Regulating terrorist content on tech platforms: A proposed framework based on social regulation

Amy-Louise Watkin

Submitted to Swansea University in fulfilment of the requirements for the Degree of Doctor of Philosophy in Criminology

Swansea University

2021

Summary (Abstract)

Scholars have been arguing for years that responses to terrorist content on tech platforms have, to-date, been inadequate. Past responses have been reactive and fragmented with tech platforms self-regulating. Over the last few years, many governments began to decide that the self-regulatory approach was not working. As a result, a number of regulatory frameworks have been proposed and/or implemented. However, they have been highly criticised. The purpose of this thesis is to propose a new regulatory framework to counter terrorist content on tech platforms and overcome many of these criticisms. Scholars have argued that it is vital that future regulation be informed by past experience and supported by evidence from prior research. Therefore, a number of steps were taken.

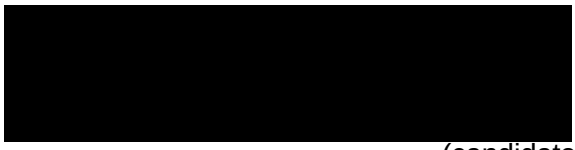
First, this thesis examines a review of literature into what platforms are exploited by terrorist organisations. Next, a content analysis was undertaken on blogposts that tech platforms publish in order to investigate the efforts that tech platforms report making to counter terrorist content on their services and the challenges that they face. Third, a sample of existing or currently proposed regulatory frameworks were examined in order to learn what was done well and what gaps, limitations and challenges exist that require addressing in future regulation. Finally, social regulation theory was identified as applicable in this regulatory context. Social regulation strategies were examined in three other regulatory contexts in order to examine whether they could be used in this regulatory context.

The findings from the above analyses were used to inform a new regulatory framework that is proposed in this thesis. In addition to proposing a new regulatory framework, this thesis also identified three compliance issues that tech platforms may face. These compliance issues are addressed alongside the proposal of the framework. Overall, it is argued that previous regulatory attempts failed to consider the diverse array of challenges that are faced by different platforms when countering terrorist content. The regulatory framework proposed in this thesis researched these challenges and identified strategies from a social regulation approach, learning lessons from how they were applied elsewhere to overcome some of the key criticisms and limitations of existing regulatory practice.

Declarations and Statements

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed 
..... (candidate)

Date ...16/04/2021.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed  .....
..... (candidate)

Date .....16/04/2021.....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed 

.....

Date .....16/04/2021.....

NB: *Candidates on whose behalf a bar on access has been approved by the University (see Note 7), should use the following version of Statement 2:*

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans **after expiry of a bar on access approved by the Swansea University.**

Signed ..... (candidate)

Date .....

Table of Contents

Acknowledgements …….. 6

List of tables and figures …... …… 7

Abbreviations …… 8

Chapter 1: Introduction ……. 10

Chapter 2: Terrorist exploitation of tech platforms …….. 24

Chapter 3: What has been done to counter online terrorist content to-date: Tech platform responses …… 50

Chapter 4: What has been done to counter online terrorist content to-date: Government responses …… 89

Chapter 5: The Regulatory Approach: Social Regulation Theory ……. 126

Chapter 6: Regulatory Framework: Objectives and Ethos ……. 161

Chapter 7: Regulatory Framework: Mandatory Regulatory Standards ……. 194

Chapter 8: Regulatory Framework: Four Regulatory Tracks …… 225

Chapter 9: Conclusion …… 259

Glossary ……. 274

Bibliography …….. 276

Acknowledgements

It is a common misconception that a PhD is a solo journey. It takes a village to get someone through a PhD, and I have been blessed with the most wonderfully loving village.

I would first like to express the utmost gratitude to each of my supervisors. I would like to thank Stuart for many years of patience and kindness. I will always appreciate the many hours that were spent reading my chapters and providing feedback. You have always been on hand with the best advice and reassurance. I could not have asked for a better supervisor and mentor. I would like to thank Lella for all the emotional support and advice you have given me. When I needed a supervisor, you were a supervisor, when I needed a friend, you were a friend. I would like to thank Patrick; I know that being my supervisor was not the original plan and so I appreciate every chapter that you took the time to read and provide feedback on.

I would like to thank Katy; you became both a friend and an unofficial mentor. Thank you for always being there to answer many random questions and especially for all the help with learning how to cite legislation and case law. I never want to cite another case in my life! Thank you to David for being my Scottish friend and showing me the ropes in Swansea. Thank you to all those in the PGR community in Swansea, especially Nia and Joe for listening to all the PhD-related dramas over the years. Thank you to Rachel for making sure my Swansea journey has been so memorable and fun, it feels like a second home. Thank you to my best friend Natalie for absolutely everything.

Finally, a very special thank you to my family. To my parents, Liz and Nick, who witnessed every panic and tear along the way, you are my rock. Thank you for everything, I love you. Thank you to my grandad, Reggie, who passed away just before my PhD began, you were my favourite person, biggest supporter, and provider of cake. I would not have gotten through my PhD without cake.

List of Tables and Figures

List of Tables

Table 1 – Platform size and income …… 62

Table 2 – Number of blogposts collected for each platform …… 63-64

Table 3 – Platform mission and values …… 65-67

Table 4 – Key terms, groups, movements, ideologies, individuals, and attacks that the platforms address in their blogposts …… 68-70

Table 5 – Regulatory Tracks …….. 242-244

Table 6 – Assortment of compliance issues into the relevant regulatory track …… 244

List of Figures

Figure 1 – Order in which compliance issues should be addressed …… 229

Figure 2 – Proposed enforcement pyramid …… 240

Abbreviations

Al Qaeda	AQ
Best practicable environmental option	BPEO
Civil Society Organisation	CSO
Counter Terrorism Internet Referral Unit	CT IRU
Digital Industry Group Inc	DIGI
European Convention on Human Rights	ECHR
Europe, the Middle East and Africa	EMEA
Environmental Protection Agency	EPA
European Union	EU
European Union Internet Referral Unit	EU IRU
European Union Internet Forum	EUIF
General Data Protection Regulation	GDPR
Global Internet Forum to Counter Terrorism	GIFCT
Human Rights Council	HRC
International Government Organisations	IGOs
Internet referral units	IRUs
Islamic State	IS
Internet Service Provider	ISP
Lesbian, gay, bisexual, transgender and queer or questioning	LGBTQ+
Germany's Network Enforcement Act	NetzDG
Non-Governmental Organisation	NGO
Organisation for Economic Co-operation and Development	OECD
Occupational health and safety	OHS
Post-Traumatic Stress Disorder	PTSD

Referral Action Days	RADs
Terrorist Content Analytics Platform	TCAP
The United Nations Counter-Terrorism Committee Executive Directorate	UN CTED
Unreasonable-Harm Prevention Principle	UPP
Uniform Resource Locator	URL
Virtual Private Network	VPN

Chapter 1: Introduction

Context and Motivation

The early 2000s saw hundreds of terrorist websites appearing all over the internet (Weimann, 2004; Conway, 2006). This was during a period of what has been termed the "open internet" (Freedman, 2012). During this period, very little thought was given to the idea of regulating the internet because it was thought of as a separate space where activities were treated differently from the "offline world". The internet was seen as a new space that allowed more speech to be heard than ever before, and provided access to more information than ever before, therefore allowing greater freedoms (Freedman, 2012). Many early internet activists were completely against the idea of internet regulation, for example, John Perry Barlow said, "I declare the global social space we are building to be naturally independent of the tyrannies you [the government] seek to impose on us. You have no moral right to rule us…" (Barlow, 1996).

However, the internet has changed dramatically since this time and so has its use by terrorist organisations. Tech platforms began to emerge in the mid-2000s and offered even more opportunities and features for exploitation than the websites and chats forums that came before them. 'Tech platform' is used as an umbrella term in this thesis to cover a number of different types of online platforms, all of which have been found in this thesis to be used by terrorist organisations: social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites. As terrorist organisations began utilizing tech platforms and benefitting from the easy access to mass mainstream audiences, anonymity, fast flow of information, community building tools, content repositories, and use of multi-media services, it became easier for them to spread their messages, attract new supporters and recruits, build communities, and plan, incite, undertake and livestream attacks (Weimann, 2004; Weimann, 2010; Wu, 2015; BBC, 2019). This was not an issue that tech platforms had given much consideration during their early days of development (Conway cited in Sahinkaya, 2019).

Nevertheless, it soon became clear that there are great dangers to an unregulated internet. It also became clear that internet regulation is no easy task (Wu, 1996; Freedman, 2012). Fast forwarding a decade, the years 2014-2015 were termed the 'golden age' for terrorists on tech platforms as the platforms were going largely unregulated (Conway Khawaja, Lakhani, Reffin, Robertson, and Weir, 2017). It was estimated that there were between 46,000 and 90,000 pro-IS Twitter accounts alone between September and December 2014 (Berger and Morgan, 2015). Around this same time, there was an increase in terrorist attacks across Europe (Europol, 2016).

As a result, governments around the world began to increase demands for tech platforms to do more to counter terrorist content on their services. For example, former UK Prime Minister Theresa May (2018) said that tech platforms need "to move further and faster in reducing the time it takes to remove terrorist content online" and that the content should be "removed automatically". However, the appearance of unlawful content on tech platforms created an issue that traditional legal instruments have struggled to solve. The global and borderless nature of the internet has meant that tech platforms are arguably better placed to remove terrorist content than state actors due to the fact that the users posting unlawful content are sometimes not able to be identified by the state or are beyond its jurisdiction (Schulz, 2018; Ardia, 2009). Therefore, many governmental responses first sought to place demands on the tech platforms via a self-regulatory approach (BBC, 2020). Self-regulation relies "substantially on the goodwill and cooperation of individual firms for their compliance" (Sinclair, 1977, p.534) A problem arose, however, that tech platforms did not cooperate with this approach in a manner that governments deemed acceptable (Yar, 2018). As a result, governments began to propose and implement regulatory frameworks where failure to comply would result in punitive actions against the tech platforms (Yar, 2018).

Due to existing frameworks in this area receiving significant criticism, this thesis sought to identify a regulatory approach that has yet to be studied in this context in order to investigate potential new strategies that can be implemented in the regulatory framework that is proposed in this thesis with the aim of overcoming as many criticisms of the existing frameworks as possible.

Two types of regulation feature prominently as alternative approaches throughout the regulatory literature: economic regulation and social regulation. Economic regulation, as its name suggests, deals solely with economic issues, primarily in industries that contain monopolistic tendencies (Ogus, 1994). Examples of economic objectives include pricing, investment and output (Baldwin, Cave, and Lodge, 2010). However, academics have argued that there are a vast range of other issues and values across many regulatory areas that also require consideration that economic regulation fails to consider (Baldwin et al., 2010), which brings us to social regulation. Social regulation is usually a response to one of two types of market failure (information failures and negative externalities) and is concerned with a much broader range of issues than its economic counterpart (Ogus, 1994). Ogus (1994) describes the areas that fall under the responsibility of social regulators as having public interest justifications, Prosser (2010) describes the responsibility as promoting human rights, social

solidarity, and social inclusion, and Wilson (1984) describes it as regulation that aims to promote a general societal good.

Social regulation was chosen for this thesis, first, because of the identification that it is applicable to the online terrorist content regulatory context. Second, because it has been applied to three other regulatory areas (environmental protection, consumer protection, and occupational health and safety) that each contain similar issues to those that are faced in regulating online terrorist content. Therefore, it was thought that there could be much to learn from how social regulation was applied in these three other areas that could be implemented in the regulatory framework that is proposed in this thesis.

Social regulation theory is a good fit for this area because terrorist content on tech platforms is not exclusively an economic issue, although many platforms will have economic interests and concerns, this issue differs from the typical economic issues faced in many other service industries. Online terrorist content affects the public interest, the promotion of human rights, social solidarity, social inclusion and general societal good because it targets vulnerable users for radicalisation and incites violence upon its intended enemies (Weimann, 2010; Putra, 2016; Welch, 2018). Further, as argued by Miller (2019), tech platforms have portrayed their services from the beginning as a social mission as opposed to a commercial enterprise (e.g., Google's "Don't be evil" and Facebook's "Bring the world closer together" mission statements). Moreover, social regulation aims to overcome information failures and negative externalities, both of which are found as market failures in chapter 4's examination of existing frameworks in this area.

Upon research into social regulation theory, and more specifically, its application in environmental protection, consumer protection, and occupational health and safety, four strategies were found to provide valuable lessons and insights that were used to inform and develop the regulatory framework put forth in this thesis with the aim of overcoming many of the criticisms of existing frameworks. These were the precautionary principle, prior approval, information regulation, and the implementation of a health and safety culture. Although these strategies are not all a perfect fit for this regulatory context, there are nevertheless valuable insights and lessons that are critical of consideration if we are to overcome the criticisms of existing frameworks.

Recent years have therefore seen a range of regulatory actions proposed or implemented by governments across the globe and the European Commission. Many of these frameworks,

however, have generated significant concern and criticism, including the effect on free speech and creating unfair burdens on certain platforms that could reduce market competitiveness (März, 2018; Schmitz and Berndt, 2018; Splittgerber and Detmering, 2017; Article 19, 2017; Echikson and Knodt, 2018; Theil, 2019; Tech Against Terrorism, 2020d; Broughton and Jaques, 2019; Bishop, Looney, Macdonald, Pearson, and Whittaker, 2019; Bogle, 2019; Oboler, 2019; Douek, 2019; Schmitz and Berndt, 2018; Alkiviadou, 2019). As a result, this thesis identified the need for a new regulatory framework that addresses the criticisms of existing regulation and is informed by past regulatory experience and research evidence of both this and other regulatory contexts.

This thesis proposes what it is terming as a new 'regulatory framework' to counter terrorist content on tech platforms. However, an explanation of what this means is required. There is the issue that there is often disagreement on definitions of regulation, with Levi-Faur (2011, p.4) noting that "regulation is hard to define, not least because it means different things to different people". Legislation, on the other hand, is easier to define. In the UK, legislation is created as an act of parliament, whereas regulation is made by bureaucracies (Kosti, Levi-Faur, and More, 2019). Legislation and regulation differ in that "while legislation sets out the principles of public policy, regulation implements these principles, bringing legislation into effect" (Kosti et al., 2019). Regulation can be in traditional command-and-control format; however, it can also use non-legislative instruments such as standards, guidance, and licensing (Kosti et al., 2019). Regulation is used for what might be called 'double edged sword' activities where the aim is to attenuate the negative aspects of an activity whilst preserving its positive aspects. For example, the regulatory framework proposed in this thesis aims to prevent terrorist exploitation of tech platforms while maintaining the many benefits that tech platforms provide more generally.

While hard law refers to legally binding obligations that are precise and delegate authority for its interpretation and implementation, often enforced by a court, soft law, is described as being weaker along one of these three dimensions (obligation, precision, or delegation) (Abott and Snidal, 2000). For example, under soft law, there may be an agreement that is not formally binding, it may be less precise than its hard law counterpart and provide discretion regarding implementation, or, it may not delegate an authority to enforce it. A soft law example that is discussed later in the thesis is the voluntary Code of Conduct on Countering Illegal Hate Speech, whereby, it is reliant on tech platforms volunteering compliance and there are not any repercussions on the platform if they fail to comply.

While sometimes hard law is more appropriate, and at other times soft law is more appropriate, on some occasions, there has been what Shaffer and Pollack (2010, p.717) have called a "complex hybrid of hard- and soft-law instruments". Sometimes, non-binding law is the beginning of the path to binding hard law. While hard law instruments are able to have direct legal effects in national jurisdictions, soft law instruments can provide greater flexibility and help develop common norms (Shaffer and Pollock, 2010).

The regulatory framework proposed in this thesis is not enforced by a court, as is typical of hard law. It has a softer legal nature, providing the flexibility it requires given how rapidly tech platforms and terrorists both evolve; however, it is not voluntary like the code of conduct that is later discussed in this thesis. This thesis adopts Black's (2001a) and (2002) definition of regulation as the intentional and sustained use of an authoritative party to attempt to change the behaviour of others which is done by setting specific, defined standards through the use of information-gathering and behaviour control. The framework is formed by 12 mandatory regulatory standards. These standards are instruments that subject the suppliers of the goods or services being regulated to specific behavioural controls in order to create a particular outcome, with those who fail to comply risking punishment (Ogus, 1994). The framework also outlines the objectives and ethos that underpin it, propose that it is monitored and enforced by an independent regulatory body, and propose that it includes a further four regulatory tracks that aim to address and overcome potential compliance issues that are identified throughout the research undertaken in this thesis.

The framework seeks to address the issue of terrorist content being accessible on tech platforms. However, it also seeks to address several key criticisms of existing frameworks in this area, most notably, the concern of over-blocking and infringing free speech, and placing unfair burdens on smaller platforms (or any platform that lacks the capacity or expertise to comply). The thesis identified that platforms can be categorised on their levels of willingness, capacity, and awareness to comply with regulation. This framework seeks to ensure that platforms that are willing to comply but struggle with capacity and/or awareness are not penalised for this, but instead, provided support and guidance through an educative approach that increases their ability to comply. It is hoped that such an approach will be received well from these platforms, increasing their abilities to comply with the regulation and counter terrorist content on their services without imposing financial burdens or negatively affecting their growth and innovation. A diverse array of enforcement actions is included and aimed at platforms that are unwilling to comply in order to incentivise compliance.

Although some of the literature throughout this thesis touches on the wider problem of extremism and hate speech (in order to learn about other approaches to regulation), this thesis and regulatory framework is focused on countering online terrorist content only. I began writing this thesis five years ago in 2016 and one of the main challenges with this thesis is that it is a rapidly evolving topic. Terrorism, the internet, and terrorist use of tech platforms have all changed significantly in the last five years. In 2016, terrorist content was mainly found on the major platforms (Facebook, Twitter, YouTube etc.). Further to this, it was mainly jihadist organisations targeting these platforms, for example, the so-called Islamic State (IS) and al-Qaeda (AQ). Much of the jihadist propaganda was easily identifiable due to the use of, for example, IS symbols and logos. However, in recent years, this landscape has increased in complexity. As a result of disruption tactics, there is a diverse ecosystem of tech platforms being targeted by terrorist organisations, including small and micro-platforms that struggle to respond to such content. Further, there has been a significant increase in content from other ideologies, such as far-right content which is not as easily identifiable as unlawful. Terrorist organisations are increasingly adapting to post content that is more nuanced and falling into what is sometimes called a 'grey area', creating contentious free speech issues.

The decision to focus on terrorist content is partly due to the online terrorist content landscape that was in place in 2016 when this PhD began. However, there is also a decision to focus on terrorist content only because expanding the regulatory scope to other types of content, such as extremist content and hate speech would create a much wider and more complex regulatory scope that would require different free speech considerations. It is therefore argued that extremist content and hate speech requires a different regulatory approach to that of countering online terrorist content.

Therefore, the decision to focus on terrorist content is two-fold. First, due to the timing of the beginning of this PhD and the online terrorist content landscape at the time. Second, because countering extremism and hate speech also, requires, it is argued a different approach due to the more nuanced and complex nature of the content and different free speech implications. This would create a much wider and more complex scope for the regulatory framework

Finally, it should be noted that this thesis aimed to propose a thoroughly researched regulatory framework to counter online terrorist content, applying social regulation strategies in order to overcome criticisms of existing frameworks, it did not set out to discuss who decides, and how

they decide, what is terrorist content or not. However, this thesis acknowledges that is these questions are crucial to the success of regulation and should be addressed in future research.

As the regulatory framework that is proposed in this thesis is aimed at the UK, this thesis adopts the UK Terrorism Act 2000's definition of terrorism, which involves the following:

- Terrorist means (serious violence against a person; serious damage to property; endangers a person's life, other than that of the person committing the action; creates a serious risk to the health or safety of the public or a section of the public; or, is designed seriously to interfere with or seriously disrupt an electronic system);
- A terrorist target (designed to influence the government or an international governmental organisation or to intimidate the public or a section of the public); and,
- A terrorist motive (a political, religious, racial or ideological cause).[1]

This framework is proposed as a UK framework for a number of reasons. The first is because it could be argued that a good approach to new regulation would be to test it on a smaller scale, in this case, in the UK, in order to examine any challenges or limitations that may require amending. After this testing phase, amendments could be made and consideration could be given to rolling out the regulation more widely, for example, regionally, across Europe, and perhaps even internationally. However, an issue with this, and the second reason as to why it is aimed solely at the UK at this stage is that such a framework will face challenges regarding geographic reach because of the transnational nature of cyberspace. There is the possibility that people will exploit the features of cyberspace to try and circumvent the regulatory framework and that the regulator will face difficulties carrying out enforcement actions across jurisdictions. It is also difficult to achieve international adoption due to the significant differences that exist between countries and their varying levels of protection of free speech (e.g., America with the First Amendment, and also authoritarian countries such as China) (Aziz, 2015; Bychawska-Siniarska, 2017; Barednt, 2007).

Thesis Statement and Objectives

The aim of this thesis is to propose a new UK regulatory framework for countering terrorist content on tech platforms. This regulatory framework will be informed by existing regulation, past regulatory experience, and supported by evidence from existing research. It will also address key criticisms of existing frameworks. In order to develop the regulatory framework

[1] United Kingdom: Terrorism Act 2000 https://www.legislation.gov.uk/ukpga/2000/11

in this manner, research was required in a number of areas. The first area is deciding which platforms should fall under the scope of the new framework. In order to make an informed decision, the following research questions are asked in chapter 2:

1) Which tech platforms are exploited by terrorist organisations?
2) How do terrorist organisations exploit tech platforms?
3) What considerations do terrorist organisations have when choosing which platform to use?

These questions aim to provide a greater understanding of what is currently termed the 'ecosystem' of platforms that are used by terrorist organisations (Frampton et al. 2017; Fisher et al. 2019, Conway, Khawaja, Lakhani, Reffin, Robertson and Weir, 2019; Macdonald, Grinnell, Kinzel and Lorenzo-Dus, 2019), in order to inform the regulatory framework proposed in this thesis. The next area of research that is required is an analysis of the efforts that tech platforms already report undertaking to counter terrorist content on their services. Chapter 3 therefore asks:

4) In their blogposts, what efforts do tech platforms report taking to counter terrorist content on their services?
5) What challenges do tech platforms face in their efforts to counter terrorist content on their services that could affect their compliance with regulation?

These questions aim to provide a greater understanding of what tech platforms report already doing to counter terrorist content on their services, what challenges they face in these efforts, and how this could affect their ability to comply with regulation. The findings are used to inform the regulatory framework in this thesis and identify potential challenges that platforms may face regarding regulatory compliance. The next area of research that is required is an examination of existing and currently proposed regulatory frameworks. Chapter 4 asks:

6) What has and has not been effective in existing regulatory frameworks that seek to counter online terrorist content?

This question provides a greater understanding as to what has been done well in existing regulation and what gaps, limitations and challenges exist that require addressing in future regulation. The findings informed the development of the new regulatory framework. Finally, this regulatory framework aims to bring a new regulatory approach; an approach that has the potential to help overcome the criticisms of existing frameworks. Chapter 5 therefore

investigates the potential of social regulation theory and its applicability to this regulatory context. Chapter 5 asks:

7) Is social regulation theory applicable to this regulatory context?
8) What is there to be learned from examining social regulation in other regulatory contexts?
9) Could these strategies be applied in this regulatory context?

The findings will reveal if there are strategies that were effective in other regulatory contexts that could be applied to this regulatory framework as a fresh approach to countering terrorist content on tech platforms.

The findings of chapters 2-5 have informed the development of the regulatory framework that is proposed in chapters 6-8. Chapters 6-8 will propose a regulatory framework that includes the frameworks objectives and ethos; mandatory regulatory standards; enforcement actions and solutions that can aid potential compliance issues that have been identified from chapters 2-5.

Originality

This thesis provides several contributions to the growing regulatory field responding to terrorist content on tech platforms. The first contribution is the investigation into whether social regulation theory is applicable to this regulatory context. Given the argument that new regulation should be informed by existing regulation, past experience and supported by evidence from research, this thesis sought to explore a regulatory approach that had not been considered in this regulatory context before and learn valuable lessons from its application elsewhere (Windholz, 2010). This thesis investigates, firstly, the applicability of social regulation theory to the countering online terrorist content context, and secondly, the application of a social regulation approach in three other regulatory contexts. These are environmental protection, consumer protection, and occupational health and safety. There is a wealth of empirically-grounded findings from studies of regulatory efforts in these other contexts, and each of these contexts are found to overlap with the terrorist content regulatory context in some form (e.g., in terms of the types of regulatory challenges that are faced). This research argues that there are many strategies across these three areas that can be applied to this regulatory context to provide a new approach. These strategies were not explored in the existing frameworks that were examined in this thesis despite their potential to overcome some of the key criticisms found in existing regulation. The regulatory framework proposed in this

thesis therefore identified and adopted a new approach to countering terrorist content on tech platforms and overcomes some of the key criticisms of existing regulations.

The next contribution is the regulatory framework itself. It is informed by the findings in chapters 2-5; therefore, it is argued that the framework is informed by existing regulation, past experience and supported by evidence from research. The decision as to which platforms fall under the scope of the framework is justified by the findings in chapter 2. The regulatory framework itself is informed by approaches in existing and currently proposed regulatory frameworks that also seek to counter online terrorist content (or hate speech and extremist content) that appear to be effective based on the findings in chapter 4. It is also informed by social regulation strategies that have been found to be effective in other regulatory contexts in chapter 5. Moreover, this regulatory framework identifies three main compliance issues from the findings of chapters 2-5 that could arise and addresses this through the implementation of four regulatory tracks (chapter 8). Each track provides a solution to a potential problem that the regulator and tech platform can use to try to overcome the issue. This was not seen in any of the existing frameworks that were examined in this thesis. It is argued that the adoption of a social regulation approach could increase compliance with the regulation and minimize the issue that other frameworks had around infringing free speech and placing unfair burdens on certain platforms and reducing market competitiveness.

The final contribution is that the blogposts published by the tech platforms are understudied in the terrorism context and the workings of tech platforms are often thought of as private, somewhat of a mystery (Pasquale, 2015). Although, there are limitations to the dataset (discussed in the methodology section of chapter 3), particularly that, the blogposts only tell us what the platforms want us to know, the blogposts are one of the few official modes of consistent communication that the tech platforms use to explain their policies, decision making, and efforts to counter terrorist content. The blogposts therefore provide insight into how tech platforms wish their efforts in countering terrorist content to be perceived, their decision-making process around these efforts, and the challenges that they face. This thesis argues that tech platforms face three main compliance issues when faced with regulation and that existing regulation does not consider or address these compliance issues adequately. Researching the tech platform blogposts aided in creating a better understanding of the platforms and their challenges, which subsequently aided in identifying the compliance issues that they face. The regulatory framework in this thesis addressed these compliance issues in chapter 8 with

proposed solutions. The existing frameworks that are examined in this thesis arguably do not give enough attention to addressing potential compliance issues.

Thesis structure and methodology

This thesis begins with a literature review in chapter 2 investigating which tech platforms are exploited by terrorist organisations, how terrorist organisations exploit these platforms, and what considerations terrorist organisations have when choosing which platforms to use. This literature review aims to investigate the argument in existing literature that terrorist organisations utilize a whole ecosystem of platforms. The findings are crucial for informing the regulatory framework that is proposed later in the thesis as to which platforms should fall under its scope. Once the aim of the literature review was established, a search strategy for identifying relevant literature was developed (Snyder, 2019). There was a particular focus on studies that were published in the last five years (although this was not an absolute rule in the search strategy), in order to provide an up-to-date review of which platforms are used by terrorists. There was also a focus on tech platforms (e.g., social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites) because the findings are used to develop a regulatory framework that aims to counter terrorist content on tech platforms. It was found during the literature review that although there are studies across a number of ideologies, terrorist organisations, and platforms, there is a disproportionate number of studies on well-known jihadist organisations, particularly, the so-called Islamic State (IS) and platforms that are used by the IS, such as Twitter and Telegram.

Overall, this is a popular area of research with an abundance of studies. However, the literature review could not discuss every study; therefore, the review included the studies that were assessed as most relevant for answering the research questions. The chapter is split into sections that are categorized per platform to provide a clear structure.

The aim of chapter 3 is to investigate the efforts that a sample of platforms (Facebook, Twitter, YouTube/Google, Microsoft, Gab and Telegram) report taking, in blogposts that they publish, to counter terrorist content on their services. The first four platforms were chosen because they have been largely targeted by terrorist organisations, they are at the centre of government attention and demands (Corera, 2017; Hope, 2017), they consistently fall under the scope of regulatory frameworks (as seen in chapter 4), and finally, they are part of well-known

collaborative initiatives such as the Global Internet Forum to Counter Terrorism (GIFCT).[2] Gab and Telegram were chosen to ensure a diverse sample of platforms. Gab is an alternative platform known for its association with the far-right, its unwillingness to remove content, and criticism of regulation (Bennett, 2018). Telegram is an encrypted messaging app that is known for prioritizing security and privacy and was at one time reluctant to regulate but has since increased its efforts (Europol, 2019c). This chapter aims to identify both similarities and differences across the platforms, as well as identify any issues that platforms may face regarding regulatory compliance. The blogposts are one of the main ways that platforms consistently communicate with their users and were collected over a period of four years (1st Jan 2016 – 31st Dec 2019). Tech platforms use these blogposts to update their users on new decisions that the platform makes, new services and tools that they implement, and how they counter bad actors on their sites. A content analysis was applied to the blogposts and several themes emerged. These themes are discussed in relation to potential compliance issues that the platforms may face regarding regulation. The findings informed the development of the regulatory framework that is proposed in this thesis. There is a detailed methodology section outlining the content analysis and limitations in this chapter.

Chapter 4 examined a sample of government responses that sought to counter terrorist content on tech platforms. The aim of this chapter is to examine the responses, which consist of existing and proposed frameworks (NetzDG, Australia Abhorrent Violent Material Act, UK Online Harms White Paper, amongst others), in order to identify their strengths and limitations. Although not all of the frameworks counter terrorist content specifically, some counter extremism and hate speech, there is still much to learn from these frameworks. This chapter also examines scholarly work that researchers have undertaken regarding the frameworks. The responses that are included were chosen because: 1) they were timely and topical at the time of writing; 2) they cover a variety of countries; 3) these countries all adhere to western democratic principles; 4) together they cover a wide variety of regulatory approaches; and 5) they received widespread attention and criticism due to their limitations which are valuable for the development of the regulatory framework that is proposed in this thesis. The examination involved going through each framework and studies that had been written about that framework and drawing out its strengths, weaknesses, gaps, limitations and challenges. This chapter then summarised all of these and provided a list of lessons to be learned that informed the development of the regulatory framework that is proposed in this thesis. The regulatory

[2] https://gifct.org/

framework in this chapter adopts some aspects of other frameworks that appear to have worked well. It also fills gaps that have been missed, identifies compliance issues that other frameworks failed to identify, and addresses these compliance issues with proposed solutions. A limitation of this chapter is that this is not an exhaustive examination of all of the relevant frameworks in the world; there are relevant frameworks (for example, France's Countering Online Hate Law and Australia's Online Safety Bill) that it was not possible to include due to word limits and therefore there is scope for a much broader examination in future research.

Chapter 5 introduces social regulation theory and the argument that it is applicable to the regulatory context of this thesis. There is an abundance of research and regulatory lessons to be learned from the application of social regulation theory to three other regulatory contexts. This chapter examines the use of social regulation strategies in these other contexts which are: 1) environmental protection; 2) consumer protection; and 3) occupational health and safety. A search was undertaken to find relevant literature and studies that examined the use of social regulation strategies undertaken in each of the three regulatory contexts. Upon reviewing this literature, this chapter draws out similarities between these three regulatory contexts and the regulatory context that is addressed in this thesis. This chapter then identifies strategies that were used in the three regulatory contexts that could be applied in the regulatory framework that is proposed in this thesis, and lessons that could be learned, from the three regulatory contexts that were examined. These findings are used to inform and develop the regulatory framework proposed in this thesis. The research undertaken in chapters 4 and 5 are necessary steps in the formation of new regulation because new regulation should be informed by existing regulation and learn from its criticisms and limitations (Windholz, 2010).

The findings of chapters 2-5 were used to inform and develop the regulatory framework that is proposed in chapters 6-8. Chapter 6 introduces the proposed regulatory framework by introducing four objectives that underpin the framework and the framework's ethos. The intention behind the chosen objectives and ethos is to ensure that the many benefits of tech platforms are not lost as a result of the proposed regulation to counter online terrorist content. Chapter 7 introduces twelve mandatory regulatory standards that tech platforms must comply with to counter terrorist content on their services. This chapter proposes which platforms should fall under the scope of the framework based on the findings of chapter 2. These twelve standards are proposed to support the four objectives and ethos of the framework and draw on strategies that were identified as working well in existing frameworks in chapter 4 and from other regulatory contexts where social regulation was applied in chapter 5. The findings of

chapters 2-5 revealed recurring concerns regarding three compliance issues. It became clear during the development of the 12 mandatory regulatory standards that these three compliance issues could be problematic for the effectiveness of the framework proposed in this thesis if not addressed. Chapter 8 therefore proposed four different solutions, termed in this thesis as 'regulatory tracks' that the regulator and tech platforms could use together to address these issues and work towards the goal of achieving full compliance with all of the mandatory regulatory standards. This chapter also proposed an enforcement approach in the case of (unwillingness) non-compliance.

Overall, the aim of this thesis is to propose a new and well-researched regulatory framework to counter terrorist content on tech platforms and try to overcome some of the key criticisms of existing regulation. The hope is that this framework will propose a new social regulation approach, overcome many of the key criticisms that have been identified in existing regulations, and identify and address the main compliance issues that tech platforms could face.

Chapter 2: Terrorist exploitation of tech platforms

Introduction

When access to the internet became available to the masses, scholars such as Rushkoff (1995) and Arquilla (1996) predicted that it would not be long until bad actors exploited such technology and the networks of postmodern society for their own gains. This was true, terrorist organisations have been exploiting the internet for a long time now. Hundreds of websites had been created by terrorists and their supporters by the early 2000s (Weimann, 2004; Conway, 2006). The easy access, anonymity, lack of regulation, mass reach and fast flow of information are some of the reasons why the internet appealed to such organisations (Weimann, 2004). Although terrorist organisations utilise the internet widely, this thesis is focused on terrorist use of tech platforms specifically. 'Tech platforms' is used as an umbrella term throughout research to cover social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites – all of which, will be discussed in this chapter, have been found to be exploited by terrorist organisations. Tech platforms began to emerge in the mid-2000s and offered even more opportunities and features for exploitation than the websites and chats forums that came before them (Weimann, 2010). Such platforms have millions (if not more) users and are used to find like-minded people, create communities and store and share information, files and audio-visual content with others in real time (Weimann, 2010; Wu, 2015). As a result, terrorist organisations can utilise these platforms to reach more potential followers and elevate the possible scale of destruction (Wu, 2015).

Scholars have argued that counter efforts, initiatives, and responses to terrorist exploitation of tech platforms have, to-date, been inadequate (Frampton, Fisher and Prucha, 2017). Frampton et al. (2017) argue that terrorists' use of tech platforms has created a situation in which we have been drawn into fighting them on their terrain and have been responding to their initiatives and innovations, as opposed to terrorist organisations responding to our initiatives and actions, putting us in the weaker position. Up until the last few years, the general response to terrorist exploitation of tech platforms was reactive. The response was also fragmented, it has not been until recently that government, tech platforms and civil society have come together to try to develop cohesive approaches (Frampton, 2017). When governments began calling on tech platforms to do more to counter terrorist exploitations of their sites, the demands were similar to this example from former UK Prime Minister Theresa May (2018) that tech platforms need "to move further and faster in reducing the time it takes to remove terrorist content online' and

that terrorist content should be “removed automatically”. A major concern with such government demands is that they are vague, general and fail to acknowledge the intricacies and complexities of terrorist use of tech platforms (Weirman and Alexander, 2020), as well as the possibility that terrorists may not exploit all platforms similarly.

Research has already established that terrorist organisations, much like any typical tech platform user, implement a multi-platform approach to their use of tech platforms (Frampton et al. 2017; Fisher, Prucha, and Winterbotham, 2019). “While tech platforms may function independently as institutions, from the user’s perspective these services coexist; they are options to choose from, or to use in tandem” (Gillespie et al., 2020, p.5). The definition of a multi-platform approach is that terrorist organisations utilise many different tech platforms to aid them in their cause and activities. Terrorist use of many different tech platforms is sometimes referred to as a whole ecosystem of platforms (Frampton et al. 2017; Fisher et al. 2019, Conway, Khawaja, Lakhani, Reffin, Robertson and Weir, 2019; Macdonald, Grinnell, Kinzel and Lorenzo-Dus, 2019). Further to this, there are a range of different objectives that terrorists seek from such an ecosystem including access to a wide audience, operational security and a place to store propaganda (Frampton et al. 2017).

In order to respond to terrorist content on tech platforms it is important that, first, there is a clear understanding of the above-mentioned ecosystem of tech platforms that terrorist organisations exploit. Further to this, there must be a clear understanding of the objectives that terrorist organisations have for such exploitation and other factors that they consider when choosing platforms.

This raises the following questions:

1) What tech platforms are exploited by terrorist organisations?
2) How do terrorist organisations exploit tech platforms?
3) What considerations do terrorist organisations have when choosing which platform to use?

Without such an understanding, it will be difficult to make an informed decision about which tech platforms should fall under the scope of the regulatory framework that is proposed in this thesis, and whether there are any other considerations that must be included in the development of the regulatory framework.

Terrorist use of tech platforms: A review of existing literature

Scholars have been researching terrorists' use of the internet for almost as long as terrorists have been using the internet. A key question among these scholars has been to ask how exactly terrorist organisations use the internet. In an early study, pre-tech platforms, Weimann (2004) undertook research examining the hundreds of terrorist websites that existed. His analysis concluded that there were eight different, though sometimes overlapping, ways that terrorists use the internet. The first was for psychological warfare, which could include spreading disinformation and threats, and disseminating violent and gory content, all with the aim of creating fear. The second was for publicity and propaganda. The internet removed the barriers of relying on the traditional media to reach their intended audiences with their messages and content. The third way was data mining; the internet is used as a "vast digital library" (p.6) providing free access to more information than was ever available before. The fourth was fundraising. They can collect donations via the internet from the mass network they are able to connect with. The next way was to recruit followers and mobilize them to increase their participation and undertake actions for the cause. A further way was for networking and staying connected with their followers. The final two were to share information with one another, for example, on how to build weapons and to plan and coordinate attacks. Research with similar findings has been undertaken by Cohen (2002), Furnell and Warren (1999), Thomas (2003), Conway (2006), and Macdonald and Mair (2015).

Following the emergence of tech platforms and the subsequent exploitation from terrorist organisations that followed, scholars began to turn their attention to researching terrorist organisations' use of such platforms. Early research undertaken by Weimann (2010) into terrorist use of such platforms reported several ways that terrorists were exploiting them. The first was to find and attract new followers. They did so, for example on platforms such as Facebook and Twitter by disseminating propaganda to users and reaching out with direct messages. They also created groups that users could join, allowing them to build online communities. Facebook was also found to be used to search for information and spy on military personnel. The research also revealed that terrorists used platforms such as YouTube as a repository for their video propaganda to share with followers. Overall, this meant that unlike previously with traditional websites, terrorist organisations were able to become more proactive in their recruitment. Instead of waiting for people to find them, they were able to directly search for and reach out to potential followers. Another early study undertaken by the Department of Homeland Security (2010) focused specifically on jihadist organisations and

supporters on Facebook. The study found that jihadist organisations used Facebook to share operational and tactical information (for example, bomb manuals); to share links to chat forums, other websites and videos on YouTube; as a media outlet for their propaganda; and to search for information on enemy military personnel. This study shows that terrorists began taking a multi-platform approach very early in their use of tech platforms. It was realised that platforms offer different functions and advantages to one another and it made sense to take advantage of this. This research, however, predated the platform censorship that is seen today.

More recent research has developed from identifying the different activities that terrorist organisations undertake on platforms generally, to investigating the purpose for which different types of tech platform are used by terrorist organisations, within the context of the wider ecosystem of platforms. The next sections of this literature review will focus on specific platforms or types of platforms and the purpose that they serve in the ecosystem. This literature also examines the effects of the disruption terrorist organisations have faced in recent years. The aim of the research is to understand not just the terrorist activities taking place on the platform but the functions that the platforms provide in the overarching ecosystem. This will provide an understanding of the differences in exploitation across platforms. The findings of this literature review inform the development of the regulatory framework proposed later in this thesis.

Twitter

Twitter is what is termed as a 'major' platform and allows users to tweet 280-character messages and attach a limited range of files to the users who follow them. Twitter was, back in the 'golden age' of 2014, the chosen platform for many jihadist groups, particularly the so-called Islamic State (IS) (Conway et al. 2019). Therefore, many of the studies in this area focus around Twitter. However, shortly after this, IS began to be heavily targeted by disruption efforts (such as the removal of content and accounts), following government pressure on the major platforms to do so. IS put in great efforts to remain on Twitter but the question asked in a study undertaken by Conway et al's (2019) was whether it was still a worthwhile effort for IS to remain on Twitter? The tweets collected in this study were split into two categories: pro-IS tweets and tweets that were pro-other jihadist groups. The research found that pro-IS accounts were suspended much quicker than the other jihadist accounts. Further, the findings revealed that IS accounts appeared to have adapted to this quick suspension by operating in a 24-hour news cycle. Every 24 hours, IS would create new accounts and disseminate a new

batch of propaganda by tweeting URLs that led their followers to content on other platforms. These accounts are referred to as throwaway accounts because IS knew that these accounts were likely to be suspended. It did not matter that these accounts would soon be suspended because the content was hosted on a mix of other social media and content-hosting platforms. The content included videos, images, short press releases, radio podcasts and other documents.

This 24-hour news cycle strategy allowed IS to provide their followers with daily links to new content (Conway et al. 2019). Twitter was clearly no longer an ideal platform to host or store content, and even when it was, it had a limited range of content formats than, for example, platforms such as Telegram (e.g., PDFs and other large files). However, it was able to be used to signpost a potentially large mainstream audience to large volumes of content elsewhere on a daily basis. When the Twitter accounts or tweets would be removed, this would not affect the content itself, as it is linked to and stored elsewhere. The limitation of this strategy, however, is that the short timeframe likely made it difficult to gain new followers. It would make sense that the followers they had on Twitter were already committed to IS. Any new sympathizers who stumbled across a Twitter account would be lost when the account was suspended unless they were quickly informed as to how they could find future accounts.

Conway et al.'s (2019) research estimated that one in every 2.5 pro-IS tweets contained a URL. Although it was found that the majority of URLs in the dataset linked users elsewhere within Twitter, around 13 percent of the URLs contained URLs that led to platforms and sites external to Twitter. YouTube was the most linked to platform outside of Twitter for both pro-IS and other jihadist accounts, indicating the importance of video in their propaganda campaigns. Overall, pro-IS accounts linked to thirty-nine different platforms as well as IS's own server. The research notes that there were six consistent platforms, termed as the "big 6": justpaste.it, IS's own server, archive.org, Sendvid.com, YouTube and Google drive.

This research reveals that Twitter came with both benefits and limitations, neither of which are static. Twitter, therefore does not appear to meet all of the needs of an organisation such as IS, and even if such a platform did exist, the disruption efforts discussed show the risks that come with putting all of your eggs in one basket. Pre-disruption efforts, Twitter allowed IS to reach a wide audience that was ideal for recruitment; people were likely to accidently stumble across their accounts and content. However, now, in a time where the major platforms are under immense pressure to counter terrorist content, organisations are less likely to maintain a persistent online presence on these sites. Although, it should be noted that Conway et al.'s

(2019) research revealed that some terrorist organisations (e.g., non-IS Jihadist groups) are less likely to be removed than others. Nevertheless, the nature of organisations such as IS is to adapt (discussed further under the Telegram section). This research shows that IS adapted from using Twitter as one of its main content-posting platforms to a place where it can signpost followers to large volumes of content elsewhere in 24-hour news cycles, where it is safer to store it. Although IS are more likely to have users stumble across their accounts on Twitter than other platforms, such as Telegram, Twitter is not as ideal for this as it previously was. A large effort will likely have to be put into ensuring any new followers are able to find them again after removal. Twitter is also not as secure as other platforms, such as Telegram, for communication purposes.

Weirman and Alexander (2020) also undertook research on Twitter. Similar to Conway et el.'s (2019) research, this study found that in the face of disruption, IS quickly tried to adapt. This was again found to happen via the use of URLs in tweets. This research used a mixed-method approach to survey 240,158 URLs that were shared by a sample of English-language pro-IS accounts on Twitter between 2016 and 2017. Within the dataset, approximately 28 percent of the unique tweets contained at least one URL, if not more. As with the previous study, the majority (64 percent) of URLs led users to other content within Twitter. Excluding these internal URLs, the authors analysed the remaining 86,537 external URLs which led users to 4,372 unique domains. The study found that file-sharing sites made up a significant number of the domains. The researchers pointed out that in addition to disruption measures on Twitter, the platform does not allow for file sizes as large as many file-sharing sites, explaining the purpose of the external URLs. YouTube was found to be a popular outlink. Other popular domains were justpaste.it, archive.org, vid.me, soundcloud.com and top4top.net. These platforms and sites have the appeal of supporting a range of file types (e.g., video, audio, image, text etc.). Social media platforms are also outlinked to with Facebook being one of the most popular. This would make sense given that Facebook has an enormous userbase and is popular in the West. Instagram, known for image-sharing, and Telegram were also ranked highly in the dataset. Other platforms and sites included curiouscat.me, tumblr.com. gofundme.com, periscope.tv and ask.fm. This displays the wide variety and diversity of platforms and sites in the ecosystem and reveals the many different specific functions and services that such terrorist organisations require to undertake their activities. The authors point out that,

> "by redirecting one's followers on Twitter to other networking sites that are better equipped to handle specific tasks, a single user can simultaneously leverage the

> efficiency of Twitter and overcome the platform's inherent limitations. When used in conjunction, social network platforms may have higher potential than individual platforms afford in isolation. Policymakers and practitioners should account for this dynamic in the design and implementation of future approaches." (Weirman and Alexander, 2020, p. 250).

An additional finding in Weirman and Alexander's (2020) research was that IS also adapted regarding the content that they could post on Twitter to try to circumvent removal strategies. In the dataset, news sources were the most frequently shared domain type. The news sources that were shared included *Al Jazeera, Reuters, The Guardian, The Washington Post* and the *BBC.* The authors argue that this strategy on a major platform allows IS to exploit the mass media to their advantage, using it to validate their stance to a mass audience. These domains may be more likely than domains leading to pro-IS content to receive clicks from those who stumble across them. Sharing these domains may also more easily circumvent automated removal than explicit IS content or domains leading to pro-IS content.

Weirman and Alexander (2020) conclude from their research that in response to terrorist use of tech platforms, policymakers tend to focus on major social media platforms and neglect or fail to acknowledge that the ecosystem is much wider and diverse than those platforms. Many of the platforms that have found to be outlinked to on Twitter are overlooked. Policymakers also fail to acknowledge the vulnerability of smaller sites and platforms that are less equipped than the major platforms regarding tools, technology, man power etc. to respond and counter terrorist organisations to the same extent. Content-based regulation, which is one of the most popular demands by governments (Weirman and Alexander, 2020) is not necessarily going to be feasible for such platforms, which is one of their main appeals to terrorist organisations. Additionally, organisations are working to circumvent content removal strategies as seen in the sharing of news sources example. As already argued, unless there is some standardization in platform responses across the ecosystem, it is going to be difficult to break the whack-a-mole pattern that has emerged over recent years (Nouri, Lorenzo-Dus, Watkin, 2021).

A study by Macdonald et al. (2019) also examined IS tweets on Twitter, however, this study focused specifically on IS attempts to disseminate its online magazine *Rumiyah*. The study found that IS used the tactic of posting URLs to a large number of file-sharing sites to access the magazine elsewhere. The outlinks in the dataset contained 244 different hostnames. The top ten hostnames included Google Drive, Justpaste.it, archive.org and Dropbox, all of which

are common sites for storing content. Smaller platforms such as justpaste.it are also not as able to remove content as well as the major social media platforms for reasons such as having less staff, resources and expertise in the area. As with the results of Conway et al.'s (2019) study, many of the accounts in the dataset ended up being suspended. The findings suggest that, again, as with Conway et al.'s (2019) study, IS had adapted to using Twitter as a gateway platform to signpost their followers and a large mainstream audience elsewhere in a similar throwaway-account style.

A study by Defence (2020) undertook research into terrorist use of Twitter, as well as Discord and Reddit. Discord was originally built for use by video gamers, however, has evolved into a chat service for anyone who wishes to use it. Reddit is a platform that allows users to post text, photos, videos and URLs. Reddit allows users to create subreddits which are based on a particular topic or theme. Subreddits are moderated by the users in their community themselves. The research was based on terrorist groups based in the Middle East. Content was collected from all three platforms. In the Twitter dataset that was based around collecting key words, the researchers collected URLs. The researchers noted that there was virtually no discussion space on Twitter, instead the focus was on linking followers elsewhere. While the URLs in the Tweets mainly led to news articles or to Twitter itself, the domains did include Dailymotion, YouTube, Gab and Periscope. Alt-right political websites such as Breitbart were linked to as well. The researchers noted that sometimes links were posted in the profile description box as opposed to in Tweets. In the Discord dataset, many of the links were to Discord itself, however, YouTube and Soundcloud were both significantly linked to. The Reddit dataset was different as it was not collected from key words, but instead selected subreddits containing Salafi content. The researchers revealed that arguments and opposing views were found in the subreddits studied with a presence of those opposed to Islamism. Although the research was focused on terrorist groups based in the Middle East, the researchers identified a strong alt-right and fake website presence in the themes that were examined. Overall, this research identified the interconnectedness of the use of the platforms studied. It found differences in that Twitter was mainly used to signpost followers elsewhere, whereas, Reddit was used for discussions.

Berger (2018) undertook research into alt-right use of Twitter. The themes in the tweets collected ranged from pro-Trump content, to white nationalist content and general far-right content. This included anti-immigrant and anti-Muslim content, trolling, and conspiracy and fake news. An analysis was undertaken on the URLs in the tweets. Popular domains included

YouTube, Facebook, Instagram and WordPress. Gab was also on the list, and Amazon was a common domain with many of the links leading to self-published books. News outlets were common including Brietbart.com and Fox News. There were also links to apps that allow a user to monitor or manipulate follower counts. This study displays the variety of domains used by the alt-right and the interconnectedness in the ecosystem.

The final study is by Phadke and Mitra (2020). This study differs from the rest in that it is a cross-platform study that researches 72 groups (covering different ideologies: white supremacy; religious supremacy; anti-Muslim; Anti-LGBT; and anti-immigration) from the Southern Poverty Law Center's extremist files across both Twitter and Facebook. Content posted by the groups were collected and annotated from both platforms. The researchers also extracted URL links from the posts. The results of the URL analysis reveal that they tend to lead to domains for the following purposes: streaming; promotion (petitions sign-ups, membership forms, merchandise etc.); information; opinion; and news. Overall, the sharing of URLs to both Twitter and Facebook allowed the groups to signpost followers to their own websites and other extremist blogs. The researchers also noted that the domains linked to often contained "toxic language and extremist propaganda" that would be at high risk of removal on Facebook and Twitter. The results revealed that news sharing was more prominent on Twitter, whereas, the provision of links to informative sources and websites, and the promotion of group opinions were both more prominent on Facebook. The groups therefore utilised the two platforms differently.

In summary, Twitter is no longer used in the same way that the major platforms were used in the early research mentioned. Platform censorship has changed the advantages that Twitter has to offer terrorist organisations. Some terrorist accounts struggle to remain on Twitter longer than 24 hours before removal. Twitter can no longer be utilized as a main way to attract new audiences, host a community, and disseminate propaganda. Terrorist organisations now use this 24-hour window to signpost followers to other platforms where they can disseminate propaganda and host a community, for example, small file-sharing sites that do not have the same capacity as the major platforms to police their platform or alternative platforms that are unwilling to remove content. This use of Twitter to signpost to a wide range of other platforms shows some support for the argument that terrorists use a diverse ecosystem of platforms to fulfil the different functions and services that they require. This raises a question about whether a single platform will be able to provide all of the necessary functions that a terrorist organisation needs. Twitter no longer provides operational security or the ability to reach a

mass audience for more than 24-hours. It can, however, provide a gateway to other platforms that can provide the services it is missing. However, organisations can try to adapt their content to be able to remain online longer, as seen in Weirman and Alexander's (2018) study.

Telegram

Telegram is defined as a cloud-based online instant messaging platform that has become increasingly popular among both jihadist (Clifford and Powell, 2019b) and radical right (Urman and Katz, 2020) organisations in recent years, particularly as the major platforms have increased their content removal efforts. Telegram can offer terrorist organisations, security and privacy with its encryption and features such as self-destruct messages and secret chats. Telegram is also unwilling to disclose data to governments from private communications on the platform (Telegram, 2021c). Furthermore, Telegram allows users to share an unlimited range of content, including photos, videos and files of up to 1.5GB each. Telegram provides users with the ability to host or participate in both public and private channels whereby content and messages can be sent to an unlimited number of subscribers (Urman and Katz, 2020). Unlike public channels, private channels are not searchable. Telegram users also have the ability to be anonymous; channel creators can send messages anonymously. All of these features and services are ideal for a range of the activities found in Weimann's (2004) eight ways that terrorists use the internet, such as communicating, coordinating, planning and mobilizing (Urman and Katz, 2020). It is therefore not a surprise to find that its use has been adopted by a range of terrorist organisations. Telegram was found to have been used as both a recruitment and coordinating tool in several attacks that were claimed by IS. IS propagandist Rachid Kassim was found to have used his Telegram channel to recruit supporters to undertake attacks in France (Meichtry and Schechner, 2016). Telegram was also found to have been used for communicating and coordinating during attacks in both Brussels and Paris (Bloom, Tiflati and Horan, 2019). Despite Telegram's unwillingness to disclose data to governments from private communications, Telegram did work with Europol in 2019 to undertake coordinated action against the dissemination of terrorist content on its platform (Europol, 2019a).

