steps (with policies acting as a prelude to HRDD): Human Rights Impact Assessments (HRIA's), integration, tracking performance, and remedy.[864] The following section will discuss each aspect of the due diligence process, first as a general principle and then consider its specific application to social media platforms, and in particular how these requirements would affect the content moderation process. It should also be noted that these processes all have a requirement of communication to the public and affected stakeholders, a theme that is embedded into the UNGP's.[865] Throughout the following processes, the platform must engage in some measure of transparent reporting about its policies, HRIA's, integration and remedies.[866] This requirement reflects the importance of transparency in achieving human rights reform, a topic that has been frequently discussed in this thesis. Social media companies in particular have been roundly castigated for their lack of transparency at every stage in the content moderation process, from the refusal to provide detailed rules (see 3.3.1), the reluctance to indicate how many moderators work for platforms (at 4.5), and the opaque appeals processes (at 5.2.3).

7.3.2: Policies

Before the due diligence process begins, companies are required to develop human rights policies to demonstrate their commitment to rights.[867] These policies must be approved by the company leadership, informed by internal and external expertise (which may necessitate consultations), and communicated directly to parties with whom they have business relationships and other parties such as states.[868] The human rights principles that will be

---

[863] *UN Guiding Principles*.
[864] Technically remedy is a separate head of responsibility in the Protect, Respect, and Remedy framework but remedy is the logical final step of any overarching scheme for mandatory HRDD. Keeping Remedy separate in the Ruggie framework was likely done to emphasise that both countries (which are covered by the Protect heading) and companies (covered under the Respect heading) had a duty to offer remedies for human rights violations. Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 18.
[865] *Report (A/73/163)*, 4.
[866] Tracking performance does not seem to have a communicative element because it is really a complementary duty (almost a sub-heading) to HRIA's and integration despite being articulated as a stand-alone principle. Integration and conducting HRIA's would be less effective if there was no underlying duty to track performance.
[867] Guiding Principles 15(a) and 16, *UN Guiding Principles.*
[868] Commentary for UNGP 16, *UN Guiding Principles.*

considered will be based on the international bill of human rights and the core conventions of the ILO because these requirements "comprise the benchmarks against which other social actors judge the human rights impacts of companies."[869]

These human rights policies should explicitly state systems of accountability (such as who to contact if stakeholders are concerned about human rights issues) and should be supported with personnel training (including the training of the policy development teams and the content moderators themselves)on the specific human rights relevant to social media platforms. Finally, it important that companies "embed" this policy commitment by ensuring that their other business policies are consistent with human rights.[870] Expressing a policy commitment to human rights is also one the least onerous requirements of the UNGP's and a Shift study found that 88% of the companies in their research sample shared a human rights policy.[871] One might question, however, if all of these companies truly "embedded" these policies into their practices such as by ensuring their other policies were coherent with their human rights principles, providing training to personnel, and consulting with experts. This suspicion is borne out by a study of UK companies, which found that despite many of them having human rights policies, these policies were not used on an everyday basis and most were developed as a response to external pressures such as negative attention from the media.[872]

In addition to the other human rights documents listed above, social media platforms may also wish to consider relevant policy documents such as the Santa Clara Principles on Content Moderation and the Global Network Initiative Principles. While "broad aspirational language" can be used to describe the importance of respecting human rights, it is also important that detailed rules be created to ensure that these policies are not just a symbolic

---

[869] This is also enshrined in UN Guiding Principle 12, the commentary of it which makes clear that in certain situations one might also want to consider "United Nations instruments…on the rights of indigenous peoples; women; national or ethnic, religious and linguistic minorities; children; persons with disabilities; and migrant workers and their families. Moreover, in situations of armed conflict enterprises should respect the standards of international humanitarian law." Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 17; *UN Guiding Principles.*

[870] Commentary for UNGP 16, *UN Guiding Principles.*

[871] *Evidence of Corporate Disclosure*, 8.

[872] Obara, "What Does This Mean?," 267.

gesture.[873] As mentioned previously (see 6.2.3), BSR recommended that Facebook have a specific, stand-alone human rights policy instead of attempting to infuse their terms and conditions with human rights values.[874] Accordingly, all social media platforms should develop specific human rights policies and make them publicly available. It would be possible to have variations in their policies, of course, as countries adopt different approaches to some categories of content, such as Holocaust denial material.[875] These limitations, however, should fall within a margin of appreciation for human rights and should be explicitly addressed in the policies so that users have the tools to contest controversial interpretations of these principles.

When developing these policies, platforms should consult with experts on human rights and how their technologies impact on these rights. Platforms need to communicate these policies "actively"[876] (which seems to indicate that posting them on their website may not be enough) to businesses with which they have contractual relationships (such as the contracting companies they use when they outsource moderation), the countries that interact with platforms on issues of illegal content and removal, and in particularly risky situations, affected stakeholders such as groups representing the Rohingya in Myanmar.

Adopting explicit human rights policies is important because it provides a benchmark against which users and states can compare the behaviour of the platform. It shifts the discourse that occurs between platforms and users from a discourse on a self-made body of rules that platforms can change at will to a set of universal protections that businesses should respect. It also provides a measure of protection against censorial states (albeit a soft law measure) as it is more politically fraught for states to request that platforms deviate from human rights principles as opposed to their own, rapidly changing terms and conditions.[877]

---

[873] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 18.
[874] The extent to which Facebook's terms and conditions are infused with human rights values is, of course, debatable.
[875] Platforms currently deal with these variations by employing country-withheld-content, where content that is generally legal but is illegal in certain countries is inaccessible within the borders of those countries.
[876] Commentary for UNGP 16, *UN Guiding Principles.*
[877] This is also why it is important that platforms publicly share information about the requests states make on them or else this form of protection will be ineffective. Kaye, *A/HRC/38/35*, 9, 15.

Creating human rights policies is an important first step for social media companies and can help inform the next stage in the HRDD process: HRIA's.

7.3.3: Human Rights Impact Assessments (HRIA's)

One of the key features of mandatory HRDD is the performance of HRIA's.[878] The objective of an HRIA is to draw on expertise (both internal and external to the company) and consult with potentially affected groups in an effort to understand how the activities of a business can cause or contribute to human rights issues.[879] HRIA's are important regulatory tools to predict, assess, and prevent risks and the types of assessments companies are engaged in continue to grow. For example, companies regularly perform environmental impact assessments and data impact assessments. These assessments are at the core of any due diligence process. HRIA's are the fulcrum of mandatory HRDD as they require that companies take stock of the risks their business activities create and the changes that would be necessary to foster respect for human rights.

HRIA's are a diagnostic tool that can offer a tremendous amount of value to companies in identifying, preventing, and remedying risks. Drawing on expertise and consulting with stakeholders (before important decisions are made)[880] can provide new insights and different perspectives.[881] HRIA's can help solve or prevent problems that could result in negative media coverage, shareholder ire, and legal consequences. Conducting thorough assessments also signals that companies are behaving in an ethical manner (assuming they then integrate their findings) and are moving beyond verbal commitments to respect human rights and are acting on these intentions. Of course, these assessments will be more

[878] This essential component of HRDD is based on Principle 18 of the UN Guiding Principles, which states "In order to gauge human rights risks, business enterprises should identify and assess any actual or potential adverse human rights impacts with which they may be involved either through their own activities or as a result of their business relationships." *UN Guiding Principles.*
[879] Commentary to UNGP 18, *UN Guiding Principles.*
[880] Götzmann claims that too often, consultation with stakeholders only occurs after important decisions are made, which drastically diminishes their role in the process. Nora Götzmann, "Human Rights Impact Assessment of Business Activities: Key Criteria for Establishing a Meaningful Practice," *Business and Human Rights Journal* 2, no. 1 (2017): 99, https://doi.org/10.1017/bhj.2016.24.
[881] This is the strength of an HRIA but unfortunately a UN study found that too many HRIA's are done as tick-box exercises and lack "meaningful engagement with potentially affected stakeholders." *Report (A/73/163)*, 8-9.

complicated when one is considering rights-conflicts or balancing exercises (such as considering the boundaries between hate speech and free expression) but even in those situations, identifying issues early and coming up with a strategy for addressing them will produce better results than the more reactionary tactics currently employed by companies like social media platforms.

HRIA's should first be conducted every time a new product/service is introduced, or even considered[882] as this is when there is the most flexibility, before choices "become strongly fixed in material equipment, economic investment, and social habit."[883] HRIA's should then be performed every time a material change that will likely result in a heightened risk of human rights violations (such as political unrest) occurs. In all other circumstances, HRIA's should be updated on a regular basis[884] as it is intended to be a "dynamic, iterative, and ongoing management process"[885] and must become a frequent practice in a business if it is to be effective. It is also important HRIA's focus on adverse impacts to human rights (and not include information about positive impacts a company may have on human rights) as that "facilitates a space for the implicit offsetting of adverse impacts."[886]

There are three sets of factors that companies should always assess when carrying out HRIA's and these factors will help to define the scope of the HRIA.[887] The first is the geographical context in which their business activities take place. It is clear that certain countries will pose greater human rights risks than other states based on the political climate and factors like a recent or ongoing conflict. Additional care would need to be taken in countries with weak human rights protections, autocratic governments, or high degrees of

---

[882] *UN Guiding Principles*.
[883] Langdon Winner was talking more generally about how values are encoded into politics here but his suggestions are clearly applicable to HRIA's. Winner, *The whale and the reactor: a search for limits in an age of high technology*, 29.
[884] Institute for Human Rights and Business and Shift, *ICT Sector Guide on Implementing the UN Guiding Principles on Business and Human Rights* (Brussels: European Commission, 2013), 16, 19.
[885] John F. Sherman III, "UN Guiding Principles: Practical Implications for Business Lawyers," *In-House Defence Quarterly*, December 21 2013, 51.
[886] Götzmann, "Human Rights Impact Assessment of Business Activities: Key Criteria for Establishing a Meaningful Practice," 98.
[887] These three factors are taken from Ruggie's explanation of the Protect, Respect, and Remedy framework. It is likely that companies may choose to include more factors in their assessment but Ruggie outlines the minimum requirements for effective due diligence. See: Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 17.

intra-state strife. This is particularly important because a 2006 report by Ruggie found that corporate-linked human rights issues occurred most frequently in countries with the worst governance challenges.[888]

The second set of factors relates to what human rights impacts their own activities may cause in these countries. When considering those impacts, one must be mindful of the different roles a company may play, such as "producers, service providers, employers, and neighbours."[889] This is important because their role may affect the appropriate steps that should be taken.

Finally, companies should consider whether they might contribute to human rights violations through their relationships with states, businesses, and non-state actors in that country.[890] The goal in HRDD is, to the extent that is possible, for corporations to "assume the responsibility to respect human rights for the entire corporate group, not atomise it down to various constituent units that may operate in poorly regulated contexts."[891] It is important that companies aggregate risks across the whole enterprise, which is why Ruggie argues that companies should integrate human rights concerns into company risk management systems that already exist at the company.[892] In short, while HRIA's are a complex process, the corresponding reward is an identification of risks that could be damaging to both the public and the company, which should offset some of the natural resistance that will occur after mandatory HRDD is introduced.

---

[888] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 6.
[889] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 17.
[890] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 17.
[891] Ruggie, "Global Governance and 'New Governance Theory'," 13.
[892] Ruggie, "Global Governance and 'New Governance Theory'," 14.

In relation to social media platforms, HRIAs could have helped to predict and prevent some of major human rights scandals these companies have experienced.[893] An HRIA would need to be conducted for every country a platform is available in and for every product they introduce. This may seem daunting but after the initial outlay of resources in creating the first assessments, the HRIA's would become easier to revisit and update. Platforms would want to assess not only their policies, enforcement processes, but also the curation of feeds and any envisioned transitions from human to algorithmic processes.[894] Social media companies should also assess the human rights issues posed by the companies they outsource moderation labour to and any risks inherent in the activities those individual moderators conduct for the platform.[895]

For social media companies, geographical context would be especially important when considering what relationship the platform may have with that government and whether this context will require a high degree of country-withheld content or specific moderation changes. Social media companies operate in a variety of different markets and cultures and the content moderation challenges they experience would vary across those contexts. Social media companies would accordingly need to consider the impacts they would have in a country as a platform for users and an employer of general staff (such as moderators). Another important consideration is ensuring that platforms consider their "supply chains" when conducting HRIA's. Applying the notion of a supply chain to a social media company is interesting because while it is clear that the moderators that are contracted to work for platforms are part of the supply chain, there is some ambiguity whether the people who produce content (users) would also be suppliers. At the very least, content-creators who are

---

[893] These controversies have been discussed throughout the thesis but some notable scandals have included: Twitter's issues with extremism and abusive content, Facebook's problems with fake news and being used as a platform for genocide in Myanmar, Instagram's uncertain position on self-harm content, and most recently, YouTube discovering that the comments feature on videos by underage content-creators was being used by paedophiles for grooming. See: Samuel Osborne, "YouTube disables comments on videos featuring children after paedophile ring scandal," Independent, last modified March 1, 2019, https://www.independent.co.uk/life-style/gadgets-and-tech/news/youtube-comments-disable-children-videos-paedophile-ring-a8802316.html; Cynthia M. Wong, "Social Media's Moral Reckoning: Changing the terms of engagement with Silicon Valley," Human Rights Watch, accessed October 29, 2019, https://www.hrw.org/world-report/2019/essay/social-medias-moral-reckoning.

[894] Kaye, *A/HRC/38/35*, 17.

[895] For example, these moderators may experience serious psychiatric harms at work.

financially linked to the platforms (such as members of YouTube's Partner Program) should be considered as part of the supply chain for the purposes of HRIA's.

Crucially, platforms would want to assess their role as a company that might be called upon to cooperate with the national government. An HIRA would be a useful opportunity to consider the requests that a specific government might make of a platform and how they would respond to those demands. For example, platforms may be comfortable removing or withholding certain types of content[896] but might be unwilling to restrict other categories of content or hand over information about users to that government. Social media can legitimise and de-legitimise issues based on who they permit to use their platform and what content can be shared (such as how they apply the "terrorist" label) so it is essential that they consider how their relationships contribute to alleviating or exacerbating human rights issues. Anticipating these challenges ahead of time will allow platforms to devise appropriate responses to these requests.[897] When assessing the impacts a company has on human rights, it is important to adopt the perspective of potential victims, who will likely have a significantly lower tolerance for human rights risks than the company.[898]

When conducting HRIA's, it is essential that platforms do not solely identify external risks to human rights that are mediated by the platform but also evaluate the risks that the infrastructure and activities of the platform create for users. Social media platforms have a tendency to perceive threats to human rights as originating from governments that pressure the companies to disclose information or enforce controversial laws.[899] While these actions do pose a serious risk to human rights, platforms fail to recognise that moderating content based on their own internal criteria will also have human rights implications. This

---

[896] In a country-withheld-content approach where the content would still be accessible outside the borders of that particular state.

[897] Of course, the platform may choose to simply comply with all government requests. This poses a serious threat to international human rights and should be discouraged. Unfortunately, platforms provide little information about government requests, perhaps sensing that this issue will likely cast the companies (which purport to be platforms that help facilitate expression and connectivity) in a negative light.

[898] This is often referred to as the Ford Pinto fallacy: where a company's shareholders are willing to tolerate the negative outcomes from their products (in Ford's case, wrongful death lawsuits) but these outcomes are considered intolerable from the perspective of the public. Lynn Sharp Paine, *Value shift: why companies must merge social and financial imperatives to achieve superior performance*, 1st ed. (New York: McGraw-Hill, 2003), 220-22.

[899] Jørgensen, "Framing Human Rights," 352.

misperception is likely informed by "the strong belief in the liberating power of technology" and "echoes the United States online freedom agenda, which largely focuses on threats to the free and open internet from repressive governments" while failing to consider the possibility of threats from the infrastructure itself.[900] It is interesting that a country with such robust free speech protections, a country that was "the first nation ever to be argued into existence in print,"[901] has created the necessary environment for the rise of these platforms, some of the most powerful speech regulators in the world today. Conducting HRIA's can help platforms to understand their own role in respecting human rights and how their activities or platform design can have serious consequences.

An example of a strong HRIA is the report conducted by BSR on Facebook's presence in Myanmar. There is a detailed methodology section that identifies both actual and potential human rights impacts and prioritises the risks by considering the scope of the risk, the scale (or severity) of the risk, remediability, the likelihood of the risk, attribution (how closely is Facebook connected to the impact) and leverage (how much power does Facebook have to influence the risk).[902] BSR also considered how Facebook was related to a particular human rights impact by applying UN Guiding Principle 19 and questioning whether the platform had caused the impact, contributed to the impact, or was linked to the impact by its products, services, operations, or business relationships.[903] This relationship between Facebook and the impact would then dictate what suggestions BSR would make, with the company having a greater obligation if they caused the impact as opposed to merely being linked to the impact.[904] It is likely that if social media platforms were to conduct HRIA's for all their products and services then they would see a diverse range of impacts that vary in the factors BSR identified and how closely connected the platform was to the risk. A social media

[900] Jørgensen, "Framing Human Rights," 352.
[901] Postman, *Technopoly*, 66.
[902] BSR, *Facebook in Myanmar*, 9.
[903] *UN Guiding Principles*; BSR, *Facebook in Myanmar*, 33-34.
[904] If Facebook caused the impact, "the company should take the necessary steps to cease or prevent the impact. "If the platform contributed to the impact, "the company should take the necessary steps to cease or prevent its contribution and use its leverage to mitigate any remaining impact to the greatest extent possible." Finally, if the platform was linked to the impact, "the company should determine action, based on factors such as the extent of leverage over the entity concerned, how crucial the relationship is to the enterprise, the severity of the abuse, and whether terminating the relationship with the entity itself would have adverse human rights consequences." *UN Guiding Principles*; BSR, *Facebook in Myanmar*, 33-34.

company's role as a platform means that they will often be linked to impacts around the world and it should be noted that the UN Guiding Principles expects all impacts to be addressed although the remedy may vary.[905]

It is important to complement country-specific reports with assessments of how the platform's enforcement activities and business models more generally could pose risks for human rights. The BSR did make some limited recommendations concerning definitions in the terms and conditions and the lack of local moderators but the Myanmar-specific mandate of the report meant that larger issues could not be identified and examined. The specific situation in Myanmar would not likely affect the majority of Facebook's global population of users. Accordingly, digital activists have been calling for HRIA's on content regulation and have argued that platforms should share this information publicly.[906] Platforms occasionally share the number of government requests to remove content but do explain not how their own content moderation policies may affect free expression and other rights.[907] Admittedly, HRIA's that identified serious human rights issues in Facebook's general activities around content moderation and advertising would be much more damning for the platform. Once again, platforms seem more comfortable sharing information about how governments use them as tools for speech-regulation without facing up to the risks that are inherent in their current methods of content moderation. It has also been found that platforms are more likely to share information about the risks they pose to the right to privacy rather than the right to free expression.[908] It is possible that platforms are more comfortable with privacy issues because directors like Mark Zuckerberg perceive it as a changing norm, one that is growing less important.[909] Conversely, the raison d'etre of social media is ostensibly expression so a platform's role in censoring content is much more difficult to reconcile with its public image. Mandatory HRDD would ensure that not only would more information be available to the

---

[905] *UN Guiding Principles.*
[906] Ullman, Reed, and MacKinnon, *Content Regulation in the Digital Age*, 2.
[907] Ullman, Reed, and MacKinnon, *Content Regulation in the Digital Age*, 4.
[908] Ullman, Reed, and MacKinnon, *Content Regulation in the Digital Age*, 2, 6.
[909] Bobbie Johnson, "Privacy no longer a social norm, says Facebook founder," The Guardian, last modified January 11, 2010, https://www.theguardian.com/technology/2010/jan/11/facebook-privacy.

public but also that platforms would not be able to selectively share less controversial information in a bid to appear transparent.

It is possible that after conducting an HRIA, a platform may conclude that the risks to human rights in a particular country outweigh the benefits of their presence in that jurisdiction.[910] In particular, they may determine that complying with national laws on content would result in greater restrictions of freedom of expression (such as forced cooperation with authorities to identify rule-breakers) than if that social media company did not operate in that jurisdiction.[911] Currently, there is little incentive for social media companies to stay out of those problematic jurisdictions, with even Google considering whether it should create a censored version of their search engine for the Chinese market.[912] It has been demonstrated, however, that not only do platforms cooperate with laws that violate Article 19 of the of the International Covenant on Civil and Political Rights (free expression) but that they even take pre-emptive measures, adjusting their terms and conditions to comply with repressive countries before they have even been asked.[913] Mandatory HRDD, however, could act as a counter-balance to the inclination of platforms to expand into authoritarian countries, requiring that platforms explain why they entered or continued to operate in jurisdictions that posed serious human rights risks. Failure to demonstrate effective safeguards for human rights could then lead to legal consequences. This might mean that there are certain countries these platforms could not operate in (or may have to cease operating in if new developments occur) but at least social media would not become a tool for repressive regimes while still wearing the guise of free expression and connectivity.

---

[910] Questions have been raised, for example, whether Facebook's continued presence in Myanmar has done more harm than good, with one commentator to the BSR saying "Maybe Myanmar isn't ready for Facebook." BSR, *Facebook in Myanmar*, 24.