The first study that will be discussed is Bloom et al.'s (2019) study examining IS's use of Telegram over a two-month period. The authors observed channels that were focused on three main accounts, all of which were semi-official IS news outlets. The dataset included 3 channels and 16 chatrooms. The research examined channel growth, lifespan and the member flow of channels. The authors observed three types of Telegram users: users seeking information; users

wishing to engage with the organisation; and propagandists who also participate in both of those activities. One of the findings was that channel administrators were very quick to remove users that they did not believe were true followers. There was a lot of suspicion around infiltration of law enforcement. This suggests that security is a priority on Telegram. Another finding was channel administrators managed multiple chatrooms and whenever a new piece of content was shared it was quickly disseminated across dozens of other channels as well. This shows coordination and suggests a high degree of centralization within the network. The content that was shared contained a range of recruitment content and a mix of photos, beheading videos, audio files and other links. There were also memes that appear to be intended as funny and light-hearted, for example, cute photos of kittens and babies. The authors observed that backup channels were created days before any content was posted to them. This shows that the administrators of the channels were savvy to the risk of removal if they post before many users have had a chance to find and access the channel. This study therefore shows that IS used Telegram as a platform to disseminate content. Telegram allows users to have complete control over which users could and could not participate in channels, allowing security to be prioritised and tightly controlled.

Al Darwish (2019) researched IS use of Telegram. This research revealed that initial content dissemination takes place on Telegram, however, results in a much wider dissemination, reaching major platforms such as Twitter. The research walks us through the stages that occur in between. The research argues that it is through the use of "fanboys" that propaganda is able to take such a cross-platform journey. Fanboys spend a lot of time and effort spreading the propaganda to as large an audience as possible. Whenever there is a new release, for example, a trailer for a video, fanboys will create a large number of Telegram channels named after the trailer. Next, they will create Twitter accounts for IS supporters to use to undertake a Twitter invasion to disseminate the new release. They will publish a list of trending or popular hashtags to hijack to help further the spread of the new propaganda. This research tracked 221 IS fanboy Twitter handles that disseminated propaganda for four days. These accounts were found to use several strategies to maximise dissemination including retweeting their own tweets, retweeting each other's tweets and replying to tweets with links or instructions as to how to download IS propaganda. This research reveals that both Telegram and Twitter were seen as crucial to propaganda dissemination.

Another relevant study is Frampton et al.'s (2017) research into IS's use of Telegram. Frampton et al. (2017) analysed 300,000 IS posts on Telegram. The research revealed that IS would

typically use a small number of channels to disseminate their original mix of audio-visual content. After content has been distributed on these Telegram channels, some of that content would then be further disseminated into other channels to ensure resiliency if anything were to happen to the main channel. These messages were also filled with URLs to other external platforms; 76 percent of the verified jihadist channels in the sample led users to other platforms. Including subdomains (such as wordpress blogs), the URLs led to over 400 distinct domains. The use of this range of domains fluctuated over time, most likely because of attempts to disrupt the organisation across many different platforms. The range of platforms that were linked to from these channels included Archive.org, Google Drive, YouTube, Sendvid (all of which are particularly useful for storing videos), as well as Twitter and Facebook which are used to reach a wide audience. The researchers found that in particular, "Twitter is a key and in many ways dominant means of delivering content to those not already accessing it via Telegram. Twitter, in other words, is a crucial gateway to the uninitiated…" (Frampton et al., 2017, p.53).

Frampton et al. (2017) concluded from their research that there is a clear jihadist ecosystem of tech platforms that groups such as IS require. These groups simply cannot rely on just a small handful of platforms to remain online, distribute a range of different types of content, reach a mass audience and provide them with their complex security needs. As certain platforms have become better at countering terrorist content on their sites, a multi-platform approach has become increasingly important to the organisations. Jihadist groups have, as a result, been observed as continuously having to reconfigure themselves online. This phenomenon has been termed as the 'swarmcast' due to the similarity with the way a swarm of bees or flock of birds constantly re-organise themselves mid-flight (Frampton et al. 2017; Fisher et al. 2019). The swarmcast is defined by the following features. First, speed; the swarmcast is able to quickly disseminate and transfer content to a wide network. This speed prevents disruption to the network when one platform successfully removes their content. The second is agility; this is the ability to switch between platforms with speed and ease, adopting and taking advantage of new technologies as they go, even if it is only short lived. The reposting of copies of content across many platforms also minimizes disruption to the network whenever content is removed from one or two platforms. Finally, resilience; this is the ability to survive or circumvent takedowns and suspensions. The aim is to create a distributed network instead of relying on one central node in the network. This research found that Telegram is clearly a core node in the ecosystem. Telegram allows for the three features found in the swarmcast phenomena. However, the reposting to other channels and linking to hundreds of other platforms ensure the

organisation the resiliency that it requires in case it suddenly faces disruption efforts on Telegram, as well as the reach that it cannot acquire on Telegram alone.

Research by Fisher et al. (2019) resulted in similar findings to those in Frampton et al.'s (2017) study. This research also collected IS messages posted in Telegram channels with a focus on the use of Telegram to signpost followers elsewhere. The study collected outlinks that were posted in the messages. There was a total of 1,648 unique outlinks in the dataset. The findings in this study were that over 50 percent of the outlinks collected led to self-hosted websites, 18 percent led to archive.org servers and 14 percent led to major platforms such as Facebook and Twitter. These platforms are where jihadist groups store large volumes of their content, sometimes termed as 'content stores', for their followers to directly access via a simple URL. The self-hosted websites and smaller platforms are chosen specifically because they are less likely to have systems in place to rapidly locate and takedown the content. The major platforms were used to signpost users to the locations where the content could be accessed as well as to communicate and coordinate with followers. This makes sense given that IS have always acknowledged that the major platforms are best for reaching a mass audience. Clifford and Powell (2019b) cite a quote from an IS Telegram post in a different study in which IS tell their followers that their main platforms should be where the general public is found.

Fisher et al. (2019) compared the source, which is where the link is posted, with the target, which is where the link takes you to, and came up with three different categories of platform in the ecosystem. The first are beacons. Beacons are sources of traffic and provide a signposting function that points followers to the locations in which they can find content. The second are the already mentioned content stores. Content stores are platforms that host large volumes of content that followers can access using a link supplied by either beacons or the final category, content aggregators. Finally, content aggregators are platforms that gather a range of material and provide followers with links to specific pieces of content. This revealed that domains provided different functions. Some were used for content dissemination, while others were used for content collection.

Fisher et al. (2019) put forth the argument that diversity is crucial for the resiliency of any ecosystem and that terrorist organisations such as IS have realised this. This is shown in the large diversity of tech platforms that they utilise. The purpose of the tech platform is dependent on factors such as accessibility, security, reach of audience, and how good the platform is at removing content. Disruption on one platform will not necessarily cause significant ripple

effects on the other platforms because they utilise multiple platforms for content storing and content aggregating, and each platform in the ecosystem is connected to at least three other nodes creating wide dispersity. Therefore Telegram, although crucial to the ecosystem for IS, is just one platform in a large range of platforms that they rely on.

Clifford and Powell (2019b) have also researched just over 600 English-speaking IS channels and groups on Telegram. The research revealed that IS used these channels to communicate with followers, disseminate official and unofficial propaganda and provide instructional operational materials. The authors identified three particular tactics used by IS in the dataset to ensure community resiliency. These were using joinlinks (this is where a user requires a URL invitation to gain access to a private group or channel), file-sharing, and observing cybersecurity measures. In addition to this, as with the other studies, many of the channels posted outlinks. External sites that were outlinked to were mostly made up of file-sharing sites to store content and major social media platforms with the already mentioned purpose of reaching a wide audience and building resiliency to disruption. A total of 731 unique domains were outlinked to in the dataset; 15 of the top 20 were file-sharing sites. All major media releases such as news videos were disseminated amongst these file-sharing sites, the most popular being archive.org, justpaste.it and top4top.net. Therefore, as with the other Telegram studies, this one found that Telegram is important for communicating with followers and disseminating propaganda and materials. However, it is also a useful node in the ecosystem for URL sharing. Sharing links to file-sharing sites is required to ensure resiliency against disruption efforts and to major platforms to provide the mass reach that is desired.

Although much of the Telegram research has been focused on jihadist organisations, Telegram has become attractive to radical right organisations also, given the increase in disruption efforts from major and mainstream tech platforms to counter them (Urman and Katz, 2020). Although this is an understudied area at the time of writing, research by Urman and Katz (2020) undertook a network analysis of radical right Telegram channels that were found via a Telegram bot starting with one seed account chosen for having a large number of subscribers among known-political channels and then using snowball sampling. Overall, data was collected on just over 53,000 channels, and public and private groups. One of the findings was that there was an increase in users around the same months that well-known radical right organisations were banned from platforms such as Facebook and Twitter. Another finding is that the radical right channels had managed to create a strong community structure on Telegram, with multiple distinct groups, spanning many languages, forming a decentralized structure around ideology

and nationality, similar to what many groups had achieved on the major platforms prior to disruption. Finally, similar to the jihadist research, the number of followers on Telegram tend to be less than what organisations were able to achieve on major platforms. Although this research did not provide the insights regarding URL links like the jihadist research did, meaning that it is not clear whether Telegram is used by the radical right in the same interconnected way jihadist organisations do with other platforms, what is clear is that as soon as radical right groups are disrupted from one platform, they will migrate elsewhere in the ecosystem. In particular, they will migrate to platforms such as Telegram that can offer them the protection from disruption that they faced on the major platforms, even if they may not necessarily have the same reach that they did on the major platforms. Therefore, as long as there is a large ecosystem of platforms with different levels of disruption and responsiveness there will likely always be the issue of whack-a-mole.

In summary, this body of research identified Telegram as core to organisations such as IS because of the operational security and privacy advantages it offers. This makes it difficult to disrupt their activities and communication (Urman and Katz, 2020). However, with the security advantages also come disadvantages. One of the main issues with Telegram is that due to its security-oriented architecture, it is unlikely that non-followers are going to stumble across terrorist channels and content in the same way that they might on the major platforms (Frampton et al. 2017; Urman and Katz, 2020; Rogers, 2020; Clifford and Powell, 2019b). Therefore, although Telegram is used to communicate, disseminate propaganda and signpost followers to other platforms, it is not an ideal platform for attracting new followers. Those who have found IS on Telegram are almost certainly already followers of the organisation. Telegram has therefore been found as a key platform for operational security and resilience against disruption, however, cannot offer organisations the whole package. Without utilising other platforms alongside Telegram, organisations would struggle to reach a wide audience and attract new followers.

YouTube

YouTube is an online platform that was created to share and store video content. Registered users can upload videos and anyone can search for videos. YouTube is appealing to terrorist organisations because video links can be shared and published on other tech platforms. Several studies have been undertaken into terrorist use of YouTube. Conway and McInerney (2008) undertook an early study of jihadist use of YouTube. The videos in their dataset were focused

on martyrdom. They found that YouTube was primarily used as a repository for large volumes of jihadist video content. They also found that jihadist organisations were utilising the comment section on the videos as opportunities to pursue the radicalisation of their followers. Kayode-Adedeji, Oyero and Aririguzoh (2019) also collected and analysed jihadist YouTube videos. The themes found in the videos included claims of attacks, threats and recruitment. Klausen, Barbieri, Reichlin-Melnick and Zelin (2012) similarly collected jihadist videos on YouTube and found themes of promoting violent acts, threats and demonstrations. They also found that one single organizing entity was behind a network of YouTube channels in their dataset revealing a coordinated effort to maintain an online presence.

Research has also been undertaken into far-right organisations use of YouTube. Ottoni, Cunha, Magno, Berdardina, Meira and Almeida (2018) collected far-right videos on YouTube and examined the topics and discriminatory bias in the videos and comments. One of the findings was that there was a high percentage of negative word categories, for example, aggression and violence. There was also a bias in the videos and comments against Islam, immigrants and the LGBTQ community. Rauchfleisch and Kaiser (2020) also analysed German far-right videos on YouTube. This research found similar themes to the previous study with the main theme being the refugee crisis. The results showed a distinct tight-knit community in the dataset on YouTube.

Overall, YouTube appears to be used by organisations as both a repository for video content as well as an opportunity to communicate and make recruitment efforts via the comments section. As can be seen from other sections in this chapter, YouTube is a common domain in the URLs that are posted elsewhere in the ecosystem. A particularly problematic aspect of YouTube has been highlighted by a number of scholars. O'Callaghan, Greene, Conway, Carthy and Cunningham (2015) found evidence that users who view extreme right-wing content on YouTube are likely to be recommended more extreme right-wing content in the recommendation system. Reed, Whittaker, Votta and Looney (2019) also found evidence that YouTube recommends extreme right-wing material after one has interacted with similar content already on their platform. The latter research investigated whether this was the case on Gab and Reddit, however, there was no evidence to support that this is the case on either platform. This is an issue that could assist terrorist organisations in their missions.

Reddit

Reddit is referred to as a 'social news aggregation' platform where users can create communities in the form of subreddits to discuss topics and share content (Gaudette, Scrivens, Davies and Frank, 2020). Reddit has a voting algorithm that allows users to promote the content that they want to see or demote what they do not within their subreddit. Reddit is also unique in that its users can become voluntary moderators of subreddits. Gaudette et al. (2020) collected the 1000 most highly-upvoted comments in r/The_Donald subreddit which is known to be home to a variety of right-wing movements. The research found that a common theme in these comments were hateful towards Islam and immigration. Many of these comments had not been removed by the moderators of the subreddit. Another theme was spreading fear around future survival of the in-group and rising numbers of the out-group. There was, however, very little evidence in these comments of incitement or encouragement to do anything about this. This could signal an intentional attempt at carefully avoiding removal on the subreddit. A different theme in the comments was the threat of The Left who appeared to be perceived as violent. Overall, the researchers conclude that the far-right use Reddit to form a collective identity and discuss their hate for their perceived enemies. There was a lack of dissenting views found in the comments, possibly due to the downvoting system. The dataset was based on the top 1000 upvoted comments therefore suggesting that many users in the subreddit, including the moderators, support the extreme views that were present in the comments.

Research by Mittos, Zannettou, Blackburn and De Cristofaro (2020) undertook a study examining the discussion of far-right subreddits regarding genetic testing. A dataset was also collected on 4chan. The results found that this was a highly discussed topic in the far-right on both platforms with hateful, racist and misogynistic comments. A common discussion was the use of genetic testing to marginalize or eliminate minority groups with calls for genocide. Another study by Baumgartner, Zannettou, Keegan, Squire and Blackburn (2020). These two Reddit studies suggest that Reddit is used primarily to discuss content that may be at risk of removal on major platforms such as Facebook and Twitter. The voting system allows dissenting content to be demoted and less likely to be viewed, and the moderation system may also work in their favour.

Small file-sharing sites and alternative platforms

The findings of this literature review revealed a small number of studies examining terrorist use of small file-sharing sites and alternative platforms. Similar to Urman and Katz (2020) Telegram research, a study by Nouri et al. (2021) examined the issue of whack-a-mole in the

far-right. This research examined UK far-right group Britain First as they were removed from Facebook and migrated to Gab which is an alternative platform known for attracting far-right groups and individuals because of its lack of willingness to regulate and remove content. Visual content (images) that were posted by the group were collected from Facebook prior to their removal and then again from Gab after their migration. This research noted several visual changes in the content that the group posted over the course of their migration. The main finding was an expansion of 'othering' practices from that of Islamic extremism (on Facebook) to practices of Islam more broadly (on Gab). These findings were interpreted as resulting, at least in part, from the group setting up home in a less regulated space and therefore warns of the appeal of such platforms and the potential for content to become more extreme there.

Trujillo, Gruppi, Buntain and Horne (2020) undertook research into the use of the alternative platform Bitchute. The Southern Poverty Law Center has described Bitchute as a "low-rent YouTube clone that carries an array of hate-fuelled material" (Hayden, 2019). Bitchute is a peer-to-peer video hosting platform. The researchers collected a corpus of video metadata and comment data in 2019. According to the researchers, video engagement on Bitchute has a heavily skewed distribution with only a small number of videos receiving high views. The most viewed video in the dataset contained hate speech against homosexuality and Islam and came from a user who claimed to have been banned from YouTube for the same video. The third most viewed video in the dataset came from Infowars and the fifth most viewed was a QAnon video. Upon analysis of the channels, the research found that the top viewed channels claim to be news services, journalist or political commentary. The researchers show that in the top 40 domains that were linked to in video descriptions, six were social media platforms, four were for fundraising/monetary support purposes, and thirteen were news-oriented. 25 percent of the links were to YouTube as some of the users were still able to maintain less extreme content on YouTube. Users were found to be linked to Bitchute via Infowars, Gab and Voat. Overall, this research revealed the primarily news-sharing and political content dissemination purpose of this alternative platform. It also, however, showed the interconnectedness that is seen in many of the jihadist studies in this chapter. Finally, this Bitchute research, as with the previous Gab study, revealed the whack-a-mole consequences when platforms in the ecosystem do not adopt an industry-wide response.

A study by Shehabat and Mitew (2018) studied IS's use of three specific sites that they label as "anonymous sharing portals", using actor network theory to better understand their role in the ecosystem. These were sendvid.com, justpaste.it and dump.to. These sites have been

increasingly used since the disruption on major platforms and can provide advantages that major platforms cannot. These platforms are less likely/able to remove content than the major platforms. Further, these platforms allow for easy sharing and storing of content whilst providing anonymity for users, therefore, acting as "automated message amplifiers, playing the role of black boxes in the wider information network" (Shehabat and Mitew, 2018, p. 82). The authors identified three main categories of users on these platforms: 1) those who produce and aggregate content; 2) those who act as intermediaries which involves retranslating and curating content to other platforms; or 3) those who passively consume content.

The first platform in the study is Justpaste.it which is a free content sharing site that allows users to share and store files and other content (text, images etc.) without having to register an account. Links to content on justpaste.it have been found on Twitter, Facebook and Telegram amongst other platforms and sites (Shehabat and Mitew, 2018). The site is not searchable and content can require a specific link to access it. According to the authors, IS began to use justpaste.it in early 2014 to disseminate videos and online magazines amongst other content. Justpaste.it offers a free folder share feature that IS used to store, share and disseminate their online magazines to mass audiences. Justpaste.it also offers features that are missing from major platforms such as exporting to PDF and password protected access to content. Although justpaste.it has made efforts to try to remove IS content, the site is run solely by its one creator and despite help from Tech Against Terrorism[3] and the Global Internet Forum to Counter Terrorism[4] (both of which are defined and discussed in later chapters) it is difficult for the platform to have the capacity to remove all terrorist content, particularly in a timely manner (Tech Against Terrorism, 2017c).

Sendvid.com, described as an instant upload portal, is the next platform examined in Shehabat and Mitew's (2018) study. This is a predominantly video-sharing platform that has been used increasingly since YouTube's disruption efforts. IS have used sendvid.com to share a range of videos, which can then be linked to via other platforms. The authors claim that a search for the platform on Twitter results in mostly links to IS propaganda videos. The videos can then also

[3] Tech Against Terrorism is a capacity-building organisation that is supported by the United Nations Counter Terrorism Executive Directorate (UN CTED). The aim is to work with tech platforms to counter terrorist use of their services whilst respecting human rights. The initiative revolves around three pillars: outreach, knowledge-sharing and practical support. The outreach pillar involves working to promote constructive relationships between governments and tech platforms. The knowledge-sharing pillar focuses on working with tech platforms to share best practice. The final pillar offers tech platforms practical support with implementing tools to respond to terrorist content. https://www.techagainstterrorism.org/

[4] https://gifct.org/

be stored on Google Drive or Dropbox for future use. According to the authors, Sendvid.com make very little effort to remove such content from its site, adding to its appeal. Finally, Dump.to is the last online sharing platform examined in the research. Similar to justpaste.it, this platform does not require user registration, and similar to sendvid.com, the platform makes little effort to identify and remove content. The platform allows users to upload and share links from other sharing sites. The content that can be stored and shared includes files, video and audio material. Content on this platform can also be password-protected and edited by anyone anonymously.

The three platforms examined in Shehabat and Mitew's study offer many advantages to terrorist organisations. One is that they are less likely to actively search for and remove the content than major platforms, either due to a lack of capacity or for other reasons such as unwillingness. Second, they offer anonymity, in some cases, users do not even have to register with the platform to use it. Finally, they offer a range of features such as hosting different types of content, sharing to other platforms, and password protected content. All of this builds a level of resiliency and security that is valuable for such organisations. The organisations can use major platforms to signpost followers to these platforms, as mentioned with the other studies as well, thus enhancing connectivity in their network. The authors estimate that these three platforms have contributed to approximately 20 percent of the information that IS have disseminated to Twitter alone.

The final study found in this literature review is by Zannettou, Blackburn, Cristofaro, Sirivianos and Stringhini (2018). This study seeks to understand the misuse of web archiving services on social media platforms. This study differs from the others in that it is not focused on a terrorist organisation. However, some of the content in the dataset can be classed as content that is often shared by the alt-right and fringe communities. Despite this, the study still provides an insight into the use of archiving sites regarding content that is potentially at-risk of removal. The authors argue that due to the ephemerality of online content, archiving sites play an important role to many people. They allow an individual to create a URL to share across platforms that take people to a snapshot of a web page or piece of content. For example, archive.is has been known to be used to preserve "controversial blogs and tweets" that may later be removed or deleted for a variety of reasons (Zannettou et al. 2018). The authors claim that there is anecdotal evidence to suggest that alt-right communities use archive services to boycott domains that they disagree with. For example, they will encourage users to share archive URLs instead of the domains on URLs.

The study collected 21 million URLs from archive.is as well as 391,000 URLs from Wayback Machine that were shared across four tech platforms: Reddit, Twitter, Gab and 4chan's Politically Incorrect board (/pol/). The findings were that, first, news and social media posts are the most common types of content that are archived which is likely due to their vulnerability regarding removal. Second, the URLs in the dataset are extensively shared on fringe communities on Reddit and 4chan, most likely to protect it in case of removal on the major platforms. The authors claim that archive.is is known for its use in fringe web communities and found that it was favoured in the dataset over Wayback Machine on /pol/ and Gab. The most popular domain in the dataset was the Wayback Machine's archive.org. On Reddit, Twitter and 4chan, the most popular domain archived through archive.is was the platform itself. Both mainstream and alternative news sources were heavily archived and shared on Reddit, /pol/ and Gab. URLs to chat forums were found on both Gab and Reddit. Overall, both the archiving sites in the study were used on all four platforms to disseminate URLs to news sources, social media platforms and marketing sites. This suggests that archiving sites play a crucial role in preserving content that is subject to disruption efforts, however, as already mentioned, the content collected in the dataset was not terrorist content nor was it explicitly linked to any proscribed organisation, therefore, the findings of this study cannot be interpreted in the same manner as the other studies in this chapter.

The findings of this Zannettou et al.'s (2018) suggest that archive services are likely to be utilised by internet users who wish to preserve content that they fear may be removed from major platforms for one reason or another and to circumvent removal strategies. Archive sites can be used to share links to a variety of other platforms, including social media platforms, alternative platforms, mainstream and alternative news sources and chat forums. Archive services are therefore ideal for maintaining connectivity in a network and ensuring resiliency to disruption efforts. Such services were found to be used across platforms that were missing from the other research included in this chapter (Reddit, Gab, /pol/). The platforms in this study have a tendency to be used by radical right organisations and the alt-right (Gaudette, et al., 2020; Zannettou, Bradlyn, De Cristofaro, Kwak, Sirivianos, Stringini, and Blackburn, 2018; Colley and Moore, 2020). As already mentioned, there is a need for more research regarding non-jihadist ideologies and examining a wider variety of platforms than Twitter and Telegram. Zannettou et al.'s (2018) study found that in addition to being used in similar ways to other platforms and sites in the jihadist research, archive sites were used to boycott mainstream news

sources that the users disagreed with. The authors believe one reason for this is to affect the advertising revenue these news sources receive.

Overall, these studies provide further support for the argument that a whole diverse ecosystem of platforms are required for actors, such as terrorist organisations, in order to ensure a persistent online presence and undertake a wide range of activities. This highlights the already mentioned need for policymakers to include such a wide range of platforms and consider the intricacies and inter-connected use of platforms in policy-making, as opposed to focusing solely on major platforms. The next section in this chapter will discuss the key points that should be taken from the examination of this literature review into how terrorist organisations exploit tech platforms and the differences in this exploitation across platforms.

Key considerations from literature review findings

The purpose of the literature review undertaken in this chapter was to answer the following questions:

1) What tech platforms are exploited by terrorist organisations?
2) How do terrorist organisations exploit tech platforms?
3) What considerations do terrorist organisations have when choosing which platform to use?

The literature review establishes that terrorist organisations do, as the research argues, adopt a multi-platform approach; they utilise a whole ecosystem of platforms to fulfil their online activities. Tech platforms began to emerge in the mid-2000s and terrorist organisations had adopted a multi-platform approach by 2010 (Department of Homeland, 2010). As shown in the above research, it was not long before the multi-platform approach evolved from a mix of major social media platforms, chat forums and traditional websites, to an interconnected web of social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites. This is for two main reasons, the first and most critical is to remain online. Given the pressure that tech platforms face from governments around the world, it is too risky to rely on a small handful of platforms for an online presence. Relying on one platform as a place to communicate and store content, in the face of disruption, would mean losing all of the organisation's followers and content, undoing all of the work the organisation would have undertaken to build a community and attract new followers. Terrorist organisations therefore have realised that they must back up their content on platforms that are less likely to remove it and spread their activities across several platforms. They realised that it is easier to circumvent

disruption strategies if they post URLs to their less-censored content stores on the major platforms than if they posted the content itself to the major platform.

The second reason why terrorist organisations adopt a multi-platform strategy is that it appears that no single platform can provide all of the services, protections, and audiences that they would require to undertake all of the activities that they need to fulfil their cause. It is apparent from the above research that terrorist organisations utilise tech platforms for many different activities. These include, but are not limited to: communication, community-building, content storing, content dissemination, recruitment, and signposting followers elsewhere. Tech platforms are each built with a unique architecture and different priorities, values, capacity and functions. As a result, each platform offers a limited but unique set of services and userbase, as well as different levels of protection regarding security and privacy. For example, Telegram was built to prioritize security and privacy and is unwilling to disclose data to governments, thus offering terrorist organisations protection regarding communication, coordination, planning, mobilizing and content dissemination. However, this also comes with limitations. As discussed, this creates difficulties in attracting new followers. The major platforms (Twitter and Facebook etc.), on the other hand, were built with the intention of connecting people and tend to have a very large mainstream userbase. They lack the security protections afforded by Telegram but bring the benefit of reaching a mass mainstream audience that is ideal for finding new followers and recruitment. In addition to this, some platforms are ideal for the dissemination or storage of a particular type of content, for example, YouTube and Sendvid were built primarily to store and disseminate video content, and Telegram allows the uploading and sharing of large files. Other platforms are ideal because of features such as providing anonymity where users do not even have to register (e.g., justpaste.it). Tech platforms also differ in terms of the capacity and/or willingness they have to identify and remove content, making some platforms more attractive than others for activities such as content storage. An example was justpaste.it. It struggles with having the capacity to remove large volumes of terrorist content because it is run solely by one person. Other platforms such as sendvid and dump.to make very little efforts to counter terrorist exploitation, however, it was not clear from the research exactly why this is.

The key points arising from these two findings are that, first, the more platforms an organisation utilises the more resiliency it builds against disruption. Without a wide spread of platforms, they risk losing their communities, followers and content. Second, different platforms offer different functions, protections, and levels of access to users. Terrorist organisations require

the use of many different platforms to utilise all of the functions and reach all of the audiences that they require. They therefore require a whole ecosystem of diverse platforms to survive online and carry out all of their desired activities.

Another finding from the above literature review is that, although different platforms are used for different functions, a major strategy of terrorist organisations is to use these platforms in an interconnected manner. This is done via sharing URLs on one platform to signpost followers to their accounts and content on other platforms. This allows for a level of connectivity in their network that is difficult to disrupt. It also insures the organisation against losing communication with their followers in the face of disruption on a small number of platforms. The constant dissemination of URLs leading to external sites ensures followers always have access to the organisations and their content. As mentioned by Weirman and Alexander (2020), terrorist organisations have realised that when used in conjunction with one another, platforms have higher potential and offer more benefits than they do in isolation.

Key considerations for the proposed regulatory framework

The first consideration is that the regulatory scope of the framework must include the whole ecosystem of platforms that are used by terrorist organisations (e.g., social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites) because they are all vital in one way or another to the success of their operations. If regulation is only aimed at, for example, platforms that are considered "beacons" then this would only disrupt part of an organisations operations and will likely leave other parts of the operation unaffected. Similarly, if the "content stores" are disrupted then this again would likely only affect this part of their operations. This would barely disrupt the network if content or followers were backed up elsewhere. A holistic approach is required if disruption is to take place to the whole operation that terrorists exploit tech platforms to undertake. Terrorist organisations such as IS have spent a lot of time creating an interconnected network in the ecosystem to try to limit the damage of disruption efforts and ensure that their followers will always be able to find them and access their content. Therefore, without such a holistic approach it is likely that the whack-a-mole effect discussed in the literature will continue.

The findings that terrorist organisations use multiple platforms in conjunction, or migrate from one platform to another in times of disruption reveal that this is an industry-wide problem that requires an industry-wide response. The literature review revealed that terrorist organisations, at one-point, utilised major platforms such as Twitter as their main location for disseminating

content. Twitter, however, put in place policies and technology that became fairly effective at quick and automated removal of certain (although not all) terrorist content (Conway et al. 2017). As a result, terrorist organisations, such as IS, migrated to Telegram and although Telegram does not provide all of their desired functions, it has been an effective platform for them in many ways. Terrorist organisations also heavily utilise smaller file-sharing sites that either do not have the capacity or willingness to respond to their exploitation. This further supports the argument that all platforms in the ecosystem should adhere to the same regulations. In addition to this, the exclusion of some platforms in a regulatory framework would likely create issues. The platforms that fall under the scope of the regulation are unlikely to view the regulation as fair given that the research shows that a wide array of platforms are exploited. These platforms may view the regulation as affecting competitive practices and de-incentivise compliance.

Another consideration is that the platforms in the ecosystem vary greatly in their capacity and/or willingness to respond to terrorist use of their site. Regarding capacity, the research revealed that while some of the platforms that are exploited are major platforms that are known to have offices all around the globe with a large staff and resources, there are also platforms in the ecosystem that are run by as little as one person with very little in the way of resources. The latter type of platform may wish to counter terrorist exploitation, however, physically struggle to do so. Further, the volume of content uploaded to tech platforms may also differ greatly. Therefore, regulation that places the same demands on all platforms without considering capacity (e.g., resources), as well as the required expertise, is unlikely to be effective across the board. There is similarly the issue of willingness. The literature review revealed that some tech platforms are far more willing to counter terrorist exploitation than others. While the previous paragraphs argued for an industry-wide response whereby platforms all follow the same regulations, there must also be consideration that some platforms may require assistance when it comes to capacity in order for them to be able to comply. Such a consideration could ease the issue of placing unfair burdens and uncompetitive practices. To tackle the issue of willingness, platforms may require an incentive to comply.

Finally, this is not an issue that tech platforms will be able to counter alone, it is likely going to take a collective effort with multi-stakeholder engagement. Many of the platforms found in the ecosystem did not consider, back in their days of early development, that their platform would be exploited by terrorist organisations (Conway cited in Sahinkaya, 2019). As such, many platforms lack the knowledge and expertise necessary to counter terrorist use of their

sites. As a result, a range of expertise will be necessary, it will require collaboration between at least governments, tech platforms, academia and civil society (Frampton, 2017). Where there is an overlap in how platforms are exploited, platforms may be able to come together and share with one another the lessons they have learned and best practice regarding their responses.

Conclusion

This chapter undertook a literature review of research that investigated terrorist use of tech platforms. The findings established that terrorist organisations utilise a whole ecosystem of tech platforms that include social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites. Terrorists use these platforms for a diverse array of functions including communication, recruitment, planning and mobilizing, content dissemination, content storing and signposting followers elsewhere. The research revealed, however, that terrorist organisations utilise such a large, diverse ecosystem of platforms for a number of reasons. One is to ensure a persistent online presence, protecting their network and communication with their followers and access to their content in the face of disruption. Another is because there does not appear to be such a thing as the perfect platform that can provide all of the necessary protections, functions and audiences that terrorists require. A third finding is the interconnected nature of the ecosystem. This is most likely a result of the previous point that some platforms have better access to mass audiences or provide better protections for content storing than others and as such terrorists have realised platforms are more useful in conjunction than in isolation. It is also another way to protect the connectivity of their network and avoid feeling the effects of disruption efforts. Finally, the research highlighted the issue that platforms do not all contain the same level of expertise, capacity and willingness to respond to terrorist use of their sites.

It is therefore important that all of these tech platforms fall under the scope of future regulation. It is an industry-wide problem that requires an industry-wide response. There are, however, additional intricacies and complexities that must be considered. These surround the capacity, expertise and willingness that tech platforms have to comply with future regulation. Without such considerations, issues regarding fairness, burdens, and competitive practices arise creating a disincentive to comply. These findings informed the development of the regulatory framework that is proposed later in this thesis.

Chapter 3: What has been done to counter terrorist content: Tech platform efforts

The aim of this chapter is to gain a deeper understanding of the work that tech platforms report taking in response to terrorist content on their services. This chapter will answer the following questions:

4) In their blogposts, what efforts do tech platforms report taking to counter terrorist content on their services?
5) What challenges do tech platforms face in their efforts to counter terrorist content on their services that could affect their compliance with regulation?

In doing so, this chapter aims to identify both similarities and differences across the platforms, as well as identify any issues platforms may face regarding regulatory compliance. This chapter applied a content analysis to blogposts published by Facebook, Twitter, YouTube/Google, Microsoft, Gab and Telegram over a period of four years. The blogposts are used by the platforms to update their users on new decisions that the platform makes, new services and tools they implement, and how they counter bad actors on their sites.

The findings that are reported in this chapter have informed the development of the regulatory framework that is proposed later in this thesis. The chapter will begin with the methodology. It will then provide an introduction to the difficulties and challenges tech platforms face when trying to counter terrorism content from their services. The chapter will then provide a brief overview of each platform and its history. The chapter will then present the findings of the content analysis thematically. The chapter will finish with a thematic summary of the key efforts and challenges that platforms have reported undertaking and facing.

Methodology

The following platforms were selected for this research: Facebook, Twitter, YouTube/Google, Microsoft, Gab and Telegram. The first four are included because they have been largely targeted by terrorist organisations, they are at the center of government attention and demands (Corera, 2017; Hope, 2017), they consistently fall under the scope of regulatory frameworks (as seen in chapter 4), and finally, they are part of well-known collaborative initiatives such as the Global Internet Forum to Counter Terrorism (GIFCT).[5] YouTube and Google have been grouped together because Google owns YouTube and there is overlap in the blogposts where Google speaks about the work of YouTube on the Google blog and vice versa. Gab and

[5] https://gifct.org/

Telegram were chosen to ensure a diverse sample of platforms. Gab is an alternative platform known for its association with the far-right, its unwillingness to remove content, and criticism of regulation (Bennett, 2018). Telegram is an encrypted messaging app that is known for prioritizing security and privacy and was at one time reluctant to regulate but has since made some efforts including cooperating with Europol to undertake action against jihadist content (Europol, 2019c). The period of 1st January 2016 – 31st December 2019 was chosen for data collection. Although platforms had policies in place prior to this, the year 2016 was chosen as the start date because this was around the time that platforms began to escalate their efforts in response to the increased government attention and demands that were being placed on them following the 'golden age' of around 2014-2015 where terrorist content had been going largely unregulated (Conway et al. 2019). A period of four years of blogposts was considered an appropriate-sized dataset and followed the platform throughout a period of many new policies and decisions in the counter terrorism area.

The blogposts were chosen as an appropriate dataset for this research because it is one of the few ways in which the platforms consistently update their users in a systematic manner. These blogposts are understudied in this context. The only academic publication that analyses these tech platform blogposts in a systematic way in this area of research is Watkin and Conway (2021). The platforms use these blogposts to communicate policy updates, decision making, countering bad actors and implementing new technologies, amongst other things. If these blogposts are an accurate representation of what the platforms are doing, then this dataset is useful for an analysis that aims to develop a greater understanding of platform efforts to counter terrorist content on their services and the reasoning, methods and challenges that are involved in this.

However, there are a number of concerns around this. One concern is that the blogposts are simply a PR exercise. Whilst there is no doubt that there is a PR element to the blogposts, that does not mean that the information in the blogposts do not provide insight into how the platforms are trying to be perceived regarding their counter terrorism efforts, decision making, and the values they claim to prioritize. This still has value in this research's aim of trying to create a better understanding of the platforms, their efforts, the challenges they face, and to identify potential compliance issues that platforms may encounter. Another concern is that the blogposts are simply rhetorical or that they are not a true reflection of what the platforms are actually doing to counter terrorist content. This is an issue identified under 'Information Regulation' which is discussed further in chapter 5 on Social Regulation Theory. Essentially,

all businesses, across many industries, tend to emphasize and publish the positives about their work, particularly in trying to overcome issues, and are vague around the negatives and issues associated with their products and services (Ogus, 1994). Therefore, this is an issue that applies to consumer services more widely than simply in this context. A finding that the platforms blogposts are rhetorical or vague regarding certain efforts is in itself an interesting finding. The final concern is that platforms may be doing a lot to counter terrorist content on their services, however, are not publicizing it. In order to overcome these limitations, the findings of this chapter should be considered in the development of the regulatory framework proposed in this thesis, however, taken with caution and consideration of the limitations.

This chapter involved the manual collection of blogposts that addressed responses to terrorist content or concerned relevant policies and regulation from each company between a time period of 1 January 2016 to 31 December 2019. The blogposts were manually collected because it was difficult to use key search terms because the platforms are not always consistent in the terms that they use, for example, some platforms appear to use 'terrorism' and 'violent extremism' interchangeably, sometimes the platforms group issues such as hate, extremism and terrorism together and address these issues in the same blogpost. This issue will be discussed later in the chapter. However, it is therefore argued that manual collection was the better collection strategy because it reduced the possibility of missing relevant blogposts, which may have happened if a small number of key search terms were used. All of the blogposts that had titles that referred to countering terrorism, bad actors, violent extremism, extremism, and hate, as well as blogposts with titles surrounding how platforms make policy decisions or regarding regulation, were read and then assessed upon reading, whether it was relevant for collection. The blogposts were then saved and the details of the blogpost were collected in a spreadsheet (e.g., the title, date published, author, URL).

The next stage was applying a content analysis to the dataset. This analysis was done inductively through the use of open coding to create categories and identify themes, as opposed to deductively, which would have tested existing categories and concepts (Elo, Kääriäinen, Kanste, Pölkki, Utriainen, and Kyngäs, 2014). Different themes emerged regarding the different strategies that the platforms mentioned using/implementing to counter terrorist content; the different terms that were used in discussing these efforts (e.g., terrorism, violent extremism etc.); the different ideologies, movements and groups that were discussed and addressed; the values that the platforms claimed were prioritized during the efforts to counter terrorist content (e.g., user safety, free speech etc.); and finally, responses to regulation. These

themes are discussed throughout this chapter regarding, firstly, a greater understanding of what each platform has reported in relation to their efforts to counter terrorist content, and secondly, in relation to the effect the findings may have on a platform's compliance with future regulation.

Difficulties and Challenges

Before discussing the findings, it is important to acknowledge the difficulties and challenges that tech platforms face in their efforts to counter terrorist content. Many of these problems were highlighted in a blogpost by Facebook but apply across the industry (Bickert, 2017a). The first challenge is the enormous scale of the platform. Many platforms have millions or even billions of users on the platform every day, creating posts in dozens of languages, and these posts include text, photos, video, audio and livestreaming. As a result, platforms require a range of different technologies that can identify violating content across all of these different types of content and languages (Bickert, 2017a). Sometimes, technology is developed and employees are trained to use it, and then, terrorist organisations learn to adapt and find loopholes (Van der Vegt, Gill, Macdonald and Kleinberg, 2019). Platforms are therefore constantly having to adapt and innovate with technology.

Another issue is that these platforms were not created with the intention to counter terrorist or any other violating content in mind. Many platforms did not foresee the extent to which they would be exploited by such organisations and therefore did not readily have the expertise required to identify and remove such content. Platforms now have a long list of violating content that they must remove, many of which require whole teams of experts (e.g., terrorism, sexual exploitation, bullying, animal cruelty etc.). Reviewing flagged content is also an enormous undertaking requiring a large number of human content reviewers. Content reviewers only have a short time to accurately identify if the content is a credible threat or likely to inspire violence even though it is not always clear (Bickert, 2017a). Sometimes words are ambiguous, the intent behind them is unknown, or the context around the post is unclear, and as language evolves, words that were not previously slurs become slurs (Allen, 2017). Regional and linguistic context can be critical in making decisions about content (Allen, 2017). Platforms also have the difficulty of following local laws or facing threats from governments of being banned in that country. Content may be legal in one country but illegal in another (Bickert, 2017a). Different countries also have different definitions of terrorism. This raises the question, should a platform use an existing definition and, if so, which? If not, does the platform

have sufficient expertise to create its own definition? There is a similar problem around designated terrorist lists (Meserole and Byman, 2019).

Removing terrorist content is therefore an enormous undertaking for platforms with many complex difficulties and challenges. Twitter once said that there is no "magic algorithm" for removing terrorist content from their platform and as such, they have to make difficult judgement calls with limited information (Twitter, 2016). Google also described the challenge as "the haystacks are unimaginably large and the needles are both very small and constantly changing" (Walker, 2017). Facebook used a quote from the Irish Republican Party to demonstrate and summarise the difficulty of countering terrorist content: "Today we were unlucky, but remember that we only have to be lucky once – you will have to be lucky always" (Bickert, 2018b).

Global Internet Forum to Counter Terrorism

Before the chapter presents a brief overview of each platform, it is important to give a brief overview of one the most relevant and well-known voluntary collaborative initiatives that several of the platforms in this sample are founding members of. Similar to the platforms when they work alone, this collaboration has collectively made efforts, faced challenges and received criticism that must be considered in any future regulation.

In 2016, Facebook, Twitter, YouTube and Microsoft joined forces in the battle to remove terrorist content from their platforms and reduce the replication of efforts and resources that was taking place among them. In December 2016, before the GIFCT was officially formed the four platforms announced that they were creating and sharing an industry database of "hashes" which are unique digital fingerprints (Facebook, 2016). When one platform removes a piece of content that they deem to be terrorist content, the hashes from that content will be added to the shared database so that the other platforms can then use those hashes to identify and remove the same piece of content from their own platforms if they deem it to violate their policies also. This is therefore not an automatic process; each platform continued to work in line with their own policies and definitions of terrorist content (Facebook, 2016). Other platforms were then later granted access to the database (Facebook, 2017).

Six months later in June 2017, the four platforms went a step further and created the Global Internet Forum to Counter Terrorism (GIFCT) (Facebook, 2017). The forum builds on work

undertaken by the EU Internet Forum[6] (which is discussed in chapter 4) (Facebook, 2017). The GIFCT currently sets out three main pillars that it is centered on (GIFCT, 2021). The first is equipping platforms and CSOs with the tools and knowledge to develop sustainable programs to disrupt terrorist and violent extremist activity. The second is to bring key stakeholders together to mitigate the impact of a terrorist or violent extremist attack. Finally, to support research that could aid understanding and responding to terrorist or violent extremist content. Since this time, other platforms have been added to the forum. A list of members can be found on the GIFCT official website. [7] In 2019, GIFCT announced that it was becoming an independent organisation led by an Executive Director (GIFCT, 2019) and announced in 2020 that it would include an Independent Advisory Council composed of government representatives and civil society members, and finally would have a series of Working Groups to allow stakeholders to raise issues that the forum should address (Bickert, 2019). GIFCT often work in partnership with Tech Against Terrorism and the Global Network on Extremism and Technology (GNET) to fulfil their objectives (GIFCT, 2019).

The collaboration was initially viewed positively. Before this, platforms were all expending time and resources fighting the same content. The ability to share resources not only reduces this but also allows smaller platforms who lack the capacity to undertake these efforts to benefit from shared tools, technology and expertise. However, this collaboration has also received criticisms. One of the criticisms is that, particularly with the sharing of the database, when biases or errors occur in one platform's systems, it could easily find its way into other platforms without users ever knowing (Douek, 2020). Additionally, once a member of the GIFCT, platforms may be able, to some extent, to avoid individual blame and accountability, falling back on the argument that their decisions and processes are in line with the Forum. This is even more problematic given the lack of transparency that users have regarding decisions made within the GIFCT (Douek, 2020, Llansó, 2019). When smaller platforms do adopt the processes of the larger platforms in such initiatives it arguably gives the larger platforms more power because they become the deciders of not only what is allowed on their platform but are setting the standards for other platforms also (Doeuk, 2020). Particularly where smaller platforms lack the capacity to go through each piece of content or decision individually to assess whether they come to the same conclusion as the larger platforms (Llansó, 2019). Llansó (2019) argues that the collaboration is one way that the platforms can try to avoid government regulation. Overall,

[6] https://www.internetforum.eu/
[7] https://gifct.org/membership/

the lack of oversight into the Forum creates a myriad of concerns regarding transparency and accountability and raises questions about whether the processes take a human rights approach. However, as a result of the benefits such a collaboration can have, particularly on smaller platforms, perhaps independent audits could help the GIFCT to overcome the issues and concerns that have been raised (Douek, 2020; Llansó, 2019).

Overview of Tech Platforms

This section will now provide a brief overview of each of the platforms in this research.

Facebook

Formed in 2004, Facebook is the most widely used tech platform and as of 2020 had 2.7 billion users (Statista, 2020a). It was early 2014 when Western intelligence agencies began to express concern about the Islamic State's (IS) use of platforms such as Facebook (Waters and Posting, 2018). Facebook have long had what they term 'Community Standards' where they outline what content and activity is and is not accepted on the platform. Facebook claims that one of the reasons they have Community Standards in place is because they know that people will not go to Facebook if it is not a safe place (Facebook, 2018b). As of January 2021, the community standards read,

> "We recognize how important it is for Facebook to be a place where people feel empowered to communicate…Our policies are based on feedback from our community and the advice of experts in fields such as technology, public safety and human rights. To ensure that everyone's voice is valued, we take great care to craft policies that are inclusive of different views and beliefs, in particular those of people and communities that might otherwise be overlooked or marginalised…The goal of our Community Standards has always been to create a place for expression and give people a voice…building community and bringing the world closer together depends on people's ability to share diverse views, experiences, ideas and information…" (Facebook, 2021).

Facebook further speak of prioritizing a commitment to the following specific values: voice, authenticity, safety, privacy and dignity (Facebook, 2021).

In relation to terrorist content, the Community Standards prohibit "language that incites or facilitates serious violence", as well as content that glorifies violence (Facebook, 2021). Content is removed when it is believed to be a credible threat and that there is a risk of "physical

harm or direct threats to public safety" (Facebook, 2021). Under its Dangerous Organisations policy, Facebook do not allow "organisations or individuals that proclaim a violent mission or are engaged in violence" including those involved in terrorist activity and organized violence and hate, nor does Facebook allow support or praise for those involved in such activities (Facebook, 2021). Facebook further prohibits organizing and promoting activities that intend to harm people, businesses, property and animals. In addition to terrorism, Facebook prohibit hate speech, which the platform defines as a "direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability" (Facebook, 2021). An attack is defined as violent or dehumanizing speech, harmful stereotypes, statements of inferiority and calls for exclusion and segregation (Facebook, 2021). Facebook introduced efforts to address "white nationalism" and "white separatism" in their policies in March 2019.

Twitter

Twitter was founded in 2006 and as of 2020 has 330 million users (Lin, 2020). Twitter created what they have termed the 'Twitter Rules' which state that,

> "Twitter's purpose is to serve the public conversation. Violence, harassment and other similar types of behaviour discourage people from expressing themselves, and ultimately diminish the value of global public conversation. Our rules are to ensure all people can participate in the public conversation freely and safely" (Twitter Rules, 2021).

The Twitter Rules prohibit threats of violence and the glorification of violence. It has a Violent Organisations policy which includes both affiliating and promoting terrorist and violent extremist groups that are designated on national and international terrorist designation lists (Twitter Rules, 2021). In addition to this the Twitter Rules prohibit targeting or inciting harassment and abuse of someone because Twitter believes in "freedom of expression and open dialogue, but that means little as an underlying philosophy if voices are silenced because people are afraid to speak up" (Twitter Rules, 2021). The Twitter Rules promote an idea of "healthy dialogue" although it is not necessarily clear in the Twitter Rules what this means. Further to the above, the Twitter Rules also prohibit 'Hateful Conduct' which is violence or an attack threatened on someone because of their race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability or serious disease.

Hateful imagery is also prohibited from profile details (e.g., profile pic, username, bio etc.). Twitter explains its rationale is "free expression is a human right" (Twitter Rules, 2021).

YouTube and Google

Google-owned YouTube launched in 2005 and was bought by Google in 2006. As of 2020, it has approximately 2 billion users (Mohsin, 2020). YouTube have 'Community Guidelines' which are "developed in partnership with a wide range of external industry and policy experts, as well as YouTube creators. New policies go through multiple rounds of testing before they go live to ensure our global team of content reviewers can apply them accurately and consistently" (YouTube, 2021a). These guidelines aim to "strike a balance between keeping the YouTube community protected and giving everyone a voice" (YouTube, 2021a).

YouTube has a 'Violent Criminal Organisations' policy which prohibits content that praises, promotes or aids violent criminal organisations including terrorist organisations (YouTube, 2021a). This includes insignia, logos and symbols affiliated with such organisations. The policy states "If posting content related to terrorism or crime for an educational, documentary, scientific, or artistic purpose, be mindful to provide enough information in the video or audio itself so viewers understand the context" (YouTube, 2021a). The 'Violent or Graphic Content' policy prohibits inciting violent acts. Finally, YouTube have a 'Hate Speech' policy whereby content is prohibited if it promotes violence or hatred toward others based on age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims and kin of a major violent event and veterans (YouTube, 2021a). Four principles that YouTube published are: *remove* content that violates policies as quickly as possible; *raise* up authoritative voices regarding news and information; *reward* trusted eligible creators and artists; and *reduce* the spread of content that brushes right up against the policy line (YouTube, 2019).

Microsoft

Microsoft was created in 1975 and although it is not a tech platform in the same sense as Facebook, Twitter and YouTube, it is one of the founding members of GIFCT and therefore contributes significantly to the fight against terrorist use of tech platforms. Microsoft own a number of services such as LinkedIn and Skype. Microsoft write on their website that they "recognize that we have an important role to play to curtail use by terrorists and terrorist organisations of our hosted community service" (Microsoft, 2021a).

Gab

Gab, commonly classed as an 'alternative platform' was established in 2016 and as of 2020 had over 1.1 million users. However, after the events surrounding the deplatforming of former United States President Donald Trump and platforms such as Parler, Gab has reported a steadily increasing number of new users in early 2021 (Brandt and Dean, 2021). Gab has gained a lot of attention for hosting content from the Pittsburgh Synagogue attacker (Archer, 2018) and is associated with many radical right public figures (Bennett, 2018). The tagline on the Gab homepage as of July 2019 reads that Gab is "A social network that champions free speech, individual liberty and the free flow of information online. All are welcome." (Gab Homepage, 2019).

> "Gab is a free speech software company. We build open-source software with the primary purpose of conserving and exporting the uniquely American value of free speech to the world online. Our most popular product is Gab.com, our social media platform with over one million users from around the world. On Gab the rules are simple: if it's legal speech, it's allowed. Illegal content and activity is not allowed" (Torba, 2019).

Gab has been responding to the issue that they, on a number of occasions, "have been no-platformed from dozens of service providers, domain registrars, payment processors, and even banks. We've overcome these challenges by building our own infrastructure and by leveraging censorship-resistant technology" (Torba, 2019). This has been a result of the content that they allow to remain on their platform. Gab views itself as different to the major platforms discussed so far. Their CEO stated that, "the rise in the censorship of American citizens by American corporations and the subversion of democracy by those same US corporations is harrowing to me as a Christian American. Silicon Valley companies do not share my Christian values and indeed have a consistent history of silencing Christians online" (Torba, 2019).

In 2019, Gab's 'about' webpage stated:

> "We believe that the future of online publishing is decentralized and open. We believe that users of social networks should be able to control their social media experience on their own terms, rather than the terms set down by Big Tech…Gab's codebase is free and open source…you, the user, have a choice when using Gab Social: you can either have an account on Gab.com, or if you don't like what we're doing on Gab.com or simply want to manage your own

> experience, you can spin up your own Gab Social server that you control, that allows you to communicate with millions of users on their own federated servers from around the world, including users on Gab" (Gab, 2019a).

Gab has a policy termed "Prohibited Uses" that states that the platform cannot be used in any way that violates federal, state or local law of the United States of America or is not protected by the First Amendment (Gab, 2021a). This policy further states that users cannot "engage in any other conduct which, as determined by us, may result in the physical harm or offline harassment of the Company, individual users of the Website or any other person (e.g., "doxing"), or expose them to liability". It also has "Content Standards" that state user contributions must not aid, abet, assist, counsel, procure or solicit any attempt of an unlawful act. Unlawfully threatening or inciting lawless action is also prohibited. It states that,

> "Although our Content Standards, following the First Amendment, do not proscribe offensive speech, we strongly encourage you to ensure that your User Contributions are cordial and civil. The foundation of a free society requires people to peacefully settle their differences through dialogue and debate. Gab exists to promote the free flow of information online. It is our view that the responsible exercise of one's free speech rights is its own reward and, as a general rule, the most well-respected online publishers tend to be the ones who behave most civilly and put forward their arguments most intelligently" (Gab, 2021a).

Regarding monitoring and enforcement, Gab states that they "strive to ensure that the First Amendment remains the Website's standard for content moderation. We will make best efforts to ensure that all content moderation decisions and enforcement of these terms of service does not punish users for exercising their God-given right to speak freely" (Gab, 2021a). Gab further states that they do not review material before it is posted on the platform and "cannot ensure prompt removal of unlawful material after it has been posted" (Gab, 2021a).