[911] The same calculus should apply to the introduction of new features on the platform. Products would then be more carefully vetted before they are introduced to the public. Association for Progressive Communications, *Content Regulation in the Digital Age*.

[912] Google seems to have shuttered this product due to public outrage but it was still developed and the project could very well be re-activated in the future. For more on this issue, see: Ryan Gallagher, "Google's secret China project 'effectively ended' after internal confrontation," Intercept, last modified December 17, 2018, https://theintercept.com/2018/12/17/google-china-censored-search-engine-2/.

[913] These changes are often specific to a particular country instead of a global adjustment. Association for Progressive Communications, *Content Regulation in the Digital Age*, 2.

In conclusion, one might question whether platforms are capable of creating insightful HRIA's. It must be acknowledged that there will be some growing pains as companies acclimatise to the new perspective they must adopt when conducting business. However, there is a wealth of resources available online to assist companies in conducting HRIA's and implementing the results. Of particular use is the ICT-specific guide on implementing the UNGP's that Shift and the Institute for Human Rights and Business created for the European Commission.[914] It would also be possible to create a self-assessment template for companies to use when engaging in HRDD, something that could be modelled on the checklists the ICO provides to help small to mid-sized companies in a range of different sectors to comply with the GDPR.[915] Indeed, the European Commission also included additional guidance for SME's in their ICT Sector report on applying the UNGP's.[916] Perhaps platforms could create databases that would assist them in maintaining a wide range of HRIA's, update them with ease, and solicit contributions from affected stakeholders. Currently, there are no major incentives for companies to conduct HRIA's in the UK. Conversely, there is a social *disincentive* as companies that "that are transparent about risks and challenges are criticised for not doing enough whereas less responsible competitors go below the radar of NGOs and journalists."[917] This is why it is important to pass laws *requiring* HRDD so that all social media platforms are on a level playing field and companies that take human rights issues seriously can be incentivised. The HRIA, however, is only part of the HRDD framework and effective reform will not occur if platforms stop after that stage.

7.3.4: Integration

Once the risks for human rights have been identified, action must be taken to prevent these risks from occurring (or reoccurring in some scenarios).[918] The results of the HRIA

---

[914] Institute for Human Rights and Business and Shift, *ICT Sector Guide on Implementing the UN Guiding Principles on Business and Human Rights*.
[915] "Data Protection Self-Assessment," Information Commissioner's Office, accessed June 12, 2019, https://ico.org.uk/for-organisations/data-protection-self-assessment/.
[916] Institute for Human Rights and Business and Shift, *ICT Sector Guide on Implementing the UN Guiding Principles on Business and Human Rights*, 16.
[917] A UN Working Group referred to it as a "first-mover challenge." *Report (A/73/163)*, 10.
[918] The UN Working Group on human rights and business put it best when they explained that integration meant that "if the enterprise is causing the impact, it should take steps to cease or prevent it; if it is contributing to the impact, it should take steps to cease or prevent its contribution and use leverage to

must be integrated into the business by assigning responsibility for addressing these impacts, allocating resources to the task, and creating oversight processes.[919] The actions such integration will require will vary depending on how the company is related to the risk (did it cause it, contribute to it, or is it exposed through a business relationship?) and how much leverage they have in addressing the risk.[920] Human rights policies and responses must be implemented throughout the company so that human rights concerns are not siloed off into a specific department. The same issue exists at the government-level, where human rights concerns are treated as separate species ("segregated within its own conceptual and typically weak institutional box")[921] from other governmental considerations that shape economic policy. The left hand must know what the right hand is doing for true reform to occur, and this can be challenging in large companies. Failure to coordinate activities will likely lead to contradictory actions at every level of business.[922] Accordingly, Ruggie has called integration "the biggest challenge in fulfilling the corporate responsibility to protect." [923] This has been confirmed by studies of multinational companies where integration of human rights assessments into core business practices was perceived as the most difficult aspect of HRDD.[924] It is relatively easy to create a human rights policy and even to conduct assessments, but responding to the results and changing the processes at a company requires more time, effort, and resources. Integration, however, is the factor that transforms due diligence from a paper tiger into an actionable model for protecting human rights.

---

mitigate the remaining impact; if it has not contributed to the impact, but that impact is directly linked to its operations, products or services by its business relationships, it should take steps to gain and use leverage to prevent and mitigate the impact, to the greatest extent possible." *Report (A/73/163)*, 4.

[919] Principle 19, *UN Guiding Principles.*

[920] Principle 19, *UN Guiding Principles.*

[921] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 8.

[922] Ruggie has argued that a failure to integrate human rights policies will mean "product developers may not consider human rights implications; sales or procurement teams may not know the risks of entering into relationships with certain parties; and company lobbying may contradict commitments to human rights." Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 18.

[923] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 18.

[924] John Morrison and David Vermijs, *The state of play of human rights due diligence: Anticipating the next five years* (London: Institute for Human Rights and Business, 2011), 27.

When carrying out reforms, priority should be given to risks based on severity not on likelihood of occurrence.[925] The rationale behind this is that companies might be tempted to focus on quick-fixes and easy wins when it is more important to ensure that the worst harms can be avoided. Focussing on severity also ensures that the assessment assesses risks from the rights-holder's perspectives instead of considering what constitutes a risk to the business. [926] Sherman gives a number of examples of corporate disasters that were considered extremely unlikely and did have a serious impact on human rights when they occurred. These examples include the 2008 recession, the 2010 Deepwater Horizon leak in the Gulf of Mexico, and the Fukushima nuclear power plant accident in 2011.[927]

Another mistake in the integration stage is to prioritise the risks that are getting the most public attention (as opposed to the risks that pose the greatest threat to human rights).[928] Some of the due diligence legislation explored above inadvertently contributes to this because specific legislation on conflict minerals, child labour, or modern slavery will likely result in companies prioritising these issues above other human rights issues, regardless of their actual level of risk. Integration, therefore, must be responsive to the risks that the HRIA identified rather than the risks that the company may find the easiest or the most politically pressing to address.

Platforms must be mindful of how they have prioritised risk as in the past; platforms have often engaged in the common regulatory mistake of "random agenda selection" whereby "regulators focus on high-salience political issues rather than the issues that pose the greatest threat to public safety."[929] In the world of social media, the clearest example of a severe risk was when it became clear that Facebook was being used as a platform to

---

[925] Principle 24, *UN Guiding Principles.*
[926] Götzmann, "Human Rights Impact Assessment of Business Activities: Key Criteria for Establishing a Meaningful Practice," 106.
[927] Sherman III, "UN Guiding Principles," 54.
[928] *Report (A/73/163)*, 8.
[929] Platforms have often made another regulatory mistake: tunnel vision. Tutt defines tunnel vision as when parties do not engage in cost-justified regulation because they are unduly focused on carrying out their narrow mission without attention to broader side effects of regulatory choices. It is this very issue that mandatory HRDD might be able to correct as platforms have been too focussed on their narrow mission of content moderation without considering how their practices impact society. Andrew Tutt, "An FDA for Algorithms," *Administrative Law Review* 69, no. 1 (2017): 113, https://doi.org/10.2139/ssrn.3293577.

manipulate voters in the 2016 fake news scandal. That being said, this thesis is replete with other examples of risks that were left unattended by platforms and became serious controversies such as Twitter's challenges with hate speech and extremist content. There have been many issues identified with how platforms moderate content but integration requires more than just an investigation, it requires protracted effort and a commitment to true reform. Integration is likely where one would see the sharpest division of outcomes between voluntary and mandatory HRDD processes as integrating the results of HRIA's is resource-intensive and complex.[930] It is therefore necessary to devise regulatory schemes with incentive structures and enforcement bodies to encourage true reform in how social media companies address human rights issues in their content moderation processes.

7.3.5: Tracking Performance

When integrating the results from HRIA's, it is important for companies to track the effectiveness of their actions.[931] This tracking should draw on feedback from a variety of different sources, including people in the business and affected stakeholders.[932] While tracking performance is treated by the UNGP's as a distinct (potentially chronological) stage in the due diligence process, it is better to treat it as an underlying duty that runs throughout the stages of the due diligence process like transparency. There is an element of tracking in the assessment stage, the integration stage, and the remedy stage.

Companies should consider how progress could be tracked and use both qualitative and quantitative indicators.[933] Choosing the appropriate metrics or Key Performance Indicators (KPI's) is important as companies may inadvertently gather information about irrelevant factors or fail to identify best practices in their field.[934] This might involve some

---

[930] In the Shift study, only 16% of companies disclosed "relatively strong supporting" evidence of integration such as providing examples or giving examples of how their business decisions have human rights elements. Of course, this may be an issue with self-reporting or how much information companies chose to divulge with Shift so it is unclear whether this statistic displays an integration problem or a transparency problem. *Evidence of Corporate Disclosure*, 12.

[931] Principle 20, *UN Guiding Principles.*

[932] Principle 20, *UN Guiding Principles.*

[933] Principle 20, *UN Guiding Principles.*

[934] This has been identified as a common reason that due diligence processes can fail. *Report (A/73/163)*, 10-11.

trial-and-error but consulting with experts with experience in methodology could assist companies in identifying key parameters and measurement methods. Over time, it is likely that tracking performance will become less onerous as companies improve their assessment tools, standardise metrics, and disseminate best practices through industry initiatives.[935]

A study conducted by Shift, however, has indicated that companies may not be tracking this information. Only 9% of companies within the research sample disclosed appropriate supporting evidence that they were tracking the effectiveness of their responses to human rights impacts.[936] A failure to track performance may signal that the company's commitment to human rights is cursory at best. It is reasonable to assume that activities that companies consider a high priority (such as profits, project costs, and personnel management) will be monitored in order to identify patterns and make recommendations. A failure to track the performance of human rights measures must, therefore, be interpreted as judgement that it is a low priority. It is likely that the introduction of mandatory HRDD will make tracking performance a higher priority for companies.

Social media companies would have an advantage over more traditional businesses when it comes to tracking performance. Digital tools would streamline evidence gathering, automate pattern identification and trend forecasting, and make it easier to identify ineffective reforms. Platforms already gather information about the content moderation process, a fact evident in their transparency reports.[937] Information about government requests for removal, how long moderation decisions take, and what categories of content are being flagged is already being tracked by social media companies. Realigning these tools to track compliance with a mandatory HRDD law would, therefore, be relatively easy for a set of companies that are frequently implicated in surveillance scandals.[938] Social media companies could also help to introduce best practices and develop tools that would then be used by other companies that are less developed in their approach to tracking performance.

---

[935] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 19.
[936] *Evidence of Corporate Disclosure*, 13.
[937] An example of a transparency report is: "Facebook Transparency Report," Facebook, accessed February 20, 2020, https://transparency.facebook.com.
[938] See, for example, the Edward Snowden controversy.

7.3.6: Remedies

Remedies are defined broadly in the UNGP's as "apologies, restitution, rehabilitation, financial or non-financial compensation and punitive sanctions (whether criminal or administrative, such as fines), as well as the prevention of harm through, for example, injunctions or guarantees of non-repetition."[939] Having a system of redress is essential as corporations may not be able to identify and prevent all human rights issues before they manifest. They comprise one of the three Pillars in the Protect, Respect, and Remedy framework and both states and corporations have duties to provide remedies under the UNGP's. Remedies are perceived as a set of interlocking mechanisms by the UNGP's, with operational-level mechanisms providing opportunities for early recourse and state-based mechanisms offering solutions to broader, more systemic issues. It should be reiterated that states have the primary duty to ensure access to remedy for business-related human rights abuses.[940] Remedies are an example of what Waldron terms the "waves of duties" that rights entail, where states have duties of protection, enforcement, and remedy (among others) and a failure to fulfil these duties will only generate more duties.[941] States must ensure that victims of abuses in their jurisdiction have access to an effective remedy.[942] This can entail both judicial and non-judicial mechanisms such as courts (civil and criminal), labour tribunals, national human rights institutions, ombudsmen and government-run complaints offices.[943] This section, however, has examined the obligations a business would have under a mandatory HRDD scheme and so the obligations of businesses to provide remedy will be the focus here.[944]

Companies must create industry-level or company-level grievance mechanisms to allow people who feel their human rights have been violated to register complaints. Smaller companies may benefit from creating collective grievance mechanisms allowing them to

[939] Commentary to Principle 25, *UN Guiding Principles.*
[940] Principle 25, *UN Guiding Principles.*
[941] For example, a failure to protect people or provide remedies could lead to duties to conduct internal investigations or take further action inside the government. Jeremy Waldron, "Rights in Conflict," *Ethics* 99, no. 3 (1989): 510.
[942] Waldron, "Rights in Conflict," 510.
[943] Commentary to Principle 25, *UN Guiding Principles.*
[944] Remedies at the state-level will be discussed at 7.4.5.

share resources.[945]Another option would be for companies to outsource the creation of grievance mechanisms to consultants who could assist them with development. Companies should value complaints systems as they can act as an early-warning system for issues and help them to identify systemic issues.[946] These issues should then be fed into HRIA's which can then help to inform further developments in the remedy process.[947] A grievance mechanism should be a place where people can register their complaints even if they haven't morphed into full-scale human rights abuses as it could be possible to remediate the issue before it escalates.[948] Remedies may entail paying compensation to people or communities affected by corporate human rights abuses. The UN draft Bill on mandatory HRDD includes a provision explaining that "due diligence may require establishing and maintaining financial security, such as insurance bonds or other financial guarantees to cover potential claims of compensation."[949]

The very fact that a grievance mechanism exists is not sufficient, it must meet certain criteria, which serve to ensure that stakeholders do not feel disempowered or disrespected by the process, thus "compounding a sense of grievance."[950] A grievance mechanism, therefore, must be legitimate (enabling trust from stakeholders and ensuring accountability), accessible, predictable, equitable, transparent, rights-compatible, a source of continuous learning, and based on engagement and dialogue.[951] This requirement dovetails with another principle of good regulation: the due process requirement (that procedures are fair, accessible, and open).[952] Baldwin and Cave argue "Attention is paid to equality, fairness, and consistency of treatment but also to the levels of participation that regulatory decisions and policy processes allow to the public, to consumers, and to other affected parties."[953] In fact, many of the principles that form a strong grievance mechanism

[945] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 26-27.
[946] Perrin and Woods, *Online harm Reduction*, 50.
[947] Götzmann, "Human Rights Impact Assessment of Business Activities: Key Criteria for Establishing a Meaningful Practice," 104.
[948] Commentary to Principle 29, *UN Guiding Principles.*
[949] Article 9(h), UN Human Rights Council, "Zero Draft Bill."
[950] Commentary to Principle 31, *UN Guiding Principles.*
[951] Principle 31, *UN Guiding Principles.*
[952] Baldwin and Cave, *Understanding regulation*, 79.
[953] Baldwin and Cave, *Understanding regulation*, 79.

have been discussed throughout the thesis as being key components of a social media content moderation process.

Finally, it must be reiterated that while the provision of remedies is essential in a scheme that fulfils both the state and corporate duties towards human rights, preventative measures should always be preferred to remedial actions. Buhmann explains:

> "human rights damage is rarely fully remediable: an arm lost in an occupational health and safety accident cannot be replaced; a childhood lost to factory labour cannot be relived; lethal chemicals polluting drinking water or agricultural land cannot disappear overnight; and the impacts of environmental damage to the possibilities for farmers or fishermen to provide for themselves and their families may persist for a long time."[954]

It is important to remember that these are the stakes in HRDD. The core of HRDD must be the identification and prevention of adverse impacts with remedies treated as a last resort when more proactive measures have failed. It is imperative that companies do not treat remedies as a Ford Pinto style cost-benefit analysis: assessing how much it would cost to pay out compensation compared to resolving the problem at an earlier stage and choosing the cheaper course of action.[955] That being said, a mandatory HRDD scheme, where companies are penalised not only for human rights abuses but also for failing to engage in due diligence processes (even if no harm has occurred yet) should help to rectify this problem by making compliance more advantageous than non-compliance.

For social media platforms, the germ of a grievance-mechanism already exists: the appeals process. It would be possible to expand the purview of this process and allow users or civil society organisations the ability to provide more nuanced feedback on the platform's policies and processes (see 3.5 and 5.4.2) and make formal complaints when they suspect there is an issue with human rights. The content moderation process can therefore be

---

[954] Buhmann, "Neglecting the Proactive Aspect of Human Rights Due Diligence? A Critical Appraisal of the EU's Non-Financial Reporting Directive as a Pillar One Avenue for Promoting Pillar Two Action," 40.
[955] The Ford-Pinto situation involved a decision by the Ford company to keep the Ford Pinto car on the road even when they discovered it had a design flaw that meant it burst into flames when rear-ended. Their decision was based on a cost-benefit analysis that demonstrated that it was cheaper to pay compensation to victims than do a product recall on all Ford Pintos. Mark Dowie, "Pinto madness," *Mother Jones*, September, 1977.

improved and expanded to address the inherent human rights issues that exist on platforms that host content. Complaints would pass through a set of interlocking mechanisms at the platform in a similar way to the extended appeals process proposed earlier (at 5.4). Moderators would be given special human rights training on issues that are relevant to social media and would triage complaints, weeding out vexatious claims, addressing individual problems that are easily resolved and passing on legitimate concerns to the policy teams at the platforms for collation, consideration, and identification of systemic issues.

One also needs to consider what methods of redress should be available for the remediation of human rights issues in the content moderation process. As discussed earlier, the UNGP's provide a list of potential remedies and while not all of these methods would be appropriate for social media networks, it is clear that simply removing or reinstating content should not be treated as the only options when a human rights issues occurs. Chapter Five provided a number of suggestions of other remedies such as compensation, public acknowledgement of error, and newsfeed optimisation to rectify unfair removals (at 5.4.1). This is not an exhaustive list and social media platforms would likely develop a new range of creative remedies if incentivised to do so by a law mandating HRDD.

This section has explored the bundle of processes that make up the HRDD model in the UNGP's. These processes include policies, HRIA's, implementation, tracking performance, and the provision of remedies. All of these processes would need to be undertaken in a transparent way with some disclosure to the public (although certain information would be anonymised to protect stakeholders). While these processes may be novel in their application to human rights, the core of these activities are already undertaken by companies in relation to other issues such as business risks, environmental damage, and, most recently, data protection. Human rights, therefore, can inform the development of important new processes as companies are provided with "a globally recognised framework for designing those tools and a common vocabulary for explaining their nature, purpose and application to users and States."[956] The UNGP's offer a starting point for legislators when mandating due diligence but it is likely that further refinements will be required when these

[956] Kaye, *A/HRC/38/35*, 15.

obligations are implemented. One will also have to consider what compliance would look like in each sector, as the challenges and capabilities of social media companies may differ markedly from, say, a large agri-business.

7.4: The implementation of mandatory HRDD at the national level

7.4.1: Introduction

This chapter has thus far explained the justifications and the requirements of HRDD. Exploring how this idea would be implemented is just as essential. It would be myopic to argue that procedural due process is as important as substantive concerns and then only focus on the substantive content of the proposed solution. Implementing a national law mandating HRDD should be construed as the first step in encouraging corporate respect for human rights. It also helps the UK to meets its obligations under the UNGP's, which require that states protect against abuses by businesses in their territory and "set out clearly the expectation that all business enterprises…respect human rights throughout their operations."[957] Admittedly, a national law will always be bounded by its jurisdictional limitations and this can seem particularly nonsensical in relation to transnational technologies like social media. An international treaty, therefore, would be the ideal solution for addressing companies that have a global impact. But it is likely that the UN draft treaty (at 7.2.3.i) will be a long-term project as reaching a conclusion about the substantive content will require extensive consultation and negotiations. A national law, therefore, can act as a more streamlined, responsive precursor to any subsequent multilateral efforts.[958] Palombo concurs, arguing that while domestic models provide "sub-optimal solutions," they do offer a real opportunity to "provide victims of abuses with effective remedies in domestic courts"

[957] Principles 1 and 2. *UN Guiding Principles.*
[958] Bilchitz terms this as having an eye on the short-term possibilities of business and human rights law as well as the long-term. National laws, therefore, represent a short-term gain that should not be overlooked just because an international treaty would be the ultimate achievement. Bilchitz, "The Necessity for a Business and Human Rights Treaty," 223.

while international law continues to debate over the role of non-state actors in human rights issues.[959]

Implementing a national law also makes sense in a social media context as it allows for a faster response in a sector where reform is urgently needed. It is inevitable that governments will regulate social media (they are already introducing strong laws such as the German Network Enforcement Act) and if advocates want human rights considerations to be included in these laws then they cannot wait for international deliberations. Successful national attempts to impose human rights obligations on platforms can also help to inform the development of international law as treaty-makers begin to understand how social networks can be regulated. This is characterised as a "continuous upward-downward cycle of norm creation" and means that domestic law should not be seen as an inferior substitute for international treaties.[960]

This section will explain how a mandatory HRDD law could be implemented in the UK, the scope of the law, and how it could be enforced through a regulator. It will introduce the requirements of such a regime while investigating the implications of this approach for social media companies. This approach is actionable and offers a necessary antidote to punitive plans to regulate social media on the basis of harm or the speed of content removal, both of which side-line important human rights issues. While the focus of this section is the British context, many of these elements could be applied in other jurisdictions or inspire other regulatory initiatives. The objective of this regime is to force social media companies to prioritise human rights and to identify the complex ways that platforms both help and hinder the enjoyment of human rights.

7.4.2: Establishing the new law of mandatory HRDD

The UK would need to introduce new legislation creating a general obligation for companies to carry out HRDD. Businesses would owe this duty to "classes of persons whom a reasonable exercise of due diligence identifies as likely to be at risk from the business

[959] Palombo, "Towards a New Treaty," 286.
[960] Ronald Dworkin, *Taking rights seriously* (Delhi: Universal Law Publishing 1999), 81; Deva, "Human Rights Obligations of Business," 3.

activity."[961] This obligation would apply to all businesses (not just social media companies) because all companies can violate human rights. It is also important not to focus only on tech companies as the barrier between what is and isn't a tech company is growing weaker as more companies embrace digital technologies. That being said, the focus of this section will be on how this law will impact on social media companies. As indicated previously, this law would be modelled on the UNGP's so companies would be required to enact human rights policies, carry out HRIA's, implement the results, track performance, and provide remedies. There would need to be regular disclosures about all of these activities both to the designated regulator and the public. The emphasis would be on ensuring that companies have the appropriate frameworks in place and are engaging in due diligence processes.

If a social media company is taken to court because human rights violations have occurred, any good-faith attempts to engage in due diligence (so long as it resulted in a reasonable response) would be rewarded by a reduced likelihood of liability (at least under the proposed law).[962] Companies that could display sound due diligence procedures may mitigate or even escape liability for certain actions by their subsidiaries on the grounds that these human rights issues were not reasonably foreseeable.[963] That being said, if a human rights abuse did occur, a failure to exercise due diligence could create a rebuttable presumption of causation and therefore liability.[964] The burden, therefore, rests on the

[961] Doug Cassel, "Outlining the Case for a Common Law Duty of Care of Business to Exercise Human Rights Due Diligence," *Business and Human Rights Journal* 1, no. 2 (2016): 200, https://doi.org/10.1017/bhj.2016.15.
[962] A similar point was made by Kinley and Chambers although they were making it in relation to the UN Norms, which they argued "provide a framework for decision-making that allows companies a reasonable margin of discretion in what they decide**."** This thesis argues that a mandatory HRDD scheme based on the UNGP's instead of the UN Norms would also offer this but would provide more certainty to all affected parties. David Kinley and Rachel Chambers, "The UN Human Rights Norms for Corporations: The Private Implications of Public International Law," *Human Rights Law Review* 6, no. 3 (2006): 476, https://doi.org/10.1093/hrlr/ngl020.
[963] This would likely be rare as most human rights issues should be foreseeable if thorough HRIA's are conducted. De Schutter explains that "Only where the parent company can demonstrate that it was unable to effectively avoid the contested behaviour of the subsidiary company from occurring, despite having exercised due diligence and despite its best efforts to seek information about such behaviour and to react accordingly, should its liability be excluded." Olivier De Schutter, "Towards a New Treaty on Business and Human Rights," *Business and Human Rights Journal* 1, no. 1 (2016): 53, https://doi.org/10.1017/bhj.2015.5.
[964] Cassel, "Outlining the Case for a Common Law Duty of Care of Business to Exercise Human Rights Due Diligence," 180.

company in those scenarios to prove that it has met its due diligence obligations.[965] Palombo argues that this is essential as victims of human rights abuses will often have limited resources and information and accordingly may be unable to meet the standard of proof regardless of the validity of their case.[966] This approach incentivises companies to monitor and assess human rights issues in both their business activities and those of their subsidiaries because the more engaged the company is, the easier it will be to show that it has engaged in due diligence. It can be contrasted with the "piercing the corporate veil" approach where due diligence of a subsidiary by a parent company is disincentivised because it could be used as evidence that they are sufficiently close to impose liability on the parent company for a subsidiary's actions.[967]

This approach would afford platforms a measure of discretion over what content is permitted on the platforms while ensuring that the processes and activities of social networks be open to judicial and regulatory oversight. This discretion, of course, would be bounded by the requirements of the law, so Child Sexual Abuse Material (CSAM), extreme pornography, and terrorist content would not be permitted because of the laws that are already in existence about that content. This proposed HRDD law has a significant advantage over more censorial approaches because it encourages platforms to consider all relevant human rights issues instead of concluding that compliance must take the form of excessive content removal. Of course, the government might also pass more substantive laws in the future (within the bounds of their human rights obligations of course) but a residual discretion would exist for platforms in areas that weren't the target of specific legislation. There would still be a diversity of platforms with different environments but all platforms would be more responsive to human rights issues. This form of regulation would be principle-based and allow flexibility, which Information Commissioner Elizabeth Denham argued was a necessary feature of any attempt to successfully regulate social media.[968]

---

[965] This reversed burden of proof has been proposed in both of the Swiss proposals for due diligence. Palombo argues that this is an innovative approach and likely to remedy some of the deficits in previous due diligence laws which were too onerous for victims. Palombo, "Towards a New Treaty," 284.
[966] Palombo, "Towards a New Treaty," 283-84.
[967] De Schutter, "Towards a New Treaty," 52.
[968] Denham did not explicitly discuss mandatory HRDD, the comment was about principle-based regulation. Select Committee on Communications, *Regulating in a digital world*, 14.

Mandatory HRDD would also preserve some of the advantages of self-regulation (efficiency, scalability, a large amount of resources) while emphasising the expectation that "the private sector develops and enforces rules in an accountable and transparent way."[969] It is similar to Murphy's argument that there must be a "code for codes" that would provide a "quality control template or standard reference points" for stakeholders concerned about corporate behaviour.[970]

Critics of mandatory HRDD might argue that such a requirement would burden small tech companies and thereby represent a significant barrier to innovation. As we have seen, the duty of vigilance law introduced in France only applied to larger companies. Innovation is important but tech start-ups also have the technical ability to create streamlined HRDD processes that can expand as the company develops. Some commentators also argue that the burden of due diligence on SME's may not be disproportionate in any event as SME's are likely to have less products and services to track, which would "decrease their compliance cost." [971] Take, for example, Snapchat, which has a very narrow range of offerings as compared to Facebook, whose platform is constantly adding new functionalities. While Facebook has more resources to conduct HRDD, it also has more services it needs to assess. The UNGP's also state that the responsibility of companies to respect human rights "applies to all enterprises regardless of their size, sector, operational context, ownership and structure." [972] Whether there may be smaller start-ups with more complex HRDD requirements, instead of providing exemptions for smaller companies, the UK government could instead introduce concessions (such as the GDPR permitting a group of small companies to share a single data protection officer) and provide educational resources to assist companies with compliance. [973] After all, smaller companies can still cause or

[969] Mifsud Bonnici and de vey Mestdagh, "Right Vision, Wrong Expectations: The European Union and Self-regulation of Harmful Internet Content," 145.
[970] Murphy, "Taking Multinational Corporate Codes of Conduct to the Next Level," 426.
[971] *Conflict Minerals: Final Rule (Release No. 34-67716; File No. S7-40-10)*.
[972] Principle 14, *UN Guiding Principles.*
[973] The UK has already created some useful tools for companies to access when considering human rights issues. These include the Overseas Business Risk Service which provides information about countries (including human rights issues) where UK Trade and Investment has a presence, the Business and Human Rights Toolkit, training courses, and an online hub where information and best practices can be shared. Many of these tools could be further adapted to meet the needs of a mandatory due diligence scheme and could ease the transition for companies. Secretary of State for Foreign and Commonwealth Affairs, *Good Business*.

contribute to serious human rights abuses.[974] Making exceptions for smaller companies is also out-of-touch with the realities of the digital sector. One need only think of the damage the company Cambridge Analytica left in its wake to understand the problem with exempting smaller companies just at the moment in history when technology has enabled small companies to have large impacts. It is also important that companies integrate human rights considerations into their activities from the beginning because it can be more difficult to make changes later if the company suddenly becomes extremely successful and then meets the threshold for a law that only applies to larger companies. HRDD, therefore, must be characterised as simply the cost of doing business in the UK.

7.4.3: Jurisdiction

As a domestic provision, it would only be applicable to British companies and to companies that direct their services into the UK. British companies would be expected to conduct due diligence for all of their processes (including activities carried out in other countries or through affiliates) while foreign companies would only be expected to carry out due diligence for the activities they conduct in, or directed towards, the UK. Despite certain limitations of jurisdiction, imposing obligations on companies that are either based in the UK or direct their services into the UK would still have a significant impact on human rights issues. For example, 83 of the top 2000 largest companies in the world are British. Imposing HRDD obligations on these companies would therefore have positive consequences in all the countries where these companies conduct business or have subsidiaries or supply chains.[975] If the mandatory HRDD law is successfully implemented, it can also help to inform other laws around the world or provide useful information for the UN draft treaty so it is possible that the normative influence of the law could extend far beyond its legal boundaries.

In short, when it comes to companies established outside the UK, the state's authority claim is "partial and thus restricted to those business activities which directly affect the country." [976] This is a similar approach to the GDPR, which applies to all data

[974] A similar point was made by the *Report (A/73/163)*, 19.
[975] "The World's Largest Companies," Forbes, accessed November 9, 2019 https://www.forbes.com/global2000/list/#country:United%20Kingdom.
[976] Reed and Murray, *Rethinking the jurisprudence of cyberspace*, 22.

controllers/processors based in the European Union and (crucially) all processing of data subjects from the EU where the processing is related to the offering of goods or services and the monitoring of behaviour.[977] This provision embodies the notion that one can impose obligations on how businesses conduct their activities within the UK regardless of where they are established[978] and that one can impose further obligations on British companies. The commentary for UNGP Two also states that states are not prohibited from regulating the extraterritorial activities of businesses domiciled in their territory "provided there is a recognised jurisdictional basis."[979] Coupled with UNGP One, which requires that states protect against human rights abuses that occur within their territory,[980] it seems clear that this approach to jurisdiction complies with the spirit of the UNGP's, which takes an "agnostic approach to the extraterritoriality issue."[981]

The basis for this approach to jurisdiction would be the notion of targeting, where it is considered appropriate to assert legal jurisdiction over companies based outside the territory if they direct (or target) their services to customers in the territory. This is relatively uncontroversial as laws have always claimed to regulate activities that occur outside the jurisdiction but which have effects within a state territory.[982] A similar approach is taken by American courts when dealing with jurisdiction in the context of e-commerce.[983] The 'sliding scale' approach envisions websites based outside of America as existing on a spectrum.[984] On one side are companies clearly doing business with Americans over the Internet. As they are 'purposefully availing themselves' of activities in that territory (a similar notion to voluntarily directing data into a jurisdiction), exercising jurisdiction would be proper. On the other end of the spectrum would be passive websites that simply post

---

[977] Article 3: Territorial Scope, General Data Protection Regulation.
[978] Although for companies that do not have assets in the UK, these laws would need to include measures to excludes companies from the UK marketplace if they do not comply with sanctions. These methods would include controls on imports and technical measures such as blocking for online companies.
[979] UN Guiding Principle 2, *UN Guiding Principles.*
[980] UN Guiding Principle 1, *UN Guiding Principles.*
[981] Marco Fasciglione, "The Enforcement of Corporate Human Rights Due Diligence: From the UN Guiding Principles on Business and Human Rights to the Legal Systems of EU Countries," *Human rights and international legal discourse* 10 (07/01 2016): 106.
[982] Chris Reed, *Making laws for cyberspace*, 1st ed. (Oxford: Oxford University Press, 2012), 30.
[983] Elissa Okoniewski, "Yahoo!, Inc. v. LICRA: The French Challenge to Free Expression on the Internet," *American University International Law Review* 13 (2002).
[984] *Zippo Mfg. Co. v. Zippo Dot Com, Inc.*, 952 F. Supp. 1119, 1124 (W.D. Pa. 1997).

information accessible in any country and exercising jurisdiction over these websites would be deemed improper.[985] The middle of the spectrum, like any legal test that employs a scale, is the least clear and the source of judicial debates. The application of jurisdiction to these websites depends on the level of interactivity between the user and the host computers (and whether they are exchanging information).[986] Interactivity is measured by considering the intended uses of the website and other case-by-case features.[987] In the European context, a similar approach to targeting was used by the French court in the *Yahoo! v Licra* case when they ruled that that it was proper to exercise jurisdiction because Yahoo! was aware that the auction could be accessed from France, they had the technical capabilities to identify French users, and had even responded to the French presence by displaying French advertisements on the website.[988] Of course, targeting does have critics[989] but it seems a particularly appropriate tool to use in considering the responsibilities of social media companies, who have benefitted from their technological abilities to identify where users are located and serve them localised advertisements.

Targeting, would therefore, be a factual analysis. One essentially looks at the facts to determine whether a company is "participating in the commercial life of that foreign country."[990] It is therefore important to identify "those who have continuous and persistent communication with residents of the state."[991] In relation to social media platforms, one might consider, for example localised advertisements, advertising revenue generated in the UK and whether the platform has the technical means to identify where their users are based.

---

[985] This test also bears some resemblance to the "minimum contacts" doctrine, where foreign defendants must have a sufficiently close connection to a jurisdiction to be subject to judicial proceedings. In the case of *CompuServe v Patterson*, the court stated that it had three criteria for making this determination. These could be summarised as the defendant must choose to act within that territory or cause a consequence, the cause of action must arise from the defendant's actions in the territory, and the acts must have a "substantial enough connection" to the territory to make exercising jurisdiction reasonable. *CompuServe v. Patterson*, 89 F 3d 1257(6th Cir 1996).

[986] Christopher Wolf, "Standards for Internet Jurisdiction: An overview," Find Law, last modified March 3, 2008, http://corporate.findlaw.com/litigation-disputes/standards-for-internet-jurisdiction.html.

[987] *Zippo Mfg. Co. v. Zippo Dot Com, Inc.*, 952 F. Supp., 124.

[988] *UEJF et LICRA v. Yahoo! Inc. et Yahoo France* Tribunal de Grande Instance de Paris [TGI] [High Court of Paris] RG 05308 May 22, 2000, 18, 20, 30.

[989] Uta Kohl, for example, argues that interactivity is not a relevant factor to determining jurisdiction over websites. See: Uta Kohl, *Jurisdiction and the Internet: a study of regulatory competence over online activity* (Cambridge, UK: Cambridge University Press, 2007), 84.

[990] Reed and Murray, *Rethinking the jurisprudence of cyberspace*, 22.

[991] Reed and Murray, *Rethinking the jurisprudence of cyberspace*, 24.

Since these platforms are already subject to the GDPR and must comply with those requirements, it is clear that companies are aware that they have users in the UK (as well as across Europe) and since these companies already comply with national laws (using, for example country-withheld-content to address different hate speech legislations) it is unlikely that these companies will contest that they target UK users.

Critics may argue that social media companies may simply choose to make their services unavailable in the UK rather than comply with these requirements. This seems unlikely for a number of reasons. First, platforms do obey many local laws that are much more repugnant to an American company than a HRDD law such as laws requiring that they censor content within a particular country that would in the US have been protected by the First Amendment.[992] It is also possible that these companies would perceive this regulatory bid as legitimate as it does not excessively disrupt the online environment. Reed and Murray explain that the online environment is made up of the technologies that are used, social and business practices, and the current normative landscape. Laws that are too disruptive to this environment are met with a high degree of resistance and perceived as less legitimate.[993] This due diligence law seeks only to align some of these social and business practices along a human rights perspective so it is unlikely to merit sustained adversity.

Second, the introduction of the GDPR has shown that when faced with a choice of compliance or making their services unavailable, companies will comply so long as the requirements are not too onerous. The due diligence law, as outlined in this chapter, would likely benefit companies in the long-term as they prevent larger controversies so it is likely that social media platforms would choose to comply. Third, a decision to refuse compliance with an HRDD law and withdraw from the UK would likely generate negative publicity and provide an opportunity for other, more ethical social media companies to assert dominance.

---

[992] An example of this is Facebook's compliance with Thailand's "lese-majeste" laws which strictly prohibit any negative content about the royal family. Facebook even blocked a video of Thailand's king wearing a crop top and showing off his tattoos in a shopping centre. One cannot think of a form of speech that reflects the sprit of the American Constitution more than speech criticising the leaders of a country. Gabriel Samuels, "Thailand threatens to sue Facebook over anti-monarchy posts," Independent, last modified May 12, 2017, https://www.independent.co.uk/news/world/asia/thailand-facebook-anti-monarchy-posts-lawsuit-sue-military-government-king-maha-a7731846.html.

[993] Reed and Murray, *Rethinking the jurisprudence of cyberspace*, 84.

Finally, it is unlikely that social media companies would make themselves unavailable in the UK because it is a valuable market for social media companies. As of 2019, 67% of the British population (or 45 million people) are considered "active social media users" with the average British person spending 1 hour 50 minutes a day on social media.[994] This is quite a valuable market, a wealthy country with high social media penetration so it seems unlikely that social media platforms would choose to withdraw rather than comply with a due diligence law, just as social media companies chose to comply with Germany's Network Enforcement Act rather than withdraw their services from the country.

7.4.4: The Business and Human Rights Regulator

A regulator would need to be identified when the mandatory HRDD law is introduced. There are two viable options in the UK when choosing a regulator for business and human rights: create a new regulator or endow an existing regulator with an additional mandate. The advantage of creating a new regulator is that it would be a bespoke solution to the issue of business and human rights and would offer the possibility of a blank slate to tailor policies and processes. There is, however, a significant disadvantage to creating a new regulator for human rights. Since 2010, regulatory bodies have been targeted for closure, with many regulators being dissolved or conglomerated by the Conservative government.[995] New regulators have been introduced, such as the Anti-Slavery Commissioner and the Gangmasters and Labour Abuse Authority but their budgets have been so low that they have been unable to carry out broad investigatory and enforcement activities.[996] In this climate, it might be difficult to convince policymakers to create a new regulator or to endow one with sufficient resources to do the job.

The other option would be to designate the Equality and Human Rights Commission as the appropriate regulator in the statute introducing mandatory HRDD. This is an obvious choice as the EHRC does already provide information and consult on issues of business and

[994] Simon Kemp, "Digital in 2019: Global Internet Use Accelerates," We Are Social, last modified January 31, 2019, https://wearesocial.com/uk/blog/2019/01/digital-in-2019-global-internet-use-accelerates.
[995] Polly Curtis, "Government scraps 192 quangos," The Guardian, last modified October 14, 2010, https://www.theguardian.com/politics/2010/oct/14/government-to-reveal-which-quangos-will-be-scrapped.
[996] Funding of regulators will be discussed in greater detail below.

human rights. In the past the EHRC has been plagued with allegations of in-fighting[997] and some academics have criticised the EHRC for being timid in their approach to business and human rights and excessively focussing on issues of equality in workplaces and not human rights more generally.[998] Lately, however, the EHRC has released a number of useful guides on human rights in business and seem to have taken a more active stance.[999] A statutory HRDD obligation would give the Commission new powers in relation to companies and could infuse the organisation with a new energy and assertiveness. Of course, it is very important that the EHRC budget be increased accordingly as it was also slashed ten years ago during David Cameron's so-called "bonfire of the quangos."[1000] Both approaches (creating a new regulator or choosing the EHRC) have their advantages and disadvantages but it seems more politically palatable to identify the EHRC as the appropriate business and human rights (BHR) regulator and then empower them stronger enforcement tools. The rest of this section, however, would be applicable whichever approach was chosen.

Currently, there are many issues with social media companies and one might be tempted to argue that the obvious solution is to designate a specific "Social Media Regulator" for all of the problems that could be construed as social media issues. This would be a mistake however because it creates a false distinction between issues in social media and other related issues that are already managed by regulators. Competition issues for tech companies are best managed by the Competition and Markets Authority whereas data protection problems fall under the remit of the Information Commissioner's Office (ICO). The government has also announced that Ofcom will likely receive new powers to address safety issues on social media (such as cyber-bulling and self-harm content).[1001] Therefore, the human rights issues of social media companies should be assigned to a business and human

---

[997] Joint Committee on Human Rights, *Equality and Human Rights Commission: Thirteenth Report of Session 2009–10* (London: Stationery Office, 2010), 8, 11, 15.
[998] Laidlaw, *Regulating speech in cyberspace*, 276.
[999] See, for example: "Guidance for small businesses and human rights," Equality and Human Rights Commission, last modified July 15, 2019, https://www.equalityhumanrights.com/en/advice-and-guidance/guidance-small-businesses-and-human-rights#h3.
[1000] In 2007, the EHRC's budget was £70 million but it dropped to £26.8 million by 2015. In 2018, it was even lower, with a budget of £19.47 million. *Strategic Plan 2012-2015* (London: EHRC, 2012); *Business Plan 2018-19* (London: EHRC, 2018).
[1001] "Regulator Ofcom to have more powers over UK social media," BBC News, last modified February 12, 2020, https://www.bbc.co.uk/news/technology-51446665.

rights regulator, because human rights issues in the online sphere are still human rights issues, even if they differ in their scale, method, or impact. This belief is supported by the UN Special Rapporteur on Freedom of Expression, who stated in 2015 that the role of private actors is one of the most pressing human right issues in the digital age.[1002] The digital world is no longer a discrete universe, it is *our world*, and a regulator that fails to deal with the online and offline aspects of any subject is doomed to become obsolete.