Telegram

Telegram was established in 2013 and as of the start of 2021 has approximately 500 million users (Cuthbertson, 2021). Telegram describes itself as a "cloud-based mobile and desktop messaging app with a focus on security and speed" (Telegram, 2021a). Prior to the finding that Telegram was used in the planning of the Paris attack in 2015, Telegram's creator Pavel Durov showed very little concern with countering terrorist use of his site. He stated that in his eyes,

privacy was more important than such issues such as terrorism (Counter Extremism Project, 2017). However, after the Paris attack, Telegram somewhat changed its stance. Telegram does not have lengthy set of community standards or guidelines but state in the Terms of Service that users must not "promote violence on publicly viewable Telegram channels, bots, etc." (Telegram, 2021b). On the Telegram FAQ webpage, the answer to the question "There's illegal content on Telegram. How do I take it down?" is,

> "All Telegram chats and group chats are private amongst their participants. We do not process any requests related to them. But sticker sets, channels, and bots on Telegram are *publicly available*. If you find sticker sets or bots on Telegram that you think are illegal, please ping us at abuse@telegram.org. You can also use the 'report' buttons right inside our apps…" (Telegram 2021c).

The FAQ page also states that whenever Telegram receives a report via its relevant email accounts that public content is thought to be illegal, Telegram undertakes "the necessary legal checks and take it down when deemed appropriate", however,

> "this does not apply to local restrictions on freedom of speech. For example, if criticizing the government is illegal in some country, Telegram won't be part of such politically motivated censorship. This goes against our founders' principles. While we do block terrorist (e.g., ISIS-related) bots and channels, we will not block anybody who peacefully expresses alternative opinions" (Telegram, 2021c).

Due to Telegram's hard stance on privacy and security, governments have forced them into taking action with threats of banning the app in their country. The first was the Russian government in June 2017 because of the apps role in enabling terrorist attacks (Tan, 2017). The Indonesian government also blocked the web version of Telegram in 2017 and threatened to block access to the app version due to its role in enabling terrorists to carry out attacks (BBC, 2017). In response Telegram removed all public terrorist-related channels that the government had reported to them (BBC, 2017).

Platform Size and Income

Before discussing the blogpost findings, there are some notable differences between the platforms. These include the platform size, the ways in which the platform earns revenue and

an overview of their publishing of blogposts which is the main way that they communicate their efforts to counter terrorist content with their users.

The platforms range in size. When referring to size this could be defined by either the size of the userbase or the number of employees a platform has. Platforms also differ in how they earn revenue. The capacity a platform has to counter terrorist content (or any other violating content) on its site will likely be affected by both the size of the userbase, number of employees and ability to earn revenue. Table 1 shows that there is a large variance across the platforms for both number of users, number of employees and method of earning revenue.

Table 1. Platform size and income

	Number of users	Number of employees	Primary Source of Revenue
Facebook	2.7 billion[8]	44,500[9]	Advertisements[10]
Twitter	330 million[11]	4,600[12]	Advertisements[13]
YouTube	2 billion[14]	10,000[15]	Advertisements[16]
Microsoft	Unknown[17]	163,000[18]	Office Consumer, Devices, Gaming, and non-volume licensing of Windows operating system[19]
Gab	1.1 million[20]	"has only a handful of employees"[21]	Donations, selling premium accounts and merchandise[22]
Telegram	500 million[23]	330[24]	N/A[25]

[8] Statista (2020a)
[9] Statista (2020b)
[10] Johnston (2021)
[11] Lin (2020)
[12] Twitter Investor Relations (2019)
[13] Reiff (2020)
[14] Mohsin (2020)
[15] Owler (2021a)
[16] Beattie (2020)
[17] Microsoft has various different services and companies (e.g., Microsoft Office, Microsoft Teams, LinkedIn etc.). It was difficult to find a total number of users for all of their services and companies.
[18] Statista (2020c)
[19] Visnji (2019)
[20] Brandt and Dean (2021)
[21] Timberg, Harwell, Dwoskin and Brown (2018)
[22] Zannettou, De Cristofaro, Sirivianos, Stringhini, Kwak and Blackburn, 2018
[23] Cuthbertson (2021)
[24] Owler (2021b)
[25] Telegram do not currently generate revenue (Iqbal, 2021).

People “often conceptualize Silicon Valley companies as behemoths with vast resources” (Fishman, 2019), but as seen in chapter 2, terrorist organisations exploit a diverse range of platforms, “the smallest of which can count their employees on one hand and do not have the resources to hire counter-terrorism specialists or dedicate large engineering and operational teams to counter-terrorism” (Fishman, 2019). Although not necessarily the case, it is likely that platforms with larger userbases could have much higher volumes of content being posted to their site every day. This becomes more complex as more countries and languages are represented in the userbase. Even with the use of technology, a large number of employees with a wide range of expertise is required to counter terrorist content on a platform, particularly a platform with an enormous volume of content across many languages and cultures. In order to attain the number of staff required to counter terrorist content, as well as other resources (e.g., technologies), a platform will likely require great and stable financial resources. Therefore, platforms that do not have stable ways to earn revenue may not always have the capacity required to finance what they need to counter terrorist content on their sites. Without significant financial and staffing resources, a platform may struggle to comply with regulatory demands to remove terrorist content, particularly, as seen in the previous chapter, when regulatory frameworks provide strict timeframes. These platforms may rely on other platforms sharing technologies and best practice which they may not be willing to do.

Blogposts

Table 2 displays the number of blogposts that were collected for each platform. The table reveals that some platforms publish blogposts that address terrorist content, regulation and policy changes more frequently than others.

Table 2. Number of blogposts collected for each platform

Platform	**Number of blogposts collected**
Facebook	64
Twitter	43
YouTube/Google	42
Microsoft	15
Gab	61

Telegram	7

Content Analysis: Thematic blogpost findings

Platform mission and values

The first theme that emerged during the analysis was platform values (the ideals that the platform aim to serve its users). Platforms repeatedly described the values that fuel the work that they claim to do both in fulfilling the mission of their platform and in countering terrorist content from their sites. While there are subtle differences, the major platforms all share missions along the lines of connecting people, giving people a voice and being a place where people can share ideas, however, at the same time keeping their users safe (see Table 3 below). The major platforms all repeatedly mention a wide selection of values, many of which are seen across the platforms. However, having such a long list of values raises some concerns and questions. For example, it is arguably difficult to ensure that such a wide array of values is implemented, particularly given that some of them are likely to compete with one another at times. For example, user safety and freedom of speech sometimes conflict. This suggests that neither value is absolute and therefore is likely context-dependent. This is demonstrated in platform policies; when something should be taken down (prioritizing user safety), and when something can remain (prioritizing free speech). Some platforms also have their own 'buzz' words that are particularly prominent on their platform. For example, Twitter repeatedly discusses prioritising the promotion of 'healthy' conversation and trying to maintain the 'health' of the conversation on their platform, however, it is not clear what this means or how it could be measured.

The differences in platform missions and values become starker when examining Telegram and Gab. Telegram has a much more niche mission and set of values based around privacy and security. Telegram do not have the issue that the major platforms have of balancing potentially competing values and therefore it is likely much clearer to their users what they can expect from the platform. Gab, on the other hand, communicates many values. Some of these are different to the major platforms, for example, "American values", "Western values", "patriotism", "anonymity" and "individual sovereignty". Whilst some of these values are also emphasised by the major platforms, Gab's interpretation of what they require is distinct. For example, freedom of speech. When the major platforms discuss valuing freedom of speech, it is clear, as mentioned, that this is not absolute. Freedom of speech sometimes competes with

other values, for example, user safety and will not always take precedence (evidenced by their policymaking decisions and implementation of removal/enforcement mechanisms). Gab, however, states that it will prioritise freedom of speech unless a piece of content is unlawful under the First Amendment. Additionally, the CEO of Gab mentions his Christianity in the blogposts and in the terms of service refers to users "God-given right to speak freely" (Gab, 2021a). None of the other platforms in this research refer to religion in their blogposts or terms of service in this way.

Therefore, these findings reveal that in a small sample of platforms, there are many similarities as well as stark differences regarding the reported platform missions and the values that the platforms report using to guide their efforts and services, as well as how these affect a platforms likeliness to remove content. It is likely that these differences are even greater in the very diverse ecosystem of platforms that are exploited by terrorist organisations.

Table 3. Platform mission and values

	Mission statement	Values discussed repeatedly in blogpost
Facebook	"Facebook's mission is to give people the power to build community and bring the world closer together. People use Facebook to stay connected with friends and family, to discover what's going on in the world, and to share and express what matters to them"[26]	user safety freedom of speech transparency connecting people privacy fairness accountability diversity authenticity civil rights giving voice human rights
Twitter	"The mission we serve as Twitter, Inc, is to give everyone the power to create and share ideas and information instantly without barriers. Our business and revenue will always follow that mission in	giving voice freedom of speech accountability transparency user safety civility diversity privacy

[26] Facebook Investor Relations (2019)

	ways that improve – and do not detract from – a free and global conversation."[27]	openness connect people authenticity human rights healthy conversation public conversation
YouTube/ Google	"Our mission is to give everyone a voice and to show them the world. We believe that everyone deserves to have a voice, and that the world is a better place when we listen, share and build a community through our stories."[28]	Tolerance user safety freedom of speech diversity giving voice connect people transparency openness accountability human rights civil rights accountability privacy fairness
Microsoft	"Our mission is to empower every person and every organisation on the planet to achieve more". [29]	trust transparency user safety human rights health digital civility respect dignity tolerance freedom of speech privacy kindness accessibility inclusion empathy respect
Gab	"A social network that champions free speech, individual liberty and the free flow of information online."[30]	Freedom religion patriotism democracy truth freedom of speech individual liberty safety connect people western values giving voice transparency fairness openness privacy American values human rights civil rights dignity anonymity civil liberties christian values

[27] Twitter Investor Relations (2021)
[28] YouTube (2021b)
[29] Microsoft (2021c)
[30] Gab (2021b)

		free flow of information individual sovereignty
Telegram	"Our mission is to provide a secure means of communication that works everywhere on the planet".[31]	trust security privacy democracy openness freedom of speech

Mention of key terms, groups and movements

A noticeable finding in the analysis of the blogposts was that the platforms in this research use a wide array of terms, sometimes interchangeably (incorrectly – for example, "terrorism" and "violent extremism"), or grouped with other issues (such as "extremism" and "hate") when addressing their efforts to counter terrorism. As well as specific terms such as "terrorists", vague terms, such as "bad actors", and broader terms, such as "dangerous organisations and individuals" were also used. Table 4 shows a list of terms that each platform used throughout the blogposts in the dataset. This, combined with some of the platforms creating their own definitions for what is identified as "terrorism" and "terrorist content", provide insight into how the platforms view and perceive terrorist content, particularly highlighting the challenges and complexity that platforms can face in identifying it. This highlights the need for platforms to have access to experts in the terrorist field. While some platforms may have the financial capacity to hire terrorism experts, other platforms may not, and will therefore have to seek access to the expertise via other avenues (e.g., capacity building organisations such as Tech Against Terrorism, a regulator, or other platforms who are willing to share or collaborate etc.)

The findings revealed that the ideologies, groups, movements and individuals that platforms mention or address in the blogposts is varied (see Table 4). Facebook and Gab mention a variety of ideologies, groups, movements and individuals. Twitter only mentions ISIS and YouTube only mention white supremacy and Nazi ideology. Neither Microsoft or Telegram mention any. This is surprising given that a range of ideologies and groups have been studied by scholars across these platforms, as seen in chapter 2. The context of these discussions on the major platforms was what the platforms were doing to counter these groups, the removal of the groups/individuals, or discussing which ideologies and movements would be included in the policies. Although the major platforms often claim to remove a large number of terrorist

[31] Telegram (2021c)

organisations and a huge volume of terrorist content, neither YouTube, Twitter, or Microsoft mention an overly wide variety of ideologies, groups, movements or ideologies in their blogposts. This does not mean, however, that other efforts are not taking place. As mentioned in the methodology section, platforms may not report all of the decisions and efforts that they are making in this area. Gab, on the other hand, does not mention addressing terrorism very often, however often refers to, for example, terms such as hate speech in quote marks (e.g., "hate speech"), signalling a lack of belief in the term or a lack of respect for the term. Questions are raised in the Gab blogpost such as what is hate speech and who gets to decide this. The blogposts explain that Gab does not remove "hate speech", "hate organisations" and "hate groups" because such groups or speech are not illegal under the First Amendment. Regarding attacks, Telegram is the only platform that does not mention any, and the Christchurch attack is the only one that is discussed/addressed by each of the other five platforms. The platforms tend to publish in the blog any involvement that their services had with the attacker/s, and what the platform is doing in response. It is unknown why platforms would address some attacks and not others.

Overall, the findings show that some platforms report more key terms in their blogposts than others. This highlights the complexity surrounding the terms and definitions that are used, and the need for guidance and expertise in this area. The findings also reveal that the ideologies, groups, movements and attacks that are mentioned or addressed in the blogposts varies greatly, however, it is unclear why this is. This could be related to expertise. The platforms may have employees who are experts in one ideology but lack expertise in others. It may be part of the PR exercise. Platforms may not want to report too many removals because they want to be seen as having a balance between safety and free speech. If platforms report and discuss all of their removals in their blogposts, they may be worried that they will lose users who are anti-censorship. Platforms may focus on ideologies and attacks that have the most publicity or virality. Overall, explanations for these findings are not clear and would benefit from future research.

Table 4. Key terms, groups, movements, ideologies, individuals and attacks that the platforms address in their blogposts

	Terms	Groups/movements/ideologies/individuals	Attacks
Facebook	Terrorism Violent extremism	White supremacy Separatism	Christchurch attack

	Hate organisations Dangerous individuals and organisations Hate speech Hate groups Extremist organisation	White Separatism Jihadists White nationalism Militant environmental groups Irish Republican Party Tommy Robinson Alex Jones ISIS Al Qaeda Britain First Arakan Army Myanmar National Democratic Alliance Army Kachin Independence Army Ta'ang National Liberation Army	Halle attack
Twitter	Terrorism Extremism Violent extremism Hateful conduct Hate speech Bad actors	ISIS	Christchurch attack
YouTube/Google	Terrorism Violent extremism Extremism Hate speech Hateful conduct Bad actors	Supremacy White supremacy Nazi ideology	Christchurch attack Charlottesville attack Strasbourg attack
Microsoft	Terrorism Violent extremism Extremism Hate speech Bad actors	None	Christchurch attack

Gab	Hate speech Political violence Terrorism Extremism Political extremism Hate groups Hate organisations Extreme organisation Dangerous individuals	Far-right White nationalism Anti-Semitism Neo-Nazi White supremacy Alt-right Fourth reich Alex Jones Laura Loomer	Charlottesville attack Christchurch attack Pittsburgh attack Dayton attack El Paso attack Halle attack Jacksonville attack
Telegram	None	None	None

Policymaking

Many platforms change their policies frequently given the constantly evolving nature of the internet, and adaptability and changing nature of terrorist organisations. Throughout the blogposts, Facebook discussed several key efforts that they have made regarding their policymaking. One example is the Civil Rights Audit that took place in 2018 after encouragement from the civil rights community, in order to advance civil rights on the platform (Murphy, 2019). The audit aimed to create a forum for dialogue between the civil rights community and Facebook, and examined key areas of concern. The audit was led by Laura Murphy who is a civil liberties leader.[32] The work of the audit was undertaken by the Audit Team which included individuals from a civil rights law firm. During the first 6 months of the audit, Murphy conducted interviews with over 90 civil rights organisations to discover what the main issues are. The overall goal of the audit was to ensure that important civil rights laws and principles are respected and inform the work that is happening at the platform.

As a result of the audit, Facebook expanded its policies against white supremacy and hate to explicitly ban praise, support and representation of white nationalism and white separatism. It additionally updated its violence and incitement policies to prohibit posts from users who made it clear that they intend to use weapons to intimidate or harass others, or encourage others to do so. Facebook also adopted a policy designed to stop attempts of users to organise events

[32] https://lwmurphy.com/about.html

that aim to intimidate or harass a targeted minority or vulnerable group. Finally, the audit made Facebook aware of concerns with removal errors. An investigation was then conducted examining examples of when this occurred and revealed several changes that could be made to review tools and training (Murphy, 2019). The investigation found that there had not been sufficient attention to captions and context. Sometimes the tool that was being used did not display the caption. The audit suggested more training for human reviewers on how to consider context (Murphy, 2019).

Another key effort mentioned throughout the blogposts was that Facebook created the Data Transparency Advisory Group (DTAG) in 2018. They define this as

> "an independent body made up of international experts in measurement, statistics, criminology and governance. Their task was to provide an independent, public assessment of whether the metrics we share in the Community Standards Enforcement Report provide accurate and meaningful measures of Facebook's content moderation challenges and our work to address them" (Plumb, 2019).

In 2019, Facebook reported that the DTAG concluded that their approach to content moderation was adequate given the scale and amount of content. They also found that the way Facebook audit the accuracy of their content review is well-designed. They found the metrics to be reasonable ways of measuring violations and in line with best practice. The groups made suggestions for Facebook to implement in the future.

Facebook frequently discussed the measures of harm that they use when enforcing these policies. Facebook argued that time-to-take-action is a less meaningful measure of harm than metrics that focus on the amount of exposure that content receives (Bickert and Fishman, 2018). Facebook argued that some content receives high numbers of views within minutes, whereas other content can remain up for a longer period with very little views and that is why it is important to use the metric of 'prevalence', prioritising content that is the most viral.

Finally, Facebook addressed policy changes as a result of high-profile attacks and incidents throughout the blogposts. For example, in the wake of the Christchurch attack, Facebook decided that the violation of policies, such as, the Dangerous Organisations and Individuals policy, will result in the user being restricted from being able to use Facebook Live (Rosen, 2019). Finally, Facebook announced investing $7.5 million in new research partnerships with

leading academics from three universities to improve image and video analysis technology (Rosen, 2019).

One of the main initiatives that Twitter discussed regarding policymaking throughout the blogposts was asking users for feedback on a policy before it became officially part of the Twitter Rules (Gadde and Harvey, 2018). Twitter claimed that it would ask for user feedback on a new policy to ensure that they had considered global perspectives and how the policy would impact different communities and cultures. The other main theme that occurred in the Twitter blogposts was how their policies were enforced against world leaders. This is thought to be a result of the prominent and controversial use of Twitter by the Former President of the United States, Donald Trump. Twitter found themselves under a lot of pressure to explain what happens if a world leader violates its rules. The first time this was done was in 2018 where they said that,

> "Twitter is here to serve and help advance the global public conversation. Elected world leaders play a critical role in that conversation because of their outsized impact on our society. Blocking a world leader from Twitter or removing their controversial Tweets would hide information people should be able to see and debate…we review Tweets by leaders within the political context that defines them, and enforce our rules accordingly" (Twitter, 2018).

However, in 2019, Twitter updated this with a blog explaining that "world leaders are not above our policies entirely" and provided a list of instances in which they may take action against a world leader, including the promotion of terrorism and threats of violence (Twitter, 2019). In 2021, Twitter became the first platform to permanently suspend former President Donald Trump (who was president at the time) due to the incitement of violence surrounding the Capitol Hill attack (Twitter, 2021a). This finding highlighted that some platforms face unique challenges that other platforms may not have to deal with to the same extent.

Throughout the YouTube blogposts, YouTube referred to four Rs regarding their policymaking. These were: *remove* ("content that violates our policy as quickly as possible"), *raise* ("up authoritative voices when people are looking for breaking news and information"), *reward* ("trusted eligible creators and artists"), and *reduce* ("the spread of content that brushes right up against our policy line)" (YouTube, 2019). Another focus in the YouTube blogposts was their Intelligence Desk which is a team that monitors the news, social media and user reports in order to detect new trends surrounding inappropriate content, and aims to ensure that

teams are prepared to address the content before it becomes a bigger problem (Kantrowitz, 2018). Finally, YouTube and Google blogposts discussed policy decisions to be stricter regarding content that does not *clearly* violate their policies, for example, inflammatory religious or supremacist content (Walker, 2017). Instead of removing such content, a decision was made that the content will receive a warning and will not be able to be monetised, recommended or allowed to have comments or user endorsements. This decision was made in order to make the content more difficult to find and limit the engagement that it can receive.

Neither Microsoft nor Telegram had overly prominent themes in their blogposts regarding their policymaking process. Gab, on the other hand, rather than discussing efforts like the major platforms did, discussed throughout the blogposts the fact that they do not have such restrictive policies and therefore, aside from content that is illegal under the First Amendment, will not remove content or infringe on anyone's free speech. Gab frequently criticised other platform's policies. These findings reveal again, the differences across the platforms. It highlights that some platforms report putting in more effort than others (however, this should be taken with caution), that platforms can face different unique challenges, and can prioritise different approaches.

Technologies

The use of technology to remove terrorist content was mentioned frequently throughout the blogposts of Facebook, Twitter and YouTube. Facebook explained that some of this technology removes terrorist content at the point of upload, therefore resulting in the content not being able to reach the platform at all, while other technology proactively tries to find violating content on the site (Bickert, 2017b). Facebook claim that their technology works across a range of languages and reported in November 2017 that 99 percent of Islamic State and al Qaeda related content was removed by their technology before receiving any human flags or even reaching the platform (Bickert, 2017c). Facebook further claim that once a piece of content is detected, they are able to remove 83 percent of successive copies of that content within one hour of upload (Bickert, 2017c). In November 2018, Bickert and Fishman wrote in a blogpost that they are now using machine learning to assess posts that may signal support for ISIS or al-Qaeda. The technology creates a score that indicates how likely it is that the content violates their policies. This allows reviewers to prioritize posts that have the highest scores. Content that has extremely high scores can also be removed automatically. Facebook discussed how some of their technology works in the blogposts, for example, one technology they discussed was

‘image matching’. They explained that image-matching is a process that compares uploading content with a database of previously removed terrorist-related content and if there is a match then the uploading process is interrupted (Bickert, 2017b). This means that a lot of terrorist content is unable to reach the site (Bickert, 2017b).

In 2019, Facebook reported in the blogposts that the use of technology had led to the removal of more than 26 million pieces of content that was related to global terrorist groups, such as ISIS and al-Qaeda, in the previous two years. Further, 99% of this content was proactively identified and removed before it was flagged by a human (Facebook, 2019a). The blogposts also discussed that Facebook expanded the use of technologies to a wider range of dangerous organisations. They reported banning more than 200 white supremacist organisations. In March 2019, after the Christchurch attack, Facebook addressed what they have in place to counter the abuse of their Livestream feature. They said that they use artificial intelligence to detect and prioritize livestream videos that are likely to contain harmful acts and have improved the context they provide to their reviewers. They also built systems to quickly help contact first responders (Rosen, 2019).

Twitter also mentions throughout the blogposts that it utilizes automated and artificial intelligence technology as well as the January 2021 transparency report revealing that 94 percent of the terrorist and violent extremist content removed was proactively identified (Twitter, 2021b). One of the main methods that Twitter has reported using to identify terrorist-related accounts is using spam fighting tools. This is useful when terrorist organisations attempt to manipulate something as ‘trending’ which is a strategy that the Islamic State have been known to try. Twitter also realized that its ‘who-to-follow’ algorithms unintentionally allowed terrorist accounts to easily find other terrorist accounts through its personalized suggestions on what accounts their users may wish to follow (Berger and Morgan, 2015) and subsequently utilized these algorithms to identify and remove such accounts (Twitter, 2016). Twitter does not go into as much detail as Facebook regarding how the technology works.

Another theme that emerged in the Twitter blogposts was that Twitter has also been creating technologies that users can use to manage their user experience. For example, in 2019, Twitter published a blogpost that said users were going to be able to have more control over their conversations (Xie, 2019). In addition to all the user controls that Twitter (as well as some other platforms) already have, for example, allowing users to filter, block and mute content and accounts that they do not wish to see or engage with, Twitter was implementing technology

that would allow users the option to hide replies to their tweets. They said that they would roll this out in one country and get user feedback before rolling it out globally. The blogpost said that Twitter "will be looking at how this feature gives more control to authors while not compromising the transparency and openness that is central to what makes Twitter so powerful". Twitter's research into this reveals that people mostly hide replies that they consider to be irrelevant, off-topic or annoying; it is a new way to shut out noise – most of the people using this feature are not engaging in blocking or muting; public officials have not largely been using this feature; in Canada, 27 percent of users who had their Tweets hidden said that they would reconsider how they interact with other users in the future (Xie, 2019). Some people, however, said that they would not use this feature because of the fear of backlash that they had used it. Therefore, Twitter discussed the technology they employ but also discussed their efforts to extend the range of technologies that their users could use to moderate for themselves if they find content and accounts that they do not want to see.

The Google and YouTube blogposts also discussed the use of technology. They discussed similar image-matching technology to that discussed by Facebook, technology that prevents re-uploads of content that they have already removed, and comment moderation tools. As with Facebook again, these blogposts gave some detail as to how the technology works, for example, through the use of content-based signals. This highlights that some platforms are being more transparent than others regarding the technology that they are using. In 2017, YouTube reported that 98 percent of the terrorist content being removed was removed by this technology before receiving any human reports (Wojcikci, 2017). In 2017, YouTube reported that the volume of content that had been removed since the implementation of machine learning technology would have required 180,000 employees working for 40 hours a week (Wojcikci, 2017). Similar to Twitter, YouTube also discussed technology that users could use to review and moderate comments on their videos as opposed to reporting it or waiting for it be flagged by technology.

Microsoft has taken a more reactive approach. While the other major platforms have both proactive and automated technologies, and user reporting mechanisms, Microsoft relies primarily on user reporting mechanisms. Microsoft have a web form for users "to report content posted by or in support of a terrorist organisation that depicts graphic violence, encourages violent action, endorses a terrorist organisation or its acts, or encourages people to join such groups." Microsoft also have a "notice and takedown" process whereby either users or governments who find content that they believe violates Microsoft policies can report it and Microsoft will review it and remove it if a violation has occurred (Microsoft, 2016). However,

Microsoft invested funding and technical support to Dartmouth College to create PhotoDNA[33] which is technology that can scan the internet and flag violating content (although it should be noted that this was primarily developed to find images of child exploitation) (Microsoft, 2016).

Telegram does not discuss any implementation of the proactive technologies discussed by the major platforms in their blogposts. Telegram does, however, have an official Telegram channel that is used to report-Islamic State-related content removals to their users called @ISISwatch. This was created in 2016 and updates its followers every day with how many ISIS bots and channels were removed from Telegram on the previous day. Although not mentioned in their blogposts, Telegram have also worked with Europol's European Internet Referral Unit (EU IRU) to undertake Referral Action Days (RADS) which coordinates action that seeks to remove as much terrorist content as possible. In response to this collaboration, Europol said that Telegram,

> "has put forth considerable effort to root out the abusers of the platform both bolstering its technical capacity in countering malicious content and establishing close partnerships with international organisations such as Europol. Thanks to this collaboration, the already-existing content referral tools available to Telegram's users have been strengthened and expanded. Now, any user is able to refer and classify the content they find inappropriate and violent via the referral feature in public groups and channels. In addition, new technical developments, such as the advanced automated content detection system, continue to strengthen Telegram's effort in obliterating extremism on the platform even further" (Europol, 2019a).

Telegram's response is heavily focused around the Islamic State whereas other platforms mentioned aiming their technology at a wider array of terrorist organisations and ideologies. Gab do not mention using technologies to search for, identify or remove content in any of their blogposts. They state in their terms of service that, "we do not review material before it is posted on the Website and cannot ensure prompt removal of unlawful material after it has been posted" (Gab, 2021a).

Human review

[33] https://www.microsoft.com/en-us/photodna

Facebook frequently mentioned throughout their blogposts that technology is not yet at the stage where it can be used alone to counter terrorist content. Facebook reported that their algorithms are still not as good as humans at understanding the context behind some content, giving the example that some terrorist-related photos are used in news stories as opposed to being used to praise or support terrorism and AI struggles to tell the difference (Bickert, 2017b). In many instances, the technology flags the content but much of it is viewed by a human reviewer for the final decision. Human reviewers are also necessary to inform law enforcement when content or activity suggests a credible threat. Therefore, they make the point that human reviewers are crucial.

Facebook have reported that the overall number of employees on teams covering safety and security grew to 30,000 in 2019 (Facebook, 2019c). In 2016-2017, Facebook hired more than 200 terrorism and safety experts that encompassed a range of academic counter-terrorism experts, former prosecutors, former law enforcement agents and analysts, and engineers (Bickert, 2017b; Zuckerberg, 2018). They said that this was to bring their knowledge together and focus on countering terrorism (Bickert, 2017b). In April 2019 at the Home Affairs Committee oral evidence session, Facebook claimed to have a growing team of 15,000 content reviewers based in over 20 different offices globally, covering around 50 languages. Additionally, Facebook published some blogposts that address safety issues and the well-being of their human content reviewers.

YouTube also use a combination of machine learning and human content moderators to detect violating content (YouTube, 2021a). YouTube claim that their employees manually reviewed over a million videos in order to provide more examples of extremist content and improve the accuracy of their machine learning technology (YouTube, 2017a). In addition to human reviewers that are employed by the company, YouTube are known for their 'Trusted Flagger' program (YouTube, 2016). This came about when they noticed in 2012 that some users were very active in flagging videos for human review with very high accuracy. This led to the program which provides tools for those who are exceptionally interested in and highly accurate at voluntarily reporting content (YouTube, 2016). YouTube reported in a blogpost that their Trusted Flaggers have been reporting content that is accurate in over 90 percent of cases, which is three times more accurate than the average flagger (YouTube, 2016). Due to the success of this program, YouTube decided that they would create another program called YouTube Heroes to support the flaggers who help them make their platform a safer place (YouTube, 2016). Trusted Flaggers in the Heroes program have access to a community site which is

separate from the main YouTube site, where they can help each other and learn from each other. They are able to work their way through different levels by earning points and they can also unlock rewards (YouTube, 2016).

Although Twitter also briefly mentions using human content reviewers, the blogposts do not go into any detail about how many reviewers they have, how this work is undertaken or the well-being of the reviewers. Neither Microsoft, Gab or Telegram explicitly mention or discuss the use of human reviewers in their blogposts. However, some of the other strategies they discuss, such as user reporting mechanisms suggest that they do use human reviewers. Once again, this finding highlights that platforms vary in terms of transparency, also in what they prioritise communicating about in their blogposts, and in the moderation strategies they report using.

Collaborations, counter-speech and re-direction

All four of the major platforms discuss their participation in the GIFCT throughout the blogposts. As well as this, Facebook frequently mention their different efforts to form collaborations with others in the industry, NGOs, CSOs and academia, as well as efforts to support the use of counter-speech and other CVE strategies. Some examples include working with the Institute for Strategic Dialogue on an Online Civil Courage Initiative which responds to the proliferation of online hate, violence and terrorism across Europe with a range of efforts to help NGOs and CSOs create counter-speech (Bickert, 2017b). Another collaboration was with Affinis Labs to host hackathons around the world where community leaders and tech experts could brainstorm solutions to online extremism content (Bickert, 2017b). Further, Facebook partnered with the Edventure Partners to create the P2P Global Digital Challenge that gave university students the opportunity to create and disseminate online CVE campaigns (Bickert, 2017b). Finally, in 2019, Facebook published a blogpost explaining that they were starting to connect people who searched for terms associated with white supremacy with resources from an organisation called Life After Hate which tries to help people leave hate groups and provides crisis intervention, education, support groups and outreach (Facebook, 2019b). This was then reported as being expanded to Australia and Indonesia in partnership with Moonshot CVE to measure the impact of these efforts to combat hate and extremism (Facebook, 2019a). When people search for terms associated with hate and extremism, they will be directed to local organisations that help individuals leave the path of violent extremism and terrorism.

Twitter also discussed similar efforts throughout their blogpost. These collaborations include People Against Violent Extremism (PAVE) (Australia), the Institute for Strategic Dialogue (UK) (Twitter, 2016), Parle-moi d'Islam (France), Imams Online (UK), Wahid Foundation (Indonesia), The Sawab Centre (UAE) and True Islam (US) (Twitter, 2016). As well as this, Twitter has been involved in government-convened summits on CVE with Australia, the UK, France, the European Commission and the United Nations (Twitter, 2016). In a 2016 blogpost, Twitter announced that their Media team trained more than 300 Members of the European Parliament on their Twitter Coaching Program on counter-narratives and capacity building, and 60 members undertook one-to-one sessions to learn how to use Twitter as a tool for amplification (Twitter, 2016). This training involved learning how to create campaigns to counter hate such as the campaign NotInMyName, and also campaigns to unite communities in the aftermath of terrorist attacks such as #JeSuisCharlie (Twitter, 2016). The training allowed the members to ask questions, discuss their own campaigns and receive a copy of Twitter's NGO Handbook: *Campaigning on Twitter* (Twitter, 2016). In 2019, Twitter announced that it was launching a new handbook for educators called "Teaching and Learning with Twitter" (Costello, 2019). The handbook aims to help educators teach students media literacy skills which will enable them to think critically about the content that they engage with and information they consume. It includes best practice guidelines on media literacy from UNESCO and was published in 9 languages. Finally, throughout the blogposts, Twitter discussed the creation of their Trust and Safety Council. This was established in 2016 and brings together more than 40 experts and organisations to help advise Twitter on their products, programs and the Twitter Rules (Pickles, 2019). The members have expertise across media literacy, digital citizenship and grassroots advocacy organisations (Cartes, 2016). However, just over two years after its establishment, it came to light that Twitter stopped communicating with the council, leading to the council members reporting that they were unsure as to whether or not Twitter still valued their input (Matsakis, 2019).

Google and YouTube discuss efforts in this area as well. The Redirect Method is mentioned throughout the blogposts. Google's think tank Jigsaw partnered with Moonshot CVE (YouTube, 2017b) and used information obtained in interviews with ex-Islamic State members to create the Redirect Method (RM). How the RM works is, when an individual searches for terrorist content, they are redirected to a YouTube video that aims to discredit and debunk the terrorist content that the individual searched for. The RM was mainly targeted at those searching for content produced by IS. It used Google's advertising service Adword's targeting

tools to find the target audience. The videos that the individuals are redirected to are curated YouTube videos that have been uploaded by YouTube users worldwide in an attempt to counter radicalisation (Jigsaw, 2017). These videos mainly include "citizen testimonies, on-the ground reports, and religious debates" (Jigsaw, 2017). They do not include government-produced content in order to maintain credibility and allow those searching for IS content to be exposed to alternative views that they may not be exposed to otherwise (Jigsaw, 2017). Jigsaw tested the RM in an IS-focused campaign in both Arabic and English over the course of 8 weeks. The results found that during this period, 320,000 people watched over 500,000 minutes of the 116 YouTube videos that people were redirected to (Jigsaw, 2017). However, it is difficult to measure what impact this was able to have, if any. In 2017, YouTube stated that they plan to expand these search terms into more languages; use machine learning to update the search terms; work with NGOs to create more content with a diverse range of counter-radicalisation messages; and collaborate with Jigsaw to expand the RM across Europe (YouTube, 2017b). In 2017, Microsoft started a similar program on Bing. This has resulted in Jigsaw and Bing sharing best practices and expertise (Walker, 2017). Further to these efforts, Google reported in a blogpost that they would provide $5 million to a fund for creating tech-driven solutions and grassroots efforts to counter hate and extremism (YouTube, 2017a).

Microsoft also discussed efforts in this area. One example is their partnership with the Institute for Strategic Dialogue which worked on counter-speech when people search for terrorism on their search engine Bing which is similar to YouTube's Redirect Method (Microsoft, 2017). As with the RM, the videos used as counter-speech include testimonials of former terrorists and aim to deconstruct and discredit terrorist narratives. This was first piloted in the UK but then expanded to other countries and languages. Microsoft also discuss in a blogpost that they are a founding member and financial sponsor of a public-private partnership that develops activities to counter terrorist abuse of tech platforms. This brings together the United Nations Counter Terrorism Committee Executive Directorate, civil society, academics, government and others in the industry to work together (Microsoft, 2016). Finally, Microsoft discuss in their blogposts their work on media literacy. Microsoft have a YouthSpark Hub which provides access to online safety resources for young people. These "resources include material designed to help young people distinguish factual and credible content from misinformation and hate speech as well as tools for how to report and counter negative content" (Microsoft, 2016). Microsoft also have a Teen Council for Digital Good where teenagers are selected to join a

program where they are able to share their thoughts and feedback about the work Microsoft is doing to promote digital civility and online safety (Beauchere, 2017).

Neither Telegram nor Gab discuss such efforts in their blogposts. These findings reveal again that there are differences in what the platforms are reporting in the blogposts regarding their efforts in this area. While some platforms report implementing a range of different efforts in this area, others report much less.

Transparency reports

Facebook, Twitter and YouTube post blogposts fairly regularly that summarise what they have reported in their latest transparency reports. This includes discussing what is included in transparency reports as well as different statistics such as how much content is removed. In 2019, the first GIFCT transparency report was published involving Facebook, Twitter, YouTube and Microsoft. Neither Telegram nor Gab discussed anything regarding the publishing of transparency reports in their blogposts.

Attitude Towards Regulation

The final theme that emerged from the analysis was platform response and attitude towards regulation. This differed across the platforms. The major platforms discussed voluntary willingness to sign up for the EU Code of Conduct on Countering Illegal hate speech in their blogposts. Aside from this, Facebook, mentioned regulation throughout the blogposts. For example, one blogpost written by Mark Zuckerberg (CEO) said,

> *"I believe we need a more active role for governments and regulators. By updating the rules for the internet, we can preserve what's best about it – the freedom for people to express themselves and for entrepreneurs to build new things – while also protecting society from broader harms."*

Another blogpost read,

> *"Mark warned that we're increasingly seeing laws and regulations around the world that undermine free expression and human rights. He argued that in order to make sure people can continue to have a voice, we should: 1) write policy that helps the values of voice and expression triumph around the world, 2) fend off the urge to define speech we don't like as dangerous, and 3) build new*

institutions so companies like Facebook aren't making so many important decisions about speech on our own."

The majority of Facebook blogposts that fall under this theme were in relation to the creation of the Oversight Board.[34] The first of its kind, created after consultations with various stakeholders around the world (Harris, 2019), it is an independent board made up of 40 members from a diverse array of backgrounds and expertise. Facebook already had an appeals mechanism; however, it was not independent to the platform. The board members will select and review some of the most contentious content cases the platform has to respond to. The board is funded and supported by an independent company. Facebook claim that the board will be obligated to the people who use Facebook and not to Facebook itself (Clegg, 2019). The board will make decisions on the content that it reviews, provide explanations for the decisions and can provide recommendations for Facebook to consider regarding its policies (Clegg, 2019). In a report published by Darme, Miller and Steeves (2019), Facebook explains that they spent six months holding online and in-person consultations around how this board should be created. They heard back from more than 2,000 people from more than 85 countries including academics, grassroots organisers, and everyday people. They additionally hosted workshops all around the world. The results revealed that people wanted the board to exercise independent judgement and have a set of higher-order principles that are informed by free expression and international human rights law. Secondly, that the board should consult with experts with specific cultural knowledge, technical expertise, and an understanding of content moderation. Finally, people wanted a board that is as diverse as the userbase on Facebook.

The Oversight Board has been praised because it arguably provides users with more participation than they have ever had before regarding appeals (Klonick, 2019). Users will have more rights and clarity than they would if their content was to be removed by automated technology, particularly since they will be given a detailed explanation of why the decision was made (Douek, 2019). Subsequently, it is hoped that the board will allow users voices to be heard, and greater transparency and accountability from the platform regarding decisions. However, while the discourse in the blogposts appear supportive of an increase in regulation, Facebook's creation of the Oversight Board has been criticised by scholars as a way for Facebook to avoid government regulation whilst diverting responsibility and blame away from themselves with the added benefit of positive PR (Clegg, 2019; Douek, 2019; Klonick, 2019).

[34] https://oversightboard.com/

Further, although classed as independent from the board, Facebook did select the selectors of the members of the board (Sonnemaker, 2020) and could disband the board if they chose to (Douek, 2019). It was further highlighted that "by adopting a structure of government, [Facebook] is essentially admitting it has some of the powers of a government" (Roberts, 2019). Therefore, it is unclear exactly what Facebook's true motivations were for the Oversight Board and raises doubts about the platform's supportive stance on government regulation.

Twitter did not refer to regulation in their blogposts to the same extent, however, when they did, they did not seem as supportive of the idea as Facebook. For example,

> *"With the passage of new legislation and ongoing regulatory discussions taking place around the world about the future of public discourse online, we are seeing a potential chilling effect with regards to freedom of expression. According to Human Rights Watch, the wave of regulatory pressure in Europe and beyond is setting an emerging precedent and creating a "domino effect" as "governments around the world increasingly look to restrict online speech by forcing social media companies to act as their censors".*

Regulation was mentioned throughout the Google/YouTube blogposts. The platform appeared to have a supportive stance and mainly published suggestions as to how future government regulation could be improved in their opinion. For example,

> *"It's important for oversight frameworks to recognize the different purposes and functions of different services. Rules that make sense for social networks. Video-sharing platforms, and other services primarily designed to help people share content with a broad audience may not be appropriate for search engines, enterprise services, file storage, communication tools, or other online services, where users have fundamentally different expectations and applications."*

> *"We and other tech companies have pushed the boundaries of computer science in identifying and removing problematic content at scale. These technical advances require flexible and legal frameworks, not static or one-size-fits-all mandates. Likewise, legal approaches should recognize the varying needs and capabilities of startups and smaller companies."*

> *"It's important to start with a focus on a specific problem and seek well-tailored and well-informed solutions, thinking through the benefits, the second-order impacts, and the potential for unintended side-effects".*

The suggestions made by Google/YouTube align with the findings of the previous chapter of this thesis, which raised the idea that a one-size-fits-all regulatory approach is unlikely to be the most effective.

Microsoft did not mention regulation outside of the platform voluntarily signing the EU Code of Conduct mentioned previously and there was no mention of regulation from Telegram. Gab discussed regulation throughout their blogposts. Gab portrayed a stance that is completely against the idea of any kind of regulation. For example,

> *"…Congress is drafting a bill to create a federal social media task force to police speech on the internet and target pro-free speech websites like Gab. They are framing this around "protecting users from harmful online content," when in reality users already have all the tools they need as individuals to protect themselves. The federal government doesn't belong in the position of policing online speech anymore than Big Tech companies do".*
>
> *"What these lawmakers fail to realize is that the signal can not be stopped. The open source and decentralized version of the internet is already here. Anyone, anywhere in the world can now create and host their own Gab Social server with their own rules. We can't stop them. The government can't stop them. Big Tech can't stop them."*
>
> *"Facebook has offices around the world. Because of this Facebook has to bend the knee to foreign governments. Case in point, a new landmark privacy ruling from the EU Court of Justice could force Facebook to patrol its platform and remove all "offensive" and "hateful" posts by users in the EU…you should also be on Gab.com, which only has offices in the United States of America and only answers to American privacy and speech laws".*

Gab repeatedly reminded the readers of its blogpost that it is based solely in the United States and therefore only follows U.S. law, in particular, the First Amendment. It repeatedly stated that if content is lawful under the First Amendment then it is allowed on Gab.

These findings reveal that there are different stances ranging from platforms supporting regulation, to taking regulation into their own hands (Oversight Board), to criticising regulation, being unwilling to comply with regulation, as well as Telegram's lack of communication regarding regulation at all. If there are such diverse differences amongst these six platforms, then it is likely that there is an even bigger range of differences in the whole ecosystem of platforms that are exploited by terrorist organisations. Complete unwillingness to comply with any regulation will create enormous difficulties for regulators, particularly where there are jurisdictional issues that may require the regulator to have powers to implement significant incentives to comply and severe penalties in the face of non-compliance. Platforms that are willing to comply with regulation, however, face challenges that were raised earlier in this chapter, such as lacking the necessary resources to comply. These platforms may require different types of guidance and assistance in order to be able to fully comply.

Summary

This chapter aimed to answer the following questions:

4) What efforts to tech platforms report taking to counter terrorist content on their services in their blogposts?
5) What challenges do tech platforms face in their efforts to counter terrorist content on their services that could affect their compliance with regulation?

Overall, this chapter aimed to gain a deeper understanding of the work that a sample of tech platforms report that they take, in their blogposts, in response to terrorist use of their platform and identify the similarities, differences, and unique challenges that are faced by the platforms in this work. This also aimed to identify any potential regulatory compliance issues that could result from these differences and challenges. The main themes that arose from the analysis of the blogposts across the platforms were in relation to attitudes towards regulation, policymaking, implementing technologies, the use of human content reviewers, counter-speech, collaborations and CVE efforts, as well as the ideologies, movements and terrorist attacks that the platforms address. This chapter also examined the size of a platform's userbase, the number of employees it has, and the means of earning revenue.

There are several key findings to take away from this chapter. First, tech platforms are often accused of not doing enough to counter terrorist content on their platforms (see earlier example Theresa May (2018)). These accusations are often vague and assume that platform efforts are consistent across the enormous ecosystem of platforms that are exploited by terrorist

organisations. This chapter has revealed, that although there is still definitely more that needs to be done in this area, efforts to counter terrorist content are not being reported consistently across platforms in their blogposts. Some platforms have reported implementing a wide array of strategies, such as the use of many different technologies, hiring thousands of human content reviewers, collaborating with NGOs, CSOs, government, academia and others in the industry to exchange knowledge and best practice, as well as partnering on projects regarding counter-speech, media literacy and other CVE strategies. Other platforms, on the other hand, have reported doing more in some areas than others, or very little, if anything at all. There are also differences in how transparent platforms are willing to be, with some posting more blogposts and communication, and providing more detail about, for example, how the technology works. Finally, some platforms even report it as core to their mission and values that they will not proactively search for violating content. However, it must be noted that platforms may be doing work in this area and not reporting it in blogposts, or the blogposts may be more rhetorical than practical.

An important question to ask, regarding the findings, is why this is. Is this an issue around having the capacity to counter terrorist content? This chapter has demonstrated that platforms can widely differ regarding how they earn revenue and how stable or lucrative this is. This then snowballs onto whether a platform has enough financial resources to hire the number of employees that are required to build technology, provide the expertise that is required and manually review content. This chapter found that the platforms in this research differ greatly in the number of employees and users that they have, and that while the major platforms appear to have stable advertising revenues, other platforms do not. This supports the finding in the previous chapter that platforms may not all have the same levels of capacity and expertise to make efforts to counter terrorist content on their platform. However, there could also be other issues. For example, the findings that platforms report different attitudes towards regulation, as well as their mission and values, raise a question around willingness. Are platforms willing to make the effort that is required to counter terrorist content on their platform and comply with regulatory attempts to counter terrorist use of platforms? Regarding attitudes towards regulation, Facebook's creation of the Oversight Board leads to many that are cynical and critical of the platform questioning whether this a genuine move to aid in the removing of violating content whilst balancing other competing values, such as free speech, or whether this is an attempt to appear to be cooperating with the removal of violating content whilst avoiding government regulation. However, this sceptical perspective could be applied to the blogposts

generally, for example, are the blogposts a PR exercise, are they rhetorical? It is difficult to know whether the reports in the blogpost are accurate representations of the efforts actually being undertaken. However, they nevertheless provide value in how the platforms want to be perceived. Gab reports prioritizing free speech and do not appear to be willing to comply with any regulation. Gab will only remove content that is unlawful under the First Amendment. Telegram do not directly address how they would respond to regulation in their blogposts, however, communicate prioritizing privacy and security above all else. Another finding that was highlighted was that platforms can face unique challenges. An example is Twitter and their world leader policy. Such challenges can be unforeseen and are going to shape a platform's priorities in ways that other platforms may not have to deal with. Another example of unique challenges is Gab's issue of being no-platformed by service providers, domain registrars and app stores. This has led to Gab building their own versions of these services and becoming decentralised. This is going to create enormous issues for future regulators and most likely lead to regulators requiring tools such as ISP blocking (which is discussed in chapter 8).

Conclusion

In conclusion, the findings of this chapter have revealed that in a small sample of six platforms, there are, firstly, differences in the number of employees a platform have, the size of its userbase, and how it earns revenue. These factors could affect the capacity of a platform to make efforts to counter terrorist content on their services. Secondly, there are stark differences the following areas, based on what platforms report in their blogposts: the efforts and strategies that are used to counter terrorist content on their services; the ideologies, movements and attacks that the platform responds to and address; the platform mission and values; attitudes towards regulation; transparency; and the unique challenges the platform faces. This raises the issue that platforms may not be willing to comply with regulation that seeks to counter terrorist content online. Future regulation should acknowledge these findings and consider ways in which they could be minimized or overcome. The findings of this chapter were considered and influenced the development of the regulatory framework that is proposed later in this thesis. Limitations of the research in this chapter are that it is not known whether the information reported in the blogposts accurately reflect the work and efforts that the platforms are actually making to counter terrorist content on their services. It is unknown how much of what is reported in the blogposts is a PR exercise or rhetorical. However, despite these limitations, the blogposts provide insight into how the platforms wish to be perceived and what they want the public to know about their efforts and there is still much that can be taken from this dataset.

The dataset provides insight into the differences between the platforms and highlights unique challenges that are faced by each platform, both of which should be considered during the development of future research because the findings could be useful in trying to counter potential compliance issues that could arise.

Chapter 4: What has been done to counter online terrorist content to-date

Government responses

Introduction

This chapter will provide an overview of several key governmental responses to terrorist exploitation of tech platforms in recent years. This chapter aims to answer the following question:

6) What has and has not been effective in existing regulatory frameworks that seek to counter online terrorist content?

The chapter will identify what these frameworks have done well and where there are gaps and limitations. The chapter will end with a thematic summary of the key points raised and lessons to be learned. These points have informed the development of the regulatory framework that will be proposed later in this thesis. It is argued that new regulation should be informed by existing regulation and learn from its criticisms and limitations (Windholz, 2010). The regulatory framework proposed in this thesis aims to overcome these limitations where possible.

The responses examined in this chapter include Germany's Network Enforcement Act, Australia's Abhorrent Violent Material Act, UK Online Harms White Paper, as well as a number of European Commission responses and collaborative initiatives. This is not an exhaustive list of relevant government responses because there are too many to possibly cover in one chapter. The responses that are included were chosen because: 1) they were timely and topical at the time of writing; 2) they cover a variety of countries; 3) these countries all adhere to western democratic principles; 4) together they cover a wide variety of regulatory approaches; and 5) they received widespread attention and criticism due to their limitations which are valuable for the development of the regulatory framework that is proposed in this thesis. Additionally, some of these frameworks have influenced the creation of other frameworks, for example, NetzDG, highlighting the need for further study and to learn from their mistakes (York and Schmon, 2021).

Two things should be noted before beginning. The first is that the frameworks discussed throughout the chapter use a variety of terms to cover 'tech platforms'. This includes social media companies, social media platforms, social networks, hosting service providers, and intermediaries amongst others. This thesis uses the term 'tech platform', however, any of these

terms may be used in quoting from the frameworks that are discussed. Secondly, not all of the frameworks discussed seek to counter terrorist content specifically, some counter hate speech or extremist content. Despite not being concerned with terrorist content directly, there is still much to learn from the frameworks.

Regulatory challenges

Before the responses are examined, it is important to highlight the challenges and complexity of regulatory responses to terrorist use of tech platforms. As seen in the chapter 2, use of the internet is constantly and rapidly evolving, however, the law struggles to match this pace. Pre-existing laws are often insufficient to tackle the illegal online activity that takes place today (House of Lords (UK), 2019). This is particularly so surrounding terrorist exploitation of tech platforms and illegal online communication more generally. There are many great difficulties in countering such illegal activity. The sheer volume of online content and communication is an enormous structural challenge - to have a court decide the illegality of it all would be extremely impractical - and jurisdiction issues add to the complications of deciding the illegality of content (Schulz, 2018). Further, there is the issue that national laws vary in what they identify as terrorism and many of the tech platforms exploited by terrorists were established in the United States where there is strong protection for speech under the First Amendment (Schmitz and Berndt, 2018). There is the issue with the internet that content can become viral very quickly, reaching an enormous global audience and once posted it can be very difficult to ensure that it does not resurface elsewhere (Schulz, 2018). Moreover, people are able to see and access content and communication that they never would have without the internet (Ardia, 2009; Schulz, 2018). This is a problem because not everyone has the digital literacy skills to view this content critically (Schulz, 2018). This is further problematic because research has shown that people tend to be more disinhibited in their communication and responses to online activity than they are offline (Pötzsch, 2010).

Under private law in Germany, intermediaries are viewed as a “source of danger” because they provide the space for infringements, without which, the content in question could not be disseminated (Schmitz and Berndt, 2018, p.9). This is a view that other governments appear to have taken as well, for example, former UK Prime Minister Theresa May called on tech platforms on several occasions to take responsibility for facilitating the activities of terrorist organisations (Stewart and Elgot, 2018). While there is this argument that tech platforms must take responsibility for the consequences of providing a space that has become home to illegal

content, it has also been highlighted that such a response leaves the tech platforms more vulnerable to legal threats than the users who post the illegal content, leading to the problem that platforms may have a "fragile commitment to the speech that they facilitate" (Kreimer, 2006, cited in Ardia, 2009). Where regulation fails to consider this, there is a risk that platforms will over-block out of fear of facing penalties.

When a decision needs to be made regarding content removal,, a conflict arises between reviewing the content quickly enough that is does not become viral but not so quickly that mistakes are likely to be made and subsequently infringe on user's freedom of expression (Schulz, 2018). Since this is an issue that traditional legal instruments struggle to solve, many governmental responses have been to place demands on the tech platforms. It could be argued that tech platforms are better placed to remove terrorist content than state actors (Schulz, 2018; Ardia, 2009). Ardia (2009) argues that,

> "Intermediaries often are capable of exercising authority over wrongdoers who are otherwise unreachable because these wrongdoers are not capable of being identified, are beyond jurisdiction of the state, or are simply not amenable to legal pressure" (p378-379).