The BHR regulator would monitor companies for compliance, provide educational resources, hear complaints from stakeholders, carry out investigations, and impose sanctions. They would also maintain a publicly available central registry of the HRIA's and human rights policies disclosed by companies.[1003] The mandate of this regulator would include all British companies and companies directing their services into the UK. A regulator is important because it can offer free, alternative redress to complainants and is an attractive alternative to costly court procedures.[1004] This is especially the case in situations involving technology companies, which can be extremely difficult to litigate due to jurisdictional issues and can prove costly for advocates to pursue.[1005] State-based non-judicial mechanisms (including national human rights institutions) are also lauded by Ruggie for providing a more "more immediate, accessible, affordable, and adaptable point of initial recourse."[1006] Despite the benefits to be gained from a regulator, some of the previous due diligence laws, such as the French duty of vigilance, rely on complainants taking companies to court and while this would still be possible, a regulator ensures that more cases are investigated and victims are not denied justice because of a lack of resources. Social media companies have extensive

---

[1002] Kaye, "Keynote speech."

[1003] A lack of a central registry was one of the issues with both the French duty of vigilance law and the Modern Slavery Act. It is, however, mandated by the Dutch due diligence law. Anneloes Hoff, "Dutch child labour due diligence law: a step towards mandatory human rights due diligence," Oxford Human Rights Hub, last modified June 10, 2019, https://ohrh.law.ox.ac.uk/dutch-child-labour-due-diligence-law-a-step-towards-mandatory-human-rights-due-diligence/.

[1004] Ursula Smartt, *Media and entertainment law*, 2nd ed. (Abingdon, Oxon: Routledge, 2014), 529.

[1005] See, for example, the protracted litigation in both France and America over the issue of Nazi memorabilia being made available on Yahoo! auction sites. *UEJF et LICRA v. Yahoo! Inc. et Yahoo France* Tribunal de Grande Instance de Paris [TGI] [High Court of Paris]. *Yahoo! Inc. v. La Ligue Contre Le Racisme et l'Antisemitisme, et al*, 433 F.3d 1199(9th Cir. 2006); *Yahoo Inc. v. LICRA*, 145 F. Supp. 2d 1168(N.D. Cal. 2001).

[1006] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 22. Principle 27 also states that "states should provide effective and appropriate non-judicial grievance mechanisms, alongside judicial mechanisms, as part of a comprehensive State-based system for the remedy of business-related human rights abuse." *UN Guiding Principles.*

resources to fight litigation and it is inappropriate to place the onus on the public to protect their human rights from widespread violation. A BHR regulator would become a repository for complaints and could help identify systemic issues at social media platforms. One need only think of a situation such as Twitter's challenges with the harassment of women on its platform[1007] to consider how a dedicated regulator could help to push for systemic reform. These complainants would be encouraged (but not required) to approach the company first but if a reasonable amount of time has passed without a satisfactory response, they should notify the regulator.[1008] This is a similar approach to Ofcom, which prefers complaints are first made to the broadcaster but will accept complaints made directly to it in the first instance.[1009] In relation to social media companies, this means that individuals would be encouraged to first contact platforms (ideally through the dedicated forums for participation suggested at 3.5 and the enhanced appeals system at 5.4) to seek resolution of their issue before making a complaint to the BHR Regulator.

7.4.5: Enforcement

This chapter has emphasised that it is not enough to devise a mandatory HRDD scheme, there must be adequate enforcement mechanisms to encourage corporate compliance and accountability. A failure to introduce consequences for non-compliant companies (and thus creating unfair burdens on companies that do comply) has been identified as one of the common weaknesses in any business and human rights regime.[1010] Enforcement will distinguish a mandatory HRDD law from the many voluntary or laissez-faire schemes that have existed in the past and have failed to catalyse real change. Ensuring an appropriate scheme for accountability is also one of Baldwin and Cave's key tests for good

---

[1007] "Toxic Twitter-A Toxic Place for Women (Report)," Amnesty International, last modified March, 2018, https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/.
[1008] In the Dutch child labour due diligence law, stakeholders can only file a complaint with the regulator after a company has failed to respond to the complaint after six months (or dealt with the complaint in an unsatisfactory fashion). This law, however, only addresses child labour and six months may be too long in a general due diligence law depending on the nature of the harm. A discretionary approach by the regulator should instead be adopted. It would also be open to the regulator to follow up with a company and then disengage if the company began to interact with the complainants or remedied the situation. For a discussion of the Dutch law, see: Hoff, "Dutch child labour due diligence law."
[1009] Thomas Gibbons, *Media law in the United Kingdom*, 2nd ed. (Alphen aan den Rijn: Kluwer Law International, 2014), 15.
[1010] *Report (A/73/163)*, 19.

regulation, a useful set of criteria for assessing regulatory proposals.[1011] This chapter has already explored a number of laws that have lacked enforcement and produced unsatisfying results such as the French duty of vigilance law (where companies were not being penalised for failing to publish plans) and the Modern Slavery Act, (where 40% of companies did not publish statements and faced no penalties). Numerous sources have cited the importance of enforcement mechanisms that are "rigorously applied"[1012] so that they can "incentivise behavioural change in those who are tempted to breach regulators requirements."[1013] Enforcement is where actual change occurs, where governments transcend "being seen to do something" and catalyse real reform in how businesses approach human rights.

The BHR regulator should have enforcement powers modelled on the ICO, another regulator designed to monitor and encourage compliance from companies (among other parties). The first enforcement tool is the information notice, which simply demands that the party provide information to the ICO when they are investigating a specific issue.[1014] This disclosure requirement would be important in a mandatory HRDD scheme as companies may not be particularly forthcoming with the regulator about their perceived failures. This thesis has consistently criticised platforms for a lack of transparency and platforms must be required to disclose more information about their human rights processes. The second tool is the assessment notice, which allows a regulator to enter the premises, examine documents and equipment, observe processing, and interview staff.[1015] The assessment notice is clearly an intrusive method but it might be necessary if a company refuses to cooperate with a regulator. It is likely, however, that this form of notice would be used infrequently against social media platforms as the large platforms do not maintain a physical premises in the UK (although there are smaller start-ups in the UK). The third tool is an enforcement notice, which is a written notice that either requires a party to "take steps specified in the notice" or "refrain from taking steps specified in the notice."[1016] In a mandatory HRDD scheme, this

[1011] Baldwin and Cave, *Understanding regulation*, 77.
[1012] Select Committee on Communications, *Regulating in a digital world*, 15.
[1013] *Regulatory sanctions and appeals processes: an assessment of the current arrangements* (London: LSB, 2014), 12.
[1014] Section 142, *Data Protection Act, 2018*, c. 12 (UK).
[1015] Section 146, *Data Protection Act, 2018*.
[1016] Section 149, *Data Protection Act, 2018*.

notice could be used to compel platforms to engage in particular processes such as prioritising the moderation of extremist content after a terrorist attack or it could order them to refrain from certain activities such as providing information about political dissidents to an autocratic country. The final tool is the penalty notice, which notifies the party that they are being fined a specific amount because of their failure to comply with the regulatory regime.[1017]

In relation to financial penalties, it is important that sanctions be "effective, proportionate, and dissuasive."[1018] Sanctions may vary depending on the culpability of the company (was this a deliberate breach or is there evidence of recklessness or negligence) and the nature of the breach. For example, it is likely that a regulator would impose a larger sanction on a platform that was complicit in serious human rights violations (such as the Cambridge Analytica scandal) as compared to a platform that failed to post their human rights policies in their terms and conditions. A prudent approach would be to echo the fines embedded in the GDPR's scheme as a harmonised approach would be clearer for companies. Mirroring the approach of the GDPR also signals to companies that human rights are as important as data protection. For infractions of certain provisions, companies can face fines of up to ten million euros or 2% of the company's annual global turnover (whichever is greater). For infractions of other provisions (articles that could be termed particularly essential) companies can face fines of up to twenty million euros or 4% annual global turnover.[1019] These fines must be large enough that they act as a disincentive to expand into new countries or introduce new services without first considering and mitigating any adverse human rights impacts. Platforms often roll out services without identifying potential issues first, such as the live-streaming of crimes (at 4.3.2). It is likely that the threat of fines would incentivise companies to evaluate their potential offerings more seriously instead of using society as a test kitchen. Other potential sanctions could include publication on a non-compliant list and an adverse publicity order.[1020] These measures should ensure that the UK

[1017] Section 155, *Data Protection Act, 2018*.
[1018] UN Human Rights Council, "Zero Draft Bill."
[1019] "Penalties," Information Commissioner's Office, accessed June 12, 2019, https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-law-enforcement-processing/penalties/.
[1020] An adverse publicity order requires a company to display a message on its homepage (or another highly accessible online location) detailing its offence. Perrin and Woods claim that "a study on the impact of

has fulfilled its duty to ensure that those who suffer human rights abuses within their jurisdiction "have access to effective remedy."[1021]

Companies would need to provide evidence of human rights-oriented policies, robust due diligence processes, and the provision of remedies to regulators. Companies should also detail how these processes work and confirm that they are consulting with stakeholders and that their procedures are fair, accessible, and open.[1022] Another way of phrasing these requirements is that these processes should have input legitimacy (include participation in the norm-creation process), throughput legitimacy (procedural due process) and output legitimacy (are likely to achieve their desired outcome).[1023] It is important that companies are not just assessed on their HRIA's (or on their disclosure of such assessments) but rather on their processes at each stage of due diligence, such as implementation and remedies. Social media companies would therefore need to reform every stage of the content moderation process, from the issues in Creation (Chapter Three), Enforcement (Chapter Four), and Response (Chapter Five). There must be consequences that apply for a range of different actions, from failing to disclose information about human rights policies and due diligence processes, to failing to address serious human rights abuses such as the Facebook/Myanmar genocide controversy (see 6.2.3). A similar approach has been used in the Dutch child labour due diligence law, where companies may face legal consequences for failing to produce a statement on child labour, failing to carry out an investigation, failing to set up an action plan, or carrying out an inadequate investigation.[1024] Of course, child labour is less of a concern for social media companies so they have not been caught by the Dutch laws.

---

reputational damage for financial services companies that commit offences in the UK found it to be nine times the impact of the fine." However, they provide no reference for that study so it is hard to assess the validity of those findings. Perrin and Woods, *Online harm Reduction*, 52.

[1021] Principle 25, *UN Guiding Principles.* This also reflects Article 2(3)(a) of the ICCPR, which states that countries must undertake to "ensure that any person whose rights or freedoms as herein recognized are violated shall have an effective remedy."

[1022] This is another one of Baldwin and Cave's tests for good regulation. They term it the "due process claim" and explain that regulators should pay attention to the equality, fairness, and consistency of treatment as well as to the levels of participation accorded to the public and stakeholders. Baldwin and Cave, *Understanding regulation*, 79.

[1023] Paiement, "Paradox and Legitimacy in Transnational Legal Pluralism," 174.

[1024] Hoff, "Dutch child labour due diligence law."

In relation to social media companies, enforcement is necessary because too many problems have resulted from relying on the platform's goodwill to address persistent human rights issues. The law must be enforced "to the level that the prospect of further enforcement influences the behaviour of cyberspace actors."[1025] There is a danger that companies will treat infrequent sanctions as symbolic penalties if they perceive noncompliance as a better commercial decision. Therefore, it is important that these laws be enforced robustly. There is an extra challenge in enforcing these regulations against social media companies as these platforms do not have assets in the UK and may choose simply not to comply with any orders or financial sanctions. However, platforms are unlikely to take this approach because non-compliance would justify technical measures being taken against them, such as blocking access to the platform from within Britain. Platforms are often very sensitive to countries threatening to block them unless they comply with national laws even if those rules would not be acceptable in America. The HRDD law has the added benefit that compliance would give these platforms an ethical appearance unlike compliance with laws that require surveillance or the censorship of blasphemous or political content. These enforcement methods should, therefore, ensure a high degree of accountability in the new regulatory regime, an essential factor to any successful scheme in regulating the digital world.[1026]

7.4.6: Funding

The details of a funding scheme would need to be worked out by the legislature but it is likely that the business and human rights regulator will be funded by a combination of sources. Many of the proposals for new digital regulators have offered up a similar formula; a combination of government investment, an industry tax, and voluntary contributions from larger companies.[1027] This seems like a logical approach when considering businesses and human rights in general as businesses should provide some of the funding for a regulator dedicated to mediating the impacts of corporate activities. In relation to social media

[1025] Reed, *Making laws for cyberspace*, 54.
[1026] Accountability is even one of ten principles the House of Lords Select Committee on Communications recommended to guide the development of digital regulation. Select Committee on Communications, *Regulating in a digital world*.
[1027] See, for example: Miller, Ohrvik-Stott, and Coldicutt, *Regulating for Responsible Technology: Capacity, Evidence and Redress: a new system for a fairer future*, 21; Perrin and Woods, *Online harm Reduction*, 57.

companies in particular, social networks have generated quite a lot of income from British users (in terms of ad revenue and data) so a requirement that some of their profits be used to reform the industry seems fitting.

It is hard to estimate how much this regulator would cost as regulators have a wide range of annual budgets with Ofcom controlling £124.2 million whereas the Equality and Human Rights Commission (EHRC) has a budget of £19.47 million.[1028] It is likely that if the EHRC is designated as the BHR regulator then a significant increase in the budget would be required. While it is difficult at this time to provide a financial estimate, it should be emphasised that regulators must be provided adequate funding to achieve their mandate. Failure to do so will result in disappointing results and the waste of thousands or even millions of pounds on a half-completed mission. Recently, a number of regulators (which were created to deal with hot-button issues) have been provided such low budgets that they have been unable to make an impact. The Gangmasters and Labour Abuse authority, for example, was originally allocated an annual budget of £4.96 million.[1029] Even more distressingly, the Anti-Slavery Commissioner was allocated an annual budget of only £500,000, which will surely stymie the objectives of the Modern Slavery Act.[1030] Creating a regulator but failing to provide it sufficient funding is like going to the doctor for antibiotics but only taking them for a day. The problem will not be resolved, resources will be wasted, and it may become even more difficult to remedy in the future because of this earlier, haphazard attempt. In relation to social networks, after so much discussion about how these companies should be regulated (and, initially, *if* they should be regulated) it would be very disappointing if one of the early attempts to explicitly regulate their practices and policies

[1028] *Business Plan 2018-19*, 38.
[1029] This has now been raised to £7.78 million but it is likely that this is still insufficient as the GLA has a remit across all industry sectors to prevent slavery and exploitation. House of Lords, House of Commons, and Joint Committee on Human Rights, *Human Rights and Business 2017: Promoting responsibility and ensuring accountability (Sixth Report of Session 2016–17)*, 44.
[1030] The Office of the Anti-Slavery Commissioner even said that that a lack of resources meant that the Commissioner "not able to maintain the sustained engagement with the business sector that he would hope for in order to develop projects and partnerships to reduce labour exploitation in the UK and internationally." House of Lords, House of Commons, and Joint Committee on Human Rights, *Human Rights and Business 2017: Promoting responsibility and ensuring accountability (Sixth Report of Session 2016–17)*, 45-46.

failed. Therefore, funding is just as essential as enforcement tools to ensure that social media companies engage with these obligations in a genuine, concerted way.

7.4.7: The Predicted Results

Implementing a mandatory HRDD law in the UK would likely have a number of consequences in the social media field. First, users will likely benefit from sweeping reforms and enhanced procedural protections such as robust appeals systems and remedy mechanisms at the platforms. This will offer social media users more opportunities to engage in discourse with companies about how platforms should be governed and what changes could benefit users. These benefits may be felt beyond British borders as platforms may choose to apply these changes more widely because of their perceived utility, positive feedback, or because of consumer demands. Citizen journalists and influencers who derive income from platform (both of whom were discussed as having specific needs at 5.2.2) will particularly benefit from procedural changes and opportunities to engage with the platforms. NGO's focussed on digital rights such as the Open Rights Group and Privacy International will also be empowered to hold platforms accountable through the new laws when in the past they could only generate adverse publicity for social media companies.

Second, as best practices are identified, there will be an increasing amount of standardisation in the HRDD process. Social media companies can be the leader in this process as they have been effective at developing content moderation tools (such as algorithmic flagging) that are capable of handling a high volume of content and managing complex platforms. Platforms would be able to create systems to make it easier for them to comply with their HRDD obligations such as databases where HRIA's can be accessed and step-by-step frameworks for due diligence.[1031] Platforms may also create software to help update their HRIA's and monitor trends across their offerings or by geographical region. It is likely that an industry for HRDD will develop as human rights consultants help social

[1031] In fact, this framework already exists. The Human Rights Reporting and Assurance Frameworks Initiative (RAFI) (a multi-stakeholder initiative) created a reporting framework that helps companies identify human rights issues and explains what information they should disclose about these risks. It is detailed, user-friendly, and would be an excellent place for any company conducting HRDD to start. See: Shift and Mazars LLP, *UN Guiding Principles Reporting Framework with Implementation Guidance* (New York: UN Guiding Principles Reporting Database, 2015).

media companies comply with the new requirements, just as data protection experts have been useful as companies updated their practices to comply with the GDPR.

Third, as organisational change occurs, social media companies will likely incorporate more proactive and preventative aspects into their business practice. This will ease the cost of compliance as problems can be averted before they become entrenched and egregious. Over time, the legal obligations of companies may need to be refined as their actions change and any weaknesses in the HRDD law are identified. One wonders what would have happened if social media had been introduced into a world where corporate human rights obligations were already a legal reality. If platforms like Facebook and Twitter had known from day one that they would be held accountable for their human rights violations, would they have identified and addressed some of the issues that have caused such controversy before their platforms became widespread? What would the digital world look like today if platforms had perceived themselves as global citizens with obligations to the public instead of coders creating digital products in a vacuum? It is impossible to speculate on alternative histories but it is certainly not too late to enact such reforms.

Finally, societal expectations of social media platforms will likely change as companies begin to either comply with the law or face sanctions for their activities. The last six years have been a reckoning for social media platforms as a series of scandals about ISIS recruitment, self-harm imagery, electoral manipulations, and privacy breaches have resulted in a shift of perspective. The democratising power of social media was first seen in the 2011 Arab Spring protests but the rest of the decade has resulted in waves of societal anxiety about an unregulated social media. This is a common pattern as in the past, many issues that were seen as being beyond the purview of the law have subsequently been regulated. This can have a powerful normative effect on how these issues are perceived by society (assuming that the regulation is successful). Keats Citron, for example, explains that forty years ago, domestic violence and sexual harassment in the workplace were seen as essentially unregulatable.[1032] Subsequent activism and regulation helped to shift societal attitudes

[1032] Keats Citron, *Hate crimes in cyberspace*, 95-119.

about these issues and the current interest in regulating social media could be harnessed to enact lasting changes in the field of business and human rights.

In order to engineer this societal shift, however, it is imperative that the law be perceived as legitimate, that it results in the "collective acceptance of an authority claim by the overwhelming majority of those to whom the claim is addressed."[1033] That is why it is imperative that the law not be overly punitive of social media companies (which do offer many rights-enhancing qualities), embody elements of a procedural due process, and is enforced by an impartial and accountable regulator. If implemented correctly, the British HRDD law could become the gold standard of due diligence laws and could inspire similar legislation in other countries.[1034] This would be a much better outcome for social media companies and human rights defenders than the current trend in punitive social media regulations like the German Network Enforcement Act. Therefore, a British HRDD law could catalyse a shift in how business and human rights are perceived and the discourse between civil society and private regulators such as social media companies.

7.5: Conclusion

It is imperative that we orient discussions about digital regulation around the principles of human rights. This contention is at the heart of this thesis and this chapter has examined how social media companies could be compelled to prioritise human rights. The goal is not merely to regulate social networks, that is a relatively easy objective. Instead, the objective is to devise a method of capturing important social goods like human rights and translate them into new methods of regulating the digital world. This is an essential task because a failure to do so will diminish the effect human rights has in a world where private companies provide some of the most accessible and widespread forums for expression and

[1033] Reed and Murray, *Rethinking the jurisprudence of cyberspace*, 18.
[1034] A similar result occurred (for different reasons) when the US adopted legislation on business bribery of foreign government officials. US companies didn't want to be at a competitive disadvantage so they lobbied the OECD to adopt an anti-bribery treaty to level the playing field. K. W. Abbott, "Rule-making in the WTO: lessons from the case of bribery and corruption," *Journal of International Economic Law* 4, no. 2 (2001): 282-83, https://doi.org/10.1093/jiel/4.2.275.

participation in cultural life. Jodie Ginsberg, chief executive of Index on Censorship has criticised laws that impose penalties on social media platforms for not removing content fast enough. She has stated that any evaluation of new regulation should not only query if more harmful content is being removed but also whether "lawful speech flourished."[1035] The latter criteria is currently being overlooked in emerging ideas of regulating social media. In human rights terms, decisions about rights are being made but without any explicit engagement in any kind of coherent rights-balancing exercise. The UK has announced its commitment to being the safest place to go online but what if that objective was re-framed? What if, instead, the goal was to be the safest country for human rights? It is a more difficult objective to achieve but the possibility of such positive consequences makes it worth attempting.

This chapter has proposed the introduction of a mandatory HRDD law to require companies to introduce a framework to respect human rights and prevent abuses. It has considered how due diligence laws are being introduced in various countries, what due diligence entails, and how an effective mandatory HRDD law could be implemented in the UK. While this thesis has focussed on the human rights issues in social media content moderation, this solution is broader, applicable to all private companies. Social media, therefore, becomes a particular example of the operation of mandatory HRDD and how it could have a substantial impact on how companies address human rights issues.[1036]

Ruggie once wrote that "the business and human rights debate currently lacks an authoritative focal point. Claims and counter-claims proliferate, initiatives abound, and yet no effort reaches significant scale."[1037] One of the most successful aspects of the UN Guiding Principles (and the earlier Protest, Respect, and Remedy Framework) was that it provided a common language for countries, activists, businesspeople, and academics to discuss the impact of businesses on human rights. A related (but more specific) language could be created by introducing mandatory HRDD in the UK. Expectations of social media companies

---

[1035] Index on Censorship, "Wider definition of harm can be manipulated to restrict media freedom."
[1036] A similar point is made by Mac Síthigh, who writes "I believe that the everyday issues of cyberlaw (and new legislation in particular) can serve to illustrate rather than negate questions like: can corporations guarantee free speech?; what is the relationship between access to media and freedom of expression; and what are the cultural consequences of corporate policies?" Mac Síthigh, "The mass age of internet law." 88.
[1037] Ruggie, *Protect, Respect and Remedy (A/HRC/8/5)*, 4.

would be formalised in a regime that would create clear benchmarks and provide certainty to users and platforms.

For too long, social media companies have prospered by perpetuating the narrative that they are agents of disruption. Disruption, however, is not an unqualified good and regulators should not passively allow fundamental principles of good governance such as the rule of law and human rights to be diminished. Robin Mansell articulates this problem by arguing that it is "increasingly difficult to unambiguously define HR and responsibilities in cyber space…many of the judgments and social values that appear to have achieved a consensus are subject to misapplication as we come to rely on technology to implement the law."[1038] Instead, the impact of social media companies has become emblematic of a larger problem: the digital world is difficult to effectively regulate if we treat these companies as having no human rights responsibilities. It is time, therefore, to rectify the schism between "free speech as a legal concept and an experienced concept"[1039] online and demand that platforms adhere to human rights law. Regulation, of course, is possible, but it is easy to pass laws that act as command-and-control, it is much harder to determine how regulation can help maximise human flourishing in the digital sphere. If social media companies are going to truly disrupt the status quo, then we should treat their activities as disrupting the traditional view that businesses have no human rights responsibilities, while being careful not to destroy the potential of platforms to be a positive force in society. Creating a digital world (and a corporate world more generally) that respects human rights principles would be a revolutionary and meaningful project for the twenty-first century.

[1038] Mansell, "Introduction and Equity in Cyberspace," 10.
[1039] Laidlaw, *Regulating speech in cyberspace*, xi.

Chapter Eight: Conclusion

This thesis was written from 2016-2020, a period of time replete with high-profile social media controversies including ISIS recruitment, far-right groups, fake news, electoral manipulation, and ongoing concerns about harassment and hate speech. These issues have transformed the public conversation from "should social media companies be regulated?" to "how do we regulate social media companies?" What was an esoteric topic in early 2016 has become one of the most widely discussed societal challenges only a few years later. Today, it is likely that many of our assumptions about the internet, assumptions that have been codified into discussions about regulating these spaces must be reviewed and updated.

While there are many issues at social media companies that could be discussed (such as competition or data protection concerns), this thesis has focussed on the content moderation process. This is a fascinating area of study because it combines a host of typical content problems (balancing competing values, delineating the margins of permissibility, privileging certain narratives) with new challenges concerning the volume of content, the globalised nature of social media platforms, and the introduction of new regulating technologies such as algorithms. Chapters Three, Four, and Five considered each stage of the content moderation process, from creation of content guidelines, to enforcement of those rules, to the formal and informal methods of responding to those rules. Content moderation is at the heart of many of the social media controversies listed above, with critics arguing that platforms are either over or under-regulating certain forms of expression.

Early optimism about social media companies and their potential for encouraging human rights and democracy has been replaced with a "techlash" where tech companies are "coming under greater scrutiny."[1040] It is imperative, however that the regulations we introduce are not overly reactionary and do not incentivise censorship, which was the case with some of the issues with the proposals outlined in Chapter Six. This thesis has consistently argued that the procedural protection of human rights and the rule of law is

[1040] Daithí Mac Síthigh, "The road to responsibilities: new attitudes towards Internet intermediaries," *Information and Communications Technology Law* 29, no. 1 (2020), https://doi.org/10.1080/13600834.2020.1677369. 3.

uniquely at risk from the private ordering of online speech but these risks are also inherent in the *governmental response* to the private ordering of online speech.

In many ways, this thesis has had two objectives: to investigate the way social media companies moderate content (and the problems that exist in their approach) and to critically analyse how countries are responding to these issues with regulation. This thesis has found that there is not enough respect for human rights and the rule of law by either of those two groups. Platforms are moderating content in a way that lacks transparency, certainty, accountability, remedies, and accordingly, legitimacy. Governments, though, are demanding that platforms act as proxy censors, incentivising over-regulation, and using these platforms as scapegoats for larger societal problems. Both groups are ignoring the value that the rule of law and human rights principles offers for addressing content moderation issues. In fact, the recent social media controversies have laid bare the fact that neither private actors nor states have been particularly committed to protecting human rights on social media platforms. It is time, therefore, to rectify the schism between "free speech as a legal concept and an experienced concept" [1041] online and introduce the human rights due diligence requirements outlined in Chapter Seven. This will ensure that the human rights obligations of platforms obligations are publicly recognised, a situation that Chapter Two argued was important for the continued relevance of human rights principles in our changing world.

Collingridge's dilemma states "when change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult, and time-consuming."[1042] Unfortunately, social media has now become so engrained in everyday society that change will definitely be expensive, difficult and time-consuming but the imperative to reform the practices of social media companies (and the costs of doing so) will only grow with every passing year. It is clear that there is a need for more regulation but the objective should be to ensure that they align with human rights and rule of law objectives without sacrificing the positive aspects of the moderation systems these companies have constructed such as their innovation and scalability. We are starting to have a dialogue about

[1041] Laidlaw, *Regulating speech in cyberspace*. 233.
[1042] David Collingridge, *The Social Control of Technology* (London: Francis Pinter, 1980), 11.

what we can and should expect from social media platforms, a conversation that will likely continue for many years to come. The term ‘dialogue’ is key because for too long, technology companies have treated this as a monologue. It is time for that to change.

Laws and Court Decisions

Ahmet Yildirim V. Turkey (Application No. 3111/10), ECHR 458 (2012).

Bland V. Roberts, 730 F.3d 368 (4th Cir. 2013).

The Civil Rights Cases 109 U.S. 3 (1883).

Parliament of the United Kingdom. *Companies Act, 2006*, c. 46 (UK).

Compuserve V. Patterson, 89 F 3d 1257 (6th Cir 1996).

Convention on the Rights of the Child,

Cunningham V. Reading Football Club Ltd, 153 Times LR (1991).

Cyber Promotions Inc. V. America Online Inc, 948 F. Supp. 436 (E.D. Pa. 1996).

Parliament of the United Kingdom. *Data Protection Act, 2018*, c. 12 (UK).

Davison V. Randall, 912 F.3d 666 (4th Cir. 2019).

Government of France. *Décret N° 2015-125 Du 5 Février 2015 Relatif Au Blocage Des Sites Provoquant À Des Actes De Terrorisme Ou En Faisant L'apologie Et Des Sites Diffusant Des Images Et Représentations De Mineurs À Caractère Pornographique*,

European Parliament and Council of the European Union. *Directive 2014/95/Eu*, O.J. (L 330).

Everett V Comojo (Uk) Ltd (T/a Metropolitan), EWCA Civ 13 (2011).

Bundestag. *Gesetz Zur Verbesserung Der Rechtsdurchsetzung in Sozialen Netzwerken [Act to Improve Enforcement of the Law in Social Networks]*, BGBl. I S. 3352.

Parliament of the United Kingdom. *Human Rights Act 1998*, c. 42 (Eng. and Wales).

International Covenant on Civil and Political Rights,

International Covenant on Economic, Social and Cultural Rights,

Jd V East Berkshire Community Health Nhs Trust and Others 2 WLR 993 (2005).

Knight First Amendment Institute V. Donald J. Trump, 928 F.3d 226 (2nd Cir. 2019).

Lamont V. Postmaster General, 381 U.S. 301 (1965).

Marsh V. Alabama, 326 U.S. 501 (1946).

Parliament of the United Kingdom. *Modern Slavery Act, 2015*, c. 30 (UK).

Parliament of the United Kingdom. *Occupiers Liability Act, 1957*, c. 31 (Eng. and Wales).

Packingham V. North Carolina, 137 S. Ct. 1730 (2017).

European Parliament and Council of the European Union. *Regulation (Eu) 2016/679*, O.J. (L 119).

Reno V. Aclu, 521 U.S. 844 (1997).

Right to Organise and Collective Bargaining Convention, 1949, International Labour Organisation,

Robinson V. Hunt City, Tx, 921 F.3d 440 (5th Cir. 2019).

Uejf Et Licra V. Yahoo! Inc. Et Yahoo France Tribunal de Grande Instance de Paris [TGI] [High Court of Paris] RG 05308 May 22, 2000.

Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework (Hr/Pub/11/04),

The Universal Declaration of Human Rights,

Yahoo Inc. V. Licra, 145 F. Supp. 2d 1168 (N.D. Cal. 2001).

Yahoo! Inc. V. La Ligue Contre Le Racisme Et L'antisemitisme, Et Al, 433 F.3d 1199 (9th Cir. 2006).

Yl V. Birmingham City Council, 95 1 AC (2008).

Zippo Mfg. Co. V. Zippo Dot Com, Inc., 952 F. Supp. 1119 (W.D. Pa. 1997).

Reference list

7amleh. "7amleh Releases New Racism Index Exposing Heightened Israeli Online Incitement against Palestinians." Last modified March 5, 2018. https://7amleh.org/2018/03/05/7amleh-releases-new-racism-index-exposing-heightened-israeli-online-incitement-against-palestinians/.

Abbott, K. W. "Rule-Making in the Wto: Lessons from the Case of Bribery and Corruption." *Journal of International Economic Law* 4, no. 2 (2001): 275-96. https://doi.org/10.1093/jiel/4.2.275.

Adalah. "Israel's 'Cyber Unit' Operating Illegally to Censor Social Media Content." Last modified September 14, 2017. https://www.adalah.org/en/content/view/9228.

Alexander, Harriet, and Nick Allen. "Youtube Hq Shooting: Father of Dead Female Suspect Warned Police on Day of Attack She 'Hated' Company." Telegraph. Last modified April 4, 2018. https://www.telegraph.co.uk/news/2018/04/03/gunshots-heard-outside-youtube-office-california/.

Ammori, M. "The 'New' New York Times: Free Speech Lawyering in the Age of Google and Twitter." *Harvard Law Review* 127 (2014): 2259-95.

Amnesty International. "Toxic Twitter-a Toxic Place for Women (Report)." Last modified March, 2018. https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/.

Amnesty International, and Business and Human Rights Resource Centre. *Creating a Paradigm Shift: Legal Solutions to Improve Access to Remedy for Corporate Human Rights Abuse.* London: Amnesty International, 2017.

April and Fack. "Kendall and Kylie Young Funny Video." Youtube. Last modified May 5, 2018. https://www.youtube.com/watch?v=kzFmb5wjM38.

Arnold, Denis G. "Corporations and Human Rights Obligations." *Business and Human Rights Journal* 1, no. 2 (2016): 255-75. https://doi.org/10.1017/bhj.2016.19.

Asher-Schapiro, Avi. "Youtube and Facebook Are Removing Evidence of Atrocities, Jeopardizing Cases against War Criminals." Intercept. Last modified November 2, 2017. https://theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya/.

Association for Progressive Communications. *Content Regulation in the Digital Age: Submission to the United Nations Special Rapporteur on the Right to Freedom of Opinion and Expression.* Geneva: Office of the United Nations High Commissioner for Human Rights, 2018.

Baldwin, Robert, and Martin Cave. *Understanding Regulation: Theory, Strategy, and Practice.* Oxford: Oxford University Press, 1999.

Barclay, Scott. *An Appealing Act: Why People Appeal in Civil Cases.* Evanston: North-Western University Press, 1999.

Barnett, Emma, and Ian Hollinshead. "Dark Side of Facebook." Telegraph. Last modified March 2, 2012. http://www.telegraph.co.uk/technology/facebook/9118778/The-dark-side-of-Facebook.html.

Barry, Annie-Marie. "The Uk Modern Slavery Act and Corporate Responsibility: Progress and Challenges." Centre for the Study of Modern Slavery. Accessed October 26, 2019. https://www.stmarys.ac.uk/research/centres/modern-slavery/articles/corporate-responsibility.aspx.

BBC News. "Regulator Ofcom to Have More Powers over Uk Social Media." Last modified February 12, 2020. https://www.bbc.co.uk/news/technology-51446665.

———. "Upskirting Now a Crime after Woman's Campaign." Last modified April 12, 2019. https://www.bbc.com/news/uk-47902522.

Beaumont, Peter. "The Truth About Twitter, Facebook, and the Uprisings in the Arab World." The Guardian. Last modified February 25, 2011. http://www.theguardian.com/world/2011/feb/25/twitter-facebook-uprisings-arab-libya.

Bebbington, Jan, Elizabeth A. Kirk, and Carlos Larrinaga. "The Production of Normativity: A Comparison of Reporting Regimes in Spain and the Uk." *Accounting, Organizations and Society* 37, no. 2 (2012/02/01/ 2012): 78-94. https://doi.org/10.1016/j.aos.2012.01.001.

BEIS. *Consumer Green Paper: Modernising Consumer Markets.* London: HM Government, 2019.

Belli, Luca, and Jamila Venturini. "Private Ordering and the Rise of Terms of Service as Cyber-Regulation." *Internet Policy Review* 5, no. 4 (2016): 1-7. https://doi.org/10.14763/2016.4.441.

Bellis, Maurizia De. "Public Law and Private Regulators in the Global Legal Space." *International Journal of Constitutional Law* 9, no. 2 (2011): 425-48.

Bengani, Priyanjana, Mike Ananny, and Emily J. Bell. *Controlling the Conversation: The Ethics of Social Platforms and Content Moderation.* New York: Columbia University - Tow Centre for Digital Journalism, 2018.

Bennett, Owen. "Building on the Uk White Paper: How to Better Protect Internet Openness and Individuals' Rights in the Fight against Online Harms." Open Policy and Advocacy Blog (Mozilla). Last modified July 2, 2019. https://blog.mozilla.org/netpolicy/2019/07/02/building-on-the-uk-online-harms-white-paper/.

Berger, J. M. *Nazis Vs. Isis on Twitter: A Comparative Study of White Nationalist and Isis Online Social Media Networks.* Washington, DC: George Washington University, 2016.

Bickel, Alexander M. *The Least Dangerous Branch: The Supreme Court at the Bar of Politics.* 2nd ed. New Haven: Yale University Press, 1962.

Bickert, Monika. "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process." Facebook Newsroom. Last modified April 24, 2018. https://about.fb.com/news/2018/04/comprehensive-community-standards/.

Bilchitz, David. "Corporate Obligations and a Treaty on Business and Human Rights: A Constitutional Law Model?". In *Building a Treaty on Business and Human Rights: Context and Contours*, edited by Surya Deva and David Bilchitz, 185-215. Cambridge, UK: Cambridge University Press, 2018.

———. "The Necessity for a Business and Human Rights Treaty." *Business and Human Rights Journal* 1, no. 2 (2016): 203-27. https://doi.org/10.1017/bhj.2016.13.

Bilchitz, David, and Surya Deva. "Human Rights Obligations of Business: A Critical Framework for the Future." In *Human Rights Obligations of Businesses: Beyond the Corporate Responsibility to Respect?*, edited by Surya Deva and David Bilchitz, 1-26. Cambridge, UK: Cambridge University Press, 2013.

Bing, Jon. "Code, Access, and Control." In *Human Rights and the Digital Age*, edited by Mathias Klang and Andrew Murray, 203-19. London: Cavendish Publishing, 2005.

Bingham, Tom H. *The Rule of Law.* London: Penguin, 2010.

Black, Charles L. "The Supreme Court, 1966 Term." *Harvard Law Review* 81, no. 1 (1967): 69-262. https://doi.org/10.2307/1339220.

Black, Julia. "Constitutionalising Self-Regulation." *Modern Law Review* 59, no. 1 (1996): 24-55. https://doi.org/10.1111/j.1468-2230.1996.tb02064.x.

———. "Constructing and Contesting Legitimacy and Accountability in Polycentric Regulatory Regimes." *Regulation and Governance* 2, no. 2 (2008): 137-64. https://doi.org/10.1111/j.1748-5991.2008.00034.x.

Blimes, Jack. *Discourse and Behaviour.* Boston: Springer, 1986.

Board, Facebook Oversight. "Bylaws." 2020. https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf.

Bowden, Tom. "Blacklists Are Not Censorship." Ayn Rand Institute. Last modified March 23, 1999. https://ari.aynrand.org/issues/government-and-business/individual-rights/blacklists-are-not-censorship/.

Bradley, A. W., and K. D. Eving. *Constitutional and Administrative Law.* 15th ed. New York: Pearson Longman, 2011.

Brand South Africa. "Coca-Cola Moves Africa Hq to Jozi." Last modified August 21, 2006. https://www.brandsouthafrica.com/investments-immigration/africa-gateway/coke-210806.

———. "South African Cola Wars Iii." Last modified May 30, 2006. https://www.brandsouthafrica.com/south-africa-fast-facts/media-facts/pepsi.

Brenkert, George G. "Business Ethics and Human Rights: An Overview." *Business and Human Rights Journal* 1, no. 2 (2016): 277-306. https://doi.org/10.1017/bhj.2016.1.

———. "Google, Human Rights, and Moral Compromise." *Journal of Business Ethics* 85, no. 4 (2009): 453-78. https://doi.org/10.1007/s10551-008-9783-3.

Brin, David. *The Transparent Society: Will Technology Force Us to Choose between Privacy and Freedom?* Reading, MA: Perseus Books, 1998.

Brown, Ian, and Christopher T. Marsden. *Regulating Code: Good Governance and Better Regulation in the Information Age.* Information Revolution and Global Politics. Cambridge, MA: MIT Press, 2013.

BSR. *Human Rights Impact Assessment: Facebook in Myanmar.* New York: Business of a Better World, 2018.

———. *Human Rights Review: Facebook Oversight Board.* 2019.

Buhmann, Karin. "Neglecting the Proactive Aspect of Human Rights Due Diligence? A Critical Appraisal of the Eu's Non-Financial Reporting Directive as a Pillar One Avenue for Promoting Pillar Two Action." *Business and Human Rights Journal* 3, no. 1 (2018): 23-45. https://doi.org/10.1017/bhj.2017.24.

Buni, Catherine, and Soraya Chemaly. "Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech." Verge. Last modified April 13, 2016. http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

Business and Human Rights in Law. "Chronology on the Law on the Duty of Vigilance." Accessed April 26, 2019. http://www.bhrinlaw.org/law-duty-of-vigilance-2-versions-en-october-2018.pdf.

Business and Human Rights Resource Centre. "Amazon.Com Stories." Accessed March 5, 2020. https://www.business-humanrights.org/en/amazoncom.

———. "Examples of Government Regulations on Human Rights Reporting & Due Diligence for Companies." Last modified September 30, 2017. https://www.business-humanrights.org/en/examples-of-government-regulations-on-human-rights-reporting-due-diligence-for-companies.

———. "German Development Ministry Drafts Law on Mandatory Human Rights Due Diligence for German Companies." Accessed July 1, 2019. https://www.business-humanrights.org/en/german-development-ministry-drafts-law-on-mandatory-human-rights-due-diligence-for-german-companies.

Cai, Adelin. "Putting Pinners First: How Pinterest Is Building Partnerships for Compassionate Content Moderation." Tech Dirt. Last modified February 5, 2018. https://www.techdirt.com/articles/20180205/10143639158/putting-pinners-first-how-pinterest-is-building-partnerships-compassionate-content-moderation.shtml.

Calo, Ryan. "Robotics and the Lessons of Cyberlaw." *California Law Review* 103, no. 3 (2015): 513-63.

Carrier, Patricia. "2017 Modern Slavery Statement." Modern Slavery Registry. Last modified April 30, 2019. https://www.modernslaveryregistry.org/companies/19013-google-inc.

———. "Facebook's Anti- Slavery and Human Trafficking Statement." Modern Slavery Registry. Last modified April 26, 2018 https://www.modernslaveryregistry.org/companies/18576-facebook-inc/statements/27127.

———. "Twitter Uk Anti-Slavery Statement for the 2017 Financial Year." Modern Slavery Registry. Last modified March 07, 2019. https://www.modernslaveryregistry.org/companies/26502-twitter-uk-limited.

Cassel, Doug. "Outlining the Case for a Common Law Duty of Care of Business to Exercise Human Rights Due Diligence." *Business and Human Rights Journal* 1, no. 2 (2016): 179-202. https://doi.org/10.1017/bhj.2016.15.

Cassidy, Ciaran, and Adrian Chen. *Moderators.* [documentary short] New York: Field of Vision Films, 2017.

Castells, Manuel. *Communication Power.* New York: Oxford University Press, 2009.