However, there is the opposing argument that giving tech platforms this responsibility makes the powerful even more powerful (Douek, 2020). This is increasingly problematic the more tech platforms are able to make decisions without the inclusion of independent oversight.

Whilst there are two legislative instruments, one in Europe and one in the United States, that protect tech platforms from being liable for the content posted to their sites by their users, this by no means absolves tech platforms of the responsibility of countering terrorist content on their site. The E-Commerce Directive 2000/31/EC states that intermediaries in Member States are protected from liability for illegal third-party content in cases where the intermediary does not have knowledge of the illegal activity and content. However, once aware of this activity or content, the intermediary must act expeditiously to remove or disable access to the content. The E-Commerce Directive (Article 5) also prohibits states from imposing general obligations on intermediaries to monitor their platforms for illegal third-party content. The E-Commerce Directive, however, is, at the time of writing, soon to be replaced by the proposed Digital Services Act due to the fact that the Directive was implemented before tech platforms became commonplace and is thus deemed outdated. However, at the time of writing, the Digital Services Act does not appear to fundamentally depart from the approach of the directive on

intermediary liability. The United States has a similar Act: section 230 of the US Communications Decency Act which also provides intermediaries with broad immunity from liability for user-generated content posted to their sites.

The final challenge is ensuring that all regulation is based on a human rights approach. The UN Human Rights Council (HRC) said in 2012 that the "same rights that people have offline must also be protected online". While international human rights law places a responsibility to protect human rights on States, it is recognised that private companies also have a responsibility to respect human rights (Article 19, 2017). This is especially difficult when different rights conflict with one another. For example, free speech and autonomy, or free speech and the prevention of harm.

Government responses

This chapter will now provide an overview of a number of relevant government responses and end with a thematic summary of the key lessons and challenges that are identified and considered in the development of the regulatory framework proposed later in this thesis.

1. Network Enforcement Act

Overview

Prior to the implementation of Germany's Network Enforcement Act (NetzDG), tech platforms had largely been encouraged to undertake self-regulation (Schulz, 2018). However, in 2017 the German government decided that self-regulation was not enough and that the platforms' responses were too slow and limited (Schulz, 2018). This view was supported by the research findings of a 2017 study undertaken by Jugendschutz.net (Cited in Schmitz and Berndt, 2018) which found widespread inconsistency across the major platforms Facebook, Twitter and YouTube. While YouTube was found to delete 90 percent of flagged criminal hate speech, Facebook deleted 39 percent and Twitter only 1 percent. The platforms also saw a rise in hate speech and disinformation across platforms after the 2015 decision to let over a million asylum seekers into Germany (Heldt, 2020). As a result, Germany created the Network Enforcement Act (NetzDG). It should be noted that the Act was implemented despite a number of experts at the judicial committee of the Bundestag expressing major concerns around the risk of platforms over-blocking and impacting freedom of speech (Schulz, 2018).

The NetzDG law came into force on 1 October 2017 to improve the enforcement of the law in social networks, with a specific focus on clamping down on hate speech and fake news

(Network Enforcement Act, 2017). The Act is aimed at "social networks" and applies to "telemedia service providers which, for profit-making purposes, operate internet platforms which are designed to enable users to share any content with other users or to make such content available to the public" (Network Enforcement Act, s.1(1)). This excludes news platforms that host journalistic content and platforms that have less than two million registered users in the Federal Republic of Germany. The 22 different statutes to which the Act applies include categories such as "incitement to hatred", "forming terrorist organisations" and "the use of symbols of unconstitutional organisations".

Under the Act, platforms must designate an employee responsible for ensuring the regulator can access the required information. One of the main aspects of the Act is that platforms must create "an easily recognisable, directly accessible and permanently available procedure" for users to submit a complaint about illegal content on their platform" (Network Enforcement Act, s.3(1)). Any "manifestly unlawful" content must be removed or blocked within 24 hours of the platform receiving the complaint, unless the platform has reached an agreement with the relevant law enforcement on an extension. Other "unlawful content" must be blocked or removed within 7 days of the complaint. Any social networks that fall under the scope of the Act that receive more than 100 complaints per calendar year of unlawful content must create bi-annual German-language reports on how the complaints were dealt with. The reports must cover: the efforts the platform has been taking to try to eradicate illegal content; a description of its complaints process and the criteria that is used to make decisions regarding illegal content; the number of complaints received in that reporting period and who they were received from; the reason for the complaint; the expertise of the employees responsible for processing complaints, and the training and support they receive; whether the platform consults external bodies during the decision-making process; how many complaints resulted in the deletion or blocking of content; how long it takes the platform to take action; and finally, what measures are in place to inform the body/user that made the complaint of the outcome.

Failure to comply with the Act can result in a fine of 5 to 50 million Euros depending on the severity of the offence. Platforms will not be fined for minor breaches such as the failure to remove a small amount of illegal content, only for systematic failures. Examples include failing to implement a complaints mechanism that works as intended (Schulz, 2018). Another example would be failing to publish bi-annual reports correctly and on time. A platform can be fined even if the regulatory offence was not committed in the Federal Republic of Germany.

Discussion of the Act

The NetzDG Act has received enormous criticism (März, 2018; Schmitz and Berndt, 2018; Splittgerber and Detmering, 2017; Article 19, 2017; Echikson and Knodt, 2018) however, there are aspects that have been praised. The Act requires that platforms implement accessible and transparent complaints procedures. Firstly, this allows user participation. Scholars have previously written about the lack of user participation and the issue that user interests have not been prioritized (Klonick, 2017). Second, the Act ensures that the complaints procedures keep both the person who complained and the user that posted the content in question informed on the decision and the reasons behind it (Article 19, 2017). Research by Mysers-West (2018) has shown the importance of providing users with explanations as to why their content was removed. The research found that users report a lack of understanding around removal processes, and where explanations are missing, spread folk theories as to why the content was removed. This often leads to users continuing to violate the policies. Providing reasons for the removal may therefore help users who unintentionally violated the policies to self-police their content, thus reducing the amount of content that has to be reviewed by the platforms. Providing reasons also helps with overall transparency and could help users to feel heard and involved.

Another positive aspect of the Act is the bi-annual reports (Tworek and Leerssen, 2019). This ensures platform accountability and increases transparency. An issue with consistency across the reports has been highlighted by scholars, for example, some platforms provide more details than others (Tworek and Leerssen, 2019). Perhaps a template that states what information the platforms are expected to report would aid this. Although, it must be noted that transparency reports are likely to differ across platforms due to the various differences that exist across platforms. Also, it should be acknowledged that producing and publishing such reports requires substantial resources and expertise from the platform. Some platforms may require guidance in this area and may benefit from other platforms sharing their tools, if they are willing to.

Finally, the Act states that platforms must provide training and support to their staff. Tech platforms have been reporting for a while that the removal of violative content would not be possible without employees undertaking this work (Bickert, 2017b). Despite this, several cases have emerged where former platform employees have reported developing mental health conditions such as post-traumatic stress disorder (PTSD) as a result of insufficient support during their time working at the platform (Article 19, 2017; Boran, 2020; The Guardian, 2018;

Gilbert, 2019a). Given the often gory and traumatizing nature of terrorist content, and the importance of these employees in the battle against terrorist content, greater training and support is vital to the functioning of content removal processes and could be argued as a moral duty as well.

The Act, however, is highly criticised. First, the Act states that it only applies to platforms that have two million registered users in the Federal Republic of Germany. A user is defined in the fining guidelines that supplement the NetzDG as natural or legal persons who use the platform's infrastructure for purposes of accessing content and information (März, 2018, Cited in Schmitz and Berndt, 2018). The user is classed as registered if they have participated in a registration procedure that provides them with a username and asks them to read and accept terms of service (März, 2018, Cited in Schmitz and Berndt, 2018). Therefore, a user who consumes information from a platform without undertaking this procedure will not be identified as a "user" (März, 2018, Cited in Schmitz and Berndt, 2018). This allows some of the most-used platforms in the ecosystem to evade the Act. One example is Gab which has less than 2 million registered users globally and whose site can be accessed without becoming a registered user (Gilbert, 2019b). Another example is JustPaste.it. In other words, this Act is only aimed at the major platforms. This is problematic because, as seen in chapter 2, terrorist organisations exploit a large and diverse ecosystem of tech platforms. This Act therefore neglects the inclusion of many key platforms that are vital to the running of terrorist operations. Many of the platforms that are neglected by this Act are those that currently do not do very much to police such content (e.g., Gab). Whereas, although far from perfect, the platforms that are likely to fall under this Act (e.g., Facebook, YouTube and Twitter) have already reported implementing some of the demands of the Act, such as complaints mechanisms, as seen in chapter 3. This is, therefore, at best, only a partial response to countering unlawful content.

One explanation for the decision has been that platforms with more than 2 million registered users have "great perpetuating effect", suggesting that the number of users has a correlation with the rate of the site's exploitation (Splittgerber and Detmering, 2017). However, the number of registered users a platform has does not necessarily correlate with the amount of terrorist or extremist content on the site, particularly given the pressure the major platforms have received in recent years to counter this and the finding that many groups utilise small and micro-platforms (Tech Against Terrorism, 2019a). Theil (2019) argues that one upside to the "2 million users" rule is that it might protect smaller/newer platforms from being pushed out of the market place if they do not have the resources to comply with the Act. However, one

could argue that 2 million users is perhaps too large to ensure that it captures the smaller platforms that terrorist groups are exploiting. It is therefore recommended that future regulation does not pinpoint a specific number of users that a platform must have to fall under the regulation. Future regulation should include the full ecosystem of platforms that terrorists exploit under its scope and should consider the compliance challenges that will be faced by smaller platforms.

The Act contains several terms which can be argued to be unclear. The first is what is classed as a "social network". A consequence of this is that there can be uncertainty as to who the Act applies to (Article 19, 2017; Schmitz and Berndt, 2018). There is the risk that platforms may not realise that they fall under the scope of the Act and fail to comply. Under this broad term, it is expected that the Act will only apply to 10 social networking platforms in Germany (Splittgerber and Detmering, 2017). This is just a dent in the number of platforms that are used globally by terrorist and extremist groups. Tech Against Terrorism found that between 2016-2019 alone, terrorist groups made use of at least 330 different platforms (2019a). A further issue is that there is currently no universally accepted definition of "hate speech", nor does the Act include one, instead it points to pre-existing definitions that fall under the German penal code (Schulz, 2018). This is often also an issue with terrorist content because there is no universally accepted definition of what is 'terrorist content'.

It could be argued that there is not a clear enough distinction between content that is "manifestly unlawful" and content that is simply "unlawful" (Article 19, 2017; Schmitz and Berndt, 2018; Echikson and Knodt, 2018). The Act does not provide an adequate set of criteria for identifying manifestly unlawful content; therefore, platforms will have to try to gauge the distinction for themselves (Article 19, 2017; Schmitz and Berndt, 2018). Platforms may get this wrong. They may also decide against distinguishing between the two and class everything as manifestly unlawful, taking a cautious, better safe than sorry approach. Platforms have reported not always having enough context to determine differences such as this (Bickert, 2017a). The line between speech that is uncomfortable or offensive and speech that is illegal is sometimes thin (Schmitz and Berndt, 2018). This can be especially difficult where strict timeframes are given or where technology is used to automatically remove content at upload despite not yet being good at detecting irony, sarcasm and wordplay (Schulz, 2018; Schmitz and Berndt, 2018). The combination of the strict timeframe and threat of fines raises concerns of over-blocking (Tworek and Leerssen, 2019). According to the Act, however, if the platform has doubt, then the content should not be categorised as manifestly unlawful. This aims to prevent platforms

from erring on the side of caution; however, platforms could take advantage of this and claim doubt to get out of the 24-hour timeframe of removing manifestly unlawful content.

The Act did not originally state that platforms must provide users with measures to appeal content-removal decisions, despite it being well-known that errors can occur (Article 19, 2017). However, on 1 April 2020, Germany's federal government published a new draft bill to amend this, as well as stating that complaints mechanisms must be easier to access and use (Hardinghaus, Kimmich and Schonhofen, 2020). This is an important amendment because appeals mechanisms are vital from a human rights perspective; without one there is no way to overturn erroneous infringements on freedom of speech. Further, the availability of an appeal mechanism should encourage better decision making on the part of the platform.

Finally, the Act fails to encourage platforms to work with NGOs, academia, civil society, and capacity-building organisations (such as Tech Against Terrorism). Platforms could gain a lot of expertise from these kinds of partnerships. Tech Against Terrorism, in particular, can provide platforms with expertise and resources in areas such as transparency reporting and policymaking. These partnerships could also aid digital literacy programmes, as well as the creation and dissemination of counter-speech.

Research findings

Research revealed that YouTube, Twitter and Facebook handled the implementation of the NetzDG Act differently. YouTube included an option to flag content that violates the NetzDG in their existing content flagging process and created a new online form to go with it (YouTube 2018a, Cited in Schmitz and Berndt, 2018). Twitter created options for reporting content either on the Twitter app or the "report a violation" button in the help centre (Schmitz and Berndt, 2018). Whereas, Facebook created a NetzDG notification form which can be accessed on Facebook's help website (not through already existing flagging procedures like the other two platforms) (Schmitz and Berndt, 2018). The platforms each had a different number of options for users to select from when reporting content that was believed to relate to the offences of the NetzDG. The platforms would first evaluate the content regarding their own community guidelines and only if it did not violate them would it review the content against the offence listed in the NetzDG (YouTube, 2018, Cited in Schmitz and Berndt, 2018). Between January and June 2018, YouTube removed 42,052 of the 144,836 items it had received complaints about from users. They took action for 54,199 items in less than 24 hours but many complaints took more than 48 hours (YouTube, 2018, Cited in Schmizt and Berndt, 2018). In an interview

with Echikson and Knodt (2018), Google said that it supports the 24-hour rule for obviously illegal content such as child pornography, however, would prefer a more flexible timeframe for content that is more difficult to determine, and would rather courts held the final decision on difficult content. Between January and June 2018 Twitter received 244,064 notifications by users, 27,112 of which were removed. In most cases Twitter made their decision within 24 hours of receiving the complaints but in 600 cases did take more than 24 hours (Twitter, 2018, Cited in Schmitz and Berndt, 2018). Finally, Facebook received a significantly lower number of complaints, receiving only 886 complaints altogether. In most of the cases Facebook made their decision within 24 hours but in 26 cases took more than seven days (Facebook, 2018, Cited in Schmitz and Berndt, 2018).

This evaluation of these platforms did not appear to show any evidence for the predicted fear of over-blocking (Schmitz and Berndt, 2018; Heldt, 2020). However, this research appears to be focused on the platform responses to user complaints and does not provide insight into how the platforms have used automated and proactive technology to remove content which has not been the subject of a complaint. Greater clarity and transparency around the volume of content that is blocked at the point of upload is required to assess whether there is an issue of over-blocking. A second round of reports were published in January 2019 with similar results (Tworek and Leerssen, 2019). It is thought that Facebook received significantly fewer complaints because their NetzDG form is not made as accessible as the other two, with users having to go to the help website to find it (Schmitz and Berndt, 2018; Tworek and Leerssen, 2019). Especially when users can more easily flag content against the community standards by clicking a button next to the post they want to complain about. It would be better if the Act required consistency in how the platforms create their forms and where they can be accessed. Overall, the number of NetzDG complaints are low for all three platforms compared to the flagging against their own rules or community standards (Schmitz and Berndt, 2018).

Despite a lack of evidence of over-blocking, it could be argued that there is not enough scrutiny of the content reported, or sufficient transparency regarding individual content decisions (Schmitz and Berndt, 2018; Tworek and Leerssen, 2019). The platform's own rules and community standards still seem to be the dominant criteria that flagged content is evaluated against (Schmitz and Berndt, 2018). It is unclear whether or not users fully understand the list of offences they are choosing from when making complaints (Schmitz and Berndt, 2018).

Summary

Strengths:

- Platforms must implement an accessible and transparent complaints procedure
- Platforms must publish bi-annual transparency reports (if they meet the criteria to do so)
- Platforms must provide training and support to their employees

Limitations:

- It only targets a fraction of the tech platforms that are exploited in the ecosystem by terrorist organisations due to the restriction that platforms must have two million registered users in Germany
- There are issues with definitional clarity that create concerns that platforms will err on the side of caution and overblock- it should be noted, however, that at present, there is no evidence to support this concern
- There is also a concern around potential under-blocking
- There is a heavy focus on content removal - the Act fails to consider other options and strategies that could work alongside content removal
- The Act fails to encourage platforms to work with NGOs, academia and capacity-building organisations

Sharing of Abhorrent Violent Material Act (Australia)

Overview

In 2019, in the aftermath of the Christchurch attack in New Zealand, Australia introduced the Sharing of Abhorrent Violent Material Act.[35] The Act made amendments to the *Criminal Code Act 1995* that introduce new offences that require internet, hosting or content services (this excludes instant messaging services and search engines) to proactively refer 'abhorrent violent material' to law enforcement within a reasonable time of becoming aware of its existence, and expeditiously remove said material from being accessed in Australia. Schedule 1 of the Act states that a content service is either a social media service (an electronic service with the sole or primary purpose of the service to enable online social interaction between two or more users)

[35] There is also more recently the Australian Online Safety Bill 2021 which was proposed 24 February 2021 https://www.legislation.gov.au/Details/C2021B00018

or designated internet service (a service that delivers to users or allows users to access material using an internet carriage).

Schedule 1 of the Act states that abhorrent violent material is material that reasonable persons would class as offensive, and is engaging either in a terrorist act, the murder, attempted murder, torture, rape or kidnapping of a person. This material can be audio, visual or audio-visual content that is recorded by the perpetrator or an accomplice. The Act confers on the eSafety Commissioner a new power to issue a written notice to a provider of a content service or hosting service informing them that abhorrent violent material is accessible on their site. If these new offences are not complied with, a maximum penalty of 3 years imprisonment or 10,000 penalty units, or both can be enforced on an individual. A penalty unit is a standard amount of money that is used to calculate fines when the law has been breached. In Australia a penalty unit is $222 (Australian Government, 2020). A maximum of 50,000 penalty units or 10 percent of a company's annual turnover can be enforced. Defences in the form of assisting law enforcement, reporting news, public policy advocacy, good faith artistic work, research, court or tribunal proceedings and performance of public official duties can be acceptable. The fault element in this Act is recklessness as it argues that basing it on knowledge could incentivise the platforms to be "wilfully blind" to the abhorrent violent content posted to their site. However, this means that a platform does not need to have actual knowledge of a specific piece of abhorrent violent content on its site, it only has to be aware that there is a "substantial risk" that abhorrent violent content is accessible on its site to be held accountable. Although the definition for "substantial risk" is not clear, the platforms will be presumed to have been reckless if they are given a notice that abhorrent violent material is on their site (Douek, 2019).

For material to be included under this Act, it must have been produced by a person who is engaged in the abhorrent violent conduct, conspired to engage in the abhorrent violent conduct, or is engaged in or in any way knowingly concerned in (e.g., aided or abetted) the abhorrent violent conduct. This seeks to ensure that only material recorded or streamed by the perpetrators and their accomplices are included in the Act. The aim of this is to prevent perpetrators from exploiting platforms to glorify what they have done to a large audience, subsequently reducing their ability to spread fear and panic, ultimately taking the tools away from the perpetrators (Oboler, 2019). If the content is recorded by anyone else then it will not fall under the Act. Section 20 of the Act ensures that edited footage of the original material will continue to be classified as abhorrent violent material and therefore fall under the Act. This is

important given the number of edited copies of the Christchurch footage that appeared online in the days following the attack (Hern and Waterson, 2019).

Discussion of the Act

Similar to the NetzDG, this Act has received a lot of criticism and has been accused of being rushed (Bogle, 2019; Oboler, 2019; Douek, 2019). "The Act made its way through both houses of Parliament in less than two days and came into force two days later. The government did not consult with experts, civil society or industry" (Douek, 2019, p.2). The offences in the Act apply not only to content depicting abhorrent violent material in Australia but anywhere in the world if it can be accessed in Australia. The Digital Industry Group Inc (DIGI) have expressed concern that the Act will require the private companies whose sites are accessible in Australia to breach American law. For example, the Act requires the tech platforms to report abhorrent violent material to Australian law enforcement. However, there are laws in the U.S. where the platforms are based that forbid them from sharing specific types of information and data with law enforcement agencies in other countries (Bogle, 2019).

Concerns have been raised around the rule that content must have been produced by a person or two or more persons who are engaged in the abhorrent violent conduct. Although this would encompass the footage of the Christchurch attack, this requirement could be problematic. In some cases, the content could be abhorrent but not recorded by the perpetrator/accomplice. This could provide a loophole for ensuring the content remains online. It may also be difficult in some cases to identify who recorded the footage. Another criticism is that the Act does not appear to have any safeguards in place to prevent the removal of legitimate sharing of abhorrent violent content. The context in which content has been recorded or posted is not always clear. For example, documenting human rights abuses. Finally, the Law Council raised that it could "silence and criminalise whistle-blowers trying to bring attention to violent atrocities occurring overseas" (Bogle, 2019).

Although the Act states that upon receiving notification of abhorrent violent material the service providers are required to remove the content expeditiously, it does not give an actual timeframe. This is positive in that the platforms do not have to comply with very short, perhaps impossible timeframes; however, some have suggested that there is a lack of clarity as to what expeditiously means (e.g., hours or days) (Douek, 2019; Bogle, 2019). As with many of the other frameworks in this chapter, there is the fear that arbitrary time limits could result in over-blocking to err on the side of caution (Keller, 2018; Douek, 2019). Oboler (2019) points out

that it is more important for timeframes to be achievable as opposed to aspirational. Regarding the enforcement actions, particularly individual liability, it is unclear whether the employees and companies will be penalized for individual pieces of content or systematic failures only. If the former, then in cases such as the Christchurch attack, where users attempted to upload millions of edited copies of the footage, the penalties could be grossly disproportionate, particularly for smaller companies (Douek, 2019).

Douek (2019) criticizes the fault element of recklessness with the example that more than 60 violent incidents were broadcast on Facebook Live between December 2015 and April 2017. Thus, she asks whether it would be considered reckless to allow Facebook to continue to be accessible in Australia under this Act? Douek (2019) also argues that the Act ignores many issues with platforms allowing Livestreaming services. It did not require platforms to instil safeguards or put more resources into finding solutions to the exploitation of livestreaming despite being created in the aftermath of the Christchurch attack. This should have been given more thought by the Act. Regulation should do more to address the lack of safeguarding and consideration of exploitation by platforms when new features and services are designed and implemented. Finally, the Act failed to require platforms to publish transparency reports which means that there is no way of telling what efforts the platforms are making and reduces the ability to hold the platforms to account (Douek, 2019).

Summary

Strengths:

- There is not an arbitrary timeframe for content removals that could lead to erring on the side of caution

Limitations:

- Concern that platforms may have to breach other laws (such as U.S.) to comply
- The rule that content must be uploaded by the perpetrator or accomplice could be circumvented by bad actors seeking to exploit loopholes in the legislation
- Lack of safeguards or thought for user safety at the design stage of new services
- Lack of safeguards to prevent the removal of legitimate abhorrent violent content such as content that documents human rights violations
- Issues surrounding the liability of individuals

- Issues surrounding the fault element of recklessness could mean that it is reckless, for example, to continue to allow livestreaming on platforms
- No requirement of transparency reports – lack of transparency and accountability

UK Online Harms White Paper (Online Safety Bill 2021)

Overview

In April 2019, the UK Government announced that it would like the UK to be the safest place in the world to use the internet. It was argued that previous voluntary initiatives have not been sufficient in tackling online harms, therefore, a new regulatory framework is required. Upon drafting the white paper, the Government ran a public consultation to gather feedback. There were 2,400 responses from a range of stakeholders including tech platforms, academics, civil society and governments. Legislation is currently expected to be introduced to Parliament in 2021 (Tech Against Terrorism, 2020d).

The aim of the regulatory framework set out in the White Paper is to strike a balance between user safety and freedom of expression, regarding both illegal and "harmful" content. The new regulatory framework will apply to companies that "host user-generated content which can be accessed by users in the UK; and/or facilitate public or private online interaction between service users, one or more of whom is in the UK" (HM Government, 2020). This will apply to companies that provide services to UK users, even if it is based outside of the UK (HM Government, 2020). It includes a large range of online harms beyond just terrorist or extremist content (for example, child sexual exploitation and abuse; harassment and cyberstalking, and encouraging suicide).

The regulatory framework will take a two-tiered approach whereby platforms will be categorised as either category 1 or category 2 services (HM Government, 2020). It is thought that the majority of platforms will be category 2 services. This means that they "will need to take proportionate steps to address relevant illegal content and activity, and to protect children" (HM Government, 2020). Whereas, "a small group of high-risk, high-reach services will be designated as 'category 1 services'" (HM Government, 2020). While all platforms will be required to take action on relevant illegal content, only platforms in category 1 services will be required to take action on content that is classed as legal but harmful and accessible by adults (HM Government, 2020). The decision of which category a platform belongs to is a three-step process. The first consideration will be setting out high-level factors that could lead to significant risk of harm to adults regarding legal but harmful content. These factors include the

size of the platform's audience and the functionalities of the platform. Then the government will decide on thresholds for these factors with advice from the regulator. The regulator will then assess platforms against these thresholds. The justification of the two-tiered approach is to protect freedom of expression, mitigate the risk of placing disproportionate burdens on smaller platforms, and to ensure that platforms with the largest online presence are held accountable.

The Government announced in February 2020, after going back and forth between the ideas of implementing a new or existing regulator that UK broadcasting regulator Ofcom would become the regulator for this framework. The Home Office said,

> "Ofcom is a well-established independent regulator with a strong reputation internationally and deep experience of balancing prevention of harm with freedom of speech considerations. It has a proven track record of taking evidence-based decisions, which balance robust consumer protection with the need to ensure the regulatory environment is conducive to economic growth and innovation" (HM Government, 2020).

It is also likely that appointing Ofcom would be less costly and time consuming than developing a brand-new regulator, particularly since Ofcom will cover the costs from industry fees (HM Government, 2020). The core duties of Ofcom will be to set out the responsibilities of private companies under the new duty of care, create codes of practice, form a framework based on transparency, trust and accountability, provide support and guidance to start-ups and smaller platforms, oversee the appeals process, take enforcement action in cases of non-compliance, and finally, promote the adoption of technologies and education to counter online harms. An announcement in December 2020 revealed that Ofcom will be able to levy fines of up to £18 million or ten percent of a platform's annual global turnover, whichever is higher, when a platform has failed in their duty of care responsibilities (HM Government, 2020).

One of the main objectives is to establish a new statutory duty of care that will result in platforms taking greater responsibility for the safety of their users and countering harmful content. Under the duty of care, platforms should update their Terms of Service (ToS) to explicitly outline what content is inappropriate, publish annual transparency reports, implement a complaints mechanism that is easy to access, and finally, respond to user complaints in an "appropriate timeframe" which will be decided by Ofcom (Tech Against Terrorism, 2020d). Ofcom will create the codes of practice which should include a requirement that platforms take

reasonable steps to prevent new and known terrorist content on their site, including links to content. The consultation outcome update in December 2020 stated that the regulator will be able to require platforms to use technology to identify illegal terrorist content on public channels (HM Government, 2020). Further, the regulator should provide guidance on technology that proactively identifies and removes terrorist content where necessary. Under the duty of care, the companies will be required to have an easily accessible appeals process that is responded to within an "appropriate timeframe" and which will be overseen by the regulator. The regulator will not determine the outcome of appeals, the regulator will simply assess whether the process is being carried out appropriately. The white paper acknowledges the importance of including independent oversight in this process so that users will feel confident that their complaints are dealt with fairly. There will also be a super-complaints function that allows organisations to represent users and alert Ofcom to any concerns regarding systemic issues.

One of the steps the regulator must take is a duty of innovation. Platforms are given the opportunity to use alternative approaches to what the regulator sets as long as they can show the regulator that the alternative approach can also provide the same results. The aim is that this should provide flexibility. Flexibility is required because of the enormous range, size, and capacity of platforms that will fall under the framework's remit (as seen in chapter 2). This should allow platforms to be innovative and work within the scope of their budgets. It should help to protect smaller or newer platforms from being pushed out of the market place if they lack the capacity to do something in a specified way.

The December 2020 consultation response revealed that the framework will implement a 'safety by design' approach that has been developed with industry, subject and technical experts" (HM Government, 2020). This will "set out clear principles and" practical guidance on how companies can design safer online products and services" (HM Government, 2020). "The regulator will use its position to drive the development of new technologies and support the sharing of tools and best practice across companies" (HM Government, 2020). This is likely to be useful for smaller platforms that lack capacity and/or expertise in countering terrorist content. The consultation response reported that a key finding was that platforms reported differing levels of capacity and expertise. The safety by design framework will therefore be tailored to meet this different range of needs (HM Government, 2020). In addition to this, the government, Ofcom and platforms will work together to promote media literacy skills that

provide users with the skills they need to stay safe online. Ofcom have previous experience in the media literacy field (HM Government, 2020).

Category 1 services will have to publish transparency reports (HM Government, 2020). The consultation response stated that the information that should be included in the reports are likely to differ depending on the platform. It says,

> "to ensure that the transparency framework is proportionate and reflects the diversity of services in its scope, the transparency reporting requirements will differ between different types of companies. Ofcom will consider companies' resources and capacity, service type and audience in determining what information they will need to include in their reports" (HM Government, 2020).

This will help to minimize unnecessary burdens on smaller platforms.

There are a number of enforcement tools that have been consulted on that could be used by the regulator in the case of non-compliance under this framework (HM Government, 2020). One is issuing fines of up to £18 million or 10 percent of the platform's annual turnover (whichever is higher). The regulator will also be able to disrupt the business activities of any platform that provides services to UK users, including blocking access in the most serious cases. Finally, "the government will reserve the right to introduce criminal sanctions for senior managers if they fail to comply with the regulator's information requests", however, this would only be used as a last resort and will not come into power until two years after the regulatory framework is implemented (HM Government, 2020).

Discussion of the white paper

In contrast to the NetzDG, this framework does not exclude platforms from its regulatory scope based on the number of users. Widening the scope is likely to cause wider disruption to terrorist operations. On the other hand, however, it is a massive task for Ofcom to regulate such a large number of platforms, particularly given the diversity and complexity of the platforms, as seen in chapter 2. A similar argument can be made for the number of harms that fall under the scope of the proposed framework (Theil, 2019). This is also an enormous and complex undertaking for a regulator, particularly an existing regulator with other responsibilities as well.

It has been argued that new regulation should be informed by existing regulation and learn from its criticisms and limitations (Windholz, 2010). The white paper demonstrated that this took place; it researched regulatory strategies and best practice in other areas, for example,

what has been tried in health and safety legislation. Another strength of the white paper is that it acknowledges trends and research findings in the field. For example, it refers to the problem of terrorist groups migrating to newer, smaller and less censored platforms. Finally, the paper states that the regulator will undertake empirical research to understand the content it is regulating and the measures it has in place: "it will run a regular programme of user consultation, in-depth research projects, and horizon scanning activity (HM Government, 2019, p.56).

The statutory duty of care that is proposed aims to provide tech platforms with a high level of guidance across the duties that they must fulfil. This may be particularly useful for smaller/newer tech platforms or any platforms that lack the necessary expertise to counter terrorist content because such challenges are often overlooked in the regulatory landscape (Tech Against Terrorism, 2020f). The duty of care also aims to ensure that tech platforms adopt an approach whereby they continually review and improve their efforts. This is crucial as the findings of chapter 2 revealed the adaptable and constantly evolving nature of terrorist organisations. Finally, the mandatory publishing of transparency reports for category 1 services should increase transparency and accountability, particularly given the flexibility of what is required in the report depending on the services the platform run, minimizing unnecessary burdens on smaller platforms.

The proposal, has, however, received a number of criticisms. One is that it does not provide a clear distinction between what content is 'unlawful' and what is 'harmful' (Broughton and Jaques, 2019). The inclusion of both types of content was done with the intention of ensuring the proportionality of the legislation, leading to, for example, terrorist content requiring "further action" from the tech platforms (Tech Against Terrorism, 2020d). Issues with this, however, include a lack of clarity around what going "further" entails (Tech Against Terrorism, 2020d). Using the term 'unlawful' is difficult because sometimes content is obviously unlawful but in other cases further assessment may be required (Mac Síthigh, 2019) which is difficult with the high volume of criminal offences that exist (Chalmers and Leverick, 2018). An updated definition of 'harmful' was announced in December 2020 as online content and activity that "gives rise to a reasonably foreseeable risk of a significant adverse physical or psychological impact on individuals" (HM Government, 2020). The inclusion of 'psychological impact' has also been criticised because of the subjective nature of psychological harm. Several questions arise: what counts as psychological impact? Does the psychological harm have to be medically recognised? Does it have to be considered with the

most readily upset user in mind? (Smith, 2020). Broughton and Jaques (2019) argue that unclear definitions will result in companies having to determine for themselves what is harmful and what is unlawful and a fear of making mistakes could impact freedom of expression, particularly with the threat of the enforcement powers that will be given to the regulator. There are concerns that this may cause tech platforms to err on the side of caution with removals (Article 19,). Users may also experience confusion around what is and is not harmful and there may be coordinated abuse of this vulnerability with users intentionally targeting and reporting speech because they do not agree with it (Theil, 2019).

Scholars have suggested that a new independent body would have been preferable to extending the remit of a regulator. It was argued that a new body would create a positive symbolic effect that would showcase the Government's commitment to countering harmful online content (Bishop et al., 2019). It was further argued that the creation of a new body would minimize confusion among the public as to who to contact regarding the issues in the white paper (Bishop et al., 2019). As the proposed framework and duties of this regulator are very widespread and challenging, it may be too much for an existing regulator to take on with existing responsibilities. Finally, an existing regulatory body may not have the fresh eyes that a new one would and could approach the issues with strategies from its pre-existing duties that may not be a very good fit for its new responsibilities. On the other hand, it has been argued that the experience a body such as Ofcom could result in it being already attuned to the institutional culture and balancing different rights and interests (Broughton and Jaques, 2019). Additionally, a new body could create fragmentation and unnecessary additional expenses (Broughton and Jaques, 2019). Therefore, there are strengths and weaknesses to either approach.

There are concerns around competition, innovation and the capacity of some tech platforms to comply with this proposal. For example, the regulator can require platforms to use technology to identify and remove terrorist content, however, some platforms may not have the capacity to implement such technology. If smaller platforms lack the capacity to comply with any areas of the framework, this could create issues that reduce market competitiveness (Tech Against Terrorism, 2020d). This is particularly concerning given the severity of the enforcement actions at the regulator's disposal. ISP blocking should only be used as a last resort because of how integral social media has become in people's lives particularly with running businesses, it would have an enormous socio-economic impact that would cause mass uproar across the country. This enforcement power has been heavily criticised as disproportionate and very detrimental to freedom of expression, particularly the freedom of expression of many users

who have never violated the law or the platform's terms of service (Broughton and Jaques, 2019). Additionally, users may be able to circumvent this by using VPNs. Regarding the disruption of business activities, platforms, such as Gab, are starting to adapt to this and build their own alternative services to replace the third parties that would be able to do this (Tech Against Terrorism, 2019b). Once these alternative services are fully running and available this will not be such an effective power for a regulator.

Summary

Strengths:

- The all-encompassing approach to the tech platforms that fall under its scope (however, the later decision to make two categories allows smaller platforms to have less responsibility which is argued as a limitation)
- It is informed by research findings and existing regulatory strategies
- The duty of care aims to ensure that tech platforms receive guidance from the regulator – this may be particularly beneficial to smaller platforms that lack the necessary knowledge and capacity to comply without such guidance
- The decision to implement an existing regulator is less time-consuming, requires less resources and means that the regulator has a range of experience to draw on
- The framework will implement a 'safety by design' approach that has been developed with industry, subject and technical experts

Limitations:

- Unclear distinction between content that is 'unlawful' and content that is 'harmful' – which could result in platforms erroneously removing/failing to remove speech
- It could be argued that a new regulator would have been a better choice than the decision to implement an existing regulator as it would have created a positive symbolic effect, it would avoid confusion among the public as to who to contact and would have brought a fresh pair of eyes
- There is a concern that smaller tech platforms may not have the capacity to comply
- There are concerns around the severity of some of the enforcement actions, their socio-economic impact and their potential impact on free speech, as well as user's ability to potentially work around them (e.g., VPNs)

European Union Code of Conduct on Countering Illegal Hate Speech Online

Overview

The European Commission has created proposals, Codes and legislation in recent years for both terrorist and hate speech online content in an attempt to create consistency across Europe and try to prevent a fragmented framework of national rules that are confusing for platforms to comply with (Echikson and Knodt, 2018; European Commission, 2018). The first European Commission response that will be discussed is the 2016 Code of Conduct on Countering Illegal Hate Speech Online which is a voluntary instrument created in response to a number of attacks occurring in Europe (Portaru, 2017). The Code uses the Framework Decision 2008/913/JHA 2008 definition of illegal hate speech: the "public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin" (p.55). The Code brought four platforms together – Facebook, Twitter, YouTube and Microsoft – to commit to trying to combat illegal hate speech on their platforms in Europe. It has been described as putting tech platforms at the forefront of the fight against hate speech (Portaru, 2017).

The four platforms agreed to: put in place a clear and effective process to review notifications of illegal hate speech on their platforms and to respond by removing or disabling access to the relevant content within a timeframe of 24 hours; provide information on the procedures they have in place for reporting content; have community standards or rules that prohibit hateful content and content that incites violence; provide regular training for their employees on societal developments; and share best practice with other platforms. The first point of review is against the companies own policies, and only after that against national laws "where necessary". This Code also encourages counter-speech and working with CSOs as well as content removal. The Code was, however, criticised for not consulting any CSOs during its development (Article 19, 2017). It was further criticised for not specifying that platforms must include an appeals process (Portaru, 2017).

When examining regulatory frameworks or proposed frameworks with the purpose of using the findings to inform and develop a new regulatory framework, it is important to acknowledge if it has a voluntary nature. For example, a measure which may not turn out to be effective in a voluntary code of conduct may be effective in a different framework if backed-up with the threat of a formal sanction.

Research Findings

The European Commission established an EU High-Level Group to combat intolerance and hate and monitor the implementation of the Code of Conduct (Schmitz and Berndt, 2018). The Group combines Member States authorities, CSOs and EU agencies. The first progress report was based on a study in which twelve organisations across nine Member States took part to test the processes of receiving notifications over a six-week period between October and November 2016 (Jourová, 2016). These notifications were in regards to alleged illegal hate speech that were either reported through reporting channels or the processes available to users. The results, as with the NetzDG did not suggest any evidence of over-blocking and in 40 percent of the cases the platforms reviewed the notification the day it was reported. Portaru (2017) claims that despite little indication of over-blocking, overall, the content reviewers did not agree with many of the flagger's decisions; they did not consider the reported content to be illegal or against the broader terms of service in more than two-thirds of the cases. Trusted flaggers have been known to have accuracy rates of over 90 percent (Carter, 2016). This therefore raises a concern that tech platforms are unwilling to remove too much content out of fear that it will cause them to lose users (Portaru, 2017). It should also be noted that although the study only involved a small number of notifications, none of the platforms had a 100% rate of assessment in under 24 hours. Therefore, in cases where the volume of content will be higher, it is unlikely that platforms will manage the timeframe (Portaru, 2017). However, it is not known how much of this content could be considered 'grey-zone' content which would not necessarily be able to be decided within 24 hours.

In the second report in June 2017, the High-Level Group claimed an improvement in the efficiency, speed and quality of notifications (European Commission, 2017). The report records an improvement between the consistency in flags between trusted flaggers and normal users. There was, however, a lack of consistency amongst the feedback given to users between the platforms. Finally, the report claimed that there was progress with staff training and cooperation with CSOs. The European Commission (2018) reported in January 2018 in their third report that the companies removed content within 24 hours in 81 percent of cases. The evaluation in February 2019 provided more detail by revealing, for example, that 85.5 percent of content that called for murder or violence against specific groups was removed, however, does not give reasons as to why the other 14.5 percent of that content was allowed to remain.

Overall, the number of removals increased with each reporting period which may be the result of improvements in the procedure or spark concerns about possible over-blocking (Portaru, 2017). There is very little detail as to the assessment criteria for removal. Although the reports

give plenty of removal figures, without any contextual explanations for these figures, the reports are arguably limited in their usefulness, particularly when there is not an appeal process or any indication of how many errors are made. The reports reveal that more transparency, accountability and oversight are required to fully understand the outcomes and progress of the Code. Moreover, there was a significant lack of consistency across the platforms regarding feedback to users on notifications. The Code could have more explicitly stated that this is a commitment the companies should make. Finally, the companies were aware that the monitoring was taking place for each report (Alkiviadou, 2019). Data is not available outside of the monitoring periods which makes it difficult to know whether the reports reveal accurate figures that remain consistent outside of monitoring periods (Alkiviadou, 2019).

Summary

Strengths:

- The platforms agreed to provide regular training for their employees
- The platforms agreed to share best practice with other platforms
- The Code recognises the value of counter speech, digital literacy and working with CSOs in addition to content removal
- Research did not find evidence of over-blocking

Limitations:

- Only four platforms signed the Code even though this issue spans a whole ecosystem of platforms
- The code is voluntary, therefore, there are no formal enforcement powers if the platforms fail to comply
- The Code did not include an appeals process
- Criticised for not consulting CSOs during development of the Code
- Research found that there is very little detail as to the assessment criteria being used and whether there is an evaluation of removal accuracy
- The reports are limited in their usefulness
- Research found a lack of consistency across platforms regarding feedback to users on notifications

European Commission: Regulation on preventing the dissemination of terrorist content online

Overview

In 2018, the European Commission proposed 'regulation on preventing the dissemination of terrorist content online'. At the time of writing, the proposal is in the legislative trialogue process of negotiation between the Commission, Parliament and the Council. The main goal of the proposal is to make the content less accessible. It is a response to calls by Member States and seeks to create a clear and harmonised legal framework that provides clarity surrounding the responsibility of tech platforms in the actions that they need to take to ensure that their services are safe whilst still protecting freedom of expression. In preparation for the proposal the Commission consulted stakeholders to help guide the process and held a public consultation which received 8,961 responses from individuals, organisations and public administrations. Consultations with Member States revealed that voluntary arrangements were not providing adequate results and therefore binding obligations are required. Tech platforms, on the other hand, said that they would support a continuation of voluntary measures and raised the fact that legal fragmentation creates difficulties for them.

The regulation proposed a legally binding rule that platforms must remove "material that incites or solicits the commission or contribution of terrorist offences, or promotes the participation in activities of a terrorist group" within one hour of being notified of its existence on their site (European Parliament, 2019). Terrorist offences are those defined in Article 3(1) of "Directive (EU) 2017/541 and is defined as information which is used to incite and glorify the commission of terrorist offences, encouraging the contribution to and providing instructions for committing terrorist offences as well as promoting participation in terrorist groups" (S1(1.3)). The one-hour rule is to limit the dissemination of the content and potential harm as much as possible (Krasenberg, 2019). According to Krasenberg (2019, p.7), "terrorist content is considered most harmful in the first hours after it appears online because of the speed at which it spreads and therefore its reach." Failure to meet the one-hour deadline could result in a fine of up to 4 percent of the platform's global annual turnover. The regulation proposes that Member States effectively and proportionally penalize platforms that fail to comply with any of the obligations. When deciding the penalty, the authority must consider the nature, gravity, character and duration of the breach, if there have been previous breaches by the legally responsible individual, the financial strength of that individual and the level of cooperation the platform has shown with the authority.

The proposal puts forth removal orders which would be issued as an administrative or judicial decision by a 'competent authority' in a Member State. Platforms must put in place measures to expeditiously assess the content in removal orders. Once the order is received by a platform, they must seek to remove or disable access to it within one hour. The one-hour timeframe only applies when the content has been confirmed as illegal by an authority. The authority provides the platforms with an explanation as to why the content is considered illegal terrorist content to help speed up the decision process. Under this regulation, referrals should be evaluated against the platforms' own respective terms and policies.

The proposal implements a duty of care obligation on platforms. It also requires tech platforms to install proactive measures, where appropriate, to detect and remove terrorist content. The proposal also outlines that platforms must explain their policies against terrorist content and publish annual transparency reports including information on their actions and efforts to counter such content. The proposal requires platforms to put the necessary processes in place for users to appeal decisions where they believe that their content was wrongly removed and to supply the content provider with information as to why the content was removed.

Discussion of the proposal

Although this proposal has been heavily criticised, there are some positive features. The regulation does not limit the platforms that fall under it by size; it recognizes that smaller platforms are frequently exploited by terrorist groups. Further to this, the inclusions of annual transparency reports aim to contribute towards greater transparency and accountability from the platforms. Finally, the proposal includes the implementation of an appeals mechanism.

One of the first criticisms is that it was developed without consultation of free expression and human rights organisations (York and Schmon, 2021). Further, the proposal's definition of online terrorist content is overly broad and could result in the wrongful removal of legal content, for example, journalistic content and content posted by human rights organisations raising awareness of the actions of terrorist groups (Ní Aoláin, 2018; Brown, 2020; Tech Against Terrorism, 2020c; Denes, 2020; Khan and Ní Aoláin, 2020). Another criticism is concerned with the empowerment of competent authorities to decide what content should be removed. The concern is that content removal could be subject to political pressure (Brown, 2020). A suggestion has been that competent authorities should be defined as independent courts or administrative authorities only, independent of any political, commercial or unwarranted influences (Brown, 2020; Denes, 2020; Khan and Ní Aoláin, 2020).

In an open letter to the EU Parliament in February 2019, dozens of organisations and academics complained that the requirement of proactive measures could lead to some platforms installing poorly understood technology which could lead to big errors and biases (Hidvegi, 2019). The letter argues that there is no publicly available meaningful information about how well this technology works and the cost it has on democratic values and human rights. They give the example of The Syrian Archive (a CSO) having 100,000 videos wrongly deleted by YouTube. Another similar open letter from civil society organisations echoed these concerns, raising the point that adequate safeguards are required but lacking (Denes, 2020). The potential errors than can be made with such technology, in combination with the risk of fines, result in concerns around over-blocking (Ní Aoláin, 2018; Brown, 2020).

Finally, the one-hour timeframe has been criticised as disproportionate and insufficient for platforms to adhere to (Denes, 2020). It is viewed as particularly concerning due to the risk of financial penalty. The fear is that it leads to over-blocking and is particularly burdensome on small platforms that do not have the capacity to comply (Ní Aoláin, 2018, Denes, 2020). This criticism is in opposition with the argued lack of clarity of the Abhorrent Violent Material Act discussed earlier. One suggestion has been to use the standard of 'acting without undue delay' instead of a short, strict timeframe or the term 'expeditiously' in which no one seems to know how to interpret (Denes, 2020).

Summary:

Strengths:

- The proposal does not limit the platforms that fall under its scope by size, it acknowledges that a range of platforms are included in the ecosystem
- The inclusion of transparency reports increases transparency and accountability
- The inclusion of an appeals mechanism helps to ensure the overturning of erroneously removed content

Limitations:

- Criticised as having an overly broad definition of terrorism which could result in the wrongful removal of legal content
- Concern that the competent authorities could be subject to political pressure regarding content removal
- Concerns with possible errors of proactive measures and the potential for over-blocking

- The one-hour timeframe is criticised as disproportionate and burdensome for smaller platforms

Internet Referral Units

In 2010, the first Counter Terrorism Internet Referral Unit (CT IRU) was established in the UK, hosted in the Met's Counter Terrorism Command (NPCC, 2018). In the first eight years 310,000 pieces of terrorist-related material were removed, the majority of which was Jihadist-related but the unit has been increasingly working on getting extreme right-wing content removed too (NPCC, 2018; Counter Terrorism Policing, 2018). The European Union Internet Referral Unit (EU IRU) was then established in 2015 to provide support to Member States in detecting and investigating online malicious content (Europol, 2019b). The unit has a range of goals including producing strategic insight into Jihadist groups, flagging content and working in close cooperation with tech platforms on these referrals. When the unit flags content to tech platforms, much like many of the other responses, the platforms evaluate the content against their own terms of services and policies. In the Year One Report in 2016, the unit claimed that instead of trying to follow and remove individual pieces of content, they put in place a procedure that aims to focus on content that is linked to a high-profile terrorist attack and high-profile accounts. The Unit therefore prioritises the content that is likely to receive the most views, which is a strategy that Facebook claim to use (Bickert and Fishman, 2018). Finally, the Unit participates in Referral Action Days (RADs) which are coordinated referral campaigns that focus on flagging online terrorist content on certain platforms, usually within a one- or two-day period. The unit claims that these days are effective in improving coordination between it and its national counterparts (Europol, 2017).

The EU IRU does not only focus on content removal which is a reactive strategy. It is also involved in researching and analysing the content that it finds in order to further understanding of the groups and their strategies. They then share their findings with law enforcement in order to help with criminal investigations and to build partnerships with tech platforms. This is something that is missing in many of the other responses. Understanding the groups and their strategies allows proactive measures to be formulated.

EU Internet Forum (EUIF)

The EUIF was launched in December 2015 to bring together "EU Interior Ministers, high-level representatives of major internet companies, Europol, the EU Counter Terrorism Co-ordinator and the European Parliament" (European Commission, 2015). The Forum's goal is a voluntary

private-public partnership to counter: “harmful material online”. In sum, the two main aims are to reduce the accessibility of terrorist content online and to empower CSOs to create and spread counter-narratives online. Some of the platforms involved in the forum include Facebook, YouTube, Microsoft, Twitter, JustPaste.it, Snap, Dropbox and Telegram amongst others (Krasenberg, 2019).

The forum has a focus on trying to prevent content that has been removed from reappearing elsewhere. The first version of the GIFCT hash database was created by the forum for this purpose. The forum stated that “it is not just about removing terrorist content – it is also about making sure that less terrorist propaganda and narratives appear online in the first place” (European Commission, 2016). To do so the forum donated ten million Euros to civil society to create counter-messages. At the end of 2016, the Civil Empowerment Programme was created to support education, employment and inclusion (European Commission, 2016). The programme brought together CSOs, grassroots organisations and credible voices to work on capacity building, training, partnering CSOs with tech platforms, and spreading campaigns to those vulnerable to radicalisation and recruitment (European Commission, 2017). Eleven projects have been funded to date (Krasenberg, 2019).

In the aftermath of the Christchurch attack, the EUIF announced it would “address crisis response mechanisms by clarifying the roles and communication channels between law enforcement and the private sector following a terror attack with a significant online component” (European Commission, 2019a). Following this in October 2019, the forum committed to a voluntary EU Crisis Protocol to enable Member States and tech platforms to respond quickly and in a coordinated manner when terrorist content is disseminated during or in the aftermath of an attack. Additionally, law enforcement and online service providers will be able to share relevant information about the content with each other, on a voluntary basis in a secure process in real time (European Commission, 2019b).

In summary, the efforts of the EUIF aim to increase public-private cooperation and improve platforms’ responses to referrals from national authorities and Europol’s Internet Referral Unit (European Commission, 2018). Matt Brittin, EMEA President of Google said in 2017 that the EUIF “played a vital role in driving collaboration on these issues and helped lay the foundation for the Global Internet Forum to Counter Terrorism” (European Commission, 2017). Additionally, it aimed to improve the implementation of voluntary proactive measures to detect terrorist content automatically and increase transparency around platforms’ efforts. Much like

the EU IRU, it aims to create a coordinated effort, share information and eliminate the problems of fragmentation that comes with national frameworks. It aims to ensure that content is not re-uploaded, that there is a focus on a coordinated crisis response during attacks, and finally, that work is done with CSOs to create counter-narratives to direct vulnerable users down an alternative path. One of the downsides of the forum is that it is difficult to measure the impact of the counter-narratives. Another downside to the forum, and similarly the EU IRU, is that as it is a voluntary response; many platforms have failed to engage with it (European Commission, 2018). On the other hand, both the EUIF and EU IRU claim to have created constructive partnerships with the companies that do engage (Europol, 2015).

Christchurch Call

After the Christchurch attack in New Zealand in 2019, Jacinda Arden, Prime Minister of New Zealand, initiated the Christchurch Call to Action Summit. "The call outlines collective, voluntary commitments from Governments and online service providers intended to address the issue of terrorist and violent extremist content online and to prevent the abuse of the internet…" (Christchurch Call, 2019). The Call, however, promises to be consistent with principles of a free, open and secure internet which respects human rights and freedom of expression, as well as to promote innovation, economic development and inclusion.

A number of commitments are made by the governments involved (there are currently over 50) as well as the tech platforms (of which there are currently 10). The governments commit to provide education and digital literacy; ensure effective enforcement of applicable laws that prohibit the production and dissemination of terrorist and violent extremist content in a manner consistent with the rule of law and human rights law; encourage media outlets to apply ethical standards when reporting terrorist events; support frameworks and industry standards; and finally, consider appropriate action to prevent terrorist exploitation of tech platforms and join collaborative initiatives where possible. The tech platforms commit to undertaking transparent measures to prevent the upload of terrorist and violent extremist content; providing greater transparency regarding their community standards and terms of service in a manner consistent with human rights; provide user complaints and appeals mechanisms; implement measures to mitigate the risk of exploitation of their live-streaming services; publish regular transparency reports; review the operation of algorithms that may amplify terrorist and violent extremist content; and work collaboratively with other platforms. Both the governments and tech platforms vow to work with CSOs to promote community-led efforts; aid research into the

development of technical solutions and efforts to more greatly understand terrorist and violent extremist content; and work and cooperate with law enforcement.

This call should be praised for the collaborations and commitments that have been made. Some of these commitments have been neglected in other responses, for example, working to increase the capacity of smaller platforms. Further, the number of governments and tech platforms involved is commended. However, the Call fails to include definitions for terrorist and violent extremist content. Therefore, there is a lack of clarity and transparency over what the Call is working to counter. Another concern is that it failed to secure key players such as the United States to commit to the call (BBC, 2019b). Finally, the Call is voluntary. There are no regulatory incentives or enforcement measures if the government or tech platform decide not to.