CBC. "Facebook Clarifies Breastfeeding Pics Ok, Updates Rules." Last modified March 16, 2015. https://www.cbc.ca/news/world/facebook-clarifies-breastfeeding-pics-ok-updates-rules-1.2997124.

Chapman, Ben. "Samsung Faces Charges over 'Misleading' Ethics Claims after Alleged Labour Abuses in Factories." Independent. Last modified July 4, 2019. https://www.independent.co.uk/news/business/news/samsung-france-legal-case-child-labour-factories-a8988446.html.

Chen, Adrian. "Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where 'Camel Toes' Are More Offensive Than 'Crushed Heads." Gawker. Last modified February 16, 2016. http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads.

———. "The Labourers Who Keep Dick Pics and Beheadings out of Your Facebook Feed." Wired. Last modified October 23, 2014. https://www.wired.com/2014/10/content-moderation/.

Chicago Anti-Apartheid Movement Collection. "Coke Sweetens Apartheid." Accessed July 6, 2019. https://caamcollection.omeka.net/items/show/13.

Cohen, Julie E. *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice.* New Haven: Yale University Press, 2012.

Collin, Keith. "A Running List of Websites and Apps That Have Banned, Blocked, Deleted, and Otherwise Dropped White Supremacists." Quartz. Last modified August 16, 2017. https://qz.com/1055141/what-websites-and-apps-have-banned-neo-nazis-and-white-supremacists/.

Collingridge, David. *The Social Control of Technology.* London: Francis Pinter, 1980.

Committee on Foreign Affairs. *Report on Corporate Liability for Serious Human Rights Abuses in Third Countries: 2015/2315 (Ini).* Brussels: European Parliament, 2016.

Connor, Tracy. "Pepsi Re-Entering South Africa." United Press International. Last modified October 3, 1994. https://www.upi.com/Archives/1994/10/03/Pepsi-re-entering-South-Africa/3971781156800/.

Cossart, Sandra, and Lucie Chatelain. "What Lessons Does France's Duty of Vigilance Law Have for Other National Initiatives?" Business and Human Rights Resource Centre. Last

modified June 27, 2019. https://www.business-humanrights.org/en/what-lessons-does-frances-duty-of-vigilance-law-have-for-other-national-initiatives.

Cottrell, Stephen. "The Internet Must Be Made Safe for Children." Accessed 31st January 2019. https://www.chelmsford.anglican.org/news/article/the-internet-must-be-made-safe-for-children.

Cragg, Wesley. "Ethics, Enlightened Self-Interest, and the Corporate Responsibility to Respect Human Rights: A Critical Look at the Justificatory Foundations of the Un Framework." *Business Ethics Quarterly* 22, no. 1 (2012): 9-36. https://doi.org/10.5840/beq20122213.

———. "Human Rights, Globalisation and the Modern Shareholder Owned Corporation." In *Human Rights and the Moral Responsibilities of Corporate and Public Sector Organisations*, edited by Tom Campbell and Seumas Miller, 105-27. Dordrecht: Springer, 2005.

Crawford, Kate. "Artificial Intelligence's White Guy Problem." New York Times. Last modified June 25, 2016. http://mobile.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?0p19G=c.

Crawford, Kate, and Tarleton Gillespie. "What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media and Society* 18, no. 3 (2014): 410-28. https://doi.org/10.1177/1461444814543163.

Criminal Courts and Criminal Law Policy Unit. *Voyeurism (Offences) Act 2019: Implementation of the Voyeurism (Offences) Act 2019.* London: Ministry of Justice, 2019.

Curtis, Polly. "Government Scraps 192 Quangos." The Guardian. Last modified October 14, 2010. https://www.theguardian.com/politics/2010/oct/14/government-to-reveal-which-quangos-will-be-scrapped.

Curzan, Myron P., and Mark L. Pelesh. "Revitalizing Corporate Democracy: Control of Investment Managers' Voting on Social Responsibility Proxy Issues." *Harvard Law Review* 93, no. 4 (1980): 670-700. https://doi.org/10.2307/1340521.

Danielson, Dan. "How Corporations Govern: Taking Corporate Power Seriously in Transnational Regulation and Governance." *Harvard International Law Journal* 46, no. 2 (2005): 411-25.

David, Donald K. "Business Responsibilities in an Uncertain World." *Harvard Business Review* 27, no. supplement (1949): 1-9.

Davis, Rachel. "Beyond Voluntary: What It Means for States to Play an Active Role in Fostering Business Respect for Human Rights." Shift. Last modified February, 2019. https://www.shiftproject.org/resources/viewpoints/beyond-voluntary-states-active-role-business-respect-human-rights/.

DCMS. *Online Harms White Paper.* London: HM Government, 2019.

———. *Online Harms White Paper: Initial Consultation Response.* London: HM Government, 2020.

de Pressigny, Clementine. "Instagram Deleted Harley Weir's Account over Period Blood." I-d Magazine. Last modified September 7, 2016. https://i-d.vice.com/en_us/article/a3gxx4/instagram-deleted-harley-weirs-account-over-period-blood.

De Schutter, Olivier. "Towards a New Treaty on Business and Human Rights." *Business and Human Rights Journal* 1, no. 1 (2016): 41-67. https://doi.org/10.1017/bhj.2015.5.

Degeling, Martin, and Bettina Berendt. "What Is Wrong About Robocops as Consultants? A Technology-Centric Critique of Predictive Policing." *AI and society* 33, no. 3 (2018): 347-56. https://doi.org/10.1007/s00146-017-0730-7.

Deibert, Ronald J., and Nart Villeneuve. "Firewalls and Power: An Overview of Global State Censorship of the Internet." In *Human Rights and the Digital Age*, edited by Mathias Klang and Andrew Murray, 111-25. London: Cavendish Publishing, 2005.

Dempsey, Bernard. "Roots of Business Responsibility." *Harvard Business Review* 27 (1949): 393-404.

Deva, Surya. "Corporate Human Rights Abuses and International Law: Brief Comments." James G. Stewart Blog. Last modified January 28, 2015. http://jamesgstewart.com/corporate-human-rights-abuses-and-international-law-brief-comments/.

———. "Human Rights Obligations of Business: Reimagining the Treaty Business." Presented at the Workshop on Human Rights and Transnational Corporations: Paving the Way for a Legally Binding Instrument, Geneva, March 11-12 2014.

———. "Treating Human Rights Lightly: A Critique of the Consensus Rhetoric and the Language Employed by the Guiding Principles." In *Human Rights Obligations of Businesses: Beyond the Corporate Responsibility to Respect?*, edited by Surya Deva and David Bilchitz, 78-103. Cambridge, UK: Cambridge University Press, 2013.

Deva, Surya, and David Bilchitz, eds. *Building a Treaty on Business and Human Rights: Context and Contours*. Cambridge, UK: Cambridge University Press, 2018.

"Developments in the Law: State Action and the Public/Private Distinction." *Harvard Law Review* 123, no. 5 (2010): 1248-314.

Dinos, Sokratis, Nina Burrowes, Karen Hammond, and Christina Cunliffe. "A Systematic Review of Juries' Assessment of Rape Victims: Do Rape Myths Impact on Juror Decision-Making?". *International Journal of Law, Crime and Justice* 43, no. 1 (2015): 36-49. https://doi.org/10.1016/j.ijlcj.2014.07.001.

Dowie, Mark. "Pinto Madness." *Mother Jones*, September, 1977, 18-32.

Dworkin, Ronald. *Taking Rights Seriously.* Delhi: Universal Law Publishing 1999.

Economic and Social Council. *General Comment No. 24 (2017) on State Obligations under the International Covenant on Economic, Social and Cultural Rights in the Context of Business Activities (E/C.12/Gc24).* United Nations: Geneva, 2017.

Edmonds, Rhys. "Anxiety, Loneliness and Fear of Missing Out: The Impact of Social Media on Young People's Mental Health." Centre for Mental Health. Accessed October 8, 2019. https://www.centreformentalhealth.org.uk/blog/centre-mental-health-blog/anxiety-loneliness-fear-missing-out-social-media.

Edwards, Jim. "Yes, You Can Make Six Figures as a Youtube Star and Still End up Poor." Business Insider. Last modified February 10, 2014. https://www.businessinsider.com/how-much-money-youtube-stars-actually-make-2014-2/?IR=T.

Elks, Sonia. "The Uk Has Just Introduced a New Law to Protect Women." World Economic Forum. Last modified April 15, 2019. https://www.weforum.org/agenda/2019/04/the-uk-has-just-introduced-a-new-law-to-protect-women.

Equality and Human Rights Commission. *Business Plan 2018-19.* London: EHRC, 2018.

———. "Guidance for Small Businesses and Human Rights." Last modified July 15, 2019. https://www.equalityhumanrights.com/en/advice-and-guidance/guidance-small-businesses-and-human-rights#h3.

———. *Strategic Plan 2012-2015.* London: EHRC, 2012.

Ergon Associates. "Reporting on Modern Slavery: The Current State of Disclosure." Last modified May, 2016. https://ergonassociates.net/wp-content/uploads/2017/06/Reporting-on-Modern-Slavery2-May-2016.pdf.

Esler, Brian W. "Filtering, Blocking and Ratings: Chaperones or Censorship?". In *Human Rights and the Digital Age*, edited by Mathias Klang and Andrew Murray, 99-110. London: Cavendish Publishing, 2005.

Estepa, Jessica. "We're All Atwitter: 3 Times President Trump Made Major Announcements Via Tweets." USA Today. Last modified December 15, 2019. https://eu.usatoday.com/story/news/politics/onpolitics/2018/03/13/were-all-atwitter-3-times-president-trump-made-major-announcements-via-tweets/420085002/.

European Coalition for Corporate Justice. "Evidence of Mandatory Human Rights Due Diligence: Policy Note." Last modified May, 2019. http://corporatejustice.org/news/9189-evidence-for-mandatory-human-rights-due-diligence-legislation-in-europe.

———. "French Corporate Duty of Vigilance Law: Frequently Asked Questions." Last modified March 24, 2017. http://www.respect.international/french-corporate-duty-of-vigilance-law-english-translation/.

———. "Members of 8 European Parliaments Support Duty of Care Legislation for Eu Corporations." Last modified May 31, 2016. http://corporatejustice.org/news/132-members-of-8-european-parliaments-support-duty-of-care-legislation-for-eu-corporations.

Facebook. "Facebook Community Standards: Dangerous Individuals and Organisations." Accessed April 8, 2020. https://www.facebook.com/communitystandards#dangerous-organizations.

———. "Facebook Community Standards: Hate Speech." Accessed April 8, 2020. https://www.facebook.com/communitystandards/hate_speech.

———. "Facebook Site Governance." Accessed October 12, 2018. https://www.facebook.com/fbsitegovernance.

———. "Facebook Site Governance Post on Data Policy." Last modified April 4, 2018. https://www.facebook.com/fbsitegovernance/.

———. "Facebook Transparency Report." Accessed February 20, 2020. https://transparency.facebook.com.

———. "I Want to Appeal Support Decision' Facebook Help Community." Accessed October 24, 2018. https://www.facebook.com/help/community/question/?id=10202808946613203.

———. "Mark Zuckerberg: Vote on Facebook Site Governance." Last modified April 20, 2009. https://web.facebook.com/facebookapp/videos/mark-zuckerberg-vote-on-facebook-site-governance/186119950483/?_rdc=1&_rdr.

———. "Oversight Board Charter." 2020. https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf.

Facebook Newsroom. "Banning More Dangerous Organisations from Facebook in Myanmar." Facebook. Last modified February 5, 2019. https://newsroom.fb.com/news/2019/02/dangerous-organizations-in-myanmar/.

———. "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar." Facebook. Last modified November 5, 2018. https://newsroom.fb.com/news/2018/11/myanmar-hria/.

Farrand, Benjamin, and Helena Carrapico. "Networked Governance and the Regulation of Expression on the Internet: The Blurring of the Role of Public and Private Actors as Content Regulators." *Journal of Information Technology and Politics* 10, no. 4 (2013): 357-68. https://doi.org/10.1080/19331681.2013.843920.

Fasciglione, Marco. "The Enforcement of Corporate Human Rights Due Diligence: From the Un Guiding Principles on Business and Human Rights to the Legal Systems of Eu Countries." *Human rights and international legal discourse* 10 (07/01 2016): 94-117.

Fasterling, Björn. "Human Rights Due Diligence as Risk Management: Social Risk Versus Human Rights Risk." *Business and Human Rights Journal* 2, no. 2 (2017): 225-47. https://doi.org/10.1017/bhj.2016.26.

Flickr. "Flickr Community Guidelines." Accessed February 12, 2018. https://www.flickr.com/help/guidelines.

Forbes. "The World's Largest Companies." Accessed November 9, 2019 https://www.forbes.com/global2000/list/#country:United%20Kingdom.

Fortune. "Facebook Execs Feel the Heat of the Platform's Biggest Content Controversies." Last modified October 26, 2016. http://fortune.com/2016/10/28/facebook-media-content-controversy/.

Foxman, Abraham H., and Christopher Wolf. *Viral Hate: Containing Its Spread on the Internet.* 1st ed. New York: Palgrave Macmillan, 2013.

Friedman, Milton. "The Social Responsibility of Business Is to Increase Its Profits." In *Corporate Ethics and Corporate Governance*, edited by Walther Ch Zimmerli, Markus Holzinger and Klaus Richter, 173-78. Berlin: Springer, 2007.

Froomkin, A. Michael. "The Metaphor Is the Key: Cryptography, the Clipper Chip, and the Constitution." *University of Pennsylvania Law Review* 143 (1995): 709-897.

Fuller, Lon L. *The Morality of Law.* Storrs Lectures on Jurisprudence. New Haven: Yale University Press, 1969.

Gallagher, Ryan. "Google's Secret China Project 'Effectively Ended' after Internal Confrontation." Intercept. Last modified December 17, 2018. https://theintercept.com/2018/12/17/google-china-censored-search-engine-2/.

Galperin, Eva. "What the Facebook and Tumblr Controversies Can Teach Us About Content Moderation." Electronic Frontier Foundation. Last modified March 2, 2012. https://www.eff.org/deeplinks/2012/03/what-facebook-and-tumbler-controversies-can-teach-us-about-content-moderation.

Gearty, Conor. *Terror.* London: Faber and Faber, 1991.

Gerrard, Ysabel. "Beyond the Hashtag: Circumventing Content Moderation on Social Media." *New Media and Society* 20, no. 12 (2018): 4492-511. https://doi.org/10.1177/1461444818776611.

Gibbons, Thomas. *Media Law in the United Kingdom.* 2nd ed. Alphen aan den Rijn: Kluwer Law International, 2014.

Gibbs, Samuel. "Facebook Live: Zuckerberg Adds 3,000 Moderators in Wake of Murders." The Guardian. Last modified May 3, 2017. https://www.theguardian.com/technology/2017/may/03/facebook-live-zuckerberg-adds-3000-moderators-murders.

———. "Facebook under Pressure after Man Livestreams Killing of His Daughter." The Guardian. Last modified April 25, 2017.

https://www.theguardian.com/technology/2017/apr/25/facebook-thailand-man-livestreams-killing-daughter.

Gibson, Peter. "The Report of the Detainee Inquiry." (2013).

Giddens, Anthony. *The Constitution of Society: Outline of the Theory of Structuration.* Berkeley: University of California Press, 1986.

Gilbert, David. "Why Facebook Censored a "Racist" Video from Hungary's Government—Then Put It Back." Vice News. Last modified March 9, 2018. https://news.vice.com/en_us/article/gy87m4/why-facebook-censored-a-racist-video-from-hungarys-government-then-put-it-back.

Gillespie, Alisdair. "'Trust Me, It's Only for Me':'Revenge Porn' and the Criminal Law." *Criminal Law Review* 11 (2015): 866-80.

Gillespie, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* New Haven: Yale University Press, 2018.

Global Network Initiative. "The Gni Principles: Freedom of Expression." Accessed September 22, 2019. https://globalnetworkinitiative.org/gni-principles/.

Götzmann, Nora. "Human Rights Impact Assessment of Business Activities: Key Criteria for Establishing a Meaningful Practice." *Business and Human Rights Journal* 2, no. 1 (2017): 87-108. https://doi.org/10.1017/bhj.2016.24.

Gray, Mary L., and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Boston: Houghton Mifflin Harcourt, 2019.

Gregoire, Carolyn. "Pinterest Removes Eating Disorder-Related Content, Pro-Anorexia Community Continues to Thrive." Huffington Post. Last modified October 8, 2012. http://www.huffingtonpost.co.uk/entry/pinterest-removes-eating-disorder-content_n_1747279.

Grimmelmann, James. "Privacy as Product Safety." *Widener Law Journal* 19 (2010): 792-827.

GSMA, and LIRNEasia. *Mobile Phones, Internet, and Gender in Myanmar.* London: GSMA, 2016.

Guardian. "How Facebook Guides Moderators on Terrorist Content." Last modified May 24, 2017. https://www.theguardian.com/news/gallery/2017/may/24/how-facebook-guides-moderators-on-terrorist-content.

Hale, Robert Lee. *Freedom through Law: Public Control of Private Governing Power.* New York: Columbia University Press, 1952.

Halliday, Josh. "Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party." The Guardian. Last modified March 22, 2012. https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech.

Harris, Brent. "Preparing the Way Forward for Facebook's Oversight Board." Facebook Newsroom. Last modified January 28, 2020. https://about.fb.com/news/2020/01/facebooks-oversight-board/.

Hart, H.L.A. *The Concept of Law.* 3rd ed. Oxford: Oxford University Press, 2012.

Herman, Jenn. "How Much Time Should It Take to Create an Instagram Post?" Jenn's Trends in Social Media Management. Last modified January 27, 2016. https://www.jennstrends.com/how-much-time-should-it-take-to-create-an-instagram-post/.

Hern, Alex. "Facebook among 30 Organisations in Uk Political Data Inquiry." The Guardian. Last modified April 5, 2018. https://www.theguardian.com/technology/2018/apr/05/facebook-mark-zuckerberg-refuses-to-step-down-or-fire-staff-over-mistakes.

———. "Facebook's Changing Standards from Beheading to Breastfeeding Images." The Guardian. Last modified October 22, 2013. https://www.theguardian.com/technology/2013/oct/22/facebook-standards-beheading-breastfeeding-social-networkin.

Hinde, Natasha. "Musician Slams Facebook for Removing Post About Period Pain Labelling It a 'Disgusting Hit of Oppression." Huffington Post. Last modified July 19, 2017. https://www.huffingtonpost.co.uk/entry/melody-pool-facebook-status-about-period-pain-removed-from-facebook_uk_578dec2fe4b0885619b11978.

Hoff, Anneloes. "Dutch Child Labour Due Diligence Law: A Step Towards Mandatory Human Rights Due Diligence." Oxford Human Rights Hub. Last modified June 10, 2019. https://ohrh.law.ox.ac.uk/dutch-child-labour-due-diligence-law-a-step-towards-mandatory-human-rights-due-diligence/.

Hogan, Libby, and Michael Safi. "Revealed: Facebook Hate Speech Exploded in Myanmar During Rohingya Crisis." The Guardian. Last modified April 3, 2018.

https://www.theguardian.com/world/2018/apr/03/revealed-facebook-hate-speech-exploded-in-myanmar-during-rohingya-crisis.

Home Office. *Independent Review of the Modern Slavery Act 2015: Final Report* London: Government of the United Kingdom, 2019.

———. *Proscribed Terrorist Organisations.* London: Government of the United Kingdom, 2019.

Hopkins, Nick. "Facebook Struggles with 'Mission Impossible' to Stop Online Extremism." The Guardian. Last modified May 24, 2017. https://www.theguardian.com/news/2017/may/24/facebook-struggles-with-mission-impossible-to-stop-online-extremism.

———. "How Facebook Allows Users to Post Footage of Children Being Bullied." The Guardian. Last modified May 22, 2017. https://www.theguardian.com/news/2017/may/22/how-facebook-allows-users-to-post-footage-of-children-being-bullied.

———. "Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence." The Guardian. Last modified May 24, 2017. https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence.

House of Lords, House of Commons, and Joint Committee on Human Rights. *Human Rights and Business 2017: Promoting Responsibility and Ensuring Accountability (Sixth Report of Session 2016–17).* London: House of Lords, 2017.

Hsieh, Nien-hê. "Should Business Have Human Rights Obligations?". *Journal of Human Rights* 14, no. 2 (2015): 218-36. https://doi.org/10.1080/14754835.2015.1007223.

Human Rights Council. *Detailed Findings of the Independent International Fact-Finding Mission on Myanmar (a/Hrc/42/Crp.5).* Geneva: United Nations, 2019.

———. *Report of the Independent International Fact-Finding Mission on Myanmar (a/Hrc/39/64).* Geneva: United Nations, 2018.

Human Rights Watch. "Germany: Flawed Social Media Law." Last modified February 14, 2018. https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

Hunter, Dan. "Cyberspace as Place and the Tragedy of the Digital Anticommons." *California Law Review* 91, no. 2 (2003): 439-519. https://doi.org/10.2307/3481336.

ICANN. "Uniform Domain Name Dispute Resolution Policy." Last modified October 24, 1999.

Impress. "Guidance on the Impress Standards Code." 2020. https://www.impress.press/downloads/file/impress-code-guidance-2020.pdf.