Tech Against Terrorism

Tech Against Terrorism is supported by the United Nations Counter Terrorism Executive Directorate (UN CTED).[36] The aim is to work with tech platforms to counter terrorist use of their services whilst respecting human rights. The initiative revolves around three pillars: outreach, knowledge-sharing and practical support. The outreach pillar involves working to promote constructive relationships between governments and tech platforms. The knowledge-sharing pillar focuses on working with tech platforms to share best practice. The final pillar offers tech platforms practical support with implementing tools to respond to terrorist content.

One of Tech Against Terrorism's projects with the UN CTED and the Republic of Korea has been to create the Knowledge Sharing Platform. This was launched in 2017 and "is a collection of interactive tools and resources designed to support the operational needs of smaller technology companies" (Tech Against Terrorism 2017a). The resources available on the platform include: a list of terrorist groups and individuals on the UN sanctions list and their key terminologies; recommendations for model Terms of Service; model guidelines for transparency reports and standardised reporting formats; risk assessment tools; amongst other relevant resources. Overall, the platform helps smaller platforms to build capacity to protect their services. A Facebook representative commented that although Facebook possess a lot of knowledge, until this platform was created, they were unable to share it through a formal network (Tech Against Terrorism, 2017b).

[36] https://www.techagainstterrorism.org/

Another project run by Tech Against Terrorism is the Data Science Network which was launched in 2018. This is a "network of developers, academics, and researchers from the machine-learning and data analytics fields who will work to develop easily deployable tools to help small tech platforms fight terrorism on their services" (Tech Against Terrorism, 2018). Further, Tech Against Terrorism worked with the academic website Jihadology.net to ensure that sensitive content is only accessible to individuals who registered with an academic/research, governmental, journalistic or humanitarian email address (Tech Against Terrorism, 2019c). Finally, Tech Against Terrorism received a grant in 2019 from the Government of Canada to develop a Terrorist Content Analytics Platform (TCAP) (Tech Against Terrorism, 2019d). This is a centralised platform which aims to facilitate tech company moderation of terrorist content and improve quantitative analysis of terrorists' use of tech platforms. The platform should help platforms respond to terrorist use of their sites expeditiously through an alert function and securely examine verified terrorist content.

The work undertaken by Tech Against Terrorism highlights several issues that have been missing in some of the frameworks in this chapter. One is that smaller platforms often lack the capacity to counter terrorist content on their own. Another is that they also often lack expertise on how to counter terrorist content. Both of these limitations are likely to lead to compliance issues. Tech Against Terrorism's projects could aid them with these compliance issues.

Lessons to be learned

This chapter aimed to answer the following question:

6) What has and has not been effective in existing regulatory frameworks that seek to counter online terrorist content?

Overall, there is a lot that can be learned from the examination of existing and proposed regulatory efforts. Although not the case for all of the regulation discussed in this chapter, there seems to be a move in the direction of the inclusion of complaints procedures, appeals mechanisms and transparency reports. This is a positive move due to the issue highlighted by Klonick (2017) that tech platforms have not been held accountable enough to their users and users have been under-represented in the decisions regarding how their speech should be regulated. However, the research in this chapter has highlighted that not only do platforms need to implement these but they need to ensure that these mechanisms are easy to use and easy to access. The research from the NetzDG Act showed that where such mechanisms are more

difficult to find, they tend to be less utilised (Schmitz and Berndt, 2018). Appeals mechanisms are vital for free speech protections, particularly where automated technology is used due to the potential for errors and bias (Hidvegi, 2019). The failure to include appeals mechanisms has been condemned by the United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Portaru, 2018). There also needs to be independent oversight that both the complaints and appeals mechanisms are being used as they should. Regarding transparency reports, this is a move in the direction of greater transparency and accountability. However, this is only the case where the information reported is meaningful and explanatory; simply reporting removal figures does not provide meaningful information. It would be helpful if regulation outlined what should be included in transparency reports. It should also be acknowledged that publishing transparency reports can be a big undertaking and requires expertise and resources. Platforms that face challenges with this should be able to access relevant guidance. They may face similar challenges with complaints and appeals mechanisms. It would be helpful to them if there was encouragement in the industry to share best practice and technology, where possible.

Another positive theme seen in this chapter is that some of the regulation/proposals were informed by research findings, existing regulatory strategies that have been effective elsewhere, and consultations with CSOs, NGOs, academia and the public. It has been argued that it is vital that future regulation is informed by past regulatory experience and supported by evidence from prior research (Windholz, 2010). This can prevent repeating mistakes. The consultation processes further the argument by Klonick (2017) that users and other parties are under-represented when it comes to regulatory decision-making. Such consultation also brings together new and diverse expertise. Further, it is important that the effectiveness of the regulation is researched a period of time after enforcement in order to identify possible amendments and create a culture of continual improvement.

Although mentioned in one or two of the examined regulations/proposals, the issue of providing employees with training and support is arguably still not receiving enough attention. There has been growing concern in recent years that the viewing of such content can lead to the development of mental health conditions such as post-traumatic stress disorder. Several content moderators have come forth in recent years with law suits against, claiming to have developed a mental health condition as a result of poor working conditions, a lack of training and support and exposure to traumatic content (Boran, 2020; The Guardian, 2018; Gilbert, 2019a). Given that such employees are vital to the countering of terrorist exploitation of tech

platforms, and the number of employees of each platform is continuously growing (Jee, 2020), it is concerning that this issue has not received more regulatory attention.

The regulations/proposals in this chapter have demonstrated an increasingly diverse number of enforcement actions that a regulator could have at their disposal. Given the findings that there are often limitations and challenges to almost all of the enforcement actions available, it makes sense to have a number of options and for these options to fall on a spectrum of severity. Research by Clinard and Yeager (1980); Orland, (1979); Stone (1975); Fisse and Braithwaite (1984) all argue that some penalties work better for some companies than others. Further, Braithwaite (1990) argue that failing to comply is a less attractive option for an organisation when faced with a regulator armed with a range of enforcement actions than a regulator with one enforcement action. In cases where there is only one enforcement action, companies are usually aware that the regulator cannot use it lightly and that they will have to do something terrible for it to be sanctioned against them. As Braithwaite (1990) points out, regulators have more power and credibility "when they can escalate deterrence in a way that is responsive to the degree of uncooperativeness of the firm, and the moral and political acceptability of the response" (p. 63).

An important lesson to be learned, given the findings of chapter 2, is that the scope of the platforms that should fall under the regulation. The NetzDG Act highlighted the issue that a narrow scope would result in many of the key platforms in the ecosystem slipping through the cracks and subsequently evading regulation. This could lead to further issues such as migration to less regulated platforms (see Nouri, Lorenzo-Dus and Watkin, 2021). It appears to make more sense for regulation to be as all-encompassing as is possible, providing that the regulator has the resources necessary to oversee all of the platforms. The UK Online Harms White Paper highlighted a different issue, however, that is, the scope of the type of content that should fall under the regulation (e.g., child sexual exploitation, terrorist content, suicide etc.). It has been argued that the narrower the scope of a regulator, the higher the chances are of that regulator being able to manage its responsibilities (Theil, 2019).

Although this thesis argues that it is important that smaller platforms fall under the scope of regulation, there are challenges with this. Many of the regulatory frameworks and proposals in this chapter fail to consider and acknowledge the issue highlighted in chapters 2 and 3 that platforms are not all equal in their capacity to comply with regulation. Many smaller platforms in particular, face challenges regarding capacity and expertise. This could lead to unfair

burdens on smaller platforms. This could be alleviated by a proportionate and responsive approach by the regulator. Regulators should be able to provide these platforms with guidance where necessary and be able to point platforms to opportunities where they can access information regarding best practice and where other platforms are willing to share tools.

Another common criticism throughout this chapter was the failure to provide enough definitional clarity. Without this clarity there are concerns that platforms will not realise that the regulation applies to them. There are also concerns that where there is a lack of clarity around what type of content should be removed. Platforms could resort to over-blocking, particularly where the enforcement actions are severe. While research does not currently support the concern of over-blocking, this is a developing area of research that requires further research. On the other hand, platforms could also under-block because of a lack of definitional clarity or also out of fear of losing their users (Bloom, Tiflati, and Horgan, 2019; Conway, et al., 2017; Nouri, Lorenzo-Dus and Watkin, 2019; Robins-Early, 2019; Fussell, 2019). The criticisms of both over- and under-blocking somewhat make it difficult for platforms to appear to be doing the right thing. If a high percentage of flagged content is removed, there are freedom of speech concerns, however, if a low percentage are removed, there is the concern that they are not removing enough. Future regulation, therefore, needs to ensure that there is greater clarity regarding what content should be removed.

There was a notable failure of the regulation/proposals examined in this chapter to ensure that platforms consider safety and safeguarding at the design stage of new services and features, excluding the Online Harms White Paper. Cases such as the Christchurch Attack in New Zealand highlighted the issue that Facebook had failed to sufficiently safeguard its livestreaming feature. Although it will not be possible to safeguard against every potential situation, and regulation must be cautious of impeding innovation, more could be done by platforms to safeguard at the design stage. This is an area in which platforms should be more accountable to their users (Nash, 2019). There is also the issue that many of the regulations/proposals failed to consider approaches other than content removal. More attention should be given to digital literacy programmes, working with NGOs, CSOs, and academia, as well as collaborating with others in the industry to share best practice.

A major criticism throughout this chapter has been of either strict timeframes or a lack of clarity around a timeframe in which content must be removed. Regarding strict timeframes, there is firstly, the concern that platforms will take a better safe than sorry approach and over-block to

avoid penalties. It may also discourage platforms from focusing on removing the most harmful or viral content and pressure them to hit targets instead (Bickert, 2020). There is again, also the concern that the timeframes are particularly challenging and burdensome for smaller platforms, making it impossible for them to comply. On the other hand, not providing enough clarity around the timeframe and failing to specify whether the expectation is hours or days could result again in either the better safe than sorry approach or could result in content remaining online for too long. An alternative approach could be to implement an all-things-considered principled assessment that considers the capacity of a platform and the volume of content it has, amongst other relevant factors.

Finally, the implementation of strict timeframes will leave platforms no choice but to utilise artificial intelligence technology that automatically and proactively searches for and removes content (Kaye, 2018). With this comes the risk of over-blocking, particularly where uploading content is automatically matched with content stored in databases that do not allow the context of the post to be taken into consideration (Engstrom and Feamster, 2017; Keller, 2018). There have been cases where evidence of war crimes has been mistakenly removed as a result of this technology (Warner, 2019). The technology cannot identify old material in a new context, nor can it identify sarcasm, irony, word play or jokes (Keller, 2018). Mistakes can be made with language if the technology is not developed by native speakers of that language, as well as many types of bias depending on the diversity of the developers (Keller, 2018). Discrimination can occur, with one example by Keller (2018) being that short passages from the Koran and clerical teachings are often mistakenly removed. When this does occur, it has been argued that it can contribute to wider feelings of alienation, exclusion and frustration in society which are some of the key risk factors for radicalisation (Keller, 2018). There needs to be independent oversight into the technology and there must be, as already mentioned, appeals mechanisms that also have independent oversight in place to try to minimize the risks that come with using such technology. Further, the tech platforms should be capable of providing reasons for the decisions that are made.

Conclusion

The above examination has provided an overview of the regulatory landscape regarding online terrorist content. It has highlighted many of the complexities, challenges and the limitations of current regulatory attempts. The main challenges and lessons to be considered in future regulation are:

- The inclusion of a complaint mechanism, appeal mechanism and transparency reports
- Regulation must be informed by research, existing regulatory strategies and consultations with a range of stakeholders
- Regulation must ensure employee safety and well-being is considered
- There must be the inclusion of a diverse range of enforcement actions for the regulator
- The whole ecosystem of platforms that are exploited by terrorist organisations must fall under the scope of the regulation
- Definitional clarity must be ensured throughout
- Efforts must be put into safety and safeguarding at the design stage of new features and services
- During the development of regulation, the challenges that smaller platforms face must be considered and what can be done to provide guidance and reduce burdens
- Regulation must consider both proportionality and clarity around the implementation of timeframes for content removal

All of these lessons informed the development of the regulatory framework that is proposed later in this thesis. The regulatory framework proposed in this thesis aims to draw from what appears to be working and propose alternative regulatory solutions where limitations and challenges have been identified. It is argued that an examination of past and current approaches, such as the one in this chapter, is essential in the development of an informed and effective set of proposals (Windholz, 2010).

Chapter 5: The Regulatory Approach: Social Regulation Theory

Introduction

Many scholars have attempted to define the term 'regulation'. This thesis uses Black's (2001a) and (2002) definition: regulation is the intentional and sustained use of an authoritative party to attempt to change the behaviour of others. This is done by setting specific, defined standards through the use of information-gathering and behaviour control. Regulation is used for what might be called 'double edged sword' activities where the aim is to attenuate the negative aspects of an activity whilst preserving its positive aspects. For example, the regulatory framework proposed in this thesis aims to prevent terrorist exploitation of tech platforms while maintaining the many benefits that tech platforms provide more generally. The term 'regulation' will be used throughout this chapter to include governments or international governmental organisations (IGOs) that seek to control/influence practices.

It is often the case, particularly in highly complex and rapidly evolving industries, that governments are a step behind in regulating the practical application of technical and organisational developments (Bruhn, 2006). It is not surprising then that this is the case regarding the regulation of tech platforms, which have the added difficulty of being global organisations, creating jurisdictional challenges. Regulation is not typically implemented until after the risks have materialised; after some harms, injuries or accidents have occurred, and after their causes have been found to have been linked to the activity or technology in question (Bruhn, 2006). This is the case with terrorist use of tech platforms. Terrorists have been using the internet since the late 1990s/early 2000s (Weimann, 2004; Conway, 2006), and exploiting tech platforms since their emergence in the mid-2000s (Weimann, 2010) with Conway et al. (2017) defining 2014-2015 as the 'golden age' for terrorists and their largely unregulated use of tech platforms. It was shortly after this that governments around the world began to place increased pressure on tech platforms to make significantly greater efforts to counter terrorist use of their sites, after which, harm had been occurring for a while.

The 1980s saw major changes to regulation in the UK (Veljanovski, 2010). These changes included a wave of deregulation and privatisation, with one of the main changes being the shift from state ownership to private ownership and regulation (Veljanovski, 2010). With this change came concerns that private organisations would have different priorities to those of the state, and regulatory issues would have to compete with (and most likely lose to) commercial interests (Wilson, 1984). As mentioned, tech platforms went largely unregulated for a while, and then when they did begin to regulate, this was largely done through self-regulation.

However, this self-regulation was not consistent across platforms (Jugendschutz.net, 2017), nor in its application to different terrorist organisations (Conway et al. 2019), raising concerns around the approach to countering such content. Many governments, including the UK, expressed discontent with the efforts of tech platforms' responses to terrorist use of their sites (May, 2018). As seen in chapter 4, governments began to lose patience with platforms self-regulatory efforts and as a result, many countries began proposing or implementing their own regulatory frameworks, many of which have been criticised for a variety of reasons. Further regulatory changes included universal suffrage resulting in a pressure on governments to think about policies that would benefit wider populations in society, changes in technology that created new threats, and a general push for more rights (Ogus, 1994). All of this resulted in the need for a greater focus on one specific area of regulation: social regulation.

The aim of this chapter is to introduce social regulation theory and answer the following questions:

7) Is social regulation theory applicable to this regulatory context?
8) What is there to be learned from examining social regulation in other regulatory contexts?
9) Could these strategies be applied in this regulatory context?

This thesis argues that social regulation theory contains an abundance of relevant research and regulatory lessons that can be drawn on, learned from and applied to the creation of regulation that aims to counter terrorist content on tech platforms. At the time of writing, social regulation theory has yet to be applied in this context. Social regulation theory is used primarily in three areas: environmental protection, consumer protection and occupational health and safety. This chapter will introduce and define social regulation theory. It will then argue why social regulation theory is an appropriate one to underpin the regulatory framework that will be proposed in this thesis. It will finally examine the use of social regulation theory across the three areas it has previously been applied to. This examination will look to identify any strategies and lessons that can be applied to the regulatory framework that will be proposed in this thesis.

Social Regulation Theory

Two types of regulation feature prominently as alternative approaches throughout the regulatory literature: economic regulation and social regulation. Economic regulation, as its name suggests, deals solely with economic issues, primarily in industries that contain

monopolistic tendencies (Ogus, 1994). Examples of economic objectives include pricing, investment and output (Baldwin, Cave, and Lodge, 2010) and putting resources to "their most valuable uses" (Ogus, 1994). However, academics have argued that there are a vast range of other issues and values across many regulatory areas that also require consideration that economic regulation fails to consider (Baldwin et al., 2010). This brings us to social regulation. Social regulation is concerned with a much broader range of issues than its economic counterpart (Ogus, 1994). It emerged during the industrial revolution of the late 18th century (Veljanovski, 2010). Ogus (1994) describes the areas that fall under the responsibility of social regulators as having public interest justifications, Prosser (2010) describes the responsibility as promoting human rights, social solidarity, and social inclusion, and Wilson (1984) describes it as regulation that aims to "promote a general societal good such as clean air or water". It is important to note, however, that although social regulation is not primarily concerned with economic objectives, it often still has "economic effects, costs and benefits" that require consideration during its implementation (Veljanovski, 2010). Further, similar to its economic counterpart, it has major implications for politics (Wilson, 1984). As Wilson (1984: 203) summarises, economic regulation "is generally aimed at controlling relations between a single industry and its customers; social regulation is generally concerned with controlling the imposition by industry in general of social costs…". When introducing the many areas of regulation in their edited book *The Oxford Handbook of Regulation,* Baldwin et al. (2010) argues that a substantial effort has been made to move beyond Stigler's (1971: 3) original hypothesis that "as a rule, regulation is acquired by the industry and is designed and operated primarily for its benefit". Today, most organisations have learned to assume that any social harm caused by their business will at some point, "be subject to public censure, government action, and legal liability" (Gunningham et al. 2004, p. 308). Social regulation is central to the regulation of three key areas, each of which will be considered in detail in this chapter: environmental protection, occupational health and safety, and consumer protection (Den Hertog, 2010).

Ogus (1994) argues that social regulation is usually a response to one of two types of market failure. The first is when individuals fail to receive the necessary information about the quality of the goods or services being provided to them, and therefore are not able to make an informed decision on whether the goods or services meet their required preferences (e.g., the product being safe, the product delivering a specific outcome etc.). The outcome, therefore, is likely to be a failure by the company to meet consumers' expected preferences (Ogus, 1994). The

second is that market transactions oftentimes result in externalities that negatively affect individuals who are not even involved in the transactions (Ogus, 1994). Pollution could be used as an example here as both consumers and non-consumers of a specific product, who live close to the factory where it is produced and the consequential pollution, could end up with health problems as a result of the production (Ogus, 1994). Regulation is, therefore, one way of internalising these externalities and correcting the market failure.

According to Prosser (2010), the goals and justifications in social regulation are more complex than economic regulation. One goal is distributional justice which is concerned with regulation resulting in a 'fair' or 'just' distribution of resources (Ogus, 1994). For example, the goal of wanting everyone in society to be able to access the same minimum level of resources, as opposed to just the wealthy or powerful (Ogus, 1994). However, a major limitation of this is that what is regarded as 'fair' or 'just' is subjective within communities, and particularly across nations and beyond, which would become very complex for individual organisations and whole industries supplying goods or services nationally or internationally (Ogus, 1994). A justification is paternalism which is when an individual's decisions are controlled and overridden by those in charge of regulating. The liberty of an individual or group of individuals is interfered with solely in the interest of protecting their own welfare, safety, happiness, needs, and values (Ogus, 1994). Some examples given by Ogus (1994) are being prevented by law from travelling in a motor vehicle without fastening your seat belt and being unable to work under a certain employment contract without contributing a certain amount of money to a pension scheme. This too, however, is a problematic justification as removing a person's right to make their own choices and decide what is best for them is controversial despite arguments of bounded rationality. Bounded rationality is the idea that "the capacity of individuals to receive, store, and process information is limited" and as such, humans are not always highly skilled at making 'good' decisions, for example, decisions that are in their best interests (Simon, 1975 Cited in Ogus, 1994: 41). On the other hand, another justification, behavioural control, is when individuals' decisions are overridden with the justification not of protecting their own welfare, safety, happiness, needs and values but that of others (Ogus, 1994). Finally, another goal is that of community values. This is when regulation reflects not necessarily what an individual would want for themselves (e.g., for their own welfare, safety, happiness, needs and values) but what the individuals believe would benefit the community as a whole (Stewart, 1983). An example given by Ogus (1994) is that those who prefer to live and spend time in urban communities may nevertheless be in favour of policies that aim to preserve and maintain

rural amenities that are used considerably by others in their community. A limitation of this however could be a 'not in my backyard' attitude. This is when individuals are happy for regulation to be in place in situations that do not affect them but are less willing to have the same regulation closer to home where it could be an inconvenience for them. Conversely, regulation might be an antidote to a 'not in my backyard' attitude. For example, most reasonable people appreciate the importance of energy generation but would not want a power station to be located near them. Regulation can therefore be used to mitigate the adverse consequences to the local community.

According to Lave (1981:29) there are two main reasons to analyse social regulation; the first is to distinguish "scientific issues from values or political consensus-building and, second, to trace out the implications of proposed actions". Lave (1981) examines eight different frameworks that can be applied in the social regulation decision-making process in his book *The Strategy of Social Regulation: Decision Frameworks for Policy*. Relevant frameworks will be drawn on throughout the chapter. Decision frameworks, he argues, identify the main issues and challenges that affect the way the regulators and non-governmental decision-makers view them. Decision frameworks are especially important in situations where a substantial number of people are required to make a decision (Lave, 1981). When comparing the frameworks to achieve the best fit for a specific problem or situation, Lave suggests considering four criteria. The first is comprehensiveness: are all the relevant risks and issues addressed by the framework? The second is intellectual foundation: does it explore and assess wider considerations than the minimum necessary? Third, what resources are needed to implement the framework and is this realistic? Finally, felicitousness: does it address the main issues first? However, when it comes to public policy, Lave (1981) argues that timeliness is the most important consideration. Public policy requires the best possible decisions based on the available knowledge, time, and resources (Lave, 1981).

In summary, social regulation theory is concerned with a broad array of non-economic issues (Ogus, 1994). It is mainly concerned with regulating issues that concern the public interest, the promotion of human rights, social solidarity, social inclusion and general societal good (Ogus, 1994; Prosser, 2010; Wilson, 1984). Social regulation is necessary where consumers fail to receive the necessary information about the quality and safety of the goods or services that they are consuming or where the actions undertaken by an organisation result in externalities that negatively affect individuals who are not even involved in the transactions (Ogus, 1994). The goals and justifications in social regulation are complex and vary from distributional justice,

paternalism, behavioural control and community values (Ogus, 1994; Stewart, 1983). The two main reasons to analyse social regulation are to distinguish scientific issues from values and to examine the implications of the regulated actions (Lave, 1981). There are eight different frameworks that can be applied in the social regulation decision-making process according to Lave (1981), those of which that are relevant to this research will be drawn on throughout this chapter when discussing the three areas social regulation has previously been applied to: environmental protection, consumer protection and occupational health and safety.

Applying Social Regulation Theory to Countering Terrorist Content on Tech Platforms

Having provided an overview of social regulation, it is important to next explain why the theory should be applied to the regulation of the tech industry and its efforts to tackle online terrorist content. The first reason is that terrorist exploitation of tech platforms is not exclusively an economic issue. Although many platforms will have economic interests and concerns, this issue likely differs from the typical issues faced in profit-making organisations across other industries. Terrorist exploitation of tech platforms affects the public interest, the promotion of human rights, social solidarity, social inclusion and general societal good which are the main concerns of social regulation theory (Ogus, 1994; Prosser, 2010; Wilson, 1984). It does so because the strategies terrorists use in online propaganda are done with the aim of creating fear and inciting violence (Weimann, 2010; Putra, 2016; Welch, 2018). These strategies target specific groups of individuals, some of which are minority groups that require protection. It is therefore important that the regulatory approach applied to a framework to counter this takes these social considerations into account and looks beyond the typical economic issues that organisations usually face.

An argument put forth by Miller (2019) further supports this justification. Miller's argument is that the tech platforms have portrayed their services from the beginning as a social mission as opposed to a commercial enterprise. With mission statements that include phrases such as "Don't be evil" (Google) and "Bring the world closer together" (Facebook), the platforms themselves made the promotion of social good their mission. In doing so, Miller argues that tech platforms have portrayed themselves as prioritising moral over economic concerns. The analysis of platforms' blogposts in chapter 3 supports this argument (regarding the major platforms). The findings in chapter 3 revealed that the major platforms prioritised connecting people, protecting free speech, and user safety. The problem is that the efforts that platforms have made, to date, to rid their sites of terrorist content, have not been viewed by governments

as good enough. As discussed in chapter 4, self-regulation was neither consistent nor effective. Miller claims that this is not surprising, as the private companies have been expected to take on roles such as counter-radicalisation specialists, digital literacy trainers, and cybercrime police, none of which were the intentions of those who created the platforms. There is also a lack of accountability, democratic oversight and transparency regarding efforts to counter terrorist content. This thesis, therefore, puts forth, that applying social regulation theory may be the solution to accomplishing the goals of achieving social good whilst putting in place mechanisms for accountability, democratic oversight and transparency.

Another argument as to why this issue falls under social regulation theory is linked to market failure. As discussed in chapter 4, tech platforms failed to regulate themselves effectively under a self-regulatory approach, thereby creating a number of market failures. The first is that, platforms are not as transparent as they should be and lack accountability surrounding the decision-making process and implementation of policy decisions. This leaves some users (for example, activists documenting human rights violations) unsure as to how their content could be affected. There is a lack of inclusion and consultation of other parties (e.g., users, CSOs, NGOs) in policy decisions. Further to this, many platforms fail to publish transparency reports (Tech Against Terrorism, 2020). Other failures discussed in chapter 4 were that many platforms have failed to provide adequate and transparent appeal mechanisms. Regarding terrorist content, there are two main failures that can occur. One is when content is erroneously removed infringing on users' free speech rights. The second is when content should be but is not removed or not removed before it is widely viewed risking potential harms to other users and even third parties that do not necessarily use the platform. Social regulation has been applied to other industries that have experienced market failures concerning social issues. A social regulatory approach can therefore provide examples of lessons and strategies that have already been tried and tested in other industries where regulation had to be enforced to be effective. This will be demonstrated later in the chapter and used to inform the regulatory framework proposed in this thesis.

An important obstacle that is faced in arguing the importance of regulation in this context is the uncertainty around the causal contribution of viewing terrorist content. There is great difficulty in assessing a causal link between viewing terrorist content and undertaking terrorist activities. Before discussing this further it is important to explain the definition of causation being used in this thesis. *Sine qua non causation* is used by lawyers, sometimes known as the "but for" causation and will be the definition used here. This asks would Y harm/outcome have

occurred if X actions had not occurred? (Hart, 1985). Essentially, X is said to be a cause of Y if Y would not have occurred "but for" X. For example, under criminal law, in a murder case, it has to be shown that, but for the act of the defendant, the victim would not have died. Under the current context, for example, a platform user who has undertaken terrorist activity, it can be difficult to ascertain what materials they had accessed, whether the user actually viewed the content that they had accessed, and if the content was viewed what influence this had on the user. This makes it impossible to know, in most instances, but for viewing the materials would the activity still have been undertaken? A real-life example would be the Christchurch attack. There was little in place before the attack to counter terrorist use of the livestreaming feature offered by the platform. However, would the attacker have still committed the attacks had there not been a livestreaming service? It is impossible to know.

This does not, however, mean that action should not be taken to counter terrorist exploitation of platforms and the posting of terrorist content. There is a small body of empirical research that has found that while the internet may not cause radicalisation or terrorist activity, it can facilitate it (Von behr, Reding, Edwards and Gribbon, 2013; Gill, Horgan and Deckert, 2014; Gill and Corner, 2015; Gill, Corner, Conway, Thornton, Bloom and Horgan, 2017). This research concluded that the internet affords more opportunities for radicalisation. It makes it easier for terrorist organisations to send messages and target propaganda at those vulnerable to radicalisation. It also makes it easier for those vulnerable to radicalisation to either stumble across or go looking for information and propaganda. Moreover, the internet allows easier access to people or content that can confirm existing beliefs. In addition to this research, there are other reasons as to why regulation should be enforced to counter this issue. As already mentioned, terrorist content is often shared with the intention to recruit, incite violence, create fear and dehumanize others (Weimann, 2004; Weimann, 2010; Venkatesh, Podoshen, Wallin, Rabah and Glass, 2018; Welch, 2018; Baele, Bettiza, Boyd and Coan, 2019; Baele, Boyd and Coan, 2019). It is therefore argued that the enforcement of regulation to counter terrorist exploitation of tech platforms is a moral duty. It must be enforced to protect those who are vulnerable to radicalisation and those who are the targets of the incitement and dehumanizing content. As mentioned, it is difficult to prove that this content will directly result in a user undertaking violence or an attack. However, there is a risk that if terrorist exploitation of tech platforms can lead to terrorist activities and regulation is not enforced to prevent it, that significant harm could take place (the definition of harm used in this thesis is discussed in the chapter 6). Social regulation theory has previously been used in other industries where there

has been similar uncertainty as to a danger or risk of harm, yet a moral duty to protect people because the harm that could occur, if regulation is not enforced, could lead to catastrophic consequences. This will be demonstrated later in the chapter (see below Precautionary Principle and climate change example) with a discussion of the lessons and limitations of such enforcement.

There is an additional reason why social regulation is appropriate for protecting those who are vulnerable to radicalisation or are the targets of content that is intended to dehumanise, incite violence and inflict fear. Social regulation theory has a number of justifications for enforcement and considers phenomena such as bounded rationality when considering when enforcement is appropriate and how interventionist it should be. These include the earlier mentioned paternalistic and behavioural control approaches. These approaches can be considered during situations such as deciding between conflicting objectives which is a common issue in complex regulatory issues, examples of which that are relevant to this context are discussed in greater detail in the chapter 6. Previous applications of paternalism and behavioural control in other industries can be drawn on, learned from and considered before being adopted in new regulation. For example, the research has investigated the outcomes in situations where regulation took a predominantly paternalistic approach elsewhere and how the regulator/regulatees responded, as well as the advantages, challenges and limitations that were faced.

A number of arguments have been made for undertaking a process where the chosen regulatory approach is researched across other industries before being applied to a new context in order to investigate how it has worked elsewhere. Windholz (2010) argues that it is vital that future regulation is informed by past regulatory experience and supported by evidence from prior research. Such a process will reveal the advantages of the approach, as well as the challenges, limitations and unintended consequences that have been faced and how they were overcome. Social regulation theory has been implemented for several decades across three diverse industries and has led to an abundance of research that can be drawn on and learned from. The three industries it has been applied to are environmental protection, consumer protection and occupational health and safety. The next section of this chapter reports the findings of a literature review that was undertaken into the use of social regulation of these three industries. As will be seen in the reported findings, there are aspects of each industry that relate in one way or another to the issues and challenges that are faced by tech platforms in this context (for example, uncertainty surrounding causality is an issue found in both this context and

environment protection). This further adds to the relevance of researching these three industries with the intention of learning from their approaches.

Overall, social regulation theory is an appropriate theory to research and apply to the regulation of terrorist exploitation of tech platforms for a number of reasons. It is concerned with issues beyond the typical economic issues that most for-profit organisations face. It is concerned with a range of issues that are directly endangered by terrorist exploitation of tech platforms, for example, the promotion of human rights, social solidarity and social inclusion. It also provides an approach that overcomes the difficulty of proving a causal effect. This can be researched in other industries to learn about the lessons learned elsewhere. Given the research findings that the internet is a facilitative tool for radicalisation and terrorist activities, the failure to regulate and consider social issues could result in a failure to protect individuals vulnerable to radicalisation and individuals that are the target of inciting, dehumanising and fear inducing content. The scale of the consequences of these failures is unknown but has the potential to be catastrophic. Further, unlike a self-regulatory approach, social regulation will ensure that the platforms are held accountable if they fail to engage with the regulation. This will help to overcome a number of the market failures that have been identified as occurring during the period of self-regulation that has taken place in recent years. Finally, social regulation theory has the advantage of providing a plethora of previous research to draw on and learn from. It cannot be expected that regulation will be effective without having researched the approach in other industries. This includes how the regulator and regulatees responded to the approach, the outcomes, advantages, limitations and challenges.

Previous Research: Social Regulation Theory

This section of the chapter will examine the previous research that has been undertaken to investigate the approaches used in the main three areas that social regulation has already been applied to. A literature review was undertaken into the three industries: environmental protection, consumer protection and occupational health and safety. This section will provide an overview of each industry and then discuss approaches that have been used in that industry that were deemed as relevant for this framework. An approach was deemed relevant if it contains aspects and strategies that could be applied to the regulatory framework that is proposed in this thesis. The discussion will consider both the advantages of the approach as well as the challenges and limitations. The findings of this chapter were instrumental in the

thought-process behind the creation of the mandatory regulatory standards in the framework proposed in this thesis.

Before moving onto the three previous research areas, it is important to note that social regulation theory discusses six ways to approach the creation of regulatory standards, ranging in scale of specificity and how interventionist the standards could be. It is important because these approaches influenced the development of the mandatory regulatory standards in the regulatory framework that is proposed later in the thesis. Regulatory standards do not all need to take the same approach, different approaches can be drawn on depending on the desired outcome of the standard. The creation of the regulatory standards considered the benefits and limitations discussed in each approach and drew from them where applicable.

The first three range from low to high intervention. Target standards are low-level on the intervention scale; this is when the standard does not specify specific requirements for the organisation's processes or finished outputs regarding the goods or services they are providing, however, it does enforce criminal liability if the use of their goods or services result in harmful consequences (Ogus, 1994). An example would be not setting any specific requirements in the creation of a children's toy, however, being held criminally liable if that toy is ingested by a child. This standard provides organisations with a lot of regulatory freedom, encourages innovation and will not necessarily place financial burdens on the organisation; however, it is vague and lacks guidance that some smaller or newer companies may require in order to fully comply (Scott, 2010). Next is performance standards which are costlier to develop than target standards, and are slightly more interventionist. Performance standards set out specific expectations that goods or services must meet at the point of supply, however, allow the supplier complete freedom of choice as to how they choose to meet these designated standards (Ogus, 1994). An example of this standard would be specifying the amount of lead that can be in a product but total freedom over how that is achieved (Ogus, 1994). Aside from being a bit costlier than target standards, the pros and cons are very similar. At the higher end of the intervention scale are specification standards. These standards demand that organisations either use or do not use specific methods or materials when producing the goods or services (Ogus, 1994). An example would be demanding that specific tyres must or must not be used in the creation of a car (Ogus, 1994). These standards contain more guidance and are less uncertain than the other standards; it is very clear whether or not an organisation has complied. However, unlike the other two standards, there are risks to innovation and development, particularly of cheaper regulatory possibilities, which in turn could affect the consumer financially, and end

up being more detrimental to social interests (Ogus, 1994). These standards may contain different complexities and have different effects and consequences across different industries. For example, in some industries, the costs of determining a causal connection are high and difficult, perhaps with a lot of uncertainty, making the decision of choosing appropriate and effective standards challenging, whereas this is not the case in other industries (Ogus, 1994).

While the previous three standards ranged in levels of how interventionist they are, the final three approaches range in how specific the standards ought to be. Prescriptive standards are very specific and detailed, for example in occupational health and safety (OHS), this could be setting out very specific and detailed terms for organisations regarding how they should keep their employees safe in all aspects of their jobs. These standards provide a lot of guidance, which can be very helpful to some organisations (particularly newer/smaller companies). However, these standards are criticised as being reactive and creating standards that are difficult to understand and keep up-to-date. As a result, many issues often fall between the cracks. Inspectors are good at finding very specific breaches that are easily understood and identified, but neglect a broader array of major systemic problems that are not as easily understood and identified based on the long, complex standards (Gunningham, 2007). Next, principle-based standards, are similar to performance standards, however, are broader in their application and known for being 'all-encompassing' (Gunningham, 2007). An example would be setting standards for a job that contains multiple hazards: physical risks (such as using dangerous machinery), noise-risks, and psycho-social factors (Gunningham, 2007; Johnstone, Quinlan and McNamara, 2011). Ergonomic and manual handling risks may also be included (Gunningham, 2007). Creating broader standards helps to minimize the problems identified with prescriptive standards where laws quickly become dated, and allow the organisation more flexibility in how they choose to tackle risks, for example, tailoring standards to their specific machinery and processes. The downside, however, for both organisations and inspectors, is the difficulty with identifying whether or not enough has been done to comply with the standards. Finally, process-based standards identify and outline a specific set of steps that an organisation must implement and follow (Gunningham, 2007). The steps would typically include identifying hazards, carrying out risk assessments, and risk control, as well as monitoring the health and welfare of both working conditions and individual employees, and keeping records of any health and safety incidents and injuries. Process-based standards typically encourage the establishment of a whole programme that creates objectives and targets, processes for achieving them, and ways to measure whether they are working. These standards are more

concerned with influencing attitudes and encouraging organisations themselves to adopt a health and safety culture and systemic approach. Unlike the previous standards, they encourage organisations to go beyond the minimum or stated actions and continually improve and innovate their processes. There could, however, be problems with auditing such broad standards, resulting in potentially inadequate or ineffective processes being implemented.

Environmental Protection and Social Regulation

The first pre-existing area of social regulation that this chapter will examine is environmental protection. Meadowcroft (2005, 2012) recognised parallels between social and environmental policies, arguing that both types of problems involve market externalities, in which collective action is the solution. The Stern Review (2006) argued that climate change has been the most extreme case of market failure in history. Pollution control is regarded as an extremely important regulatory regime due to its great social and public interest consequences (Ogus, 1994). Further, the belief that unregulated growth of industrial activity could have "catastrophic consequences, for example through global warming" meant that governments have been "under considerable pressure to introduce regulatory measures which will reassure the electorate" (Ogus, 1994: p. 204).

Social regulation brought big change to environmental protection. According to Ogus (1994), in the mid-1970s, the regulatory method of 'best practicable environmental option' (BPEO) was in place. This is defined as the regulatory approach that provides the most benefit or least damage to the environment at the most acceptable cost in both the long- and short-term (Regional Comprehensive Economic Partnership, 1988 Cited in Ogus, 1994). Therefore, at that time, economic considerations still played a large role in the regulatory decisions of organisations that had major social concerns. The term 'acceptable costs' resulted in the aim being "to achieve a reasonable balance between the costs of prevention and/or dispersion and the benefits", as well taking the "public purse" into account (Regional Comprehensive Economic Partnership, 1988 Cited in Ogus, 1994, p.207). A problem with this type of framework is that the costs that are considered tend to be the costs used in the regulation and not the costs that the regulated activity could impose on others or the environment. However, by the 1990s, there had been some change towards social regulation: there was less focus on formalized standards, a greater demand for accountability, and an increase in public participation in the standard-setting process; The Aarhus Convention in 2001 is an example (Ogus, 1994).

- Precautionary Principle

The first social regulatory approach that is used in environmental protection is the precautionary principle. The precautionary principle falls under Lave's (1981) no-risk framework which states that the public should not be exposed to any unnecessary risk, regardless of how small. This framework does not require much data or analysis, for example, it would not analyse the consequences of banning something from a product or service. It therefore, would not result in a long, drawn-out decision-making process. This is an approach that may, on the surface, sound appealing to anyone who may be affected by the activities of an organisation. However, is not as straight forward as it sounds. According to the 1998 Wingspread Declaration, the precautionary principle states that if an activity creates a threat to public health or the environment, precautionary measures should be put in place, even if cause and effect relationships have not been reliably and scientifically confirmed (Baldwin et al., 2010). In other words, regulatory action is taken before a threatened harm occurs in order to protect the non-economic values in environmental protection that serve the best interests of the public and the environment, even if the regulator is not certain that the harm is inevitable (Baldwin et al,. 2010). The principle ranges, however, from strong to weak versions, with the weak versions resulting in very little effect (Sunstein, 2005). In 1998, a consensus statement listed four components of the principle: first, taking preventative action in uncertain situations; second, transferring the burden of proof to the proponents of an activity; third, researching a variety of alternatives to potentially harmful actions; and, fourth, raising the level of public participation in the decision-making process (Raffensperger and Tickner, 1999). Sunstein (2005) has described the principle as a 'better safe than sorry' approach. An important distinction to make is that if a decision was made as a result of complete certainty, then the regulation would be preventative, not precautionary (De Saledeer, 2002).

The precautionary principle recognizes and considers factors than are not typically contemplated by market-based mechanisms (Baldwin et al., 2010). It also acknowledges 'the possible intrinsic limitations of scientific knowledge in providing the appropriate information in good time' (Godard, 1997: 65), and essentially, 'reduces the scientific threshold for regulatory policy making' (Vogel, 2003: 566). This ensures that politicians cannot use scientific uncertainty as an excuse to avoid or delay the implementation of severe measures (Jordan and O'Riordan, 1995), and that non-scientific factors such as public opinion and social values are given weight in the decision-making process (Vogel, 2003). For example, the predicted catastrophic effects of environmental problems have, for some, reduced their

confidence in the abilities of environmental scientists and policymakers (Kriebel et al., 2001). According to Kriebel et al. (2001), there has been an increased dissatisfaction with risk assessments; this is due to the perception that there are many complex and secret assumptions that are not shared (with the public) by the experts. The precautionary principle aims to minimize the constraints of risk assessments in its pursuit of alternative approaches (Kriebel et al., 2001). An example is the decision made by the Los Angeles Unified School District on pesticide use (Kriebel et al., 2001). The policy is that all pesticides pose some risk to both health and the environment, and therefore, will only be implemented if all non-chemical methods have been considered. If one must be used it will be the least harmful one. The policy did not weigh the risks and benefits against one another, nor refer to any specific list of dangerous pesticides. There was a level of uncertainty as to exactly how dangerous the pesticides were and the precautious decision was made to choose an alternative method, if possible. Jordan and O'Riordan (1995) argue that precautionary measures are most likely to be implemented when public opinion is largely risk-averse, especially when it involves technology as there is a widely-held view that new technologies often contain dangers of which we are yet to become aware (Vogel, 2003).

The precautionary principle is, however, considered controversial, and is regularly reconstructed and challenged (Baldwin et al., 2010). Criticisms of the principle are that it is negative and reactive, it allows decision-makers to make emotional choices based on fear, it is not sufficiently scientifically reliable, it does not say how much precaution is the correct amount of precaution, and it may stand in the way of innovation by not allowing the creativity necessary to consider other developments or solutions if there is any uncertainty around them (Kriebel et al., 2001; Sunstein, 2005; Lave, 1981). Additionally, choosing to make decisions without all the information, or a complete picture, is a risk in itself; it is a risk that is taken to avoid other risks (Kriebel et al., 2001; Sunstein, 2005). Finally, in theory, if the outcome is the lowest level of harm possible, then one might think that this strategy sounds ideal and is worth some minor inconveniences. However, this is not the reality. People quickly change their minds when solutions cause them even a small amount of inconvenience or make a situation slightly less enjoyable, which is likely to be the case with this approach. Safety is not always a top priority (Lave, 1981).

The precautionary principle has been implemented over the last thirty years in policies regulating pollution, ozone-depleting chemicals, fisheries, climate change, and sustainable development (Raffensperger and Tickner, 1999). The principle was first enforced in various

areas of environmental protection due to the belief that the pace of efforts to reduce the effects of issues such as climate change had been too slow, and, as a result, had become too difficult to handle in the future (Kriebel et al., 2001). The principle informed many American environmental statutes. One example is the Clean Air Act Amendments (1970) which forced the Environmental Protection Agency (EPA) to set primary air quality standards that would safeguard the health of the public, even the most sensitive members, to ensure 'an adequate margin of safety' and gave them permission to 'assess risk rather than wait for proof or actual harm' before setting standards (Lave, 1981). A study on photochemical oxidants had previously shown that asthma attacks and eye irritation began to occur around the level of 0.15ppm (parts per million), and as a result they set the standards at 0.08ppm (Lave, 1981). This level of precaution caused controversy and the EPA was asked to re-evaluate the standard but they struggled to undertake new research because of the difficulties in measuring total oxidants (e.g., they could only undertake research in laboratory conditions as opposed to 'real' conditions). In the face of controversy and uncertainty, the EPA raised the standard to 0.12ppm. This was still seen as too precautious by industry groups, yet too lenient by environmentalists, and they ended up being taken to court by both sides. The Council on Wage and Price Stability undertook a benefit-cost framework (which is a more quantitative and formal balancing of risks and benefits) and concluded that the appropriate level should be set between 0.16 and 0.20 ppm. However, the EPA defended their decision by arguing that they were not allowed to use a benefit-cost analysis because the protection of the public health (even the most sensitive members) and the environment, was their main concern (Lave, 1981).

Therefore, while the precautionary principle has been acknowledged as controversial, with complex competing preferences in the environmental protection literature (seen in the Clean Air Act example), it is also important to acknowledge that such a precautionary approach has been controversial and a major criticism of counter-terrorism laws since 9/11 (Zedner, 2007b). For example, Zedner (2007a) discussed that such an approach results in a society where "the possibility of forestalling risks competes with and even takes precedence over responding to wrongs done" (pp.261-262). According to McCulloch and Pickering (2011, p.631),

> "the 'preventive' counter-terrorism framework is concerned less with gathering evidence, prosecution, conviction and subsequent punishment than in targeting and managing through disruption, restriction and incapacitation those individuals and groups considered to be a risk",

sparking a tension with notions of proportionality, due process, human rights, and potentially leading to prejudice against certain ethnicities, races and religions (McCulloch and Pickering, 2011; Zedner, 2007b). McCulloch and Pickering (2011) note that although the main rationale is to prevent harm and protect national security, particularly due to the mass causalities that terrorism aims to create, there are also a number of practical and ethical concerns that must be considered. This is particularly so, given that in the context of counter-terrorism, there is not a widely agreed definition of what terrorism is, and the process of proscription is often a political process, not a judicial process (McCulloch and Pickering, 2011).

The framework put forth in this thesis acknowledges the criticisms of the precautionary principle in the counter-terrorism context, however, it is important to note that there are also dangers in not applying such a principle that could lead to catastrophic consequences, for example, as seen with the lack of precaution taken to livestreaming services with the Christchurch attack. This framework implements the precautionary principle, however, sought to address the concerns and offset potential dangers with a number of the regulatory mandatory standards that are proposed in chapter 7. These mandatory regulatory standards have been designed to protect human rights considerations against an overzealous application of the precautionary principle. They include the implementation of a mandatory appeals mechanism, independent appeals board, clear policies outlining what will and will not be removed and why, and bi-annual and transparency reports (see chapter 7). These aspects of the proposed framework seek to ensure that any errors are identified and overturned and that platforms are as transparent with users as possible, in order for users to be as empowered as they can be regarding opportunities to protect their rights.

Social regulation of environmental protection faces many difficulties (Sunstein, 1993). The Clean Air Act highlights the differing, competing preferences across regulatory stakeholders. Environmental protection is an issue in society that causes mass divides, for example, well-funded climate change deniers, which makes regulation not only difficult in terms of decision-making and implementation, but also in seeing through effective long-term compliance (Gough, 2013). It is also an issue that sometimes crosses borders, therefore, becoming a transnational or global issue (Ogus, 1994; Gough, 2013). If nations find it difficult to solve internally, then crossing borders make it even more complex. There is also the "time-scale, cumulative nature, and intersection with numerous other environmental problems…its intersection with other policy areas…and its impact on the global socio-economic system" (Gough, 2013: p.3). Another issue is valuing the benefits of environmental protection; Ogus

(1994: p.204), argues that it is "highly speculative", particularly in situations where it can affect amenities, health, and future generations. There are still a lot of unknowns and uncertainty in evaluating the direct impacts, causal effects, and future predictions of pollution, which make spending large sums of money on abatement difficult and unappealing for businesses (Ogus, 1994; Gough, 2013).

Even if there could be agreement amongst parties and scientific certainty, there would still be difficulties in implementing standards across a number of individual organisations (Ogus, 1994). For example, organisations and states may differ on a number of characteristics, including financial capabilities, pollution control costs, and other industry or state-dependent factors (Ogus, 1994; Baldwin et al., 2010). One example is a study by Potoski and Woods (2002), which found that states with stronger environmental groups and weaker industry groups enforced tougher ambient standards programs, and larger states with more complex air pollution implemented more extensive programs. How organisations are perceived by regulators may also contribute. According to Hawkins and Hutter (1993: pp.201-202), regulators perceive some organisations to be "clean, tidy and concerned with safety and the environment, whereas others are seen as dirty, disorganized, careless, and inherently dangerous". Hawkins and Hutter (1993) raise other points that came out of their research: the first is the size of the organisation; the larger it is, the more complex the issues are likely to be (Reiss, 1985). This is supported by research undertaken by Williams and Matheny (1984) who found that size of firm was an indication of whether or not social regulation took place. This size and complexity issue will affect regulators' ability to understand the business and the length of time it takes for them to understand it, in terms of, for example, structure, management style, and technology (Hawkins and Hutter, 1993). This will be particularly time-consuming if the businesses they regulate all differ from each other in these aspects (Hawkins and Hutter, 1993). Where technology requires niche and complex expertise, regulators may have to rely on the business internally regulating to ensure compliance (Hawkins and Hutter, 1993). Also, some employees may not be monitored as heavily as others, and as such, both regulators and organisations rely on the employees to comply with all policies. An example where this is particularly difficult is when some employees are outsourced or sub-contracted (Hawkins and Hutter, 1993). Taking these differences into account, it does not seem sensible to apply uniform standards across industries.

Overall, this research has revealed several overlapping challenges and commonalities between environmental protection and terrorist exploitation of tech platforms. The first is that the

regulated activity is very complex. Additionally, the activities that require regulation in both industries have the potential to result in catastrophic harms without any regulation. Next is that the idea of enforcing regulation is controversial and faces backlash. It affects many parties with opposing views and interests. A further challenge is that the regulation may need to cross borders and be implemented transnationally. Another is that different organisations across each industry can be found to differ in levels in their ability to engage with regulation, possibly with their willingness to engage with it as well. The final challenge they share is the lacking the ability to prove a causal effect between the activity that requires regulation and the harm that it is at risk of causing. As mentioned, it is important to find research on industries with similar challenges in order to learn from it. Aspects of this approach were considered for the regulatory framework proposed in this thesis as a result of this research, for example, regarding the approach to content removal (see chapter 7 which introduces the mandatory regulatory standards).

- Key points from the precautionary principle (the key points from each of the three industries will be discussed together at the end of the chapter)

- The precautionary principle reduces the scientific threshold for regulatory policy making. It overcomes the issue of a lacking empirical causal link between the activity to be regulated and the potential harms that could result if regulation if not enforced.
- As a result of not being based solely on scientific factors, the principle considers other factors such as public opinion and social values.
- In falling under the category of a no-risk framework, in theory, its implementation should minimize the risk of harm.
- It can be used when deciding between conflicting objectives (as seen in the Clean Air Act example).
- It may be viewed as paternalistic by users. This could be argued as a positive given the issue of bounded rationality or as negative because it may infringe on users' rights. It assumes that safety is users' main priority, however, safety may not be prioritised by all users. Users may be unhappy with the implementation of such a principle if it causes them any inconvenience or interference.
- Tech platforms may find the implementation of the principle infringes on innovation as it removes the ability to do anything that may result in harm. It ignores that such an

approach is a risk in itself given it can lack scientific considerations and that some activities that can cause harm can also create a lot of good as well.

Consumer Protection and Social Regulation

According to the United States Consumer Product Safety Commission (2017), thousands of deaths and millions of hospital-treated injuries occur annually as a result of using consumer products. Consumer protection is an area that falls under social regulation because it is concerned with setting standards that will ensure that consumers of goods and services receive an appropriate level of both quality and safety (Ogus, 1994; Feintuck, 2010; Zhi, 1992). Consumer protection is also concerned with market failures that result from information asymmetries between consumers and organisations (Feintuck, 2010). According to Beck (1991), in situations of risk, modern regulators have been putting 'the public' first. There has been an increase in consumer rights throughout Europe (Burgess, 2001), and values such as transparency, accountability and individual choice have become more prominent (Clarke, Smith, and Vidler, 2005; Lunt, Livingstone, and Kelay, 2005; Needham, 2003).

Consumer protection is different from other areas of social regulation, firstly because consumers tend to have a market relationship with the organisations they consume from, however, do not always have the power to negotiate the standards, unlike other stakeholders (Ogus, 1994). Secondly, there is a very direct relationship between the good or service and the consumer, creating a wider array of quality and safety preferences than in other areas (Ogus, 1994). Finally, there are contrasting views on whether it should be consumers who decide what the standards should allow (Dardis, 1988). It is often the case that consumers make poor choices when consuming goods and services. Sometimes this is because they do not have all the information required. Other times, they lack self-control or are simply poor at assessing low-probability risks due to the already mentioned issue of bounded rationality as well as other cognitive biases (Ogus, 1994; Asch, 1998; Sunstein, 2005; Dardis, 1988). One example is the availability heuristic which occurs when people lack statistical knowledge about a particular situation, so they consider risks to be significant only if they can easily think of instances when those risks actually materialised. A second is probability neglect which is when people ignore probability and focus on the worst-case scenario, even if it is highly unlikely, which affects their decision-making. It is especially difficult for consumers when services or 'experience' goods are being consumed because quality and safety preferences can sometimes only be determined during or after consumption, and adverse consequences can take a while to show

(Ogus, 1994). In some cases, negative externalities can also occur causing the suffering of innocent third parties (Ogus, 1994).

Consumer protection has its difficulties. Unsurprisingly, the organisations will sometimes oppose the methods mentioned, particularly the more interventionist ones. What may be less expected though is the issue of consumers opposing the protection (Vogel, 1990). Regulation often places the burden of compliance on organisations which leaves freedom of choice to consumers as much as is possible (Vogel, 1990). According to Gilbert and Wilson (2000) consumers can suffer from the phenomenon of 'mis-wanting' which is when they want things that are or can be detrimental to their welfare, and they oppose things that would be beneficial to their welfare. Vogel (1990) examined four U.S. cases in which a large portion of consumers decided to fightback against consumer protection policies and laws that were put in place primarily with their welfare in mind. The first was the installation of seat belt systems with ignition interlocks. A survey of new car owners undertaken by Ford found that one-third of those surveyed had removed the safety interlock device, and within two years, approximately one million car owners had disconnected the device. One car owner said that he disliked feeling as though he was strapped into a high chair (New York Times, 1974). Congress received a lot of mail from car owners that were unhappy with this regulation. However, overall, many car owners left the device connected, seat belt usage increased, and car accident-related deaths decreased (Vogel, 1990). There was a similar backlash to the regulation that brought in mandatory motorcycle helmet use with a number of protests across state capitals, the banning of saccharin in food, drinks and cosmetic products due to the risk of cancer, and the number of lengthy tests being undertaken on drugs hoped to cure AIDs (Vogel, 1990). Vogel (1990) argues that the common factors in these cases were that they were uncommonly visible to consumers, they directly interfered in their lives, forced changes in their behaviour, and took away freedom of choice. Many of the consumers that opposed the protection believed that it made them worse off and that they were better placed to make the decision for themselves than the government. Motorcyclists, for example, argued that helmets interfered with their visibility and hearing as well as their fun despite the regulation being found to be effective (Vogel, 1990).

As well as opposing protection, consumers' responses to regulation could counteract the protection of policies and laws. For example, the protection that could come from the installation of air bags or seatbelts into cars could be counteracted if consumers respond to this protection by speeding or driving dangerously because they think that they can count on that protection (Dardis, 1988). Peltzman (1975) undertook research investigating this and found

that the response of drivers did appear to cancel out the protection that this regulation was supposed to offer. However, Crandall and Graham's (1984) research suggested that overall, the number of accidents decreased post-regulation, despite the reckless behaviour of some. Consumers may also take the view, for example, that while wearing a seatbelt could potentially save their life, they are not necessarily going to die because they are not wearing a seatbelt. Therefore, why would they pay or undertake costs to be protected from something that may never happen? (Dardis, 1988). According to research by Starr (1968) individuals are more likely to accept risks that are self-imposed and less likely to accept risk where it is imposed on them by others. This suggests that regulation may be more willingly received in situations where a risk is not self-imposed.

- Prior Approval

Prior approval is the first of two approaches that will be discussed under consumer protection. Prior approval from an agency is common in consumer protection. As with the precautionary principle, it is based on a no-risk framework (which means that the public should not be exposed to any unnecessary risk, regardless of how small) (Lave, 1981). A well-known example is that every new drug a pharmaceutical company wants to put on the market requires approval from an agency after being put through an extensive set of tests/experiments and discussion (Ogus, 1994). For a good or service to be approved, there must be a very high level of certainty that it will not cause harm (Ogus, 1994). Agencies often issue a licence to the organisation that gives them permission to undertake the regulated activity; it is quite common for licence conditions to involve only minimum standards and for these standards to be uniform across an industry (Ogus, 1994). Agencies will consist of experts who are not worried about balancing other aspects of running an organisation. It will be their sole job to investigate the potential dangers or negative outcomes that the good or service being regulated could cause consumers and third parties. This is something that private organisations are not always deemed to be very good at doing or concerning themselves with. The agency can demand that approval is conditional on amendments or safeguards being put in place or can alternatively refuse approval.

It is clear that the implementation of a prior approval approach would not be an exact ideal fit for the current context. It is too rigid and onerous for such a complex and rapidly evolving industry such as the tech industry. Due to issues discussed previously regarding scientific uncertainty of causal links and harm, there is the risk that nothing would ever receive approval.

There are other limitations, many of which overlap with the limitations discussed previously in the section on the precautionary principle. Sunstein (2005) argues that this is a slippery, intrusive slope which puts a lot of responsibility and trust in the person or party in charge of making decisions. This makes it too interventionist as it removes consumers freedom to choose and limits competition (Ogus, 1994; Sunstein, 2005). However, there are exceptions to Sunstein's argument, for example, many drugs that have both therapeutic benefits and potentially adverse side effects are made available anyway. Therefore, being overly paternalistic is a risk but not a definite feature of all prior approval schemes. This strategy is also costly, a lot of work and very time-consuming. Agencies can take a long time to get set up, and to get to know the ins and outs of an organisation, the technology and the goods and services to a point where they are in a position to decide whether or not to approve a decision (Bernstein, 1955). Each new application for approval will most likely create a lot of administrative work, potentially causing a bottleneck situation (Ogus, 1994). There is also the risk that trying to prevent harm from taking place, creates new harms and prevents activity that could bring a lot of good to the world. For example, preventing individuals from receiving drugs that could potentially cause them harm could also be preventing individuals from receiving potential relief from those same drugs (Sunstein, 2005). In summary, this approach is best suited to situations where the level of harm is potentially significant and there is great agreement around social aversion to the potential consequences (Ogus, 1994).

Despite the finding that this approach is not an exact fit for the current context and contains several limitations, it will be argued here that there is still something to be learned from this approach. This approach puts forth the idea and highlights the importance of proactivity. The process involved in prior approval – the use of experts to identify potential risks and harms of a new feature or service before it is available to consumers could be relevant to the current context. For example, Facebook received a lot of criticism for the lack of proactive safeguarding and policy measures surrounding their livestreaming feature in the aftermath of the Christchurch attack (Wong, 2019; BBC, 2019a). It is argued that it would be useful to have some sort of similar process whereby tech platforms must spend more time proactively identifying risks at the design and development stage and implementing amendments and safeguard before their features are consumed by users, as opposed to only happening after something bad happens. Although it will likely be impossible to predict and prevent every possible risk, such a process could minimize and prevent some risks. This approach increases the platform's accountability. Such a process is considered in the framework proposed in this

thesis (see chapter 7 introducing the mandatory regulatory standards). In addition to the advantages of proactivity, other advantages of such an approach are that there can be confidence that a certain level of quality and safety will be ensured (Ogus, 1994; Christoffel and Christoffel, 1989). Secondly, this strategy contains the earlier mentioned paternalistic argument that putting the decisions in the hands of experts will result in better decisions being made, (in terms of safety and public interest), than if the consumers themselves were left to make decisions (due to issues such as bounded rationality).

- Key points from the prior approval approach

- Although it was established that this approach is not an exact fit for the current context due to its rigidity and onerous nature, this did not mean that there was not still something to be learned from it.
- One aspect of the approach that could be applied in the current context is the implementation of a thorough risk assessment at the design and development stage of any new features or services the tech platform plans to implement. This would involve having a group of experts in that technology and terrorist exploitation of tech platforms assess the feature or service for potential misuses and risks and where possible, implementing changes and safeguards to minimize this risk.
- Failure to include such a process is likely to result in the continuation of situations similar to the Christchurch attack where such amendments and safeguards are not implemented until after something bad has happened.
- Advantages of this approach are that it minimizes risk of harm and increases platform accountability. Additionally, it could help to create a culture that prioritises health and safety which is discussed in more detail in the below occupational health and safety section.
- Limitations of the approach are that it could be seen as paternalistic by users and limiting innovation by platforms. It could also be time-consuming and costly. Finally, it could prevent the implementation of a feature or service that although contains risks could also bring a lot of good.

- Information Regulation

A different approach used in consumer protection is information regulation which is part of Lave's (1981) market regulation framework. The idea of 'perfect' information, that is, that

individuals have all the necessary information on the goods or services that they are considering purchasing, as well as all the alternative goods or services they could purchase instead, the consequences of purchasing any of these goods and services, and then being able to process all of that information and making a rational choice for themselves, is mentioned throughout regulatory literature (Ogus, 1994; Lave, 1981). This situation alongside no transaction costs, no economies of scale on production, and no externalities creates a Pareto optimal equilibrium; that no one can benefit without making another person worse off (Lave, 1981). It is argued that in such a situation, barely any regulatory intervention is needed because consumers will make decisions for themselves, and this will in turn push organisations to do better to gain customers and keep their customers, thus creating a more competitive market (Lave, 1981; Driesen, 1997). This approach assumes that because risk is an undesirable characteristic, anything with an above minimal level of risk is unlikely to be chosen (Lave, 1981). However, these assumptions and the idea of 'perfect' information, rarely, if ever, exist in reality (Ogus, 1994; Lave, 1981). Both buyers and sellers are often able to influence prices (Lave, 1981). There is almost always some degree of uncertainty or missing information (Ogus, 1994). Intentionally or not, organisations are not always as transparent as is necessary for this approach, or if they are, they are not always equally transparent. Sometimes, there can be instances where information is misleading, and for obvious reasons, organisations tend to only include the positive qualities of their products in their advertising, not negative qualities (Ogus, 1994). Further, there can be risks or consequences that the organisation has failed to foresee. Finally, there is the earlier mentioned issue of bounded rationality (Ogus, 1994). Ogus (1994: 39) raises an important question:

> "given that information is costly to supply and assimilate, the relevant policy question is rather whether the unregulated market generates 'optimal' information in relation to a particular area of decision-making, that is, where the marginal costs of supplying and processing the level and quality of information in question are approximately equal to the marginal benefits that are engendered".

Despite this, some still argue that consumers would benefit more from accepting minor undesirable characteristics instead of the government regulation (Lave, 1981). Therefore, the strategy 'information regulation' is used in consumer protection. It does not apply to the process of producing and supplying, but instead, to the information (that should be) supplied to consumers pre-purchase (Ogus, 1994). This form of regulation should offer insights into

aspects affecting price, quality, safety, and compliance with standards, and it should also prevent the organisation putting out any incorrect or misleading information (Ogus, 1994). As technology has developed, public policy has been increasingly reliant on "relevant, timely and, especially credible information" and "can actually *constitute* policy, even if only on a contingent or provisional basis" (Majone, 1977: 264). This method gives organisations the opportunity to go into more detail than they perhaps have previously, to explain why they make the decisions that they do, and also allow organisations to explain how they cope with competing demands and preferences. According to Ogus (1994: 121), if consumers do not receive "the 'optimal' amount of information, that is where the marginal benefit arising from that amount of information is approximately equal to the marginal cost of producing and communicating it, consumers sustain a welfare loss".

Unlike the previous methods discussed where information is used to create rules, information regulation works by trying to subtly and indirectly change consumer behaviour (Ogus, 1994; Majone, 1997). This can be through warning labels or any other forms of communicating risk (Majone, 1997). This method could be argued to be a form of libertarian paternalism which tries to influence the choices of consumers so that they end up in a better position than they would otherwise have been. However, their freedom of choice is still preserved as they carry out their own cost-benefit analysis based on their own willingness to accept risk, rather than someone else doing that for them, also rendering the market more competitive (Sunstein, 2005; Breyer, 1982; Dardis, 1988). It is especially important that information is provided, and done so in a clear and simple manner, where the product or service is highly complicated or technical, and that consumers are made aware of how and where they can find this information (Ogus, 1994). However, it is noted that providing complicated and technical information is not going to be an easy task for an organisation (Bardach and Kagan, 1982).

As with prior approval, this approach is not an exact fit for the current context. It is unlikely that information regarding safety is going to be of much concern to terrorist organisations. They are likely going to try to exploit the platform regardless of the implementation of an approach such as this. This approach may not even be that useful for the average user. However, there are specific groups and types of users that may benefit from aspects of this approach. For example, activist groups and those who utilise tech platforms for documenting human rights violations are likely to require clear information on what content will be removed from a tech platform. This is important because of the errors that have been made in the past and the consequences of these errors (see Kayyali and Althaibani, 2017). The example of YouTube

providing helpful guidance to users on how they prevent erroneous removals in chapter 3 is an example of how this approach could be applied to help such users. Transparency reports can also provide helpful insight into what type of content is removed. Advertisers may also find such information useful in deciding which platforms they want to advertise on. This latter point requires consistency across platforms so as not to disadvantage certain platforms. For example, if platform X publishes a transparency report that reports terrorist exploitation of their site and platform Y does not, then advertisers might move from X to Y creating issues around competitive practices.

This approach is welcomed by those who prefer less interventionist regulatory strategies. However, the success of this method is dependent on, firstly, the organisation having a level of credibility that will ensure consumers trust them and the information they provide (Majone, 1997). Subsequently, the information needs to be able to influence the behaviour of large groups of individuals and their consumption habits whilst remaining truthful (Majone, 1997). However, unless there can be absolute certainty that consumers will definitely read, understand and consider the information provided, this method is not suitable for activities that can cause significant harm or has a high level of uncertainty around its harm, for example, cases where injury or illness could occur (Ogus, 1994). Despite the limitations, the regulatory framework proposed in this thesis drew on aspects of information regulation in its implementation of standards regarding clear policies and transparency reports (see chapter 7 on mandatory regulatory standards).

- <u>Key points from information regulation</u>

- This approach is also not an exact fit for the current context, however, there is still something to be learned from it.
- There are certain groups of users that would benefit from aspects of this approach. This includes activists and those who document human rights violations. Such users have been to known to suffer from erroneous removals due to aspects of their content seeming similar to some aspects of terrorist content. The consequences of the erroneous removal are very damaging and can affect the prosecution of human rights violations (see Kayyali and Althaibani, 2017).

- This approach promotes consistency and standardisation across the industry regarding the information that is provided to consumers. This has benefits for ensuring fair and competitive practices, as shown in the above advertisers' example.
- This approach would be favoured by those who prefer a less interventionist approach.
- This approach does, however, lack accountability – although this could perhaps be mitigated by a regulator providing minimal requirements via-á-vis the provision of information, as seen in the Online Harms White Paper
- This approach requires users to view tech platforms as credible and relies on users reading and understanding the published information.

Overall, this research has revealed several overlapping challenges and commonalities between consumer protection and terrorist exploitation of tech platforms. The first is the challenge around wide, diverse and sometimes competing consumer preferences. A common conflict that applies to both areas is between what is in the consumer's best interest (regarding safety and well-being) and what the consumer's preference is (this is often not in their best interest regarding safety and well-being). This raises the common criticism around paternalism and infringing the consumer's freedom of choice. Both areas are also expected to be proactive in their approach and produce high levels of transparency and accountability to ensure safety. These are areas where the research into consumer protection can be drawn on and learned from. As mentioned, aspects from both the prior approval and information regulation approach were applied to the creation of the regulatory framework proposed in this thesis (see chapter 7 on mandatory regulatory standards).

Occupational Health and Safety and Social Regulation

Occupational health and safety (OHS) is an issue that has been described by Thomas (1948) as well-documented, and an exemplary system for other areas of social regulation over the years (Ogus, 1994). OHS is concerned with the development, promotion and maintenance of workplaces and the implementation of policies and programs that aim to protect employees mentally and physically (Nyirenda, Chinniah and Agard, 2015). Overall, all employees should have the right to safe and healthy working conditions (Windholz, 2010). When the U.S. first set up the Occupational Safety and Health Administration (OHSA) in the 1970s, the President of the United Steelworkers of America, I.W. Abel cautioned that the agency would only improve the situation "if it is administered with a zeal and a genuine concern to diminish death, injury and illness at the workplace in every possible way…" (Wilcke, 1971: 71). The

assessment of OHS law in the UK and subsequent report by the Robens Committee in the 1970s led to considerable reforms and a new approach to standard-setting (Ogus, 1994). The Robens Report (1972) argued that there were too many OHS Acts and Regulations, and they were too complicated and focused on physical circumstances instead of employees and safe systems. The report recommended that legislation include broader statements when setting goals for safety, specific hazards, and specific industries, and encourage organisations to adopt a health and safety culture.

Ogus (1994) discusses the loss abatement model in the context of OHS, "where the marginal benefits of the precautions are equal to the marginal costs" (Ogus, 1994, p.181). According to this model, employers agree to give employees a combination of care and pay in exchange for their work (Ogus, 1994). Based on assumptions of perfect competition and complete information about risks and their consequences, the combination of care and pay should be "equal to the marginal value to the employer of the employees' work" (Ogus, 1994, p.181). It should be noted that not all employees will value care and pay equally; some may prefer higher pay and less care, and others vice versa, but in cases with these assumptions and no externalities, the level of care should be optimal (Ogus, 1994). However, in some cases an employer will only be able to increase the level of care by reducing wages (Ogus, 1994) and as with other areas of social regulation, the assumptions are seldom met (Rea, cited in Dewees, 1983). Employees are often unaware of the risks involved in their work and even when informed cannot be relied upon to make sensible decisions (Loomes and Sudgen, 1982). There may also be circumstances where employers and employees struggle to reach agreement at all, particularly in situations where there is uncertainty surrounding the nature and extent of the risk (Ogus, 1994).

- <u>Health and safety culture (including consideration of psychosocial risks)</u>

It is common in occupational health and safety approaches to encourage an organisation to adopt a health and safety culture, to be more critical of themselves and develop a culture of self-evaluation (Parker, 2002; Gunningham, 2007; Gander, Hartley, Powell, Hitchcock, Mills and Popkin, 2011; Kim, Park and Park, 2016). A health and safety culture is,

> "the product of individual and group values, attitudes, perceptions, competencies and patterns of behaviour that can determine the commitment to, and the style of proficiency of an organisation's health and safety management system" (Health and Safety Commission, 1993, P.4).

Characteristics that are required to implement such a culture are having leaders who heavily promote health and safety practices, having health and safety practices communicated clearly as an organisational value, decentralized decision-making and clear accountability of who is responsible for different actions and their outcomes, on-going staff training and finally, integrating health and safety into all areas of the organisation (International Atomic Energy Agency, 2006). In order to succeed in such a culture, it must be believed throughout the organisation that safety is more important than production (Braithwaite, 1985). Research has found that accidents have decreased in organisations that developed a health and safety culture (Yau, 2014; Park, 2013).

An important point raised in occupational health and safety literature is that this should concern more than just the obvious physical hazards and harms. It should also consider psychosocial risks and harms that may be lesser known. Research has found that many organisations neglect psychosocial risks (Maxwell, 2004; Johnstone, Quinlan, and McNamara, 2011). A European Survey of Enterprises on New and Emerging Risks which surveyed 28,000 organisations across 31 European countries reported that psychosocial risks were found to be one of the key concerns, however, less than a third of the organisations had processes in place to support or address these risks (Leka, Jain, Lavicoli and Tecco, 2015).

> "Psychosocial risks arise from poor work design, organisation and management, as well as a poor social context of work, and they may result in negative, psychological, physical and social outcomes such as work-related stress, burnout or depression" (European Agency for Safety and Health at Work, 2020, no page number).

Examples of psychosocial risks are excessive workloads, working to conflicting demands, lack of clarity surrounding decisions or why something is done in a certain way, lack of employee involvement in making decisions that affect the employee and the way the job is done, job insecurity and a lack of support (European Agency for Safety and Health at Work, 2020).

Overall, this research has revealed several overlapping challenges and commonalities between OHS and terrorist exploitation of tech platforms. The first is the concern around employee well-being. As was found in the OHS literature, the tech industry has faced a lot of criticism around its approach to ensuring employee well-being. Moreover, much of this concern has also been around psychosocial risks. One example is the number of content moderators who claim to have developed mental health conditions such as PTSD because of their work duties/work

environment and have subsequently pursued legal action against the tech platform that they worked for (Boran, 2020; The Guardian, 2018; Gilbert, 2019a). These are areas where the research into consumer protection can be drawn on and learned from. The research into a health and safety culture influenced the creation of several of the mandatory regulatory standards (e.g., ensuring user safety is considered at the design and development stage and mandatory user flagging mechanisms). The research regarding the importance of considering psychosocial risks as well as physical risks was used in the creation of the mandatory regulatory standard that focuses specifically on ensuring employee well-being (see chapter 7 on mandatory regulatory standards).

- Key points from a health and safety culture

- Research has suggested that the adoption of a health and safety culture throughout the organisation has helped to reduce the number of health and safety incidents that have taken place.
- Given the criticisms that tech platforms have faced from governments for being reactive, both generally (May, 2018,) and in regards to specific attacks, such as earlier mentioned Christchurch attack (Wong, 2019; BBC, 2019a), this is an approach that is argued as applicable to the current context. Overlapping with some of the other approaches discussed in this chapter, there is a need for a more proactive approach to ensuring safety.
- The research suggests several characteristics that contribute to the adoption and success of such a culture. These have been considered and drawn from in the creation of the regulatory framework proposed in this thesis.
- The research highlighted that in many organisations, physical risks are prioritised with psychosocial risks being neglected. As mentioned, the psychosocial risks thought to be involved with the work undertaken by many tech platform employees has already led to claims of mental health conditions and lawsuits. Therefore, the regulatory framework proposed in this thesis will ensure that psychosocial risks of employees are not neglected.

Organisational Differences

A common argument made throughout all three of the areas of social regulation literature was that organisations across the same industry can differ on important factors that can affect their

ability to comply with regulation. For example, financial capabilities, number of staff, and other industry or state-dependent factors (Ogus, 1994; Baldwin et al., 2010; Gunningham, 2007). With these differences come different types of risks, some organisations may contain more 'regular risks', which are those that are common and should not cause any trouble for the standard-setter when setting standards because they do not need any highly technical information or training (Ogus, 1994). On the other hand, some organisations may have more 'special risks'. These risks may result in dire consequences or occur when specialised information is required to set standards (Ogus, 1994). The difference in these risks may depend on the organisation's production procedures and choice of technology. In organisations that are highly technical and complex, sometimes it is the case that the organisation's own employees will know the technology and processes better than any other regulatory party ever could (Gunningham, 2007).

Further, the willingness and motivations of the leaders of organisations can be a crucial consideration when setting standards. Gunningham (2007) gives an OHS example of four leadership categories, however these could arguably be adjusted to apply to the other areas of social regulation. The first is termed 'OHS leaders' who are leaders that excel in prioritising the identification and management of health and safety risks to their employees. Some even ambitiously aim for zero fatalities and serious injuries. The motivation is thought to be that health and safety incidents cause major disruption to the work process, result in compensation costs, cause staff absences, create negative publicity, and put the organisations licence to operate at risk. The second category is 'reluctant compliers'. This is when leaders view tackling OHS as a burden, with no moral obligation to do more than comply with the letter of the law and, as such, take a minimal approach. These leaders may be prone to taking shortcuts, particularly in cases where they are not closely monitored. The third category is 'the recalcitrant', who view OHS as a considerable business cost with very little in the way of compensating benefits. As a result, OHS is seen as secondary to production and short-term profit maximization. This category of leaders complies with regulations but only to avoid penalties and may participate in avoidance strategies such as failing to provide appropriate safety equipment and training if they can. The final category termed 'the incompetent'. This includes leaders who, in OHS, typically run small businesses, and are not completely aware of OHS risks and legislation. They are often focused on keeping their business financially afloat. If there were to be a mix of these leadership categories across an industry, the regulators would

need to get to know which organisation has which type of leadership and then proceed from there, as they will likely face differing compliance issues.

The findings of the analysis into tech platform's blogs in chapter 3 support this argument that organisations across the same industry can differ significantly in a variety of ways, for example, in size, available and accessible resources, values, and missions etc. Therefore, these differences and management styles should be an important consideration for any new framework. Simply setting standards and expecting all of the tech platforms within its scope to comply without accounting for differences between them and the challenges some may face is unlikely to be effective. The framework proposed in this thesis will make an effort to acknowledge and account for such differences and challenges when developing the approaches and strategies to achieve compliance with the mandatory regulatory standards (see chapter 8 on the four regulatory tracks).

Discussion of key points from each approach

This chapter examined the three previous areas where social regulation has been applied. These were environmental protection, consumer protection and occupational health and safety. These three areas provided a wide range of approaches all of which had aspects that were both desirable and not so desirable for an approach to regulate terrorist content on tech platforms. The first key finding was that social regulation theory provides approaches for situations whereby there is uncertainty as to a causal link between the activity that would be regulated and the harmful outcomes it is thought to produce. Although there is not complete uncertainty in the current context given the empirical research that argues that the internet facilitates radicalisation and terrorist activities, where there are gaps in the research and subsequent uncertainty (i.e., can viewing content lead to direct action), social regulation provides a solution. The precautionary principle can be applied in such situations. As part of a no-risk framework, the precautionary principle should, in theory, lead to the lowest levels of harm that is possible. It is justified in situations where a lack of enforced regulation could lead to catastrophic outcomes (e.g., global warming, a terrorist attack). The principle does face criticisms for lacking the requirement of scientific evidence, however, this can also be viewed as a benefit as it does not allow activities that should be regulated to fall between the cracks because of a lack of scientific evidence. It also faces criticism for being paternalistic, this can also be countered, however, by phenomenon such as bounded rationality. Finally, there is the fear that such an approach could lead to the impediment of innovation. This latter criticisms

regarding paternalism and innovation are shared by the prior approval approach in the consumer protection research. However, a key take-away from this approach is the need for a more proactive approach to safety. Tech platforms have faced criticism for prioritising user safety reactively only after something bad has happened. These two approaches are likely to be helpful when facing conflicts between interests such as a platforms' wish for innovation and the need for user safety. The prior approval approach is not an exact fit for the current context, however, the idea of implementing greater accountability via processes that increase focus and consideration of user safety at the design and development stage of new features and services is argued as an important consideration for the current context. This overlaps with the adoption of a health and safety culture that is implemented in OHS.

Information regulation raises the issue of transparency surrounding the efforts and workings of an organisation. The erroneous removal of content documenting human rights violations is an example of why this is relevant for the current context. Information regulation was also not an exact fit for the current context but aspects of it such as the promotion of transparency and a standardised approach are relevant for the proposed regulatory framework. This approach does however contain limitations such as the lack of accountability. Finally, social regulation raises an important point that regulatory considerations must stem wider than just consumers. Under OHS it raises the issue of employee well-being and the problems with neglecting psychosocial risks and harms. The example of the tech platform employees that have developed mental health conditions and pursued lawsuits against the platforms is an example of why such considerations are important under the current context. In addition to these key findings, another theme that emerged throughout the literature of all three areas of social regulation was that organisations are likely to differ greatly across an industry in a variety of ways that may affect their ability to comply with regulation. Therefore, it is important that any proposed regulation acknowledges these differences and the challenges that they may create and accounts for them in the creation of the regulatory strategies.

Whilst all of the above approaches have limitations that would affect their application in the current context, they all also produce useful, relevant ideas and lessons that can be carefully considered and applied to the current context. Each of the discussed approaches spanning the three areas where social regulation has already been applied were instrumental in the development of the regulatory framework that is proposed in this thesis. All of the strategies included in the regulatory framework are either based on or were influenced by the findings and evidence presented in this chapter. It is argued that the research and considerations that

have been undertaken in this chapter have strengthened the approach taken in the development of the regulatory framework and provides the framework with a strong theoretical underpinning.

Conclusion

This chapter aimed to answer the following questions:

7) Is social regulation theory applicable to this regulatory context?
8) What is there to be learned from examining social regulation in other regulatory contexts?
9) Could these strategies be applied in this regulatory context?

In conclusion, social regulation theory is concerned with a broad array of non-economic issues (Ogus, 1994). It is mainly concerned with regulating issues that concern the public interest, the promotion of human rights, social solidarity, social inclusion and general societal good (Ogus, 1994; Prosser, 2010; Wilson, 1984). Social regulation theory is appropriate for the current context for a number of reasons. The issues that it is concerned with are directly endangered by terrorist exploitation of tech platforms. It also provides an approach that overcomes the difficulty of proving a causal effect between the activity to be regulated and potential harms it is thought to cause. Given the research findings that the internet is a facilitative tool for radicalisation and terrorist activities, the failure to regulate and consider social issues could result in a failure to protect individuals who are vulnerable to radicalisation and individuals that are the target of inciting, dehumanising and fear inducing content. The scale of the consequences of these failures is unknown but has the potential to be catastrophic. Further, social regulation will ensure that the platforms are held accountable if they fail to engage with the regulation. This will help to overcome a number of the market failures that have been identified as occurring during the period of self-regulation that has taken place in recent years. Finally, social regulation theory has the advantage of providing a plethora of previous research to draw on and learn from. It cannot be expected that regulation will be effective without having researched the approach in other industries. This includes how the regulator and regulatees responded to the approach, the outcomes, advantages, limitations and challenges. Each of the approaches that were discussed in this chapter provided insight, ideas and evidence that informed the regulatory framework that is proposed in this thesis. It is argued that undertaking a process such as the one in this chapter prior to the development of new regulatory proposals can only strengthen regulatory frameworks.

Chapter 6: Regulatory Framework: Objectives and Ethos

Objectives

This chapter is going to propose four objectives that underlie the framework that is put forth in this thesis. Although previous chapters have discussed terrorist exploitation of tech platforms and the need to counter this, it should not be forgotten that tech platforms, although abused by bad actors, also have many benefits and have empowered their users globally in a variety of ways. These benefits include but are not limited to access to an abundance of information, anonymity, cheap communication in real time, the ability to make new connections, and the ability to raise awareness of global issues to a global audience (Aziz, 2015; Benson, 2014; Wu, 2015; Bertram, 2016). As argued in the previous chapter, regulation should attenuate the negative aspects of an activity whilst preserving its positive aspects; therefore, the benefits provided by tech platforms must not become lost as a result of the proposed framework. It is also important that tech platforms are not burdened by actions that could reduce market competitiveness or impede their ability to grow and develop unnecessarily. Therefore, the first three objectives in this framework aim to preserve these benefits through the promotion of innovation, freedom of speech and user autonomy. Whilst these three objectives are crucial, efforts are also needed to prevent tech platforms from being exploited by terrorists. The fourth objective – to prevent harm - is essential to counter online terrorist content on tech platforms. Whilst the first three objectives all seek to promote rights and interests of various stakeholders, the fourth objective is different because it seeks to prevent behaviour that will negatively impact these same stakeholders.

The objectives of the proposed regulatory framework are as follows:

1. To preserve and promote innovation
2. To promote freedom of speech
3. To promote user autonomy
4. To prevent harm

Before the chapter explores the four objectives, it is important to highlight several points. The first is that it is possible for synergies to emerge between the platforms. For example, one synergy that will be discussed later in the chapter is between promoting innovation and preventing harm. The tech platforms must be able to be innovative in order to tackle terrorist use of their site, one example of this is with their creation of artificial intelligence technology. A second point is that throughout the discussion, the reader will see that there can be differing

views of what each objective requires. The objectives are not black and white. For instance, there may be disagreement between what people believe should count as harm or what speech deserves protection. Finally, the four objectives may conflict with one another. Examples could be that the speech of one person (for example, racist speech) is thought to impede the autonomy of other users, or that the tech platforms' need to innovate to stay competitive may stand in the way of their ability to prevent their users from harm. The importance or weight of the objectives cannot be ranked. When a conflict occurs, a decision on which objective should be prioritised must be made depending on the context of that conflict.

Preserve and promote innovation

The first objective is to preserve and promote innovation. This objective has two dimensions: the first is to ensure that the framework does not impede innovation; and the second, is to ensure that the framework promotes innovation. Innovation is described as the ability to exploit new ideas with the main characteristic involving change (Neely and Hii, 1998; Rogers, 1983). In order to innovate, an organisation must be able to take,

> all those scientific, technical, commercial and financial steps necessary for the successful development and marketing of a new or improved manufactured products, the commercial use of new or improved processes or equipment or the introduction of a new approach to a social service (Organisation for Economic Co-operation and Development (OECD), 1981: 15-16).

A review of literature revealed that innovation has a direct effect on a company's ability to be competitive (Neely and Hii, 1998). Further to this, a study that interviewed 75 high performing organisations found that innovation was the key theme amongst the organisations (Neely and Hii, 1998). Studies by Cosh and Hughes (1996) and Geroski and Machin (1992) also suggests that innovation plays a key role in firm performance, more specifically, in long-term profitability and growth. Innovation, therefore, is necessary for an organisation to adapt, learn and exploit new ideas which are essential in a market that evolves quickly (Neely and Hii, 1998).

Research has shown that external influences, such as state regulation, is one of the three main factors that influence an organisation's ability to innovate (alongside culture and leadership, and internal processes) (Neely and Hii, 1998). If new regulations are well thought-out, it has been argued that they can be "the key precipitating events that trigger innovation" (Patanakul and Pinto, 2014, p.103). However, if done poorly, can greatly hinder an organisation. In order

to ensure that the framework does not impede innovation, the framework must allow sufficient space for tech platforms to be able to build an identity and products. For this reason, the proposed framework does not want to impede tech platforms technically, commercially or financially in their attempts to build their organisations and develop products. Innovation is complex and non-linear and therefore requires a great level of flexibility (Neely and Hii, 1998). The framework does not want to create market conditions that will reduce competitiveness or over-burden any one type of tech platform (for example, placing disproportionate financial burdens on smaller or newer tech platforms). Innovation is not only relevant to existing companies but is an important consideration in encouraging the creation of new companies. A framework that does not seek to preserve and promote innovation may deter tech 'start ups' which in turn will lead to reduced competition and subsequently, innovation. To meet this objective, the framework is going to, first, implement a mixture of principles and rules when setting the regulatory standards, to allow for a level of flexibility where possible. Second, the development of the framework has acknowledged that not all tech platforms have the same accessibility to knowledge and resources which can affect their compliance abilities. This latter point is addressed thoroughly in chapter 8.

It is important that while the framework seeks to preserve innovation, tech platforms must still be mindful of the potential for misuse of their platform. When innovating new features and services, the platform must consider how they can mitigate or prevent harm, where possible. For example, if innovating to allow users to livestream, platforms must consider the ways in which this service could be misused and how they can safeguard and mitigate potential risks. This is one example, however, of a conflict between innovation and harm prevention.

Although the framework is concerned with not impeding tech platforms' innovation from a business perspective, innovation is simultaneously required to meet the fourth objective of preventing harm. Terrorist use of tech platforms evolve at a rapid pace with many groups adapting and creating countermeasures in order to remain online (LaFree, 2017; Conway et al. 2019) and it is therefore necessary that tech platforms also evolve at a rapid pace. In order to do so, the tech platforms must be able to innovate. Due to the enormous volume of content that is published on many tech platforms on a day-to-day basis, the use of different technologies is critical to the mission of the framework. Tech platforms require considerable freedom to innovate in this highly complex area of technology.

This framework, will therefore, also work to promote innovation. Promoting innovation requires the framework to provide the tech platforms with an incentive to innovate. Research suggests that an organisation's decision to engage in innovative behaviour depends on the costs and benefits to the organisation (Gil, Miozzo, and Massini, 2012). However, according to Neely and Hii (1998), the most innovative organisations tend to regard regulation as a positive. These organisations "participate in standard setting and influence regulatory procedures" (p.26). Therefore, the framework proposes greater engagement with the regulator from tech platforms, as well as other stakeholders, regarding regulatory procedures. Research has also suggested that the sharing of knowledge and best practice among organisations increases their innovative capacities (Neely and Hii, 1998; Goh, 2005). Increasing the amount of training a workforce receives also leads to an increase in knowledge across an organisation, resulting in increased innovation (Patanakul and Pinto, 2014). Further research suggests that support in research and development, and technical assistance create a favourable business environment, and promotes and sustains innovation (Patanakul and Pinto, 2014; Nelson and Rosenberg, 1993). The regulator will therefore act as a source of knowledge and guidance for tech platforms and the framework puts forth that the tech platforms should work with a range of experts across a variety of stakeholders (e.g., NGOs, CSOs, academics, other tech platforms and collaborative ventures) to aid their efforts in countering online terrorist and extremist content on their sites. These efforts include but are not limited to gaining knowledge and training resources, research and development, and technical assistance. This will be of particular benefit to tech platforms that lack the resources to make these efforts on their own. The framework has a focus on assisting and guiding tech platforms as to how to make their existing resources go further in their efforts to counter this content, thereby putting in place the conditions required to support innovation.

Compliance with the framework therefore allows a tech platform a large degree of input, freedom and flexibility to innovate and provides access to range of expertise and resources. This innovation, should, according to the literature lead to the tech platforms increasing their competitiveness and business performance (Neely and Hii, 1998; Cosh and Hughes, 1996; Geroski and Machin, 1992). At the same time, this compliance should reduce the volume of terrorist content on the platform, thus meeting the framework objective of preventing harm and subsequently creating a safer environment for their users. Non-compliance, however, will result in enforcement action.

Promote freedom of speech

The second objective of the framework is to promote freedom of speech. This is a complex objective given that the legal protections for free speech differ greatly depending on where one is in the world. For example, under the First Amendment of the US constitution, the United States grants a greater degree of free speech than in Europe under Article 10 of the European Convention of Human Rights (ECHR) (Zoller, 2009), and countries such as China place strict restrictions on speech. The differences in free speech laws are therefore difficult to harmonize (Aziz, 2015; Bychawska-Siniarska, 2017; Barednt, 2007). Conflicts occur when one country refuses to enforce the decision of a court in another country to regulate online speech (Barendt, 2007). Despite these differences, both the First Amendment and the ECHR recognise that freedom of speech is a fundamental human right that underpins democracy and many other human rights (Mednel, 2007; Redish, 2013; Barendt, 2007).

There is an abundance of scholarly literature that discusses why free speech is so important. Meiklejohn (2000) argues that the primary purpose of the First Amendment is to ensure that people understand political issues in order to participate in a democratic society, whilst Smet (2010) argues that it is crucial to protect individuals from unwarranted interferences in their lives. Without freedom of speech, people would be unable to criticize the government without fear of censorship or prosecution (Chemerinsky, 2018) and unable to create transparency and accountability for human rights abuses (Bychawska-Siniarska, 2017). Further, people would be unable to enjoy other rights such as the right to vote, free assembly and freedom of association, and the freedom of thought, belief and religion (Howie, 2018; Bychawska-Siniarska, 2017).

An argument that has been influential in the United States (yet has also received criticism) is that freedom of speech is necessary for the marketplace of ideas to work (Chemerinsky, 2018; Barendt, 2007). “Just as liberal economists consider it is wrong to interfere with the operation of a free market in goods and services, so in Holmes J.’s view it was equally undesirable to manipulate the market in ideas” (Barendt, 2007, p.11). Holmes argument is that unregulated competition in the *actual* (not simply *Ideal)* market is conducive to revealing truth (Barendt, 2007). Although not everyone agrees with this argument, there can be a preference for the marketplace over government regulation because of suspicion of government and the concern of the suppression of speech by government (Barendt, 2007; Chemerinsky, 2018).

Without freedom of speech people would not be able to protest, bring change or make advancements in civil rights (Isler, 2001; Mendel, 2017). Freedom of speech is also thought to

be essential for tolerance which it is argued should be a basic value in society (Chemerinsky, 2018). This includes tolerance of speech that is considered to be unpopular or distasteful (Chemerinsky, 2018). Freedom of speech is also necessary to ensure freedom of the press (Howie, 2018). Without, freedom of speech, people would be restricted in their search for truth (Chemerinsky, 2018), which in turn, prevents the publication and ability to search for accurate facts and valuable opinions (Barendt, 2007). The argument of truth, put forth by John Stuart Mill, can be regarded as an autonomous and fundamental good and has utilitarian considerations around the development of society, however, it must be noted that unregulated speech does not always lead to finding truth (Barendt, 2007). According to Mill (1863) the silencing of expression robs people, including future generations, of the chance to find truth. Mill also argues that it can affect those who dissent the expression more than those who hold it because their opinions are never challenged and they therefore do not consider other perspectives. Further, he makes the point that people should hear all different kinds of arguments, including opposing arguments, from more than just their own teachers; people should be able to hear opposing arguments from those who actually believe them in order to assess for themselves what they think the truth is. He says it is often only through collision with erroneous expression that the truth has any chance of emerging (Mill, 1863).

Finally, it is regarded as an essential aspect of personhood and autonomy (Chemerinsky, 2018). People must be able to express their chosen values; this is seen as vital to one's right to self-development and fulfilment (Barendt, 2007). Scanlon (1972) argues that a person is only autonomous if he or she is able to weigh the different arguments that are advanced for themselves. This, raises the argument, however, about speech that infringes the autonomy of others (Chemerinsky, 2018). An example is hate speech. The regulation of hate speech creates tensions between many rights. In opposition to the stance taken by the U.S., some have argued in recent years that hate speech should be restricted because such expression can demean and injure others (Chemerinsky, 2018), with many countries deciding that racial harmony and the protection of ethnic groups should be regarded as more important than an absolute right to freedom of speech (Barendt, 2007).

As mentioned, freedom of speech is protected in the United States by the First Amendment and in Europe by Article 10 of the European Convention on Human Rights. The First Amendment states that,

> "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances".

The European Convention on Human Rights Article 10 states that,

> "Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.
>
> The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder of crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary."

However, it is important to discuss that in neither the First Amendment nor Article 10 of the ECHR, is free speech an absolute right; it can be limited under several conditions. Under the First Amendment, there are three relevant freedom of speech restrictions to censorship of terrorist content on the internet are: "Incitement of illegal activity, Fighting Words, and True Threats" (Aziz, 2015, pp.6-7). The leading case on the First Amendment is *Brandenberg v Ohio*[37] in which a Ku Klux Klan leader was convicted under the Ohio Criminal Syndicalism statute for "advocating…the duty, necessity, or propriety of crime, sabotage, violence, or unlawful methods of terrorism as a means of accomplishing industrial or political reform" and for "voluntarily assembling with any society, group, or assemblage of persons formed to teach or advocate the doctrines of syndicalism". The Supreme Court, however, overturned the conviction and drew a line between "mere advocacy" and "incitement to imminent lawless action" (Aziz, 2015). Under the Brandenburg test, speech cannot be prohibited unless such advocacy is directed to incite imminent lawless action and is likely to produce such action.

[37] *Brandenburg v. Ohio*, 395 U.S. 444 (1969)

Incitement of illegal activity is defined as speech that advocates acts that are illegal. Fighting words is concerned with preserving order and defined as speech that provokes an audience to use illegal force against the speaker. A well-known example is *Feiner v. New York* when Feiner publicly taunted a large crowd of African and Caucasian Americans and instigated pushing and shoving that resulted in the need for two police officers to intervene (Aziz, 2015). The final restriction True Threats are defined as statements that communicate a credible expression of intent to undertake an unlawful act of violence against a group of individuals (Aziz, 2015).

Although freedom of expression applies to ideas which are shocking, disturbing or offensive under ECHR Article 10, freedom of expression can still be limited where necessary to protect a democratic society, national security, territorial integrity, public safety, amongst others, for example *Handyside v United Kingdom*.[38] This includes incitement to violence, hate speech directed at minorities and promoting Nazi ideology (Bychawska-Siniarska, 2017). An example of hate speech is *Erbakan v. Turkey*[39] where the Court asserted that expression which spreads, incites, promotes or justifies hatred based on intolerance may need to be sanctioned or prevented (Council of Europe/European Court of Human Rights, 2020). Hate speech in particular can fall outside of the protection of Article 10 because of Article 17 that no one can rely on the ECHR to undermine the rights of others (Europe/European Court of Human Rights, 2019b). Interference with one's freedom of expression must have some basis in national law.

An example where freedom of speech can conflict with another right is Article 8 of the ECHR: the right to respect for private life, (Bychawska-Siniarska, 2017; Smet, 2010). Respect for one's private life includes a person's right to develop a personal identity, form relationships and participate in various aspects of society without fear of personal information being shared without permission unless done so in accordance with the law in the interest of a democratic society, national security, public safety, protection of health, morals or the rights and freedoms of others (Europe/European Court of Human Rights, 2018). Tunick (2017) argues that while subjective desires for the right to a private life may vary, people need privacy whether they know it or not, and Moore (2010) argues that it is essential to human flourishing. Failing to respect a person's right to a private life can damage their reputation; result in the loss of property rights, unjust punishment, dignity or blackmail; damage relations of trust; prevent the ability to forge new ties; discourage non-conformity; and undermine community (Tunick, 2017). In cases where freedom of expression and the right to respect for a private life conflict

[38] *Handyside v. United Kingdom* (5493/72/72) (1979-80) 1 E.H.R.R.737
[39] *Erbakan v. Turkey* (59405/00) ECtHR, 6th July 2006

with one another, the Council of Europe/European Court of Human Rights (2018) proposes that the following criteria, as seen in *Von Hannover v Germany*[40] are considered: the contribution to public interest, the degree of notoriety of the person affected, the prior conduct of the person concerned, and the content, form and consequences of the publication. A well-known case of a conflict between freedom of expression and respect to a private life was *Perincek v. Switzerland*[41] which concerned the criminal conviction of a Turkish politician for publicly expressing the view that the mass deportations and massacres suffered by the Armenians in the Ottoman Empire did not amount to genocide (Europe/European Court of Human Rights, 2020). The applicant claimed that his criminal conviction had breached his right to freedom of expression, however, the dignity of the victims and dignity and identity of Armenians today are protected by Article 8 of the ECHR (the right to respect for private life). In this case, the court had to consider the effects that the applicant's statements would have on the public interest, whether or not the statements called for hatred or intolerance, the context that the statements were made in, whether the statements affected the dignity of the Armenian community, the international law obligation for Switzerland to criminalise such statements, and the interference with the applicant's right to freedom of expression.

There can also be tension between freedom of expression and Article 9 of the ECHR: freedom of thought, conscience and religion. Article 9 provides a person with the right to the freedom to change their religion, belief and freedom, in public or private, and to manifest their religion or belief in worship, teaching, practice and observance. Freedom to manifest one's religion or beliefs is only subject to limitations that are prescribed by law in the interests of a democratic society, public safety or the protection of public order, health, morals or the rights and freedoms of others (Europe/European Court of Human Rights, 2019a). Therefore, under this Article, a person has the right to hold or change religious and other beliefs and to put their thoughts and beliefs into action. In the case of *Otto-Preminger-Institut v. Austria*[42] where the publication of a film created a conflict between freedom of expression and freedom of thought, conscience and religion, the Court held that,

> "The respect for the religious feelings of believers as guaranteed in Article 9…can legitimately be thought to have been violated by provocative portrayals of objects of religious veneration; and such portrayals can be regarded as malicious violation

[40] *Von Hannover v Germany* (59320/00) (2006) 43 E.H.R.R.7
[41] *Perincek v. Switzerland* (275101/08) (2016) 63 E.H.R.R.6
[42] *Otto-Preminger-Institut v. Austria* (13470/87) ECtHR 20 September 1995

> of the spirit of tolerance, which must also be a feature of democratic society" (Bychawska-Siniarska, 2017).

In this case, the Court stated that such a conflict would have to consider whether or not the film contributed to public debate that was capable of furthering progress in human affairs, whether the film is an attack on the religious beliefs of others, and whether the film reaches a high level of abuse and denies the freedom of religion of others (Bychawska-Siniarska, 2017).

Finally, this objective must discuss prior restraints on speech. There are two strategies that tech platforms can implement to remove content at present. First, they can remove content reactively whereby the content is brought to the tech platforms attention, for example, by a user, after it has already been posted to the platform (Llansó, 2020b). Reactive strategies have been criticised as identifying only a small amount of content on a given platform (Suzoer, 2018) and the content could be viewed by a lot of people before removal. For this reason, many tech platforms implement proactive strategies to detect violating content. Some proactive methods detect content that is subsequently viewed by a human, others, however, block content at upload (Llansó, 2020b). This latter strategy bears similarities to the legal concept of prior restraint on speech (Llansó, 2020b). Prior restraints take place when speech must be approved by an empowered third party, for example, a public official, before it can be published (Llansó, 2020b; Mayton, 1981; Morris, 2010; Barendt, 2005). Prior restraints on speech are often criticised because of the effects the process can have on freedom of expression (Llansó, 2020b), with Bantam Books v. Sullivan being the leading case against it. According to the United Nations Human Right Committee (2011) limitations to freedom of expression must be provided by law, in accordance with a legitimate aim, and must be necessary and proportionate. One argument is that it is better to allow freedom of expression and punish any violations, than to force people to endure prior restraints (Llansó, 2020b). One of the main reasons for this is because prior restraint systems often lack the necessary transparency that would allow public scrutiny (Emerson, 1970) and a problem with the tech industry, is that the systems will lack judicial processes, leaving it to the discretion of a private company (Mayton, 1981). It is therefore difficult to know the "accuracy" of the tools being utilised. Further, users have the right to know what rules are applied to regulating their speech for due process reasons (Llansó, 2020b; Macdonald, Correia and Watkin, 2019). Balkin (2014) raises the issue of overbreadth of coverage, arguing that, prior restraints subject a much greater volume and variety of content to scrutiny than would be scrutinised by a system of punishment after publication. The technology used in such a system treats every piece of content uploaded as a potential violation

and typically removes the users' right to appeal the decision (Llansó, 2020b; Barendt, 2005). These systems could also potentially have the opposite effect and fail to detect some kinds of violating content (Guynn, 2019), particularly since automated technology often struggles to identify content that requires understanding context (Engstrom and Feamster, 2017).

Although tech platforms must be aware of the severe human rights risks that are involved in prior restraint systems, particularly to freedom of speech, and as such not use them so readily (Blasi, 1981), there must also be awareness of the risk of harm posed by terrorist content that is not blocked at upload and is potentially able to be viewed by those who may act on what is asked of them in the content. One justification of using such a system is the legal concept *fait accompli* which means that once a communication is published, the world becomes a slightly different place, the effects of the speech cannot be undone, perceptions regarding what is tolerable can be altered (Blasi, 1981). Once terrorist content is viewed by a user who decides to act on that content, this cannot be undone and the effects could be catastrophic. A justification from the tech platforms' perspective is that of responsibility. For example, if a judge makes a decision, then any adverse consequences that take place as a result of that decision will likely result in the judge having to take some responsibility for the decision (Blasi, 1981). Therefore, if a tech platform allows content to be published, and there are adverse consequences because of this decision, then the relevant governments are likely to place some responsibility on the platform. There is also an argument that the regulatory process is potentially more trustworthy than the judgements or intentions of the audience, however, neither are likely to be fully trustworthy (Blasi, 1981). Even if most of the audience is trustworthy, a small percentage could still cause irreparable harms (Blasi, 1981).

These justifications for a prior restraint system raise the question of whether all forms of prior restraint should be viewed with the same degree of hostility or whether (Blasi, 1981; Barendt, 2005), under some circumstances, such as in cases of attempts to upload verified terrorist content, not implementing a prior restraint system could cause irreparable harm. This is an example of where this objective (promotion of free speech) conflicts with the objective of preventing harm. This framework argues that, in cases of verified terrorist content, the prevention of harm is more important, given that caution is taken to minimize at all costs erroneous decision-making and systems are as open to scrutiny as they can be without revealing loopholes for terrorists to exploit. Barendt (2005, p.128) argues that in some cases "it may be contended that the imposition of a prior restraint is the only realistic means of preventing harm".

This thesis has put forth the protection of freedom of speech as an objective because of earlier mentioned reasons that it is essential for democracy, the search for truth, personal autonomy and tolerance. However, similar to the conflict that can occur between the Articles in the ECHR, freedom of speech is likely, at times, to conflict with the framework's other objectives because freedom of speech is not absolute and not easily agreed upon. In cases where the framework objectives conflict, if the regulator decides that prioritising either X or Y is acceptable, then the choice can be made by the tech platform as to which they will prioritise. *Otto-Preminger-Institut v. Austria*[43] is an example of a case where the prioritisation of freedom of speech was not straight-forward against material that was deemed blasphemous: the courts ruled that seizure and forfeiture of the material was not a violation of Article 10. However, when the regulator decides that X *must* be prioritised over Y, then the regulator will require the companies to prioritise X. An example of this could be hate speech, whereby the prevention of harm could be assessed as more important than free speech, such as with *Pavel Ivanov v. Russia*[44]. A further issue under this objective is the justification of prior restraints. Prior restraints systems are heavily criticised as infringing freedom of speech, however, in certain circumstances, such as verified terrorist content, it could be argued that such systems are the only realistic means of preventing harm. Finally, the framework aims to ensure that there are sufficient, easily-accessible processes in place to safeguard or overturn the erroneous removal of legitimate speech and that any action taken against speech is proportionate to the level of harm that it risks causing.

Promote user autonomy

The third objective is to promote user autonomy. The word "autonomy" has Greek origins and translates literally to making one's own laws (Feinberg, 1982). Many scholars have tried to define individual autonomy: according to Scanlon, in order to be autonomous, "a person must see himself as sovereign in deciding what to believe and in weighing competing reasons for action" (1972, p.252). According to Wolff "autonomy is a combination of freedom and responsibility; it is a submission to laws that one has made for oneself. The autonomous man, insofar as he is autonomous, is not subject to the will of another" (1970, p.14). Finally, in the words of Rawls, "acting autonomously is acting from principles that we would consent to as free and equal rational beings" (1971, p.516). As highlighted at the beginning of the chapter, the definition of each objective may not mean the same thing to everyone. Dworkin (2015)

[43] *Otto-Preminger-Institut v. Austria* (13470/87) ECtHR 20 September 1995
[44] *Pavel Ivanov v. Russia* (35222/04) ECtHR 20 February 2017

argues that the only features that appear to be common throughout the attempts to define autonomy are that it is a feature of persons and it is deemed a desirable quality to have. Feinberg (1982, p.447), on the other hand, comes to the conclusion that autonomy is the "capacity to govern oneself", "the actual condition of self-government", "an ideal of character derived from the conception" or "the sovereign authority to govern oneself, which is absolute within one's own moral "boundaries".

Autonomy is deemed vital for a liberal democracy and underpins many human rights (Lakoff, 1990; Pendlebury, 2004). It is often referred to as a moral, political and social ideal (Dworkin, 2015). As a political ideal, autonomy is required to counter the design and functioning of political institutions that try to impose values, attitudes and rules upon a set of individuals (Dworkin, 2015). An argument in favour of prioritising autonomy is that these impositions must be acceptable to the individuals that they are imposed on (Dworkin, 2015; Pendlebury, 2004) and must allow those affected by them to participate in their development (Pendlebury, 2004). As a moral ideal, autonomy allows an individual to make decisions regarding their own moral code and based on individual conscience as opposed to behaving according to the moral code decided by an authority figure (Dworkin, 2015). However, it is important to note that this ideal cannot be absolute, there would be many obvious potential problems with this, for example, one may decide it is immoral to steal from a small business with small revenue, however, it is morally acceptable to steal from a large international chain of stores because they will never notice the missing income. If everyone adopted this moral attitude then there would be chaos. As a social ideal, autonomy assists an individual with the problem that many non-political institutions (e.g., the media) that are not democratically elected try to influence and affect the values, attitudes and beliefs of a society (Dworkin, 2015). Given the increasingly global environment that we live in today, the autonomy of many can be compromised by processes and regulation that take place in other parts of the world, not necessarily just the country that they live in, creating many complexities (Pendlebury, 2004). This latter ideal is particularly relevant in this framework. Tech platforms are not democratically-elected, yet hold enormous power over users and user autonomy. This power is unlike that of any other private company and largely affects the autonomy and speech of users all around the world. Overall, it is argued that in order to create a good society, those who make up that society should respect the principles that regulate it, and in all three ideals, "there is a notion of the self which is to be respected, left unmanipulated and which is, in certain ways independent and self-determining" (Dworkin, 2015, p.10).