Index on Censorship. "Wider Definition of Harm Can Be Manipulated to Restrict Media Freedom." Last modified February 18, 2019. https://www.indexoncensorship.org/2019/02/wider-definition-of-harm-can-be-manipulated-to-restrict-media-freedom/.

Information Commissioner's Office. "Data Protection Self-Assessment." Accessed June 12, 2019. https://ico.org.uk/for-organisations/data-protection-self-assessment/.

———. "Penalties." Accessed June 12, 2019. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-law-enforcement-processing/penalties/.

Innis, Harold. *Empire and Communications.* Oxford: Clarendon Press, 1950.

Innis, Harold A. *The Bias of Communication.* Toronto: University of Toronto Press, 1951.

Instagram Help Centre. "What Can I Do If My Account Has Been Disabled?" Instagram. Accessed October 24, 2018. https://help.instagram.com/366993040048856?helpref=search&sr=2&query=appeal.

Institute for Human Rights and Business, and Shift. *Ict Sector Guide on Implementing the Un Guiding Principles on Business and Human Rights.* Brussels: European Commission, 2013.

Interministerial Committee on Business and Human Rights. *National Action Plan: Implementation of the Un Guiding Principles on Business and Human Rights (2016–2020).* Berlin: Federal Foreign Office, 2017.

International Court of Justice. *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (the Gambia V. Myanmar) - Provisional Measures. Order of January 23, 2020.* The Hague: ICJ, 2020.

IPSO. "Complaints-Frequently Asked Questions." 2018. https://www.ipso.co.uk/faqs/complaints/#what-are-some-of-the-ways-that-my-complaint-might-be-resolved.

Irshaid, Faisal. "How Isis Is Spreading Its Message Online." BBC News. Last modified June 19, 2014. https://www.bbc.co.uk/news/world-middle-east-27912569.

Jackson, Benjamin F. "Censorship and Freedom of Expression in the Age of Facebook." *New Mexico Law Review* 44 (2014): 121-67.

Johnson, Bobbie. "Privacy No Longer a Social Norm, Says Facebook Founder." The Guardian. Last modified January 11, 2010. https://www.theguardian.com/technology/2010/jan/11/facebook-privacy.

Johnston, Casey. "Whopping 0.038% of Facebook Users Vote on Data Use Policy Change." ArsTechnica. Last modified October 18, 2018. https://arstechnica.com/information-technology/2012/06/whopping-00038-of-facebook-users-vote-on-data-use-policy-change/.

Joint Committee on Human Rights. *Equality and Human Rights Commission: Thirteenth Report of Session 2009–10* London: Stationery Office, 2010.

Jørgensen, Rikke Frank. "Framing Human Rights: Exploring Storytelling within Internet Companies." *Information, Communication and Society* 21, no. 3 (2018): 340-55. https://doi.org/10.1080/1369118X.2017.1289233.

Kahan, Dan M., D.avid Hoffman, Danieli Evans, Neal Devins, Eugene Lucci, and Katherine Cheng. "Ideology' or 'Situation Sense'? An Experimental Investigation of Motivated Reasoning and Professional Judgment." *University of Pennsylvania Law Review* 164 (2016): 349-439.

Kahneman, Daniel. *Thinking, Fast and Slow.* 1st ed. New York: Farrar, Straus and Giroux, 2011.

Kain, Erik. "Youtube Wants Content Creators to Appeal Demonetisation, but It's Not Always That Easy." Forbes. Last modified September 17, 2017. https://www.forbes.com/sites/erikkain/2017/09/18/adpocalypse-2017-heres-what-you-need-to-know-about-youtubes-demonetisation-troubles/#1614ffd6c267n.

Kaletsky, Anatole. "Coca-Cola 38 Per Cent up to $934m." *Financial Times*, February 20 1987.

Kaplan, Joel, and Justin Osofsky. "Input from Community and Partners on Our Community Standards." Facebook Newsroom. Last modified October 21, 2016. https://newsroom.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards/.

Kaye, David. "Keynote Speech." Presented at the Workshop on human rights and new technologies, University of Connecticut School of Law, Hartford, CT, October 23 2015.

———. *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (a/Hrc/38/35)*. Geneva: United Nations, 2018.

Keats Citron, Danielle. *Hate Crimes in Cyberspace.* Cambridge, MA: Harvard University Press, 2014.

———. "Technological Due Process." *Washington University Law Review* 85, no. 6 (2008): 1249-313.

Kelly, Erin. "Eight Shocking Deaths That Happened as Tv Cameras Were Rolling." All That's Interesting. Last modified October 17, 2017. https://allthatsinteresting.com/live-deaths-tv.

Kemp, Simon. "Digital in 2019: Global Internet Use Accelerates." We Are Social. Last modified January 31, 2019. https://wearesocial.com/uk/blog/2019/01/digital-in-2019-global-internet-use-accelerates.

Kerr, Orin. "The Problem of Perspective in Internet Law." *Georgetown Law Journal* 91 (2003): 357-405.

Kinley, David, and Rachel Chambers. "The Un Human Rights Norms for Corporations: The Private Implications of Public International Law." *Human Rights Law Review* 6, no. 3 (2006): 447-97. https://doi.org/10.1093/hrlr/ngl020.

Klonick, Kate. "New Governors: The People, Rules, and Processes Governing Online Speech." *Harvard Law Review* 131 (2017): 1598-670.

Kobrin, Stephen J. "Private Political Authority and Public Responsibility: Transnational Politics, Transnational Firms, and Human Rights." *Business Ethics Quarterly* 19, no. 3 (2009): 349-74. https://doi.org/10.5840/beq200919321.

Kohl, Uta. *Jurisdiction and the Internet: A Study of Regulatory Competence over Online Activity.* Cambridge, UK: Cambridge University Press, 2007.

Koops, Bert-Jaap. "Criteria for Normative Technology: An Essay on the Acceptability of 'Code as Law' in Light of Democratic and Constitutional Values." In *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes*, edited by Roger Brownsword and Karen Yeung, 157-74. Oxford: Hart, 2008.

———. "Should Ict Regulation Be Technology-Neutral." In *Starting Points for Ict Regulation, Deconstructing Prevalent Policy One-Liners*, edited by Bert-Jaap Koops, Miriam Lips, Corien Prins and Maurice Schellekens, 77-108. The Hague: TMC Asser Press, 2006.

Kpop Focus. "Sexy Kid-Another Troublemaker." YouTube. Last modified May 5, 2017. https://www.youtube.com/watch?v=vV4r5PV4I2c.

Kranzberg, Melvin. "Technology and History: 'Kranzberg's Laws'." *Technology and Culture* 27, no. 3 (1986): 544-60. https://doi.org/10.2307/3105385.

La Rue, Frank. *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue (a/Hrc/17/27).* Geneva: United Nations, 2011.

Laidlaw, Emily B. *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility.* Cambridge, UK: Cambridge University Press, 2015.

Langbein, John H., Renée Lettow Lerner, and Bruce P. Smith. *History of the Common Law: The Development of Anglo-American Legal Institutions.* New York: Aspen Publishers, 2009.

Langvardt, Kyle. "Regulating Online Content Moderation." *Georgetown Law Journal* 106, no. 5 (2018): 1353-88.

Legal Services Board. *Regulatory Sanctions and Appeals Processes: An Assessment of the Current Arrangements.* London: LSB, 2014.

Leta Jones, Meg. "Does Technology Drive Law: The Dilemma of Technological Exceptionalism in Cyberlaw." *University of Illinois Journal of Law, Technology and Policy* 2 (2018): 249-84. https://doi.org/10.2139/ssrn.2981855.

Levin, Sam. "Youtube's Small Creators Pay Price of Policy Changes after Logan Paul Scandal." The Guardian. Last modified September 18, 2017. https://www.theguardian.com/technology/2018/jan/18/youtube-creators-vloggers-ads-logan-paul.

Lindsay, Rae, George Kleinfeld, Jacqueline Landells, Wendy Wysong, and Antony Crockett. "Briefing Note on Us State Department Guidance on Reporting Requirements for Responsible Investment in Myanmar." Clifford Chance. Last modified October, 2013. https://www.cliffordchance.com/content/dam/cliffordchance/briefings/2013/10/us-state-department-guidance-on-reporting-requirements-for-responsible-investment-in-myanmar-october-2013.pdf.

Lomas, Natasha. "Facebook's Content Moderation Rules Dubbed 'Alarming' by Child Safety Charity." TechCrunch. Last modified May 22, 2017.

https://techcrunch.com/2017/05/22/facebooks-content-moderation-rules-dubbed-alarming-by-child-safety-charity/.

Louis, J. C., and Harvey Z. Yazijian. *The Cola Wars.* 1st ed. New York: Everest House, 1980.

Luscombe, Richard. "Amazon's Jeff Bezos Pledges $10bn to Save Earth's Environment." The Guardian. Last modified February 17, 2020. https://www.theguardian.com/technology/2020/feb/17/amazon-jeff-bezos-pledge-10bn-fight-climate-crisis.

Lynch, Marc. "After Egypt: The Limits and Promise of Online Challenges to the Authoritarian Arab State." *Perspectives on Politics* 9, no. 2 (2011): 301-10. https://doi.org/10.1017/S1537592711000910.

Mac Síthigh, Daithí. "Datafin to Virgin Killer: Self-Regulation and Public Law." *Norwich Law School Working Papers Series* 09/02 (2009): 1-22.

———. "The Mass Age of Internet Law." *Information and Communications Technology Law* 17, no. 2 (2008): 79-94. https://doi.org/10.1080/13600830802204187.

———. "The Road to Responsibilities: New Attitudes Towards Internet Intermediaries." *Information and Communications Technology Law* 29, no. 1 (2020): 1-21. https://doi.org/10.1080/13600834.2020.1677369.

———. "Virtual Walls? The Law of Pseudo-Public Spaces." *International Journal of Law in Context* 8, no. 3 (2012): 394-412. https://doi.org/10.1017/S1744552312000262.

Mac Síthigh, Daithí. *Medium Law.* Routledge Studies in Law, Society and Popular Culture. Abingdon, Oxon: Routledge, 2017.

MacCormick, Neil. *Rhetoric and the Rule of Law: A Theory of Legal Reasoning.* Oxford: Oxford University Press, 2005.

Macdonald, Stuart, and David Mair. "Terrorism Online: A New Strategic Environment." In *Terrorism Online: Politics, Law, and Technology*, edited by Lee Jarvis, Stuart Macdonald and Thomas M. Chen, 1-25. Padstow, UK: Routledge, 2015.

Mansell, Robin. "Introduction and Equity in Cyberspace." In *Human Rights and the Digital Age*, edited by Mathias Klang and Andrew Murray, 1-10. London: Cavendish Publishing, 2005.

Marchant, Gary Elvin, Braden R. Allenby, and Joseph R. Herkert. *Growing Gap between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem.* International Library of Ethics, Law and Technology,. Dordrecht: Springer, 2011.

Marshall, Peter. "A Comparative Analysis of the Right to Appeal." *Duke Journal of Comparative and International Law* 22, no. 1 (2011): 1-45.

McGlynn, Clare. "We Need a New Law to Combat 'Upskirting' and 'Downblousing' " Inherently Human. Last modified April 15, 2015. https://inherentlyhuman.wordpress.com/2015/04/15/we-need-a-new-law-to-combat-upskirting-and-downblousing/.

McGlynn, Clare, and Erika Rackley. "Not 'Revenge Porn', but Abuse: Let's Call It Image-Based Sexual Abuse." Inherently Human. Last modified February 15, 2016. https://inherentlyhuman.wordpress.com/2016/02/15/not-revenge-porn-but-abuse-lets-call-it-image-based-sexual-abuse/.

McGlynn, Clare, Erika Rackley, and Ruth Houghton. "Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse." *Feminist Legal Studies* 25, no. 1 (2017): 25-46. https://doi.org/10.1007/s10691-017-9343-2.

Mehta, Ivan. "A New Study Says Nearly 96% of Deepfake Videos Are Porn." The Next Web. Last modified October 7, 2019. https://thenextweb.com/apps/2019/10/07/a-new-study-says-nearly-96-of-deepfake-videos-are-porn/.

Metropolitan Police News. "250,000th Piece of Online Extremist/Terrorist Material to Be Removed." Last modified January 9, 2017. http://news.met.police.uk/news/250000th-piece-of-online-extremist-slash-terrorist-material-to-be-removed-208698.

Mifsud Bonnici, J. P., and C. N. J. de vey Mestdagh. "Right Vision, Wrong Expectations: The European Union and Self-Regulation of Harmful Internet Content." *Information and Communications Technology Law* 14, no. 2 (2005): 133-49. https://doi.org/10.1080/13600830500042665.

Miller, Catherine, Jacob Ohrvik-Stott, and Rachel Coldicutt. *Regulating for Responsible Technology: Capacity, Evidence and Redress: A New System for a Fairer Future.* London: Doteveryone, 2018.

Miller, Claire Cain. "When Algorithms Discriminate." New York Times. Last modified July 9, 2015. https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html.

Ministry of Justice. "Press Release: 'Upskirting' Law Comes into Force." Last modified April 12, 2019. https://www.gov.uk/government/news/upskirting-law-comes-into-force.

Morozov, Evgeny. *The Net Delusion: The Dark Side of Internet Freedom.* 1st ed. London: Penguin, 2012.

———. *To Save Everything, Click Here: The Folly of Technological Solutionism.* 1st ed. New York: PublicAffairs, 2013.

Morrison, John, and David Vermijs. *The State of Play of Human Rights Due Diligence: Anticipating the Next Five Years.* London: Institute for Human Rights and Business, 2011.

Muchlinski, Peter. "Implementing the New Un Corporate Human Rights Framework: Implications for Corporate Law, Governance, and Regulation." *Business Ethics Quarterly* 22, no. 1 (2012): 145-77. https://doi.org/10.5840/beq20122218.

———. *Multinational Enterprises and the Law.* Oxford International Law Library. 2nd ed. Oxford: Oxford University Press, 2007.

Murphy, Sean D. "Taking Multinational Corporate Codes of Conduct to the Next Level." *Columbia Journal of Transnational Law* 43, no. 2 (2005): 389-433.

National Society for the Prevention of Cruelty to Children. *Taming the Wild West Web: How to Regulate Social Networks and Keep Children Safe from Abuse.* London: NSPCC, 2019.

Newland, Erica, Caroline Nolan, Cynthia Wong, and Jillian York. *Account Deactivation and Content Removal: Guiding Principles and Practices for Companies and Users.* Cambridge, MA: Berkman Centre for Internet and Society, 2011.

Newton, Casey. "Bodies in Seats: At Facebook's Worst-Performing Content Moderation Site in North America, One Contractor Has Died, and Others Say They Fear for Their Lives." Verge. Last modified June 19, 2019. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa.

Nielsen, Vibeke Lehmann, and Christine Parker. "Testing Responsive Regulation in Regulatory Enforcement." *Regulation and Governance* 3, no. 4 (2009): 376-99. https://doi.org/10.1111/j.1748-5991.2009.01064.x.

Nissenbaum, Helen. "From Preemption to Circumvention: If Technology Regulates, Why Do We Need Regulation (and Vice Versa)?". *Berkeley Technology Law Journal* 26, no. 3 (2011): 1367-86.

O'Neill, Patrick. "Facebook's Efforts 'Not Nearly Sufficient in Genocide-Torn Myanmar,' Un Investigator Says." Gizmodo. Last modified March 4, 2019. https://gizmodo.com/facebooks-efforts-not-nearly-sufficient-in-genocide-tor-1833719999.

Obara, Louise J. "'What Does This Mean?': How Uk Companies Make Sense of Human Rights." *Business and Human Rights Journal* 2, no. 2 (2017): 249-73. https://doi.org/10.1017/bhj.2017.7.

Ogus, Anthony. "Rethinking Self-Regulation." *Oxford Journal of Legal Studies* 15, no. 1 (1995): 97-108. https://doi.org/10.1093/ojls/15.1.97.

Ohrvik-Stott, Jacob, and Catherine Miller. *Digital Duty of Care: Doteveryone's Perspective.* London: Doteveryone, 2019.

Okoniewski, Elissa. "Yahoo!, Inc. V. Licra: The French Challenge to Free Expression on the Internet." *American University International Law Review* 13 (2002): 295-339.

Online censorship. "How to Appeal." Accessed October 24, 2018. https://onlinecensorship.org/resources/how-to-appeal.

———. "What We Do." Accessed October 18, 2018. https://onlinecensorship.org/about/what-we-do.

Only Fans. "Only Fans Terms of Service." Last modified August 21, 2019. https://onlyfans.com/terms.

Open Rights Group Wiki. "Perrin and Woods Duty of Care." Last modified January 22, 2019. https://wiki.openrightsgroup.org/wiki/Perrin_and_Woods_Duty_of_Care.

Osborne, Samuel. "Youtube Disables Comments on Videos Featuring Children after Paedophile Ring Scandal." Independent. Last modified March 1, 2019. https://www.independent.co.uk/life-style/gadgets-and-tech/news/youtube-comments-disable-children-videos-paedophile-ring-a8802316.html.

Paiement, Phillip. "Paradox and Legitimacy in Transnational Legal Pluralism." *Transnational Legal Theory* 4, no. 2 (2013): 197-226. https://doi.org/10.5235/20414005.4.2.197.

Paine, Lynn Sharp. *Value Shift: Why Companies Must Merge Social and Financial Imperatives to Achieve Superior Performance.* 1st ed. New York: McGraw-Hill, 2003.

Palombo, Dalia. *Business and Human Rights: The Obligations of the European Home States.* London: Hart, 2020.

———. "The Duty of Care of the Parent Company: A Comparison between French Law, Uk Precedents and the Swiss Proposals." *Business and Human Rights Journal* 4, no. 2 (2019): 265-86. https://doi.org/10.1017/bhj.2019.15.

———. "The Future of the Corporation: The Avenues for Legal Change." *Future of the Corporation Working Paper* (2019): 1-62.

Pariser, Eli. The Filter Bubble: What the Internet Is Hiding from You. London: Penguin, 2011.

Pasquale, Frank. *Black Box Society: The Secret Algorithms That Control Money and Information.* Cambridge, MA: Harvard University Press, 2015.

Pendergast, Mark. *For God, Country, & Coca-Cola: The Definitive History of the Great American Soft Drink and the Company That Makes It.* New York: Basic Books, 2000.

Pepitone, Julianne. "How Diverse Is Silicon Valley?" CNN Money. Accessed June 8, 2018. https://money.cnn.com/interactive/technology/tech-diversity-data/.

Perrin, William, and Lorna Woods. *Online Harm Reduction: A Statutory Duty of Care and Regulator.* Dunfermline: Carnegie UK Trust, 2019.

———. "Reducing Harm in Social Media through a Duty of Care." Carnegie UK Trust. Last modified May 8, 2018. https://www.carnegieuktrust.org.uk/blog/reducing-harm-social-media-duty-care/.

———. "Whose Duty Is It Anyway? Answering Some Common Questions About a Duty of Care." Carnegie UK Trust. Last modified August 2, 2019. https://www.carnegieuktrust.org.uk/blog/duty-of-care-faq/.

Pfaffenberger, Bryan. "Technological Dramas." *Science, Technology, and Human Values* 17, no. 3 (1992): 282-312. https://doi.org/10.1177/016224399201700302.

Pinterest. "Terms of Service." Accessed April 8, 2020. https://policy.pinterest.com/en/terms-of-service.

Plank, Elizabeth. "This Photo Was Banned by Instagram – Thanks to Society's Sexist Double-Standards." Mic. Last modified January 20, 2015. https://mic.com/articles/108624/this-banned-instagram-photo-exposes-the-latest-double-standard-in-censorship#.cOQiZv3Dw.

Pliny the Elder. *Natural History.* Edited by John Bostock and H. T. Riley. Medford, MA: Trustees of Tufts University, 2004.

Plüss, Jessica Davis, and Andrea Tognina. "Responsible Business Initiative Heads Closer to a National Vote." Swissinfo. Last modified March 12, 2019. https://www.swissinfo.ch/eng/corporate-responsibility_responsible-business-initiative-heads-closer-to-a-national-vote/44818824.

Postman, Neil. "Five Things We Need to Know About Technological Change." University of California, Davis. Last modified March 28, 1998. https://web.cs.ucdavis.edu/~rogaway/classes/188/materials/postman.pdf.

———. "Reformed English Curriculum." In *High School 1980: The Shape of the Future in American Secondary Education*, edited by Alvin Christian Eurich, 160–68. London: Pitman, 1970.

———. *Teaching as a Conserving Activity.* New York: Delacorte Press, 1979.

———. *Technopoly: The Surrender of Culture to Technology.* 1st ed. New York: Vintage Books, 1992.

Przybylski, Andrew K., Kou Murayama, Cody R. DeHaan, and Valerie Gladwell. "Motivational, Emotional, and Behavioral Correlates of Fear of Missing Out." *Computers in Human Behavior* 29, no. 4 (2013): 1841-48. https://doi.org/10.1016/j.chb.2013.02.014.

Quora. "How Long Do Youtubers Take to Edit Their Videos?" Last modified July 19, 2018. https://www.quora.com/How-long-do-YouTubers-take-to-edit-their-videos.