However, in contrast to the libertarian view that autonomy and authority are incompatible, May (1994, p.133) puts forth the argument that "in a practical social context, the regularity and stability provided by the authoritative rules of society seem necessary for any meaningful sense of self-determination". The example May provides is traffic laws. Traffic laws allow an individual to travel from A to B and arguably enhance rather than threaten one's ability to determine their own actions. This is because there is a level or predictability that is required to undertake such cooperative actions in society and for self-determination to take place, to predict the consequences of one's own actions (May, 1994; Hayek, 1960). It is therefore argued, that certain external factors (such as traffic laws) can help a person to positively autonomously determine their own behaviour, instead of standing in the way of their ability to do so (Raz, 1986; Dorsey, 1953; Hayek, 1960; May, 1994). The argument of whether or not external considerations (such as authority) stand in the way of autonomy is seen in Kant's view that decisions must be pure and uninfluenced by external considerations, versus, Aristotle who argued that external considerations are at the very core of the determination of moral duty and practical wisdom (Kant, 1981; May 1994). Autonomy is argued to not necessarily need detachment from external influences. Instead, it requires that a person assesses external influences and then makes a decision for themselves based on this assessment (May, 1994).

A significant problem with the idea of autonomy in modern democratic societies, however, is that it is very complex and holds many tensions and conflicting demands (Dworkin, 2015; Lakoff, 1990). Autonomy is believed to be necessary for human flourishing (Gardbaum, 1996). A utilitarian assumption is that everyone aims to be happy and that each person is the best judge of what will make them happy (Lakoff, 1990). It could be argued that, typically, people enjoy being able to act freely, and so interfering with their ability to do so interferes with the way in which that person wants to be motivated and changes the kind of person that they wish to be and the type of life they wish to live (Dworkin, 2015; Gardbaum, 1996). By exercising autonomy, people are able to "define their nature, give meaning and coherence to their lives, and take responsibility for the kind of person they are" (Dworkin, 2015, p.14). Interfering with a person's autonomy interferes with their ability to do so (Pendlebury, 2004). As such, autonomy is regarded as essential for human well-being and those without it often resent not having it (Pendlebury, 2004).

What makes this complicated is that autonomy does not always lead to objectivity in one's decision-making, nor does it ensure an individual will make the choice that would be in their own best interest (e.g., regarding their safety) (Dworkin, 2015). Adding to this, there is the

complication of individuals who cannot judge what is in their best interest because, for example, they are a child or they are an adult suffering from a psychiatric condition that affects their decision-making abilities. It is argued that in some cases, interference by authority will prevent individuals from choosing lives that will lead them away from a path of human flourishing (Gardbaum, 1996). This is further complicated because in some cases individuals choose to make decisions for themselves even though they are very much aware that there are other parties or authorities that have more information than them and are therefore better placed to make such decisions. One reason why they may choose to make their own decisions anyway is in order to lessen the power of such authorities because the individual views it as having too much power (Mill, 1863). In other circumstances, it may be the case that an individual is the best party to make a decision that affects themselves (Lakoff, 1990) albeit they are not simultaneously best placed to judge the effects of their actions on other people (Turner, 2014). Finally, "autonomy can be in conflict with emotional ties to others, commitments to causes, with authority, tradition, expertise, leadership and so forth" (Dworkin, 2015, p.11).

Autonomy is also complicated because it does not always result in an individual wishing for liberty and freedom, in some cases, individuals wish to have their liberty and freedom limited because they recognise that doing so protects them in one way or another (Dworkin, 2015). In other cases, individuals have a conflict between what is referred to as their first and second-order desires (Gardbaum, 1996). For example, a first-order desire may be a desire to smoke, but the second-order desire is a desire to not have the desire to smoke (Dworkin, 2015). Sometimes people are able to identify with, for example, an addiction or behaviour, whilst other days, it feels alien to them. Mill (1863) puts forth an example whereby, if one saw a person attempting to cross a bridge which is deemed as unsafe, and there was no time to warn the person, they might pull him back without any real infringement on his autonomy because autonomy consists of doing what one desires and he most likely does not desire to fall from a bridge. However, in cases when there is not absolute certainty of a person's desire, no one other than the person crossing the bridge will know of their desire and motive to incur risk. In this case, it would be only respectful of his autonomy to warn him of the risk, not to take any action on his behalf. It is argued by some that what is crucial to being autonomous is having the capacity to ask oneself whether or not they identify with or reject the reasons for which they act and are choosing this for themselves rather than on the basis of authority (Dworkin, 2015; Gardbaum, 1996). Ultimately to be autonomous, it is argued that one should be able to change

their first-order desires if they wish to and not have this interfered with by someone else/an authoritative institution (Dworkin, 2015).

Overall, individual autonomy is very complex, not always agreed upon and contains some conflicting demands. John Stuart Mill attempted to define the boundaries of autonomy with his Harm Principle which will be discussed in the next and final objective and raises many possible conflicts between this objective of promoting autonomy and the preventing harm objective. The aim of this objective is to guarantee users the right to be autonomous, express themselves freely and have access to opportunities to take part in decision-making processes that affect them and have their voices represented. However, as will be discussed in the next objective, this will only be possible in cases where autonomy does not interfere with preventing harms to others because although autonomy is important and represented in this framework, similarly to the previous objectives, it is not absolute. Although autonomy is arguably a nebulous concept, it is important that it is an objective because of the enormous, global and unique power that tech platforms hold over user autonomy. Tech platforms are not democratically-elected and hold private business interests that may conflict with user autonomy. Implementing user autonomy as an objective provides users with the opportunity to protect their autonomy and to create accountability around autonomy. This framework will ensure that tech platforms implement processes to involve users in decision-making processes where the outcomes affect the autonomy of the users. The framework will also put forth the granting of certain user powers to allow users greater control over the content and accounts that they will see.

Prevent Harm

The final objective of the framework is to prevent harm. As mentioned, John Stuart Mill put forth his Harm Principle which in discussing the prevention of harm includes an attempt to define the boundaries of autonomy. Although Mill was not specifically referring to terrorism or radicalisation in this work, "he addresses a fundamental question about when governments are justified in exercising power over their citizens" (Hardy, 2020, p.22). Mill's harm principle states that as individuals, we should be able to do as we like, "subject to such consequences as many follow; without impediment from our fellow-creatures, so long as what we do does not harm them, even though they think our conduct foolish, perverse or wrong" (1863, p.28). He argues that societies cannot be free without this liberty and that each individual should have autonomy over his own health, whether this be bodily, mental or spiritual. "Mankind are greater gainers by suffering each other to live as seem good to themselves, than by compelling each to

live as seems good to the rest" (Mill, 1863, p.29). As long as harm is not taking place then advice, instruction, persuasion and avoidance are the only measures in which any person or institution can try to influence or change one's actions (Mill, 1863). However, he raises the point that actions should not be as free as opinions, and when an action causes harm to others without a justifiable reason then interference should take place, thus limiting one's autonomy (Mill, 1863). When harm takes place, social or legal punishments can be inflicted upon the individual to hold them accountable and protect society (Mill, 1863). The Harm principle can essentially be viewed as an anti-paternalism principle (Turner, 2014).

Therefore, under Mill's Harm Principle, an individual's autonomy can only be interfered with, against his will, if it is to prevent harm to others. A significant problem with this principle, however, is a lack of clarity as to where to draw the line on "harm" (Turner, 2014; Ashworth and Zedner, 2012). Further to this, the harm principle also fails to state how much liberty should be sacrificed in order to prevent harm (Ashworth and Zedner, 2012). Mill is criticised for some of his points such as justifying harm as "he must not make himself a nuisance to other people" (1863, p.108) and vagueness with statements such as "actions in general which affect others" (1863, p.190). Some scholars have proposed what they think Mill meant by "harm", for example Riley (1998) argues that Mill meant when one suffers damage against their wishes including physical injury, forcible confinement, financial loss, and damage to reputation. However, due to the lack of clarity, many scholars have since tried to propose their own restricted conceptions of "harm" (Turner, 2014). These include "injury to the vital interests of others" (Gray, 2013); actions that "violate or threaten imminent violation of those important interests of others in which they have a right" (Brink, 2008); "prejudice to fundamental interest" (Dyzenhaus, 1993); and "perceptible damage experienced against one's wishes" (Riley, 1998). Finally, Feinberg (1984, p.33), who is classed as one of the leading scholars in this area, has described harm as the "thwarting, setting back, or defeating of an interest", that are wrongs, in a circumstance which was not consented to by the person whose interests are harmed. A wrong is defined as unjustifiable behaviour that violates another's rights (Feinberg, 1984). Most definitions of harm seem to agree that harm is not just any negative consequence inflicted upon a person but that harm takes place when someone's rights are violated or damage is done to their vital interests (Turner, 2014). This supports Mill's view that "as soon as any part of a person's conduct affects prejudicially the interests of others, society has jurisdiction over it" (1863, p.145).

Therefore, one consensus is that interference can be made to one's actions if those actions are going to cause harm in the way of violating another's rights or vital interests. An example of these "rights" would be free speech. Vital interests, on the other hand, are defined by Mill as interests in autonomy and security; "these interests are satisfied when men refrain from invading one another's autonomy and from undermining one another's security" (Mill, 1998, p.157). Feinberg (1984, p.34) also puts forth a similar but slightly broader definition of interests: interests are "distinguishable components of a person's well-being" that can be "invaded" by the self or others. An example could be one's health. He argues that one way to test whether or not an interest has been set back is to think about whether an interest is in a worse condition than it would have been had there been no invasion by a human. These interests are distinctly different from ulterior interests that are similar to aspirations, such as buying your dream house or raising a family (Feinberg, 1984). In this framework, anything less than the interference of others' rights and vital interests will not be considered justifiable as outweighing the users' own right to autonomy and free speech. Vital interests are defined in this framework, based on the above definitions, as those components that contribute to a person's well-being, including but not limited to their health, autonomy and security.

This brings us onto the discussion of the offence principle and how it differs from the harm principle. Whilst harm has been defined as interference with the rights and vital interests of others, Feinberg (1988, p.1) uses 'offence' to "cover the whole miscellany of universally disliked mental states" and provides some examples of what may lead to an individual feeling offended: feelings of shame, disgust, embarrassment, anxiety, annoyance and hurt, however, only when caused because of "wrongful (right-violating)" behaviour by others. The term 'wrongful' refers to behaviour that the offended person believes has wronged him, whether or not it actually has. Feinberg argued that these feelings that are made to intentionally create unpleasant states in others, fall outside of the Harm Principle, and therefore, a new principle was required; the Offence Principle. The Offence Principle states that,

> "It is always a good reason in support of a proposed criminal prohibition that it would probably be an effective way of preventing serious offence (as opposed to injury or harm) to persons other than the actor, and that it is probably a necessary means to that end" (Feinberg, 1988, P.1).

Essentially, the principle argues that the prevention of offensive behaviour, is also, in addition to harm, the responsibility of the state and that offence is caused, firstly, when an individual

suffers from a disliked state, secondly, when that disliked state arises as the result of wrongful conduct of another, and finally, when the offended individual resents the individual who conducted the wrongful behaviour for causing him to be in the disliked state (Feinberg, 1988).

There has been scholarly debate as to whether or not offence should be included under the harm principle (Cohenalmagor, 2001; Petersen, 2016). However, there are several reasons why this framework will not remove content based on the Offence Principle, unless the content also breaches a nation's criminal law. The first is a limitation of the principle put forth by Feinberg himself,

> "offence is surely a less serious thing than harm…offences are a different sort of thing altogether…most people after reflection will probably acknowledge that a person is not treated as badly, other things being equal, when he is merely offended as when he is harmed" (1988, P.3-4).

As a result of offence being "a less serious thing than harm", when weighed against the other objectives, particularly promoting user autonomy and freedom of speech, which are deeply rooted in the First Amendment and ECHR, the framework cannot justify interfering with user's right to autonomy and free speech in the same way as it could if the content was interfering with the rights and vital interests of others. Hate speech can be used as an example that can sometimes be difficult to distinguish offence from harm. If the speech produces feelings of disgust, hurt or annoyance in others but does not actually interfere with the rights or vital interests of others, then the speech causes offence. However, if the speech results in others becoming unable to pursue their rights or vital interests (for example, because the speech threatens to impose violence if they do so), then this is an example of causing harm. Overall, the concern is whether the online content creates a direct risk of harm, not whether the content contains views that diverge from the mainstream (Hardy, 2020).

There is the argument that 'offence' can be interpreted very broadly, and thus, most actions could be argued as offensive to others in one way or another (Cohenalmagor, 1993; Macdonald, 2018). It must be noted that if a racist person becomes offended because a person of a different race spoke to them, then although the offence might be genuine, it is a reflection of the racist person's bigotry and not the content of the statement. If the framework aimed to remove every piece of content that offended another user, it could potentially have to remove the majority of posts that are posted to a tech platform. This would add to the previous argument of the unjustified interference of user autonomy and free speech, but would also affect business

interests and require an unrealistic number of employees and resources. Adding to this argument is the complexity around what is deemed as offensive to one person, is not necessarily offensive at all to other people, making it difficult to determine how offensive a piece of content must be and to how many people must it be offensive to in order to be removed (Cohenalmagor, 1993). Some people might be offended by the cultural norms of others, for example opposing political views or seeing inter-racial couples (Cohenalmagor, 1993). Therefore, removing such content could stand in the way of advancing civil rights.

In addition to the above reasons, the offence principle may not be the most suitable or effective principle to follow for the content that this framework seeks to remove specifically. For example, many types of terrorist content will fail to fall under the offence category. Explicitly violent, gory or inciting content will not fall under 'offence'. Nor will the non-violent, utopian content typical of terrorist organisations such as the so-called 'Islamic State' that seek to recruit and/or entice sympathisers to migrate to join the organisation (see Watkin and Looney, 2018; Lieberman, A; Payne, 2009). Therefore, under the offence principle, a variety of online terrorist content may not be justifiably removed. As a result of the above reasons, this framework cannot justify removing content based on offence. [45]

Now that the definitional issues in this objective have been discussed, the next problem the objective encounters is one of cause and effect. One problem with the objective of preventing harm regarding online terrorist content is that the causal chain between viewing such content and harm taking place as a direct consequence of this viewing is uncertain. In other words, at present there is not strong empirical evidence to show that viewing online terrorist content directly results in individuals causing harm only *because* they viewed that content. This does not necessarily mean, however, that viewing this content does not contribute to or influence individuals to subsequently undertake actions that are harmful to themselves and/or others. Many terrorist actors that have undertaken terrorist activities were known to have viewed online terrorist content beforehand or posted propaganda after (Von Behr, Ines, Reding, Edwards and Gribbon, 2013; Gill, Corner, Conway, Thornton, Bloom and Horgan, 2017;

[45] This framework will, however, under the objective of promoting user autonomy, provide users with other means to ensure that they are not exposed to content that while perhaps offensive or even grossly offensive, is not harmful. This framework will ensure that users are granted the powers to block or mute content and accounts that they believe to be offensive or grossly offensive. This prevents the users from being exposed to such content whilst protecting free speech. This is however, unless the offensive content is removed because it violates criminal law.

Beckett, 2018). One of the reasons that there is not a lot of empirical evidence surrounding this is because of the difficulty and complexities involved in testing such a theory.

Regardless of this, this thesis argues that proof of causation is not necessary to regulate. Despite the lack of empirical evidence in this area, this framework will justify the interference of removing online terrorist content from tech platforms with the Precautionary Principle. As already mentioned in Chapter 5 on Social Regulation Theory, the Precautionary Principle has been widely used in environmental protection regulation in order to regulate issues such as climate change in which there has not always been sufficient empirical evidence to justify regulation, however, not regulating risked catastrophic consequences, such as global warming (Ogus, 1994). According to the 1998 Wingspread Declaration, the precautionary principle states that if an activity creates a threat to public health or the environment, precautionary measures should be put in place, even if cause and effect relationships have not been reliably and scientifically confirmed (Baldwin et al., 2010). In other words, regulatory action can be taken before a threatened or potential harm takes place in order to protect those at risk, even if the regulator is not certain that the harm is inevitable because of the severity of the risked harm (Baldwin et al,. 2010;).

This thesis argues, that under the precautionary principle, there is not a requirement to demonstrate an empirical link between cause and effect to regulate. If content encourages or endorses harm or the prospect of harm then this is a sufficient reason to regulate because of the severity and catastrophic nature of harm that can occur if left unregulated. This thesis argues that the posting of content that endorses or encourages harm creates a new risk that harm could take place that did not previously exist before that content was posted (see Macdonald and Lorenzo-Dus, 2020). An example of the kind of harm that the regulation seeks to prevent through this is the Christchurch attack in New Zealand 2019.

Mill discussed the use of precautious action in his book *On Liberty*, saying, "It is one of the undisputed functions of government to take precautions against crime before it has been committed as well as to detect and punish it afterwards (Mill, 1863, p.185). There is also a lot of scholarly debate on prophylactic crimes whereby risk of harm does not necessarily happen as a direct result of the prohibited act; it happens after further human intervention from the original actor or others (Simester and von Hirsch, 2011; Sullivan and Dennis, 2012; Ashworth and Zedner, 2012), for example,

> "It has become increasingly common for legislatures to regulate harmful conduct on an anticipatory basis, very often by criminalising activity which is preliminary to that conduct, thereby seeking to pre-empt opportunities for the harmful conduct to be perpetrated at all…there is widespread agreement that convictions should not depend on the harm's occurrence…subject to the normal checks and balances supplied by considerations such as autonomy, free speech and the like" (Sullivan and Dennis, 2012, pp.59-60).

There is a line of argument that waiting until people are endangered or the harm is done is too late (Sullivan and Dennis, 2012). This is particularly so, when the potential harm arising from an action is severe (Ashworth and Zedner, 2012). In some cases, there can be strong public support for this, for example, in cases that concern security. However, some have warned that doing so denies individuals of their autonomy and assumes that they are highly likely to undertake harmful actions, thereby treating and labelling them as someone who cannot be trusted to make good decisions when this may not necessarily be the case (Ashworth and Zedner, 2012).

A significant concern that must be addressed in taking this precautionary approach is paternalism. This framework is cautious of being overly paternalistic in its attempts to prevent harm. Paternalism is a countervailing consideration to Mill's Harm Principle, although, many philosophers argue that it is more objectionable than many of the interferences that Mill justified under his principle (Husak, 1981). Paternalism is often found in the creation of social policies that interfere with an individual's actions without their permission because it is seen to be in the individual's best interest (Brock, 1988). Legal paternalism is the idea that the law can be legitimately used to prohibit an individual from undertaking actions that will cause unreasonable risks to themselves (Feinberg, 1989). A simple example is requiring citizens to wear a seat belt. There are two competing values when one considers the idea of paternalism. The first is self-determination, allowing individuals the autonomy to make their own choices regarding themselves and their lives without any interference. The second is an individuals' well-being, which is what paternalism aims to protect (Brock, 1988). These values conflict when an individual wishes to undertake actions that compromise his own well-being and a decision has to be made as to which of the two values is considered of greater importance (Brock, 1988). This decision will likely vary depending on how much danger the action poses to the individual's well-being and how strongly the individual feels about their need to take this action (Brock, 1988). Going back to the seat belt example, wearing it has significant

potential benefits for one's safety and there is very little inconvenience imposed on an individual who does so. It may also be considered that avoidable injuries caused by not wearing a seatbelt creates social costs; a person injured in a car accident will require medical attention which, for example, in the UK requires the services of the tax payer funded National Health Service. However, where a decision such as this may become complex is in other considerations such as how capable the individual is in making their own decisions (Brock, 1988).

Van De Veer (1986) argues that individuals have the right to self-direction; that a person regarded as competent in their decision-making capacities have the right to make their own life-decisions even if these decisions reflect what is considered as their baseline decision-making imperfections, limitations and impairments. However, Van De Veer supports the Unreasonable-Harm Prevention Principle (UPP) which justifies paternalistic interference with an individual's actions when the actions are seriously unreasonable and/or, if not interfered with will make that individual worse off (Brock, 1998). However, there are many cases when people might want to undertake what might be classed as 'unreasonable' action that they understand is not in their best well-being/welfare interests, such as stepping in to undertake pain on someone else's behalf because they do not want to see another suffer (Brock, 1988). Then, there are other situations in which a person would rather choose an option that is convenient even if they know it is not in their best interest (Brock, 1988). Feinberg (1989), however, argues that whenever the two values conflict, the individual's autonomy to make their own decision should always be prioritised over others deciding what is best for his well-being. Some people may take the view that if their decisions only affect themselves then what business is it of anyone else (Brock, 1988). One of the biggest criticisms of paternalism is disagreement as to what is regarded as justification for interfering with the actions, preferences and choices of an individual (Feinberg, 1989; VanDeVeer, 1986; Brock, 1988). Two common justifications are having the consent of the individual that is being interfered with or because it is in the individual's best interest, whether they know/agree with this or not (Brock, 1988). It is not always the case, however, that either of these two justifications are fulfilled. Finally, if paternalistic action is undertaken, caution should be taken to ensure that it is not the beginning of a "slippery slope" (Macdonald, 2018).

This thesis does not promote or encourage the use of paternalism in order to prevent harm. The framework will only apply the precautionary principle to justify removing online terrorist content, not when this content can prove a causal role (due to the difficulties laid out

surrounding this), but when content is believed to be endorsing or encouraging harm or the prospect of harm. The framework seeks to take caution regarding issues of paternalism and is wary of overreach and excessive inroads on individual autonomy and free speech.

The next issue that this objective must address is who the framework seeks to prevent from harm. This objective seeks to prevent harm to both the users and employees of the tech platforms, and also innocent third parties who may be harmed (for example, victims of a terrorist attack). There are several reasons why the framework seeks to remove online terrorist content. First and foremost, the objective is to prevent terrorist attacks. Terrorist attacks harm the victims as well as society more widely. Secondly, the objective is to protect vulnerable individuals from becoming radicalised and engaging in activities that will setback their own interests. Third, as a result of artificial intelligence technology not being at a stage where it is well-suited to perform certain functions, such as understanding the nuance of some content, tech platforms are still very reliant on human moderators (Macdonald, Correia and Watkin, 2019; Berthélémy and Naranjo, 2020). Human moderators are therefore exposed to a great volume of online terrorist content or content that has been flagged as potential terrorist content. The extent to which viewing such large volumes of often very distressing content affects human moderators, in terms of their mental health and well-being is not yet known, however there have been many claims of human moderators suffering from mental health conditions, for example post-traumatic stress disorder (PTSD) as a result of their work as a human moderator (Boran, 2020; The Guardian, 2018; Gilbert, 2019a). Therefore, under this objective, the framework will propose actions that must be taken to safeguard employees of the tech platforms from the potentially harmful effects of their role. If the framework is successful in terms of removing terrorist content quickly then terrorist organisations might assess that the length of time the content would remain up is not worth the effort. An example of this can be seen in research by Conway et al. (2019) regarding the Islamic State and Twitter. If this becomes the case as a result of the proposed framework, then this should alleviate some pressure on human moderators.

In sum, the final objective of this framework is to prevent harm. The framework will apply an interpretation of the Harm Principle put forth by several scholars that argue that harm is defined as interference with the rights and vital interests of others. The framework will remove content that is harmful, however, will not remove content that is simply offensive. This is because unlike content that is classified as harmful, content that is offensive does not outweigh the rights of its users to autonomy and free speech. Although there is the issue of uncertainty

regarding the causal link between content that the framework deems harmful and the undertaking of harmful actions, the framework can justify the removal of the content under the Precautionary Principle when content is believed to endorse or encourage harm to others. The framework seeks to prevent harm to the tech platform's users and employees, and also innocent third parties at risk of harm. Under this objective, the framework aims to ensure that online terrorist content is not accessible on tech platforms. The regulator will use the principles discussed in this objective in order to make decisions when this objective conflicts with the framework's other objectives.

Objectives Summary

Overall, the framework advanced in this thesis will pursue four objectives: to preserve and promote innovation; to promote freedom of speech; to promote user autonomy; and to prevent harm. As stated at the beginning of this chapter there can be synergies between the objectives. For example, innovation will be necessary to prevent harm regarding the creation of tools and technology, and there are many links between free speech and autonomy. This chapter has also shown that the objectives are complex with many differing views of what each objective requires. For example, there are many definitions of what constitutes as autonomy and harm. Finally, the four objectives may, at times, conflict with one another.

One example of a conflict that could potentially arise is between preserving innovation and preventing harm. For a long time, tech platforms went virtually unregulated (Silke, 2018) and as such have enjoyed enormous freedom in their ability to innovate their platform, grow their brand identity and attract new users without being forced to implement safeguards until something bad happens. As mentioned in the first objective, innovation is key in preventing the reduction of market competitiveness and the burdening of certain platforms (e.g., particularly smaller/newer platforms who lack the resources of the major platforms). It is likely that the tech platforms will find it challenging to comply with certain aspects of the framework without feeling as though the long-held freedom they have enjoyed regarding innovation is somewhat being impeded in order to prevent harms, some of which, lack empirical evidence of a direct causal chain.

Another likely conflict is between promoting freedom of speech and preventing harm. First of all, the global nature of the tech platforms and global differences in free speech protections make the promotion of free speech a complex objective even before conflicts with other objectives arise. The main tensions between these two objectives are either erroneously

removing non-violating content or failing to remove violating content that could cause harm to others. Tech platforms have been heavily criticised for errors in both the former and latter. When an error is made that erroneously removes non-violating content it could interfere with values and rights such as tolerance, the ability to protest, the ability to search for the truth or hinder advancements in civil rights (Isler, 2001; Mendel, 2017; Chemerinsky, 2018). When an error is made that fails to remove violating content, there is a risk that this content will be viewed by individuals who then undertake actions that cause harm by interfering with the rights and vital interests of others. A final example of a conflict is between promoting user autonomy and preventing harm. This is a conflict that is likely to occur regarding extreme right-wing content due to the convergence between mainstream political discourse and far-right narratives. Conway (2020) argues that extreme right-wing content is significantly more contentious than, for example, Islamic State content, for a number of reasons. Removing extreme right-wing content could be argued by users as impeding their ability to engage in political debate and discussion (this is an argument frequently put forth by former US President Donald Trump) (Conway, 2020), however, this content has potential to be classed as harmful. When conflicts arise, one objective will have to be prioritised over the other. As discussed earlier, deciding which objective to prioritise may sometimes be best left to the discretion of the tech platform. However, other conflicts which exceed the bounds of this will create cause for the regulator to step in.

Whilst it is vital that a framework is underpinned by objectives to establish a clear remit for the framework and provide clarity and consistency for industry and users, four objectives are not going to be sufficient to achieve the framework's aims. The three main findings of this chapter that the four objectives can have synergies, differing interpretations, and conflict with one another leave too much ambiguity and uncertainty. In order to address this, the framework will put forth in the next chapter a set of mandatory regulatory standards. These mandatory regulatory standards have been developed to try to solve some of the issues that have been raised by the three main findings in this chapter. There must be acknowledgement that it would be impossible for this framework to fix every issue and conflict, however, the mandatory regulatory standards will attempt to put forth how some of these issues should be resolved.

Before this thesis proposes these mandatory regulatory standards, it must outline the overall ethos of the framework. The objectives stated the overall aims and purpose the framework will fulfil. However, the ethos explains the ideals that underpin and guide the framework. An ethos provides further clarity as to what the framework expects from the tech industry and regulator.

Ethos

The underpinning ethos of the proposed framework is the importance of collaboration, transparency and accountability in an industry-wide effort to counter online terrorist content. Further to this, and drawing on studies of social regulation in other contexts, tech platforms are expected to adopt a health and safety culture that focuses on the safety of their users and employees, particularly at the design stage of new features and services. This section will now discuss each of these points and the challenges that they face

Collaborations

The framework puts forth that countering online terrorist content should be a collaborative effort between the regulator, industry and other stakeholders. These stakeholders include users, tech platform employees (e.g., human content moderators), academia, civil society organisations and non-governmental organisations. The regulator, tech platforms and other stakeholders all play an important role and provide meaningful expert information and insights that are required to identify and remove this type of content. They each have different expertise as well as varying, potentially conflicting interests that all require representation in the decision-making process. This framework will propose that each party and stakeholder have the right to participate and contribute to the policy decision-making processes involved in the framework. The framework will also propose that tech platforms collaborate with other tech platforms, NGOs, CSOs and academia to share tools and technology, advice and best practice where possible. It is acknowledged that this may be a big adjustment for tech platforms, or more of an adjustment for some than others. Tech platforms may find it difficult to relinquish the freedom and power that they have enjoyed previously. However, there are ways in which the tech platforms may benefit from such collaborations. For example, collaborations provide opportunities to gain deeper insights into what their users want, therefore, what they can do to ensure that they sustain their userbase.

One example of a collaborative effort within the industry is the Global Internet Forum to Counter Terrorism (GIFCT) which was established in July 2017 by its founding companies Facebook, Microsoft, Twitter and YouTube with the aim “to disrupt terrorist abuse of members’ digital platforms” (GIFCT, 2020). The GIFCT has become an independent membership organisation that is led by an Executive Director and is funded by industry contributions (GIFCT, 2020). The collaborative venture focuses its efforts around three main areas: 1) knowledge-sharing; 2) technical collaboration; and 3) shared research. These efforts

are intended to improve the capacity of tech platforms to prevent and respond to terrorist abuse of their platforms, enable multi-stakeholder engagement, encourage and empower online civil dialogue as positive alternative speech, and advance understanding around terrorist efforts both online and the intersections of online and offline efforts (GIFCT, 2020). One of the main contributions from the GIFCT has been the shared industry hash database. Member platforms of the GIFCT contribute hashes (which are unique digital fingerprints that belong to every piece of content) of the terrorist content that they have removed for violating their policies to the database. The database can be accessed by the other member platforms so that they can use the hashes to also remove that content from their platform if it also violates their policies (Facebook, 2016). Thirteen tech platforms have access to this database. Another contribution by the GIFCT is the Content Incident Protocol which "is a process by which GIFCT member companies become aware of, quickly assess, and act on potential content circulating online resulting from a real-world terrorism or violent extremist event" (GIFCT, 2020). The testing of this protocol was a joint effort with European law enforcement authorities and third-party governments. It was activated 70 times between March 2019 and March 2020 (GIFCT, 2020).

Another significant collaborative effort is Tech Against Terrorism. Tech Against Terrorism is a public-private partnership that was established in 2016 by the United Nations Counter-Terrorism Executive Directorate (UN CTED). Tech Against Terrorism work with UN Member States, global tech platforms, civil society and academia in order "to tackle terrorist use of the internet whilst respecting human rights" (Tech Against Terrorism, 2020a). Tech Against Terrorism has a particular focus on helping smaller tech platforms who do not have the knowledge and resources to tackle terrorist abuse of their site on their own. These efforts are focused around: 1) outreach; 2) knowledge-sharing; and 3) practical support. The outreach efforts involve the promotion of relationships between the tech industry and governments. The knowledge-sharing efforts arrange the sharing of best practice (which includes knowledge and resources concerning policies and tools) across the tech industry as well as the civil society, private and public sectors. Tech Against Terrorism also offer practical and operational support to tech platforms in order to implement various mechanisms to respond to terrorist use of their sites. One of the projects undertaken by Tech Against Terrorism is the Terrorist Content Analytics Platform (TCAP) which is a free centralised platform to facilitate tech platform moderation of terrorist content with the aim of improving quantitative analysis of terrorist use of the internet (Tech Against Terrorism, 2020b). The TCAP will have an alert function to help platforms expeditiously respond to terrorist content and will be able to use the TCAP to verify

content as terrorist. Academics and expert researchers will be able to access this dataset for research purposes. Both the GIFCT and Tech Against Terrorism are examples of the type of collaboration that this framework seeks to encourage.

Although it is argued that a collaborative approach should underpin the ethos of this framework due to the demonstration from the above examples of what can be achieved, this is not without challenges. Collaborations must be effective and able to function as intended if the framework is to rely on them to help counter online terrorist content. One challenge is that collaborations usually require substantial long-term funding. Working to continually secure funding takes time and effort away from the main aims of the collaboration and failure to secure it results in the collaboration folding. A lack of funding also lowers the capacity for how much the initiative can contribute to solving problems. Another challenge is the speed at which these initiatives can work. For example, although the GIFCT is working on many projects, some of these projects have taken a while to come to fruition (e.g., URL sharing database). It takes a lot of time, resources, effort and communication for a collaborative initiative to make progress and sustain it. If the framework is going to rely on these collaborations then the regulator will have to consider the problems that could hinder this and how the collaborative initiatives can be supported to overcome such challenges.

Transparency and Accountability

The framework emphasises the importance of a culture of transparency from both the tech platforms and the regulator. The right to information is a fundamental democratic right (Fox, 2007). Individual autonomy requires that users understand the reasons for decisions that affect their speech and behaviour. As mentioned previously, autonomy is vital for a liberal democracy and underpins many human rights. If one's autonomy has been deemed to be outweighed by another right or objective, one should fully understand why this is the case. This is especially so when these decisions are undertaken by institutions that have not been democratically elected but hold enormous power over the lives of citizens. It is also important for reasons of due process so that those affected are assured that the decision was made fairly and is justified (Macdonald, Correia and Watkin, 2019).

It is widely agreed that transparency and accountability are essential for 'good governance' and that the two are inextricably linked (Fox, 2007). Transparency is argued as necessary to efforts that aim to change the behaviour of powerful organisations by publicly holding them accountable (Fox, 2007). It is thought that this is due to the 'power of shame' (Fox, 2007).

Tech platforms must demonstrate their efforts to tackle online terrorist content transparently in order to be held publicly accountable to their users, as well as society more generally, that they are making sufficient efforts to respect their rights and keep them safe and to the regulator to confirm compliance with the framework. They must not only be transparent about their efforts but also the outcomes of their efforts in order to identify how effective their efforts are and whether or not changes or amendments are required. The implementation of regulatory standards and an independent regulator to oversee the transparency processes ensures users that the tech platforms do not have the freedom to cherry pick the information that they choose to be transparent about or publish only opaque information. Transparency on its own, however, is not necessarily sufficient for accountability, there must be some level of punitive sanction and answerability for accountability to take place (Fox, 2007). The framework will propose biannual transparency reports, however, a challenge that may be faced and that will be addressed more thoroughly in the next two chapters is that a lot of effort, knowledge and resources are required for tech platforms to be transparent. Some tech platforms are in a stronger position than others with regards to their ability to be transparent (e.g., this may be particularly challenging for smaller platforms).

Health and Safety Culture

Social regulation theory was introduced in chapter 5 and is defined as regulation that is concerned with issues of public interest, that promote human rights, social solidarity, social inclusion and societal good as opposed to regulation that is concerned with economic-related issues (Ogus, 1994; Prosser, 2010; Wilson, 1984; Baldwin et al., 2010). Academics have argued in social regulation theory literature that the best approach an organisation can take to prevent harm in the long-term is to introduce a health and safety culture at the root of everything that they do (Gunningham, 2007; Gander, Hartley, Powell, Hitchcock, Mills and Popkin, 2011; Kim, Park and Park, 2016). This approach is most notably recommended in occupational health and safety regulation which is one of the three main areas that currently falls under social regulation (alongside consumer protection and environmental protection). Scholars have argued that even with occupational health and safety management systems, harms will not be significantly prevented without a health and safety culture that promotes prevention (Kim et al., 2016; Hale and Hovden, 1998; International Labour Organisation, 2009). The idea of a "safety culture" first appeared in the literature when the "International Atomic Energy Agency introduced the term in its 1986 Chernobyl Accident Summary Report to describe how the thinking and behaviours of people in the organisation responsible for the safety in that nuclear

plant contributed to the accident" (Kim et al., 2016, P.89). It was argued that in addition to this disaster, the Kings Cross fire, Piper Alpha explosion and train crash at Clapham Junction were also thought to be affected by a lack of health and safety culture (Kim et al., 2016). According to Carnino (2000) a safety culture is a combination of values, standards, morals and norms that constitute acceptable behaviour. The ACSNI Human Factors Study Group defined a safety culture as,

> "the product of individual and group values, attitudes, perceptions, competencies and patterns of behaviour that can determine the commitment to, and the style of proficiency of an organisation's health and safety management system" (Health and Safety Commission, 1993, P.4).

Several characteristics of such a culture have been put forth by the International Atomic Energy Agency: heavy promotion by leadership; clearly communicated as an organisational value; decentralized decision-making and clear accountability of who is responsible for certain aspects of safety; training for all employees; and finally, being a top priority that it is integrated into all areas of the organisation (2005). The top five safety leader companies in Braithwaite's (1985) research in coal mining revealed that they all have a corporate philosophy of commitment to safety at senior management level. They believe that safety is more important than production. Four out of the five safety leaders had programmes in place to focus on training and thinking of solutions to hazards. A health and safety culture is thought to work best when multiple stakeholders are able to participate and there is a system of defined rights, responsibilities and duties (International Labour Organisation, 2006; Bluff and Gunningham, 2003). Considerations of openness, communication, employee welfare and listening to employees are all necessary for a safety culture (Carnino, 20000. Also, when it is inherent in the thoughts and actions of employees at every level of a company (Carnino, 2000). The latter should lead to problems being more likely to be anticipated and better communication between colleagues and departments (Carnino, 2000).

Research in the construction industry in Hong Kong displays a reduction in the number of accidents between 1986 to 2013 which is thought to be a result of the increase development of a health and safety culture (Yau, 2014). Similar findings were found in a study by Park (2013) regarding the prevention of noise-induced hearing loss in the workplace; the noise reduction status has been consistently maintained since the year 2000. Although the continued development of safety management systems could also have played a part in these findings,

many organisations that have continued to develop their safety management systems without also introducing a health and safety culture do not appear to have had the same level of reductions in accidents (Kim et al., 2016). Introducing such a culture is thought to assist with the prevention of new and emerging issues as well as existing issues (Kim et al., 2016). This framework proposes the adoption of such a culture as a result of the positive findings in other areas of social regulation. The framework will propose that tech platforms implement a stage in their design and development processes to assess user safety and risk and address any issues prior to the implementation of any new features or services that are developed. However, it should be considered that the adoption of such a culture in one field may result in issues that were not relevant in another field. The regulator will have to assist tech platforms with any issues in this area if and when they arise. Such a culture is also difficult to measure and regulate. Many organisations will use indicators such as the percentage of employees who received safety refresher training as an indicator of how well they are doing with implementing such a culture (Carnino, 2000).

Summary

In summary, the ethos of this framework is based on collaboration, transparency and the adoption of a health and safety culture at the industry-wide level. Tech platforms will become more consistently involved in collaborative ventures, and sharing expertise, best practice and tools with others in the industry, as well as academia, civil society organisations and non-governmental organisations. Tech platforms will also become more transparent with their users and the regulator in order to promote accountability. Accountability will be ensured by the combination of transparency, independent oversight by a regulator and the threat of punitive sanctions. In addition to this, the framework seeks to place less power in the hands of the tech platforms with the implementation of an independent regulator. Finally, the framework proposes learning from a strategy that appears to have had positive results in occupational health and safety regulation, under social regulation theory. This strategy is the implementation of a health and safety culture adopted by each tech platform at an organisation-wide level, which will involve adopting health and safety as a strong value that is heavily adopted and promoted by the leadership of each tech platform, training of all employees and clear accountability for health and safety responsibilities. However, these efforts do not come without challenges. With collaboration comes potentially conflicting interests and requiring tech platforms to give up some of the power they have experienced previously that has allowed them to make decisions without any other input. In order to be transparent and accountable,

tech platforms must have the knowledge and resources required to be transparent. This may be particularly challenging for small platforms. Finally, regarding the introduction of a health and safety culture, this has been primarily researched in a field that is very different to the tech industry and therefore may result in problems arising that were not addressed in the literature that was researched when forming the decision to propose implementing such a culture.

Chapter 7: Regulatory Framework: Mandatory Regulatory Standards

As discussed at the end of the previous chapter, the four objectives that underpin this framework (to preserve and promote innovation; to promote freedom of speech; to promote user autonomy; and to prevent harm) are crucial but not sufficient to achieve the framework's aims of countering online terrorist content from tech platforms. Objectives provide an important starting point but they leave much still to be worked out. Objectives leave us to work out (1) what each objective actually requires and (2) how conflicts between objectives should be resolved. The aim of this chapter is to embark on these tasks, in order to generate a set of mandatory regulatory standards. Mandatory regulatory standards are instruments that subject the suppliers of the goods or services being regulated to specific behavioural controls in order to create a particular outcome, and those who fail to comply run the risk of being penalized (Ogus, 1994).

The framework proposes the following industry-wide twelve mandatory regulatory standards that all tech platforms that fall under the scope of this framework must comply with. All tech platforms must:

1. Designate representatives in their organisation as the point of contact for the regulator
2. Maintain clear up-to-date policies regarding online terrorist content
3. Implement a multi-stakeholder approach to policy-making processes
4. Make all reasonable efforts to remove online terrorist content
5. Implement an appeals mechanism
6. Implement a user flagging mechanism
7. Implement user powers
8. Implement processes to ensure employee well-being
9. Publish bi-annual transparency reports
10. Engage in collaborative ventures
11. Support and engage with digital literacy programmes
12. Create appropriate mechanisms to ensure that user safety is considered in the design and development of new features

This chapter will propose and discuss the mandatory regulatory standards with the view that all tech platforms must comply with them. However, it is acknowledged that some of the proposed mandatory regulatory standards may be beyond the capacity or capability of the

platforms because the platform lacks the expertise and/or resources necessary to comply. Some platforms may also be unwilling to comply with standards. These compliance issues will be addressed in the next chapter. The next chapter proposes four tracks that will help the regulator to address a number of different key compliance issues that this thesis has identified.

What tech platforms fall under the scope of this framework?

Chapter 2 of this thesis concluded that there is a wide range of social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites in the ecosystem of platforms that are used by terrorist organisations. Chapter 4 concluded that where regulatory frameworks exclude some of these platforms, terrorist organisations online operations on tech platforms can only be partially disrupted. Therefore, any social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites that are used by terrorist organisations fall under the scope of this proposed framework and must comply. The next chapter discusses solutions to assist with compliance issues, particularly, platforms that face challenges with capacity and expertise.

New Regulator

This regulatory framework proposes the creation and implementation of a new regulator, as opposed to an existing one, because of the arguments by scholars in Chapter 4 that a new regulator is likely to create a positive symbolic effect that would showcase the Government's commitment to countering terrorist online content (Bishop et al., 2019). Further, that a new regulator would minimize the confusion among the public as to who to contact regarding relevant issues (Bishop et al., 2019). A new regulator will not have to balance regulating this new framework with existing responsibilities, unlike an existing regulatory body. This is important given that this framework involves a hands-on approach from the regulator (this is evidenced particularly in the next chapter). A new regulator is also likely to bring fresh eyes to the regulatory framework. A new regulator could be helpful in establishing relationships with the tech platforms because it does not have an existing reputation that tech platforms may not want to work with. However, there are notable limitations to this as it is a more timely and costly approach than using an existing regulatory body. Despite this, it is argued that a new regulator is best for this new and time-intensive, hands-on regulatory approach. The UK's Online White Harms consultation response revealed that the majority of people who were consulted believed that funding should primarily come from industry (HM Government, 2020). Therefore, this framework will adopt the approach undertaken in the UK Online Harms

consultation response which requires tech platforms with global annual revenue of a certain amount (which will be decided by the regulator) to pay annual fees that will fund the regulator. This ensures that smaller companies with less revenue are not unfairly burdened.

Rules and Principles

It is important to discuss that the standards set throughout this framework will be a combination of absolute rules, rules with exceptions, and principles. Braithwaite (2002a) puts forth a theory that there are certain circumstances in which rules are better placed to create desired outcomes, however, there are circumstances in which principles are more conducive to these outcomes. When a situation is simple, stable and does not hold huge economic interests, a small number of rules are likely to be the most effective choice. However, Braithwaite (2002a) argues that the more complex and likely to evolve the phenomenon being regulated is, and the more intertwined it is with economic interests, the more likely it is that principles should be used to deliver the desired outcomes. There may at times, have to be rules with exceptions in this framework in order to meet the objectives that underpin the framework. For example, not allowing any exceptions runs the risk of reducing market competitiveness and putting burdens on certain platforms which is incompatible with the objective to promote innovation.

Both rules and principles have a range of pros and cons. Rules tend to be prescriptive, specific and can create standardisation which can be good in situations that require explicit clarity. Principles on the other hand, provide space for a creative range of action possibilities that can be context-dependent, which is good for situations that require flexibility (Braithwaite, 2002a; Hilf, 2001; Arjoon, 2006; De Sadeleer, 2002). Rules set out procedural requirements, such as do's and dont's which encourage blind obedience. This is necessary in situations, for example, that seek to ensure safety, such as setting driving speed limits. Principals, however, are more outcome-focused and value-driven, encouraging what is necessary to get a specific result (Arjoon, 2006). An example of such a principle would be an institution telling its employees to take all reasonable precautions to ensure the health and safety of all visitors. This is less prescriptive than the driving speed limit example, giving more freedom and flexibility in the action taken to achieve the desired outcome.

As mentioned, when a situation is simple, stable and does not hold huge economic interests, a small number of rules are likely to be the most effective choice. However, as the complexity increases, the number of rules needed would become greater. Following a large number of rules can become difficult to keep track of and there is a risk that some of the rules begin to contradict

each other creating confusion (Braithwaite, 2002a). This subsequently creates difficulty with assessing compliance. Braithwaite (200a2) provides examples of difficulties with rule-following in care homes where it becomes impossible to follow so many and rules are sometimes broken to respond to the needs of residents. Rules are very focused on what can or cannot be done with a lack of focus as to why the rule is in place. Therefore, sometimes resulting in a lack of understanding as to why the rule is in place (Bruhn, 2006).

As a situation becomes more complex it makes sense to adopt principles instead; because principles are broader than rules, fewer principles would be needed (Braithwaite, 2002a). Simpson (cited in Braithwaite, 2002a) argues that this is particularly important in situations where technology develops rapidly and requires great flexibility. Facebook have argued that when standards are too specific, for example “put this button here” or “use this wording”, it becomes too difficult for platforms to learn what works best for their users (risking it becoming less likely that users will report violating content) (Bickert, 2020). Further, if an issue occurs that does not obviously fall under any of the rules in place, the issue may go unaddressed. Whereas, when principles are implemented, it is more likely that such an issue would be able to be addressed by a principle because they are broader (Arjoon, 2006). Sometimes principles and rules are referred to as a story book and a rule book, with stories being able to establish a sensibility from which action can flow (Shearing and Ericson, 1991). Principles are also typically less densely prescriptive than rules, leading to a better understanding of them as long as they are not too vague or difficult to interpret (De Sadeleer, 2002). Moreover, principles can encourage continuous improvement of the activities being regulated and in allowing a creative response, can allow the response to surpass the minimum expected response that can occur with rules that often limit the response to the actions necessary to meet compliance (Braithwaite, 2002a; Arjoon, 2006). Finally, principles also support the inclusion of multi-stakeholder participation better than rules because of the flexibility that they allow to influence change at a quick pace. Principles can therefore, in complex situations, encourage the most optimal outcomes (Braithwaite, 200a2; De Sadeleer, 2002).

An example of an absolute rule would be that tech platforms must publish a total of X transparency reports per year. A rule with an exception would be that all tech platforms must publish X amount of transparency reports per year, except for platforms with fewer than five employees. An example of a principle would be that tech platforms must make all reasonable effort to remove terrorist content swiftly. This is a principle because it allows for an assessment of what would constitute reasonable effort that is sensitive to the specifics of that company,

providing flexibility and considering context. It is also a principle because it is outcome-focused and value-driven. These examples demonstrate the different ways that both rules and principles can be useful in different aspects of this regulatory framework. However, it is important to note that the distinction between principles and rules can sometimes be blurred and subjective. Although rules can provide certainty, if the framework was based solely on rules, then it would be complex, difficult to understand, create burdens for certain tech platforms (such as small platforms) and lack the flexibility required to innovate and evolve quickly. Although principles provide flexibility, if the framework was based solely on principles then there is a risk that the all-things-considered approach could result in slow progress and difficulty with implementation (Arjoon, 2006).

As discussed, there are pros and cons to rules and principles. Scholarly literature has discussed the combination of rules and principles as necessary and optimal when regulating various industries (Braithwaite, 2002a; Hilf, 2001; Arjoon, 2006). This research made attempts to gauge when rules would be best and when principles would be best in the proposed regulatory standards in order to optimize the efforts to regulate online terrorist content. Using a combination of rules and principles has resulted in the creation of twelve mandatory standards, which this thesis argues is a reasonable and manageable number of standards. It is the responsibility of the regulator to ensure advice and guidance is provided and fully accessible to the tech platforms regarding help with understanding the standards and compliance.

Mandatory Regulatory Standards

1. Designate representatives

All tech platforms must designate one or more representatives in their organisation as the point(s) of contact for the regulator. The details of these representatives should be clearly accessible to the regulator. The representatives must respond to and engage with the regulator. These representatives have the responsibility to ensure the platform implements a strategy to ensure compliance with the following mandatory standards. Failure to do so will result in enforcement action.

Without such an action, it is going to be difficult for the regulator to form a relationship with the tech platform. This will also provide tech platforms with an open line of communication to the regulator should they require help or have any questions.

2. Maintain clear up-to-date policies regarding online terrorist content

All tech platforms must, if they have not already done so, create policies aimed at countering terrorist content. These policies should be amended as necessary to keep up with evolving strategies, tactics and trends related to online terrorist content. This is a rule that aims to standardize the creation of policies that prohibit online terrorist content on these platforms.

The creation and maintenance of clear up-to-date policies provides users with clarity surrounding what speech is and is not permissible, and what must be done in order to prevent harm. The importance of supplying consumers with an optimal level of information as to how a service works and how safe that service is, is discussed in the social regulation literature under consumer protection. This is to ensure that consumers are able to make an informed choice about what services they wish to use, particularly those at risk of erroneous removal, such as those documenting human rights violations (Ogus, 1994). Under the context of this framework, it is important that users have all the information they need to decide whether or not the platform they wish to use has policies and processes in place to ensure their safety preferences and any other desires/preferences that they may have. It is thought that providing such information for consumers to assess, leads to a more competitive market (Lave, 1981; Driesen, 1997).

All policies must be written clearly and in a way that is easily understood. This framework adopts the suggestion by the Santa Clara Principles (2018) that tech platforms provide users with clear examples of what is not permissible under each policy that is implemented. Further, tech platforms must explain the ways in which it implements these policies (e.g., what technology and tools are used and how content moderators make decisions). However, it is acknowledged that providing certain information on this can allow terrorist organisations to find loopholes to exploit. Tech platforms should be sensible in what information they provide on this topic. Other useful information that is encouraged is how users can minimize the number of erroneous or unnecessary removals. For example, YouTube provide users with an informative webpage on this. [46] An example of the advice YouTube provide is that it is helpful if users provide more context to their content via either clear titles or descriptions that outline the intention of the content, such as to document a human rights violation. Finally, tech platforms must inform users what will happen if content violates any of its policies (e.g., removal). The above information is important because if users are to make informed choices

[46] https://support.google.com/youtube/answer/6345162?hl=en-GB

regarding their use of a tech platform, then they need to understand the prohibitions as fully as possible.

In summary, tech platforms must, if they have not already done so, create policies aimed at countering terrorist content and keep them up-to-date as the strategies of terrorist organisations continue to evolve. Any compliance issues with this standard will be addressed in the proposed tracks in the next chapter.

3. Implement a multi-stakeholder approach to policy-making

A multi-stakeholder approach must be implemented to policy-making, regarding terrorist content, wherever possible. This will ensure that a range of rights, interests and perspectives are considered. This standard will work towards increasing user autonomy, raising free speech issues and addressing issues that also concern other regulatory standards, such as improving employee well-being.

It is the responsibility of the regulator to undertake multi-stakeholder consultations and develop "model" policies based on the findings. The regulator will share the model policies and results of the consultations with those tech platforms that are unable to/do not wish to undertake their own multi-stakeholder consultations regarding their policy-making practices and processes. This alleviates smaller companies trying to undertake consultations that may burden them in ways that it would not burden the major platforms. It would also reduce the different sectors included receiving hundreds of requests for consultations and the duplication of effort from many platforms. However, it is acceptable for tech platforms that want and are able to undertake multi-stakeholder consultations themselves to do so and act on their findings if the regulator assess that the promotion of innovation holds greater weight than risk of harm in that situation. This standard is a principle because it is outcome-focused and provides flexibility.

This approach should include a range of stakeholders including users, relevant employees (such as human content moderators), Internet Referral Units (IRUs), Non-Governmental Organisations (NGOs), Civil Society Organisations (CSOs) and academia. The following paragraphs justify why the framework proposes a multi-stakeholder approach.

Users are significantly under-represented in the regulatory process and tech platforms' decision-making (Klonick, 2017; Klonick, 2020; Gallup and Knight Foundation Report, 2020). Users are most affected by policies and related decision-making, yet, report a lack of understanding of, and insight into, these processes (Myers-West, 2018). Users must be included

in this process because tech platforms are not democratically elected, yet hold enormous power over users autonomy and rights. Users require the opportunity to ensure that their rights and interests (such as autonomy and free speech) are being prioritised and protected. The inclusion of users in these processes will bring different insights and contributions to these processes. Users can highlight where confusion lays with policies, if at all, and how this could be alleviated. For example, if clarity is required on what behaviours could violate specific policies and clarity on particular definitions. Users can express what makes them feel unsafe and identify gaps in policies. Users can explain why they post the content that they do and possibly eradicate any misunderstanding that the tech platforms may have regarding users' intentions, particularly in situations where tech platforms may feel that they are not provided with enough context to make informed decisions. Users may be able to provide insight on emerging trends and provide cultural or language-specific insights. An example of how a tech platform might include users in the policy-making process was demonstrated by Twitter in September 2018. Twitter asked users for feedback on a specific policy at the design stage before it officially became part of the Twitter Rules (Gadde and Harvey, 2018). Twitter asked users to participate via a survey that asked questions which aimed to collect global perspectives and find out how the policy would impact users in different communities and cultures. The findings from the survey were used to make amendments to the policy before it was implemented. This is an example of how tech platforms could involve users in their policy-making processes, however other methods can be used also as long as the method demonstrates to the regulator that the tech platform is making all reasonable efforts to comply with the regulatory standard.