Ramasastry, Anita. "Corporate Social Responsibility Versus Business and Human Rights: Bridging the Gap between Responsibility and Accountability." *Journal of Human Rights* 14, no. 2 (2015): 237-59. https://doi.org/10.1080/14754835.2015.1037953.

Raz, Joseph. *The Morality of Freedom.* Oxford: Clarendon Press, 1988.

Reed, Chris. "How to Make Bad Law: Lessons from Cyberspace." *Modern Law Review* 73, no. 6 (2010): 903-32. https://doi.org/10.1111/j.1468-2230.2010.00824.x.

———. *Making Laws for Cyberspace.* 1st ed. Oxford: Oxford University Press, 2012.

Reed, Chris, and Andrew Murray. *Rethinking the Jurisprudence of Cyberspace.* Rethinking Law. Cheltenham, UK: Edward Elgar, 2018.

Renaud, Juliette, Françoise Quairel, Sabine Gagnier, Aymeric Elluin, Swann Bommier, Camille Burlet, and Nayla Ajaltouni. *The Law on Duty of Vigilance of Parent and Outsourcing Companies: Year 1: Companies Must Do Better.* Montreuil: ActionAid, 2019.

Resnik, Judith. "Precluding Appeals." *Cornell Law Review* 70, no. 1 (1985): 603-24.

Responsible Business. "Shadow Eu Action Plan on Business and Human Rights." Last modified March 19, 2019. https://responsiblebusinessconduct.eu/wp/2019/03/19/shadow-eu-action-plan-on-business-and-human-rights/.

Reuters. "Facebook and Youtube Use Automation to Remove Extremist Videos, Sources Say." The Guardian. Last modified June 25, 2016. https://www.theguardian.com/technology/2016/jun/25/extremist-videos-isis-youtube-facebook-automated-removal.

———. "Youtube Serves up 100 Million Videos a Day Online." USA Today. Last modified July 16, 2006. https://usatoday30.usatoday.com/tech/news/2006-07-16-youtube-views_x.htm.

Roberts, Sarah. *Behind the Screen: Content Moderation in the Shadows of Social Media.* New Haven: Yale University Press, 2019.

———. "Commercial Content Moderation: Digital Labourers' Dirty Work." In *Intersectional Internet: Race, Sex, Class and Culture Online*, edited by Safiya Umoja Noble and Brendesha M. Tynes, 147–59. New York: Peter Lang, 2015.

———. "Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media's Waste." *Wi: Journal of Mobile Media* 10, no. 1 (2016): 1-18.

———. "Social Media's Silent Filter." Atlantic. Last modified March 8, 2017. https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/.

Robertson, Cassandra. "The Right to Appeal." *North Carolina Law Review* 91 (2013): 1220-81.

Robinson, Anthony. "Inside, Doubt Takes Root; Disinvestment from South Africa." *Financial Times*, June 16 1987.

Rosen, Jeffrey. "Delete Squad." New Republic. Last modified April 29, 2013. http://www.newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules.

Rossman, David. "'Were There No Appeal': The History of Review in American Criminal Courts." *Journal of Criminal Law and Criminology* 81, no. 3 (1990): 518-66. https://doi.org/10.2307/1143847.

Ruggie, John Gerard. "Business and Human Rights: The Evolving International Agenda." *American Journal of International Law* 101, no. 4 (2007): 819-40. https://doi.org/10.1017/S0002930000037738.

———. "Global Governance and 'New Governance Theory': Lessons from Business and Human Rights." *Global Governance* 20, no. 1 (2014): 5-17. https://doi.org/10.1163/19426720-02001002.

———. "Letter to Ms. Saskia Wilks and Mr. Johannes Blankenbach (Business and Human Rights Resource Centre)." Business and Human Rights Resource Centre. Last modified September 19, 2019. https://www.business-humanrights.org/sites/default/files/documents/19092019_Letter_John_Ruggie.pdf.

———. *Protect, Respect and Remedy: A Framework for Business and Human Rights.* Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises (a/Hrc/8/5). Geneva: United Nations, 2008.

———. "Quo Vadis? Unsolicited Advice to Business and Human Rights Treaty Sponsors." Institute for Human Rights and Business. Last modified September 9, 2014. https://www.ihrb.org/other/treaty-on-business-human-rights/quo-vadis-unsolicited-advice-to-business-and-human-rights-treaty-sponsors.

Samuels, Gabriel. "Thailand Threatens to Sue Facebook over Anti-Monarchy Posts." Independent. Last modified May 12, 2017. https://www.independent.co.uk/news/world/asia/thailand-facebook-anti-monarchy-posts-lawsuit-sue-military-government-king-maha-a7731846.html.

Sanghani, Radhika. "Instagram Deletes Woman's Period Photos – but Her Response Is Amazing." Telegraph. Last modified March 30, 2015. https://www.telegraph.co.uk/women/life/instagram-deletes-womans-period-photos-but-her-response-is-amazing/.

"The Santa Clara Principles: On Transparency and Accountability in Content Moderation." Last modified 07 May 2018, 2018. https://santaclaraprinciples.org.

Santoro, Michael A. "Post-Westphalia and Its Discontents: Business, Globalization, and Human Rights in Political and Moral Perspective." *Business Ethics Quarterly* 20, no. 2 (2010): 285-97. https://doi.org/10.5840/beq201020221.

Sassen, Saskia. *Losing Control? Sovereignty in an Age of Globalization.* New York: Columbia University Press, 1996.

Schrage, Elliot. *Promoting International Worker Rights through Private Voluntary Initiatives: Public Relations or Public Policy.* Iowa City: University of Iowa Centre for Human Rights 2004.

Schulte, Lauren. "Facebook & Google Block Period Language, Ok Video of Man Shooting Himself in the Face." Medium. Last modified July 27, 2016. https://medium.com/the-fixx/facebook-google-block-period-language-ok-video-of-man-shooting-himself-in-the-face-ac9c8c2e50d8.

Schumacher, Elizabeth. "Facebook Refuses to Censor Holocaust Denial as Social Media Sites Struggle with German Laws." Deutsche Welle. Last modified July 27, 2018. https://www.dw.com/en/facebook-refuses-to-censor-holocaust-denial-as-social-media-sites-struggle-with-german-laws/a-44855519.

Secretary of State for Foreign and Commonwealth Affairs. *Good Business: Implementing the Un Guiding Principles on Business and Human Rights.* London: HM Government, 2013.

Select Committee on Communications. "The Internet: To Regulate or Not to Regulate. Corrected Transcript of Evidence Given by Elizabeth Denham, Information Commissioner, Information Commissioner's Office (Ico)." Houses of Lords. Last modified September 11, 2018. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/oral/89766.html.

———. "The Internet: To Regulate or Not to Regulate. Uncorrected Transcript of Evidence Given by Hugh Milward, Director of Corporate, Legal and External Affairs, Microsoft; Katie O'donovan, Public Policy Manager, Uk, Google; Rebecca Stimson, Head of Public Policy, Uk, Facebook." Houses of Lords. Last modified October 30, 2018. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/communications-committee/the-internet-to-regulate-or-not-to-regulate/oral/92263.html.

———. *Regulating in a Digital World (Second Report of Session 2017–2019. Hl Paper 299).* London: House of Lords, 2019.

Shelton, Dinah. "Normative Hierarchy in International Law." *American Journal of International Law* 100, no. 2 (2006): 291-323. https://doi.org/10.1017/S0002930000016675.

Sherman III, John F. "Un Guiding Principles: Practical Implications for Business Lawyers." *In-House Defence Quarterly*, December 21 2013, 50-57.

Sherpa, and CCFD-Terre Solidaire. "Ngos Launch a New Tool to Track Companies Subject to the French Duty of Vigilance Law." European Coalition for Corporate Justice. Last modified July 1, 2019. http://corporatejustice.org/news/16294-ngos-launch-a-new-tool-to-track-companies-subject-to-the-french-duty-of-vigilance-law.

Sherpa, CCFD-Terre Solidaire, and Business and Human Rights Resource Centre. "Duty of Vigilance Radar." Vigilance Plan. Accessed October 22, 2019. https://vigilance-plan.org.

Shift. *Evidence of Corporate Disclosure Relevant to the Un Guiding Principles on Business and Human Rights.* New York: Shift project, 2014.

Shift, and Mazars LLP. *Un Guiding Principles Reporting Framework with Implementation Guidance.* New York: UN Guiding Principles Reporting Database, 2015.

Smartt, Ursula. *Media and Entertainment Law.* 2nd ed. Abingdon, Oxon: Routledge, 2014.

Smith, Graham. "A Lord Chamberlain for the Internet? Thanks but No Thanks." Inforrm's Blog: International Forum for Responsible Media Blog. Last modified October 21, 2018. https://inforrm.org/2018/10/10/a-lord-chamberlain-for-the-internet-thanks-but-no-thanks-graham-smith/.

———. "Rule of Law and the Online Harms White Paper." Cyberleagle. Last modified May 5, 2019. https://www.cyberleagle.com/2019/05/the-rule-of-law-and-online-harms-white.html.

———. "Take Care with That Social Media Duty of Care." Cyberleagle. Last modified October 19, 2018. https://www.cyberleagle.com/2018/10/take-care-with-that-social-media-duty.html.

———. "A Ten Point Rule of Law Test for a Social Media Duty of Care." Cyberleagle. Last modified March 16, 2019. https://www.cyberleagle.com/2019/03/a-ten-point-rule-of-law-test-for-social.html.

———. "Users Behaving Badly – the Online Harms White Paper." Cyberleagle. Last modified April 18, 2019. https://www.cyberleagle.com/2019/04/users-behaving-badly-online-harms-white.html.

Smith, Pamela G. "Free Speech on the World Wide Web: A Comparison between French and United States Policy with a Focus on Uejf V. Yahoo." *Penn State International Law Review* 21, no. 2 (2003): 319-41.

Solon, Olivia. "Facebook, Twitter, Google and Microsoft Team up to Tackle Extremist Content." The Guardian. Last modified December 6, 2016. https://www.theguardian.com/technology/2016/dec/05/facebook-twitter-google-microsoft-terrorist-extremist-content.

———. "To Censor or Sanction Extreme Content? Either Way, Facebook Can't Win." The Guardian Last modified May 23, 2017. https://www.theguardian.com/news/2017/may/24/facebook-struggles-with-mission-impossible-to-stop-online-extremismv.

Sorrell, Tom. "Business and Human Rights." In *Human Rights and the Moral Responsibilities of Corporate and Public Sector Organisations*, edited by Tom Campbell and Seumas Miller, 129-43. Dordrecht: Springer, 2005.

Spivey, John Kirby. "Coke Vs. Pepsi: The Cola Wars in South Africa During the Anti-Apartheid Era." Master of Arts (MA) Master Thesis, Georgia State University, 2009.

Stein, Laura. "Policy and Participation on Social Media: The Cases of Youtube, Facebook, and Wikipedia." *Communication, Culture and Critique* 6, no. 3 (2013): 353-71. https://doi.org/10.1111/cccr.12026.

Stephens, Mitchell. "History of Television." Grolier Encyclopedia, 2000. https://www.nyu.edu/classes/stephens/History%20of%20Television%20page.htm.

Stern, Jessica, and J. M. Berger. *Isis: The State of Terror.* Glasgow: William Collins Publishers, 2015.

Sterngold, James. "After a Suicide, Questions on Lurid Tv News." New York Times. Last modified May 2, 1998. https://www.nytimes.com/1998/05/02/us/after-a-suicide-questions-on-lurid-tv-news.html.

Stone, Biz. "The Zen of Twitter Support." Twitter Blog. Last modified January 15, 2009. https://blog.twitter.com/2009/the-zen-of-twitter-support.

Sunstein, Cass. *Republic.Com 2.0.* Princeton: Princeton University Press, 2007.

Surette, Raymond. "Performance Crime and Justice." *Current Issues in Criminal Justice* 27, no. 2 (2015): 195-216. https://doi.org/10.1080/10345329.2015.12036041.

Talbot, David, and Nikki Bourassa. "How Facebook Tries to Regulate Postings Made by Two Billion People." Medium. Last modified October 19, 2017. https://medium.com/berkman-klein-center/how-facebook-tries-to-regulate-postings-made-by-two-billion-people-bca9408b6b4b.

Tambini, Damian, Danilo Leonardi, and Christopher T. Marsden. *Codifying Cyberspace: Communications Self-Regulation in the Age of Internet Convergence.* London: Routledge, 2008.

Thomas, Jean. *Public Rights, Private Relations.* 1st ed. Oxford: Oxford University Press, 2015.

Thoreau, Henry David. *Walden.* London: Penguin Classics, 2016 [1854].

"Tiktok Community Guidelines." Last modified 21 August 2019. https://www.tiktok.com/community-guidelines?lang=en.

Time photo. "The Story Behind the 'Napalm Girl' Photo Censored on Facebook." Time. Last modified September 9, 2016. http://time.com/4485344/napalm-girl-war-photo-facebook/.

TLC. "Pretty Woman Toddler: Toddlers & Tiaras." YouTube. Last modified September 8, 2011. https://www.youtube.com/watch?v=QAxEt5YL8w4.

Travis, Alan. "Face-Off between Mps and Social Media Giants over Online Hate Speech." The Guardian. Last modified March 14, 2017. https://www.theguardian.com/media/2017/mar/14/face-off-mps-and-social-media-giants-online-hate-speech-facebook-twitter.

Tryhorn, Chris. "Evangelical Networker Who Wants Facebook to Open up the World." The Guardian. Last modified August 20, 2009. https://www.theguardian.com/business/2009/aug/20/facebook-ceo-sheryl-sandberg-interview.

Tutt, Andrew. "An Fda for Algorithms." *Administrative Law Review* 69, no. 1 (2017): 83-123. https://doi.org/10.2139/ssrn.3293577.

Twitter. "Twitter Rules: Glorifying Self-Harm and Suicide." Accessed April 12, 2020. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

———. "The Twitter Rules: Hateful Conduct Policy." Accessed February 1, 2017. https://support.twitter.com/articles/18311.

Twitter Help Centre. "Appeal an Account Suspension or Locked Account." Twitter. Accessed October 24, 2018. https://help.twitter.com/forms/general?subtopic=suspended.

———. "Help with Locked or Limited Account." Twitter. Accessed October 24, 2018. https://help.twitter.com/en/managing-your-account/locked-and-limited-accounts.

———. "Our Range of Enforcement Options." Twitter. Accessed October 24, 2018. https://help.twitter.com/en/rules-and-policies/enforcement-options.

Ullman, Ilana, Laura Reed, and Rebecca MacKinnon. *Submission to Un Special Rapporteur for Freedom of Expression and Opinion David Kaye: Content Regulation in the Digital Age.* Amsterdam: Ranking Digital Rights, 2017.

UN Human Rights Council. "Legally Binding Instrument to Regulate, in International Human Rights Law, the Activities of Transnational Corporations and Other Business Enterprises. Zero Draft Bill." Office of the United Nations High Commissioner for Human Rights. Last modified July 16, 2018. https://www.ohchr.org/Documents/HRBodies/HRCouncil/WGTransCorp/Session3/DraftLBI.pdf.

United Kingdom Interdepartmental Liaison Group on Risk Assessment. "The Precautionary Principle: Policy and Application." Health and Safety Executive. Last modified July 17, 2018. http://www.hse.gov.uk/aboutus/meetings/committees/ilgra/pppa.htm.

United Nations Economic and Social Council. *Interim Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises (E/Cn.4/2006/97).* Geneva: United Nations, 2006.

United Nations General Assembly. *Report of the Working Group on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises (a/73/163).* Geneva: United Nations, 2018.

United Nations Human Rights Council. *Business and Human Rights: Mapping International Standards of Responsibility and Accountability for Corporate Acts. Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises (a/Hrc/4/035).* Geneva: United Nations, 2007.

———. *Report of the Independent International Fact-Finding Mission on Myanmar (a/Hrc/42/50)*. Geneva: United Nations, 2019.

United Nations News. "Un Experts Report: Business 'Dragging Its Feet' on Human Rights Worldwide." Last modified October 16, 2018. https://news.un.org/en/story/2018/10/1023312.

US Securities and Exchange Commission. *Conflict Minerals: Final Rule (Release No. 34-67716; File No. S7-40-10)*. Washington, DC: SEC, 2012.

Valenti, Jessica. "Social Media Is Protecting Men from Periods, Breast Milk and Body Hair." The Guardian. Last modified March 30, 2015. https://www.theguardian.com/commentisfree/2015/mar/30/social-media-protecting-men-periods-breast-milk-body-hair.

van Dijck, José. *The Culture of Connectivity: A Critical History of Social Media.* Oxford: Oxford University Press, 2013.

Vincent, James. "I Can Haz Islamic State Plz: Isis Propaganda on Twitter Turns to Kittens and Lolspeak." Independent. Last modified August 21, 2014. https://www.independent.co.uk/life-style/gadgets-and-tech/isis-propaganda-on-twitter-turns-to-kittens-and-lolspeak-i-can-haz-islamic-state-plz-9683736.html.

Wagner, Ben. "Governing Internet Expression: How Public and Private Regulation Shape Expression Governance." *Journal of Information Technology and Politics* 10, no. 4 (2013): 389-403. https://doi.org/10.1080/19331681.2013.799051.

Waldow, Florian. "Conceptions of Justice in the Examination Systems of England, Germany, and Sweden: A Look at Safeguards of Fair Procedure and Possibilities of Appeal." *Comparative Education Review* 58, no. 2 (2014): 322-43. https://doi.org/10.1086/674781.

Waldron, Jeremy. "Rights in Conflict." *Ethics* 99, no. 3 (1989): 503-19.

———. "Rule of Law and the Importance of Procedure." *Nomos* 50 (2011): 3-31.

Wauters, Ellen, Eva Lievens, and Peggy Valcke. "Towards a Better Protection of Social Media Users: A Legal Perspective on the Terms of Use of Social Networking Sites." *International Journal of Law and Information Technology* 22, no. 3 (2014): 254-94. https://doi.org/10.1093/ijlit/eau002.

Weimer, David L. "Puzzle of Private Rulemaking: Expertise, Flexibility, and Blame Avoidance in Us Regulation." *Public Administration Review* 66, no. 4 (2006): 569-82. https://doi.org/10.1111/j.1540-6210.2006.00617.x.

Wettstein, Florian. *Multinational Corporations and Global Justice: Human Rights Obligations of a Quasi-Governmental Institution.* Stanford, CA: Stanford Business Books, 2009.

Whitehead, Alfred North. *Science and the Modern World.* Lowell Lectures. New York: Simon and Schuster, 1970 [1925].

Wilks, Saskia, and Johannes Blankenbach. "Will Germany Become a Leader in the Drive for Corporate Due Diligence on Human Rights?" Business and Human Rights Resource Centre. Last modified February 20, 2019. https://www.business-humanrights.org/en/will-germany-become-a-leader-in-the-drive-for-corporate-due-diligence-on-human-rights.

Williams, James. *Stand out of Our Light: Freedom and Resistance in the Attention Economy.* Cambridge, UK: Cambridge University Press, 2018.

Winner, Langdon. *The Whale and the Reactor: A Search for Limits in an Age of High Technology.* Chicago: University of Chicago Press, 1986.

Wojcicki, Susan. "Expanding Our Work against Abuse of Our Platform." YouTube Official Blog. Last modified December 4, 2017. https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html.

Wolf, Christopher. "Standards for Internet Jurisdiction: An Overview." Find Law. Last modified March 3, 2008. http://corporate.findlaw.com/litigation-disputes/standards-for-internet-jurisdiction.html.

Wong, Cynthia M. "Social Media's Moral Reckoning: Changing the Terms of Engagement with Silicon Valley." Human Rights Watch. Accessed October 29, 2019. https://www.hrw.org/world-report/2019/essay/social-medias-moral-reckoning.

Wu, Paulina. "Impossible to Regulate: Social Media, Terrorists, and the Role for the Un." *Chicago Journal of International Law* 16, no. 1 (2015): 281-311.

York, Jillian C. "Policing Content in the Quasi-Public Sphere." Open Net Initiative. Last modified September, 2010. https://opennet.net/policing-content-quasi-public-sphere.

Youmans, William Lafi, and Jillian C. York. "Social Media and the Activist Toolkit: User Agreements, Corporate Interests, and the Information Infrastructure of Modern Social

Movements." *Journal of Communication* 62, no. 2 (2012): 315-29. https://doi.org/10.1111/j.1460-2466.2012.01636.x.

YouTube. "Policies and Safety." Accessed April 9 2020. https://www.youtube.com/about/policies/#community-guidelines.

———. "Youtube Community Guidelines." Accessed February 12, 2018. https://www.youtube.com/yt/about/policies/#community-guidelines.

YouTube Policy Centre. "Child Safety on Youtube." YouTube. Accessed April 8, 2020. https://support.google.com/youtube/answer/2801999?hl=en.

YT Talk. "How Long Does It Take to Make a Youtube Video?" Last modified August 10, 2012. http://yttalk.com/threads/how-long-does-it-take-you-to-make-a-youtube-video.11340/.

Zack, Jason S. "The Ultimate Company Town: Wading in the Digital Marsh of Second Life." *University of Pennsylvania journal of Constitutional Law* 10 (2007): 225-55.

Zuckerberg, Mark. "Building Global Community." Facebook. Last modified February 18, 2017. https://m.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634.