Although the vast majority of prohibited content is removed by machines without ever being seen by a human content moderator, there is still a significant portion of content that is viewed by content moderators. As a result of this, content moderators have been argued to be "uniquely positioned as gatekeepers" (Roberts, 2019, p.3), bringing a unique perspective and expertise. They see first-hand the trends that emerge in content that violate policies and content that is created to try to circumvent violating a policy. They follow these trends in real-time and can therefore provide up-to-date feedback. Content moderators are likely to be the first to identify new slurs, symbols and behaviours which are important content classifiers that must be identified to be fed into a platforms' machine learning technology. Content moderators cover a wide range of languages and collectively possess knowledge on many different cultures. Content moderators can provide valuable feedback on how well the technology and processes work. An example of why frontline content moderation employees should be included in this

process was revealed when a Civil Rights Audit was undertaken at Facebook by a civil rights leader. A discovery was made during this audit that many content moderators had been making errors because they were not being shown the entire post they were assessing, for example, sometimes the tool being used was not displaying the caption which was crucial to understanding the context of the post (Murphy, 2019). These problems could be highlighted sooner if content moderators were provided with the opportunity to contribute to and amend the policies and content moderation process more widely. Research revealed that although many companies currently have processes in place to listen to content moderators in regards to recommended changes to policy, the content moderators reported in interviews that their recommendations were rarely considered or implemented (Roberts, 2019). A content moderator reported to the *Irish Times* that he did not know how policies were developed and amended, "they were just given the guidelines and told to follow them" (Boran, 2020). In Roberts' (2019) research, many moderators express feelings of frustration at how little they were able to make a difference in their role. In an *Irish Times* article, a content moderator reported asking himself questions such as "Is there anything I could have done differently? Could I have saved that person's life" (Boran, 2020). Therefore, the inclusion of tech platform employees in this multi-stakeholder approach will provide opportunities for policy changes from those that work with the policies on the frontline and could also contribute to enhancing employee well-being.

Some tech platforms may have thousands of content moderators which will pose a challenge in meeting this regulatory standard, whereas others may have significantly less, and others none at all. One suggestion for situations where tech platforms employ a large number of content moderators could be that employees nominate and vote for representatives amongst their colleagues. The role of the representatives will be to collate insights, feedback and suggestions and communicate them on behalf of their colleagues. Another option is to use large scale surveys available to all content moderators and follow this up with focus groups involving a small but representative sample of staff.

Internet Referral Units (IRUs) (discussed in chapter 4) are comprised of a range of counter-terrorism experts, translators and law enforcement experts in order to flag online terrorist content in close cooperation with the tech industry. Therefore, IRUs will (similarly to content moderators) possess expertise in emerging content trends and can contribute expertise in a variety of cultures and languages. In addition to this, IRUs can provide a law enforcement perspective. NGOs, CSOs and academia will also be able to contribute a unique and wide range

of expertise from their relevant work and research and could identify gaps and issues within the policies.

In summary, tech platform must implement a multi-stakeholder approach to their policy-making processes. This can either be through engaging with the regulator's model policies and consultation results or through undertaking their own multi-stakeholder consultations. Implementing such an approach is powerful evidence of fulfilling the frameworks objectives of promoting users' rights and could also lead to promoting innovation.

4. Remove violating content

Tech platforms must make all reasonable effort to ensure that terrorist content is removed from their platform as swiftly as possible. Reasonable effort includes not merely removing violating content in a reactive manner, it also requires proactive steps to prevent the upload of such content in the first place. This includes designing and implementing proactive technologies.[47] The term "reasonable effort" entails an assessment of all relevant information, including the size of the company, the number of users that have viewed the content, the nature of the service it provides, the nature of the content, the volume of violating content, the complexity of the assessment it requires, and the languages in which violating content is posted, amongst other factors deemed relevant by the regulator. This standard is a principle focused on the outcome of removing all terrorist content.

In many existing or proposed regulatory regimes, such as the NetzDG law and the EU's proposed regulation (both discussed in chapter 4), there has been a large focus on implementing timeframes for content removal, the so-called 'golden window'. The European Commission's regulation on preventing the dissemination of terrorist content online proposed a timeframe of one hour from the moment the platform is first notified of the content's existence. Although the fast removal of content is often critical to limit widespread dissemination and potential harm, and terrorist content is thought to be most harmful in the first few hours of dissemination (Krasenberg, 2019), prioritising the speed of assessment without considering other factors is

[47] Proactive technology and strategies refer to automated technology that proactively removed content. However, it also refers to strategies such as YouTube's Intelligence Desk that aims to anticipate problems before they emerge. This strategy monitors the news, social media and user reports in order to detect new trends surrounding inappropriate content, and aims to make sure the necessary teams are prepared to address violating content before it becomes a bigger problem (Kantrowitz, 2018). Tech platforms must, however, be cautious of intervening with legitimate content and impeding freedom of speech when implemented proactive and automated technologies.

problematic. One concern is that content that is proactively flagged for human review will be viewed based on time of upload as opposed to which content is receiving the most views. This could result in the swift removal of content that is unlikely to receive many views, whilst allowing other violating content, further down the queue, to go viral (Bickert, 2020). It may also disincentivise platforms to continue to search for old violating content because it will increase their average time to take action (Bickert, 2020). Although automated technology can view a far greater amount of content than a content moderator, both automated technology and content moderators can only view so many pieces of content within one hour. This is especially problematic where automated technology proactively flags content that requires human review, as well as for small and micro-platforms that do not have the resources to comply in such a timeframe (Hadley and Berntsson, 2020). Such a timeframe could incentivise an overly cautious, better safe than sorry approach. Tech platforms all have a different ratio of content to staff. Therefore, a set, standardised timeframe will likely impose an unrealistic burden on some platforms or make compliance impossible. Any errors that are made as a result of this race against the clock could also result in the unnecessary removal of non-violating content and infringe on the rights of users (Article19, 2020).

Finally, some terrorist content can be more difficult to assess than others. For example, some terrorist content contains logos that clearly associate the content with a terrorist organisation, whereas other content, may not be so easily identifiable as terrorist content for many reasons. The latter content will require careful examination that may take longer. On the other hand, it would be unacceptable for the re-upload of content that has previously been identified as violating to remain online for an hour. Therefore, such short, strict timeframes are too blunt an instrument: too strict in some respects and too lenient in others, with the additional concern of a chilling effect on freedom of expression because it could lead to erroneous removal by technology that lacks public scrutiny and transparency (Schmitz and Berndt, 2018). For these reasons, a principle-based assessment is proposed as better-suited. There should be an element of discretion and flexibility, for example, if a platform with a very small handful of employees was suddenly heavily targeted by a terrorist organisation then it would not seem reasonable to sanction the platform for failure to remove a large volume of content in a very small timeframe.

Therefore, regarding the setting of timeframes, this proposed framework will not work so rigidly. The regulator will assess and consider the ratio of content to staff for tech platforms to ensure that platforms are not working under unreasonable timeframes that would create burdens for some platforms or encourage and incentivise an overly cautious, rights infringing

strategy that will lead to errors and a chilling effect on free speech. The assessment of reasonableness will be sensitive to the capacity and capability of the tech platform and any compliance issues will be addressed by the tracks proposed in the next chapter.

As mentioned, regarding the nature of the content, it is important to acknowledge that some content is presented as more obviously terrorist than others. Other regulatory frameworks have tried to address this in arguably rigid ways, such as the NetzDG law that tries to distinguish between "manifestly unlawful" and other "unlawful content". The former must be blocked or removed within 24 hours of the platform receiving a complaint and the latter within 7 days of the complaint. This framework acknowledges the issue of the varying nature of content, however, argues that the NetzDG approach risks the distinction becoming too difficult and time-consuming to make, potentially leading to the removal of everything as swiftly as possible, creating a chilling effect. This framework does not propose such an approach. Proposing that tech platforms make all reasonable efforts to remove terrorist content, as monitored by a regulator, aims to result in the removal of all violating content without incentivising platforms to adopt a better safe than sorry approach. This is an example of where a principle is more likely to be effective than a hard and fast rule because of the complexity involved in the decision-making. It should also be acknowledged that it will be difficult for tech platforms to defend the failure to remove content that has been viewed by millions of users/has millions of engagements as making reasonable efforts.

As part of their efforts to remove terrorist content, tech platforms will be encouraged by the regulator to adopt a marginalization strategy. A marginalization strategy aims to reduce connectivity in terrorist groups' networks, therefore, assisting with the battle to make content less accessible. According to Alexander and Braniff (2018) terrorists' sustained efforts to stay connected and engaged in their networks is more concerning than the sheer prevalence of terrorist accounts. It is not necessarily the number of accounts that is concerning but the efforts of those accounts to be heard and inject their rhetoric to as large an audience as possible. Examples of efforts tech platforms could implement from a marginalization strategy would be preventing known terrorists from entering new networks, foiling hashtag-hijacking and revoking verification ticks. Many of these strategies could further serve to delegitimize the terrorist accounts (Alexander and Braniff, 2018). Alexander and Braniff (2018) argue that a second-order benefit to this strategy is that accounts that remain determined to post terrorist material, despite efforts to marginalize them will likely stand out and draw more attention. This may help law enforcement in its decision of where to allocate limited resources. Appropriate

strategies may differ across tech platforms depending on the platform's infrastructure and the ways in which terrorists are exploiting it. The main element of any effort to marginalize a terrorist group is to lower the ability to stay connected with their supporters and engage with potential new supporters.

In summary, this standard proposes that tech platforms make all reasonable efforts to take a proactive approach to content removal. The regulator will take an all-things-considered approach to the assessment of "reasonable effort" due to complexity and differences between platforms. In addition to proactive content removal, the tech platforms are encouraged to adopt a marginalization strategy to lower the ability of terrorist organisations to connect and engage with supporters/potential supporters whom they would be likely to share content with. Any compliance issues with this standard will be addressed in the proposed tracks in the next chapter.

5. Implement an appeals mechanism

All tech platforms must implement an appeals mechanism that is underpinned by due process to allow users to appeal the platform's decision to remove their content or account. This is important because of the number of errors that can occur during content removal as a result of the limitations of both automated and human decision-making, subsequently infringing on users' right to free speech (Duarte, Llanso and Loup, 2017; Macdonald, Correia and Watkin, 2019). The appeals mechanism must undertake the appeals as swiftly as possible. Users must be made aware upon removal of their content, that this mechanism exists, how to access it, and how it works because transparency is a prerequisite of due process (Klonick, 2020). Platforms must be able to demonstrate to the regulator that the appeals mechanism is easily accessible and that users are provided with sufficient information to have an effective opportunity to appeal. The platform should also be able to demonstrate that the appeals are reviewed by a different employee than the one who made the decision to remove the content (if it was removed by a human content reviewer) and that relevant safeguards are in place to monitor that appeals decisions are based on clear decision-making processes in order to ensure that the mechanism is not subject to abuse by employees.

An appeals mechanism is important for due process. There are two main reasons why due process is important. The first is substantive: due process helps to improve the quality of decision-making to ensure fair outcomes (Endicott, 2018; Ramraj, 2004). This is because it forces the decision-maker(s) to consider all of the relevant factors and to justify the outcome.

This in turn allows the individual affected by the decision to feel as though he or she has been treated with respect (Endicott, 2018). It is also in the public interest that the outcomes are fair (Endicott, 2018). The second is procedural: due process ensures that those affected by a decision are able to participate in the process of defending their rights and interests (Endicott, 2018; Ramraj, 2004). It further ensures equal treatment of those involved and that decisions are impartial and based on clearly communicated rules as opposed to personal opinions (Brems and Lavrysen, 2013; Ramraj, 2004). This has a positive impact on individuals' self-worth (Brems and Lavrysen, 2013) and makes it more difficult for an authority to act arbitrarily, thus increasing transparency (Endicott, 2018). Overall, due process is essential for responsible governance, good administration, treating people with respect, accountability and to improve decision-making (Endicott, 2018).

Decisions are more likely to be accepted and the decision-makers are more likely to be regarded as legitimate when due process has been respected (Brems and Lavrysen, 2013; Tyler, 2006; Sunshine and Tyler, 2003). Research has found that legitimacy is important in situations where a body has to make controversial or divisive decisions (Tyler and Mitchell, 1993). People tend to assess legitimacy by evaluating the processes that the decision-maker undertakes (procedural justice) (Tyler and Mitchell, 1993). This is important in the counter-terrorism context because feelings of illegitimacy can potentially be exploited by radicalisers. This can be seen in research by Pearson (2017, p.862) that found supporters of the so-called "Islamic State" tweeting that they became radicalised when "me and my people started to get suspended on every social media platform". This is also important given reports that erroneous removals can disproportionately affect certain groups, for example, Muslim and Arab communities (Windwehr and York, 2020). Feelings of legitimacy on the other hand, has been found in other contexts (e.g., with the police), to influence the likelihood of individuals compliance with an authority (Tyler, 2006). Similarly, research has shown that people assess fairness based on the procedures that are undertaken by an authority (Tyler and Lind, 1988; Tyler, 1990; Tyler et al. 1997; Tyler and Huo, 2002). Finally, research has found that understanding the motives of an authority and feeling shared bonds with that authority are antecedents of trust (Tyler and Huo, 2002). Overall, according to defiance theory (Sherman, 1993), when an authority does not acknowledge an individual's dignity and rights, the individual is more likely to be resistant to that authority and the outcome. Ensuring due process is taken in the appeals mechanism is one way to reassure users that the mechanism and processes are fair and legitimate.

Myers-West (2018) undertook research into users' understanding of content removal. This research found that in situations where users were unaware of why their content was removed, the users would create and spread folk theories that guessed why their content was removed. Further, these users failed to learn from the removal of their content and so violated the same policies repeatedly. It is therefore important to increase transparency and inform users of the reason why their content was removed and (where applicable) why their appeal has not been successful. As explained above, providing users with these reasons may add to feelings of being treated with respect, and it may also lead to some users self-policing in the future because they are now more aware of the reasons why their content violated a specific policy. Any self-policing of content will reduce the workload for automated tools and content moderators.

The regulator must not be involved in the decision-making process of whether or not content is restored. Users may lodge complaints if they believe the appeals mechanism is ineffectual. Systematic failures with the mechanism can result in enforcement action. Any compliance issues with this standard will be addressed in the proposed tracks in the next chapter.

It is the responsibility of the regulator to design and implement an independent board of appeals to undertake a limited number of appeals that are considered the most controversial and viral content appeals from all tech platforms. Such a board would be made up of a diverse range of experts that are completely independent from the tech platform, chosen by the regulator, and would undertake assessments without any input from the tech platform. Both users and tech platforms could propose appeal cases to the board and the board will decide which cases it will assess. The board would prioritise the most controversial cases. It must be highlighted that it will not be easy to assess what content is considered the most "controversial". Factors that should be considered when assessing whether content is controversial is if the content is in the public interest, the virality and reach of the content, the status of the person who has posted or is in the content, the content itself and how harmful it is considered, amongst other relevant factors. An example of a controversial case would be former President Donald Trump's tweets regarding the Capitol Hill Riots at the start of 2021.[48]

It is proposed that one independent board is created by the regulator for use of all tech platforms, as opposed to each platform designing and implementing their own because such a board can arguably only be truly "independent" if it is designed and implemented by an independent body. A board designed and implemented by the regulator is more likely to be

[48] https://www.bbc.co.uk/news/world-us-canada-56004916

viewed as legitimate, credible and accountable by users, than any board implemented or financed by a tech platform because it is truly independent. However, the board may be indirectly funded by the platforms via a levy or tax. It also is unlikely that many tech platforms would be able to design and implement their own due to the enormous resources required. To request that each platform do so would place huge burdens on smaller platforms. Further, proposing that each platform create their own board could result in many platforms only doing so as a public-relations move, appearing to care about users, however, with little intention of making big changes (Klonick, 2020). Tech platforms may also use it as a way to shift the blame on difficult decisions away from themselves and onto their board (Klonick, 2020). If this is the case then tech platforms may "remove more content because it can now refer frustrated users to the board" unless the platform deems the board too harsh in which case it might remove less content to prevent it from becoming the board's jurisdiction (Klonick, 2020, p.2488).

Facebook has already begun to create an independent board with its Oversight Board,[49] which is evidence that this is something tech platforms are beginning to consider necessary. Having an independent appeals board for the most viral and controversial decisions will alleviate tech platforms of the enormous responsibility and difficulty of such tasks, and will ensure unbiased oversight that is completely independent from the tech platforms' business interests, thus providing users with greater confidence that the outcome is fair and impartial. A justification for this is the *nemo judex in sua causa* legal principle that means no man should be the judge in his own case because the judge's impartiality may reasonably be questioned (also sometimes referred to as a rule of natural justice) (Vermeule, 2012). It would be problematic to have the very tech platform that is accused of making such a controversial error assess the appeal of said error. A second justification is that content moderators face great difficulties with appeal assessments and may be reluctant to overturn decisions made by artificial intelligence because of the tendency to believe that technology is less prone to error (Winter, 2018; Macdonald, Correia and Watkin, 2019). Finally, it was reported in a Gallup and Knight Foundation report (2020) that 8 in 10 research participants (all of which were American citizens) believe that a content oversight board is a "good idea" (54%) or "very good idea" (27%), and think it would be better placed than either tech platforms or the government to make decisions regarding the boundaries of free expression. The report highlighted the American participants favourability

[49] See Facebook's Oversight Board, accessed19 June 2020 via https://www.oversightboard.com/

towards the idea of oversight boards increased as they learned more about them with transparency and diversity voted as the most important contributors.

Such an independent board for appeals implemented by the regulator is important because it opens up a channel of communication between the regulator and users through the board. Users can coordinate large numbers of appeals for a specific type of content in order to increase the chances of the board selecting such a case (see Klonick, 2020). If the board chooses one of those cases then users have the opportunity to communicate why the relevant policies and removals are wrong. This could have great effects, contributing to policy changes.

Although it has been argued that there are many reasons why it would be better for the regulator to implement one independent board, instead of proposing that all tech platforms try to implement their own. There are criticisms to this idea as well. Klonick (2020) argues that,

> "the administrative and operational complexity of having one Oversight Board applying multiple sets of platform rules and values would be overwhelming…over time, that board's decision might standardize rules across the industry. This would likely reduce diversification in markets for different types of speech environments".

Therefore, there is not one perfect solution regarding such an independent board of appeals. However, in order to meet the objectives of this framework, one regulator-implemented board is arguably the solution that best prevents the creation of unnecessary burdens whilst creating a truly independent board that will be viewed as legitimate and credible. This board can prioritise free speech without any user concerns that business interests will play any part in the decision.

The independent board must publish an annual transparency report detailing the number and type of cases that have been reviewed, the outcomes of the cases and the timeframes for these outcomes. The individual appeal decisions do not need to be published because providing too much information around this could lead to bad actors using this to find loopholes and ways to avoid removal.

In summary, all tech platforms must design and implement an appeals mechanism. They must be able to demonstrate their appeals mechanism to the regulator and that it is easy to use, easily accessible and users are fully informed of how it works and where to find it. Systematic failures with the mechanism can result in enforcement action. This standard also proposes that the

regulator creates an independent board of appeals that can be used by all tech platforms that fall under the scope of this framework. This board would be truly independent because it would be designed and implemented by the regulator. The board would only review a limited number of cases that can be put forth by users and the tech platforms. The board would prioritise the most controversial cases.

6. Implement a flagging mechanism

All tech platforms must design and implement a flagging mechanism to ensure that users can report content that is considered terrorist content. Tech platforms must ensure that this flagging mechanism is easily accessible and easy to use. Users must be made aware of where to find this mechanism and how it works.

It is important that users are able to participate in the content removal process. Although automated technology keeps improving - with some tech platforms reporting that it is able to find more than 99% of so-called “Islamic State” and al-Qaeda content before user flagging (Bickert, 2018a) - automated technology is not perfect and terrorists adapt to try and circumvent it. Further, research has shown that automated technology does not appear to remove non-Islamic State jihadist terrorist content to the same extent (Conway et al. 2019). Therefore, it is important that users are able to flag content that may not be detected by automated technology. Flagging mechanisms also create another channel for communication between users and the platform and allow users to feel heard and express what kind of content makes them feel unsafe on the platform.

In summary, the tech platform must be able to demonstrate to the regulator that they have designed and implemented a flagging mechanism, that it is easily accessible, easy to use and that users have all the necessary information required to use it. Users may lodge complaints if they believe the flagging mechanism is ineffectual. If there is a systematic failing with the flagging mechanism then action will be taken. Any compliance issues with this standard will be addressed in the proposed tracks in the next chapter.

7. Grant user powers

All tech platforms must grant basic user powers that allow users to mute and block content and accounts that they do not wish to be exposed to.

Although the framework states that tech platforms must remove terrorist content on their platforms, there is the issue of content that falls into a ‘grey area’ where it does not clearly

violate a tech platform's policy, however, may be deemed by individual users as content that is harmful or makes them feel unsafe. Users should be empowered to mute or block such content to improve their well-being on the platform if they wish. This allows users to control their experience on the platform, thus promoting user autonomy, without infringing on free speech. These user powers should be easy to use, easily accessible and users must be told that they exist and how to use and find them.

Some tech platforms have already implemented an array of user powers. For example, in July 2019 Twitter tested implementing user powers that allow users to hide replies to their Tweets (Haq and Forks, 2019). This was tested in Canada to understand how the tool could be used and improved before being rolled out globally. With this tool, users are still able to see and engage with hidden replies by tapping on an icon that appears when replies are hidden. Twitter said that they "will be looking at how this feature gives more control to authors while not compromising the transparency and openness that is central to what makes Twitter so powerful" (Haq and Forks, 2019). Twitter's research into this reveals that people mostly hide replies that they consider to be irrelevant, off-topic or annoying; it is a new way to shut out noise (Xie, 2019). In Canada, 27 percent of users who had their tweets hidden said that they would reconsider how they interact with other users in the future. Some people, however, said that they would not use this feature because of the fear of backlash that they would get for using it.

Granting user powers is a response to arguments from scholars that there is a lack of user participation in policy and regulatory decision-making (Klonick, 2017; Gallup and Knight Foundation Report, 2020) and the finding that users want "to feel empowered to manage their safety online" in the UK Online Harms White Paper Initial Consultation Response (HM Government, 2020). This standard aims to promote the objective of user autonomy and prevent an overly paternalistic regulation. However, this standard could raise a conflict between user autonomy and harm prevention. On the one hand, there is the promotion of user autonomy, empowering users to make their own decisions regarding what content they wish to see and customising their user experience (Watkin and Conway, 2021). On the other hand, however, there is the risk that some users will choose to block opposing views and create an echo chamber and that vulnerable users may fall victim to radicalising content (Watkin and Conway, 2021). This thesis argues that in such a conflict, the former should be prioritised because of the limited empirical evidence to support the risk of echo chambers (Bruns, 2017; Garrett, 2017; Bright, Marchal, Ganesh and Rudinac, 2020). The former should also be prioritised because

there are other ways that tech platforms are encouraged to challenge extremist views and echo chambers.[50]

The framework encourages tech platforms to continue/start undertaking and engaging with research into the effects of granting such user powers.

In summary, the tech platforms must grant user powers to mute and block content and accounts. Users must be made aware of such powers and how to find and use them. Further, tech platforms are encouraged to continue or start undertaking and engaging with research into the effects of granting such user powers. Any compliance issues with this standard will be addressed in the proposed tracks in the next chapter.

8. Have due regard for employee health and well-being

All tech platforms must have due regard for the health and well-being of their employees that are exposed to terrorist content. Tech platforms must provide employees with access to relevant materials, resources and qualified health professionals to aid their well-being and mental health.

As seen in chapter 4, employee well-being is an issue that many existing or proposed regulatory frameworks have overlooked. However, the well-being of employees is both a moral imperative and business interest. Employees, such as human content moderators, are vital to the effective functioning of the content removal processes and this framework argues that tech platforms have a moral duty to ensure the well-being of their employees due to the concerns that undertaking this work can lead to mental health conditions such as post-traumatic stress disorder (PTSD). Several content moderators have come forth in recent years with law suits against the platform that they worked for, claiming to have developed a mental health condition as a result of poor working conditions, a lack of training and support and exposure to traumatic content (Boran, 2020; The Guardian, 2018; Gilbert, 2019a).

Historically, in occupational health and safety social regulation literature, regulation has been criticised as only considering physical risks regarding safety and preventing harm and neglecting psychosocial risks (Maxwell, 2004; Johnstone, Quinlan, and McNamara, 2011). This thesis heeds this lesson and accordingly its proposed regulatory framework encompasses psychosocial risks, as well as physical risks.

[50] See, for example, The Redirect Method https://redirectmethod.org/

> "Psychosocial risks arise from poor work design, organisation and management, as well as a poor social context of work, and they may result in negative, psychological, physical and social outcomes such as work-related stress, burnout or depression" (European Agency for Safety and Health at Work, 2020, no page number).

Examples of psychosocial risks are excessive workloads, working to conflicting demands, lack of clarity surrounding decisions or why something is done in a certain way, lack of employee involvement in making decisions that affect the employee and the way the job is done, job insecurity and a lack of support (European Agency for Safety and Health at Work, 2020). A number of these issues are relevant to content moderators within the tech industry, as was explained in chapter 4. While organisations in the UK have a legal duty to protect employees under the Management of Health and Safety at Work Regulations 1999 and the Health and Safety at Work Act 1974 (Prospect, 2019), the psychosocial and unique risks posed to employees of tech companies reviewing terrorist content are arguably not adequately addressed in this legislation. Further, a European Survey of Enterprises on New and Emerging Risks which surveyed 28,000 organisations across 31 European countries reported that psychosocial risks were found to be one of the key concerns, however, less than a third of the organisations had processes in place to support or address these risks (Leka, Jain, Lavicoli and Tecco, 2015).

The framework encourages more support and engagement from tech platforms regarding researching the effects of repeated exposure to terrorist content online in order to try to improve support for content moderators viewing such content.

Tech platforms must be transparent around the working conditions of employees and the training and support that is provided. Information on this should be made publicly accessible.

In summary, all tech platforms must have due regard for the health and well-being of their employees that are exposed to terrorist content. Tech platforms must provide employees with access to relevant materials, resources and qualified health professionals to aid their well-being and mental health. The tech platforms must be able to demonstrate that in doing so, they address both the physical and psychosocial risks that may occur as a result of working conditions and their employees' remit. Further to this, tech platforms must be transparent about the working conditions of their employees and the training and support they receive. Tech platforms are also encouraged to support and participate research into the effects of repeated exposure to

terrorist content and how to improve support to these employees. Any compliance issues with this standard will be addressed in the proposed tracks in the next chapter.

9. Publish bi-annual transparency Reports

Tech platforms must publish bi-annual transparency reports. This framework acknowledges and considers issues that have been raised around transparency reports requiring great resources and expertise (Tech Against Terrorism, 2020). Although solutions to compliance issues (e.g., expertise and capacity) are put forth in the next chapter, there must be a level of flexibility with this regulatory standard. The below bullet points demonstrate what platforms must report in transparency reports. However, the regulator has the power to make exceptions for platforms in cases where any of the bullet points are not relevant for the platform or would place an unfair burden on the platform (similar to what is proposed in the UK Online Harms government response to consultation).[51] For example, if a platform only had a very small handful of employees, the regulator could exempt them from carefully selected bullet points until they have received the help that is discussed in the next chapter that aims to get the platform to a position where they are able to more fully comply. Some of the below bullet points were influenced by the work that has currently been done by the Santa Clara Principles which were created by a range of experts in this field. [52]

- How much terrorist content was flagged
- How much terrorist content was removed
- How was the removed content flagged? (For example, X percent of content was flagged by automated technology and Y percent was flagged by users)
- Which policies did the removed content violate (For example, X percent of removed content violated Y policy)
- The format of the content that was removed (for example, X percent of removed content was images and Y percent of removed content was text-based)

[51] https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response
[52] https://santaclaraprinciples.org/

- How much content was blocked at the point of upload
- How many terrorist-related accounts were suspended (both individuals and groups)
- How suspended accounts were flagged (for example, X percent of accounts were flagged by automated technology and Y percent were flagged by users)
- How many content removal decisions were appealed and what the outcomes were
- The total number of employees that work on countering terrorist content, the number of content moderators working on countering terrorist content, and the languages that these employees cover
- Any new collaborations that the platform has participated in and contributions to digital literacy programmes

Platforms should provide qualitative as well as quantitative data in their transparency reports to aid understandings of the work that they have undertaken in these areas.[53] Platforms are also encouraged to provide additional information to that mentioned in the above bullet points if relevant and able to do so.

The regulator will supply a template for transparency reports for platforms who wish to use it. It is not mandatory that platforms use the template, however, it may be useful for platforms with less experience producing transparency reports.

As mentioned previously, the importance and expectation of providing consumers with important information regarding the quality and safety of a service has long been discussed in social regulation literature (Ogus, 1994). Scholars have also argued for tech platforms to be held to greater accountability regarding removals and countering violating content (Klonick,

[53] As Suggested by Wingfield, R. (2019) Approaches to content regulation - #4: Transparency reporting. *Global Partners Digital.* Accessed 17 June 2020 via https://www.gp-digital.org/approaches-to-content-regulation-4-transparency-reporting/

2017; Klonick, 2020). Transparency reporting is a way of ensuring that platforms take their responsibilities seriously and that platforms that are known to host terrorist content are not able to shy away from removal strategies.

However, according to Tech Against Terrorism (2020), only 4 out of the 70 tech platforms that publish transparency reports include information about terrorist content specifically. This is inadequate given that Tech Against Terrorism also found that terrorist content is present on at least 330 tech platforms. This highlights the need for a more standardised approach across the industry. Users should be able to access the same information for each platform (where possible and relevant) in order to assess the efforts and safety of the platform. Transparency reports help users to understand the extent of the threat around a particular type of content and provide reassurance that there is accountability (Tech Against Terrorism, 2020). Transparency reports allow all parties the chance to identify gaps in transparency and provide an opportunity to start a dialogue with users and other stakeholders about their decision-making. This enhances the promotion of user autonomy and provides the information necessary for users to make informed choices as to which tech platforms they wish to use. It also prevents harm by reinforcing standard 4. The regulator should also be able to identify from transparency reports the areas in which each platform may require work or improvement.

Although challenging, there is a crucial need for smaller tech platforms to publish transparency reports because the majority of the tech platforms used by terrorists are small or micro-platforms (Tech Against Terrorism, 2020). At the same time, this framework is cautious of placing unrealistic demands on smaller tech platforms. This is why regulators have the power to make assessed exemptions until platforms are more fully able to comply (see next chapter).

10. Engage in collaborative ventures

All tech platforms should, if they have not done so already, work towards membership of the Global Internet Forum to Counter Terrorism (GIFCT) collaborative venture (see chapter 3 for description of the GIFCT). All tech platforms should begin or continue to fulfil the GIFCT membership criteria once they have joined. This will provide the tech platforms with a means to cooperate and to share tools and best practice with one another. The framework acknowledges that there are criteria for membership of the GIFCT. For those tech platforms that require help and guidance to meet these criteria, this will be available from the regulator as explained further in the next chapter. The GIFCT, however, needs to work on the criticisms

that it lacks oversight and transparency, particularly around the hash database (Douek, 2020, Llansó, 2019)

Tech platforms are encouraged to support and engage with other relevant meaningful collaborative ventures and to work with NGOs, CSOs and academia to undertake meaningful research, work on innovative solutions to countering online terrorist content and to assist with achieving and maintaining GIFCT membership.

It is important that compliance with this standard and collaborations/engagements are meaningful and not exploited by tech platforms as ventures that can boost their public relations (PR) profile without providing meaningful outcomes. Twitter's Trust and Safety Council is an example of a tech platform starting a meaningful collaborative venture, however, allowing the venture to become neglected after gaining positive PR. In 2016, Twitter announced that they had formed a Trust and Safety Council by inviting more than 40 independent groups and experts to join together and provide the platform with advice (Matsakis, 2019). The Council reported that for two years the collaboration ran smoothly, however, the communication then began to break down, with members of the council feeling "left to wonder whether its leaders still value their input and expertise", stating that they went months without receiving updates and were unable to get in touch with their contacts (Matsakis, 2019). The council has said that it was not consulted when Twitter made big changes to its dehumanizing speech policies despite having the appropriate expertise.

In summary, all tech platforms should be members, or working towards membership, of the Global Internet Forum to Counter Terrorism to allow for easier cooperation between platforms, and create opportunities to share tools and best practice with one another. Tech platforms are encouraged to further this by supporting and engaging with other relevant meaningful collaborative ventures. Any compliance issues with this standard will be addressed in the proposed tracks in the next chapter.

11. Support and engage with digital literacy programmes

Tech platforms must support and engage with initiatives seeking to increase digital literacy. This is proposed as a result of the UK's Online Harms White Paper Consultation Response (HM Government, 2020) that was open to the general public, academia, business and civil society revealing a large preference for increased education and public awareness of online harms. Digital literacy initiatives should educate people to critically evaluate and understand the structures and syntax of the online content that they are consuming, become able to manage

new social norms and recognise online strategies that aim to be persuasive regardless of whether or not they are true (Kidron, Evans and Afia, 2018). Tech platforms are in a unique position to provide educational information regarding how bad actors are exploiting their sites and the efforts and services that they are providing to try to prevent this. Any compliance issues with this standard will be addressed in the proposed tracks in the next chapter.

12. Create appropriate mechanisms to ensure that user safety is considered in the design and development of new features

Tech platforms must create appropriate mechanisms to ensure that user safety is considered in the design and development of new features and services that the platform introduces. Tech platforms are encouraged to do this for existing features and services that operate on their platform as well.

The framework proposes the implementation of a safeguarding stage when designing and developing features and services for their platform that have the potential to be exploited by terrorists. This stage should involve considering the different ways in which the new feature or service could be exploited and cause harm, and then implement possible safeguards to try to minimize or prevent this. Although this should be considered heavily from a user safety perspective, it is important that risks are thought of from a range of perspectives, with consideration to being able to fulfil the other standards in this framework. For example, considerations should be given to how new designs and services will be moderated and consider the well-being of employees doing so. An example would be Twitter's June 2020 announcement that it was introducing audio tweets that would allow people to record and tweet sounds. Firstly, concerns have been raised as to how audio tweets are going to be able to be moderated with arguments that this is going to be difficult and easily exploited by terrorists (Koebler, 2020; Welch, 2020). This kind of content could take a lot longer to assess amongst other complexities, both by automated technology and content moderators (Koebler, 2020). Second, there are concerns about the effects that this content may have on the employees who have to moderate it. Little is currently known how this could affect the well-being and mental health of employees, nor the users that may have hate, harassment and terrorist rhetoric directed at them via these audio sounds. Therefore, without following this standard, new designs and services may create difficulties for compliance with many of the other standards in the framework and fail to prevent harm. Braithwaite's (1985) research found that not having plans

in place to deal with foreseeable hazards was one of the leading causes of disasters occurring in other industries.

This safeguarding stage has similarities firstly, with the 'privacy by design' approach used in General Data Protection Regulation (GDPR) in the sense that the anticipation of risks and subsequent amendments should be integrated into the organisation's typical design and development processes and considered a core component of the process (Information Commissioner's Office, 2019). This is considered proactive and good practice in GDRP (Information Commissioner's Office, 2019). Such a safeguarding stage is typical of the prior approval process that is often seen in consumer protection in the social regulation literature (Ogus, 1994). However, unlike with prior approval, this process does not require a decision that there is very high certainty that the new feature will not cause harm in order for implementation to be approved. This would not be possible in this industry; it would impede innovation. This framework has explained in Chapter 6 that it is cautious of impeding innovation, the speed with which tech platforms are able to evolve and grow their brand identity, and creating burdens that will inhibit competitiveness. However, the prevention of harm is justified by the precautionary principle. The importance of this safeguarding stage is to ensure tech platforms spend more time and effort considering the potential risks and dangers of their new features and services and provide a step in the design and development process where the designers can receive feedback and make amendments or implement safeguards, where possible, prior to implementation. Employees from other department should be involved in order to ensure that a comprehensive assessment of the risks is undertaken. This regulatory standard wishes to prevent harm, however, is cautious of impeding innovation. An example where this kind of safeguarding step could have proved valuable is the implementation of livestreaming. Although it is impossible to say whether or not such a stage could have prevented the virality of attacks such as the Christchurch attack, the fact that platforms such as Facebook failed to consider the range of footage that was required to train their machine-learning technology suggests that there was inadequate consideration of the risks of implementing their livestreaming services.

The regulator may be able to provide expertise to tech platforms if required and the regulator may also be able to recommend or introduce tech platforms to relevant experts if the tech platform requires assistance with any assessments where the technology or development is particularly novel, risky or complex and would benefit from external expertise.

Tech platforms must be able to demonstrate to the regulator that they undertake these assessments when launching new features and services. When a feature or service is exploited by terrorists (e.g., the Christchurch attack), the regulator will undertake an assessment to review whether the tech platform undertook a sufficient assessment during the design and development stage of the feature/service. If the tech platform failed to do so then the regulator will take enforcement action.

Obstacles and Challenges with Compliance

This framework has proposed twelve mandatory regulatory standards that all tech platforms must comply with. These regulatory standards were proposed to uphold the frameworks four objectives and to create a standardized approach across the industry. However, it is important to recognize that some of these standards may not be possible for all platforms. While some platforms may be in a position to comply quickly or with ease, other platforms may face a range of obstacles and challenges.

Three key obstacles and challenges that this framework has identified as standing in the way of compliance are: 1) awareness; 2) capacity; and 3) willingness. If a platform lacks the necessary awareness to comply then this means that the platform does not have the **knowledge and expertise** required to undertake the actions and processes necessary to comply with the framework. If the framework lacks the necessary capacity to comply then this means that the platform does not have access to the **resources** that are required to carry out the actions and processes that are necessary to comply with the framework. Finally, if a platform lacks willingness to comply then this means that the platform is **refusing to comply** with the framework or specific standards in the framework. An unwilling platform will not demonstrate reasonable efforts towards compliance. This could be for a large variety of reasons unrelated to the issues of awareness and capacity. For example, a platform may decide that it is unwilling to comply with the framework because the framework works against its missions and values and could result in losing its brand identity and userbase. A tech platform may face just one of these obstacles and challenges or it may face multiple/all three. A tech platform may also fluctuate as to whether or not it faces these obstacles and challenges. Both the tech industry and strategies of terrorist groups evolve so rapidly that tech platforms may find themselves struggling with any of the identified challenges at any time, even if they have not done so in the past.

This thesis recognizes that proposing the twelve mandatory regulatory standards without any consideration or acknowledgment of the three key challenges and obstacles identified will lead to an ineffective framework and poor outcomes for all parties involved. It could unfairly penalize tech platforms that are willing but unable to comply due to lack of awareness and/or capacity. It could incentivise actions that will jeopardize the very rights and interests that the framework seeks to protect (e.g., free speech). Therefore, in addition to the twelve mandatory regulatory standards, the framework will propose four regulatory tracks. Each regulatory track will address one of the three identified challenges to compliance. The tracks are designed to ensure that the regulator is responsive to the tech platforms' attitude to engagement, its level of resource and its degree of awareness. The fourth track will propose how the regulator will engage with tech platforms that do not face any of the three identified challenges and are therefore in a strong position to make all reasonable efforts to comply with all of the standards. This fourth track will be the gold standard that all tech platforms should aim for and will work by enforced self-regulation. The overarching aim is to have all tech platforms achieve compliance with the mandatory regulatory standards outlined in this chapter. It is acknowledged that some platforms will have difficulty complying this these mandatory standards. The next chapter will outline and discuss these four tracks proposed as solutions to this problem in detail.

Conclusion

In summary, this framework proposes twelve mandatory regulatory standards that comprise a variety of rules, rules with exceptions and principles that all tech platforms must comply with. The framework carefully applied this combination of rules and principles in order to maximize the effectiveness of the standards, provide flexibility to innovate and to minimize confusion and unnecessary burdens on the platforms. The combination of rules and principles led to the creation of twelve mandatory regulatory standards which this thesis argues is a necessary but manageable number of standards.

All of these regulatory standards are proposed to support the frameworks' four objectives and the overall ethos of the framework. The promotion of user autonomy will be advanced by implementing a multi-stakeholder approach to policy-making, allowing users, amongst other stakeholders, to ensure that their rights and interests are considered alongside business interests. This objective will be further advanced by the implementation of user flagging mechanisms and granting user powers. Both of these standards ensure that users have greater

opportunities to minimize their exposure to content that they do not want to be exposed to. Finally, proposing that tech platforms support and engage with digital literacy programmes empowers users via education, thus allowing them to make more informed decisions around their use of platforms. The promotion of user autonomy alongside the promotion of free speech is advanced by the implementation of appeal mechanisms and independent appeal boards by ensuring that users have the opportunity to fight appeals via methods underpinned by due process and in controversial cases by independent experts. The promotion of innovation and prevention of harm are both advanced by proposing that tech platforms design and implement proactive technologies to remove online terrorist content and participate in meaningful collaborative ventures with a range of stakeholders and other platforms across the industry. The proposal of creating appropriate mechanisms to ensure that user safety is considered in the design and development of new features is a standard that aims to ensure tech platforms continue to innovate but with regard to a range of safety and harm prevention considerations. Other standards that aim to advance the prevention of harm are the creation and maintenance of up-to-date policies regarding terrorist content. This standard also seeks to ensure the promotion of free speech by clearly communicating to users what speech is and is not permissible. Further to this, the proposal of publishing bi-annual transparency reports allows opportunities for public scrutiny of the efforts the platforms go to, to protect free speech and prevent harm. Finally, the prevention of harm is extended beyond users to the employees of tech platforms that are exposed to terrorist content. The framework argues that this is both a moral imperative and a business interest.

This thesis is not arguing that these twelve proposed mandatory regulatory standards are new. Many of the standards hold similarities with standards that are proposed in existing or other proposed regulatory frameworks. This thesis intentionally aimed to incorporate the best practice that was identified from existing regulatory frameworks into this framework. However, this thesis also sought to identify gaps and problems with existing regulation and address this with the proposed framework, learning where possible, from lessons in the social regulation literature and existing tech platform efforts. In order to do so, this thesis argues that this is an industry-wide problem that requires a standardized industry-wide approach (see chapter 2). Second, this thesis proposes that the framework must aim to promote innovation, free speech, user autonomy, and prevent harm, and that the framework should be underpinned by an ethos based on collaboration, transparency, accountability and a culture focused on health and safety (see chapter 6). In chapter 4, the argument was introduced that existing regulatory

frameworks have failed to consider the many challenges and obstacles that some platforms will face in complying with such regulation. This chapter has proposed twelve mandatory regulatory standards whilst consistently acknowledging that certain tech platforms may struggle to comply for various reasons. Platforms will not all be working towards compliance from a level playing field. This chapter identified three key challenges and obstacles that platforms may face regarding compliance. These are awareness, capacity and willingness. This thesis proposes four regulatory tracks in the next chapter to address the challenges and obstacles identified and provide further clarity around the efforts and responsibilities that are required by both the tech platforms and the regulator to aid the tech platforms in complying with the framework. Overall, this chapter argues that mandatory regulatory standards are crucial to achieve the standardisation that regulation for this industry requires. However, the twelve mandatory standards are not sufficient due to the three identified compliance issues. The four regulatory tracks discussed in the next chapter are proposed with the overarching goal of assisting all tech platforms with the necessary help and guidance to achieve compliance with the mandatory regulatory standards outlined in this chapter.

Chapter 8: Regulatory Framework: Four Regulatory Tracks

The previous chapter proposed twelve mandatory regulatory standards that all tech platforms must comply with. However, the previous chapter also identified three main compliance issues that the regulator and tech platforms are likely to face. These are: 1) awareness of the standards; 2) capacity to achieve compliance with them; and 3) willingness to do so. This thesis recognizes that proposing the twelve mandatory regulatory standards without any consideration or acknowledgement of these three key compliance issues will likely lead to an ineffective framework and poor outcomes for all parties involved. It could unfairly penalize tech platforms that are willing but unable to comply due to lack of awareness and/or capacity. It could also incentivise actions that will jeopardize the very rights and interests that the framework seeks to protect. For example, failing to acknowledge platforms struggle with awareness and/or capacity could lead to platforms taking an overly cautious approach to meet compliance and avoid penalties, subsequently infringing on free speech.

Many existing regulatory frameworks and regimes (discussed in chapter 4) fail to account for such compliance issues. However, this thesis will address the compliance issues by proposing four regulatory tracks that the regulator and tech platforms can work from to assist with the overarching goal of achieving compliance with the mandatory regulatory standards proposed in the previous chapter. The combination of mandatory regulatory standards and the four regulatory tracks proposed in this chapter attempts to put forth a solution that balances the need for standardization, whilst acknowledging that platforms are not homogeneous but rather vary in terms of awareness, capacity and willingness, as well as that they have different sized userbases, types of services, content and challenges.

This chapter will begin by discussing the importance of categorising platforms in order to assign them to an appropriate regulatory track. This categorisation assists in tailoring enforcement strategies. The categories are based around the key compliance issues that platforms have been identified as facing. The chapter will then introduce and propose an enforcement pyramid containing a range of tools and sanctions that the regulator will have at its disposal. Finally, the chapter will put forth four regulatory tracks as a solution to the mentioned compliance issues.

Categorising Platforms

Under responsive regulation, scholars have argued that compliance and non-compliance can be explained by a number of perspectives that are taken by a company (van Snellenberg and van de Peppel, 2002). For example, a company may be motivated financially or because of concerns for its reputation. Alternatively, compliance may occur as the result of accepting that rules must be followed or as the outcome of a learning process (Fairman and Yapp, 2005). Some companies may even comply because it is seen as a moral duty (Nielsen and Parker, 2009). Some scholars have taken an approach of categorising companies based on certain characteristics, attitudes, values and resources. Such categories allow the creation of profiles of each category of company, an assessment of the likelihood of compliance/non-compliance and the reasons for this. For example, Kagan and Scholz (1984) created categories of companies to explain non-compliance. These categories include "amoral calculators" (companies who use a risk-benefit analysis to make compliance decisions); "political citizens" (companies who decide not to comply because they do not agree with the rules); and "organisationally incompetent" (companies who fail to comply due to lacking sufficient management and systems). Hawkins (1984) and Baldwin (1995) also discuss a category referred to as irrational non-compliers, containing companies that refuse to comply out of malice or due to being ill-informed. Hawkins (1984) also categorised companies responsible for water pollution. These were: "socially responsible" (companies that acknowledged the importance of sorting the issue); "unfortunate companies" (who lacked the technical or financial means to comply); "reckless companies" (who openly defied compliance); and "calculating companies" (who tried to cover up their non-compliance).

- <u>Why Categorise?</u>

One deterrence strategy is unlikely to be effective if the companies it is aimed at all have different motivations to comply (Gunningham, 2010). An advantage of categorising companies and creating a profile for each category is that it can be useful in making decisions about how to counter non-compliance (Black, 2001b; Baldwin, 1997). One example by Black (2001b) is that companies categorised as organisationally incompetent should be provided with user-friendly guidance manuals that tell them everything they need to know to comply. Whereas, this may be a waste of resources for other companies who either do not require them or will not use them (Bardach and Kagan, 1982). Gunningham (2010) argues that basing responses on categorizations like this is good regulation. It is argued that regulators are likely to hold a company more responsible for compliance the more it appears that it is able to comply, provide advice and encouragement when companies appear willing but unable to comply, and take a

tough stance where non-compliance is intentional (Mascini, 2013). The above examples of existing categories that scholars have devised for companies in other industries all contain similar themes. One is that there are usually companies that lack the ability to comply, perhaps for differing reasons, such as resources (financial, technical or other) or expertise. There are also usually companies at both ends of the spectrum: those who follow the rules (whether this is because it financially benefits them or is seen as a moral duty etc.) and companies that are unwilling to comply (also for varying reasons). One issue, to note, however, may be when companies fall into a grey area where it is not clear which category they should be located in.

Following these insights gained from research into other regulatory contexts, this thesis creates categories of tech platforms based on compliance issues. As already mentioned, these categories are 1) platforms that lack the awareness to comply; 2) platforms that lack the capacity to comply; 3) platforms that lack the willingness to comply; and 4) platforms that are fully able and willing to comply. These categories are used to help create the regulatory tracks put forth in this chapter that are intended to provide solutions to the compliance issues that were identified in the previous chapter. The regulator must be aware, however, that assigning a company to a category is not a static process. Companies may fall into more than one category at a time or fleet between categories over time (Ayres and Braithwaite, 1992). It could also be difficult for the regulator to assess the motivations of a company (Gunningham, 2010). The incorrect categorisation of a company could create issues such as creating tensions between the company and the regulator and will be ineffective in achieving the goal of full compliance (Black, 2001b).

- The four categories

Three key compliance issues have been identified. The first is lacking *awareness*. This may be awareness of several different things. It could be lacking awareness of whether their platform is being exploited by terrorists and if so, in what ways.[54] It could be lacking awareness of the need to comply (i.e., they are ignorant of the regulatory standards). Finally, it could be lacking awareness of how to comply with the regulatory standards. If a platform lacks the necessary awareness to comply then this means that the platform does not have the **knowledge and expertise** required to undertake the actions and processes necessary to comply with the framework. Platforms require knowledge and expertise across many broad areas to comply

[54] A reality for many tech platforms according to Brian Fishman, Head of Counterterrorism Policy at Facebook https://tnsr.org/2019/02/crossroads-counter-terrorism-and-the-internet/

with all the mandatory regulatory standards. The regulator would first have to assess which of the different above-mentioned areas the platform lacks awareness of.

It is important to note that terrorist groups constantly adapt and evolve, and because of this compliance issues are complex and on-going. A platform may find that where it once held the necessary knowledge and expertise, it suddenly requires assistance to keep up with terrorist groups ever-evolving strategies. Tech Against Terrorism (2020e) asked the tech platforms that participated in its Terrorist Content Analysis Platform (TCAP) consultation what their interest is in learning more about transparency reporting (one of the mandatory regulatory standards in the proposed framework). 56.6% of the tech platforms that participated reported being "very interested"; 33.3% were "interested" and 11.1% were "not interested". The finding that over half of the tech platforms were "very interested" in furthering their knowledge on a key area of this framework suggests that at least some platforms will welcome a regulatory track that seeks to provide an educative approach. This chapter will propose a regulatory track that outlines ways in which the regulator and tech platforms can work together to increase awareness and subsequently compliance with the mandatory regulatory standards.

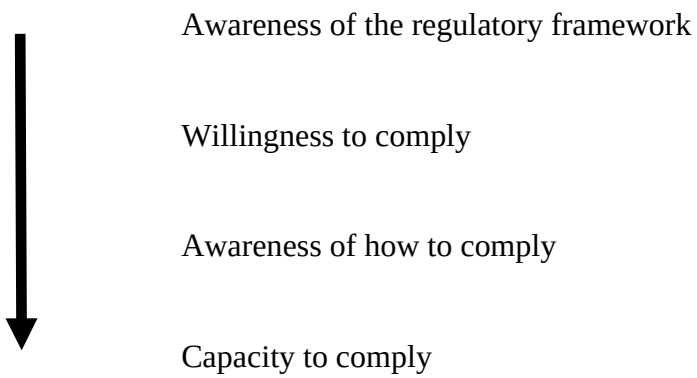
The second compliance issue is lacking *capacity.* A platform that lacks the necessary capacity to comply is one that does not have access to the **resources** needed to carry out the actions and processes required by the regulatory framework. A broad array of resources is required to comply with the twelve standards detailed in the previous chapter. However, the amount/type of resources needed will differ from platform to platform depending on, for example, the number of users a platform has, the volume of content it hosts, and how the platform is exploited by terrorist groups. As seen in chapter 2, platforms are not all exploited by terrorist groups in the same way. Tech Against Terrorism (2020e) asked tech platforms in their Terrorist Content Analytics Platform (TCAP) consultation how they would rate both their financial and technical resources for countering terrorist exploitation. Regarding financial resources, 44.4% of the platforms that participated reported that their financial resources were "available and allocated"; 11.1% "available not allocated"; 11.1% "not available"; and 33.3% "prefer not to say". Regarding technical resources, 33.3% reported "available and allocated"; 22.2% "available not allocated"; 33.3% "lacking"; and 11.1% "prefer not to say". So, only a third of the platforms that participated have allocated and available technical resources and less than half have allocated and available financial resources. This illustrates the necessity of a regulatory track that is focused on capacity-building. Moreover, as with issues of awareness, the position here is fluid. A platform may at one time have sufficient resources to comply with

the standards, however, suddenly become targeted by a terrorist group more heavily than it had been or become exploited in a new way, and find that it no longer has sufficient resources. This chapter will propose a regulatory track that will outline ways in which the regulator and tech platforms can work together to increase capacity and make existing resources go further, and subsequently aid compliance with the mandatory regulatory standards.

Finally, a platform could lack willingness to **engage** with the framework or specific standards in the framework. An unwilling platform will not demonstrate reasonable efforts towards compliance. This could be for a large variety of reasons unrelated to the issues of awareness and capacity. For example, a platform may decide that it is unwilling to comply with the framework because the framework is at odds with its mission and values and could result in it losing its brand identity and userbase. It may also be possible that a platform is unwilling because it does not adequately understand terrorist use of tech platforms/the harms, or even because of laziness. This compliance issue differs to the previous two. For the previous two, the platforms and regulator were willing to work together to increase awareness or capacity to boost compliance. With this compliance issue, however, the platforms are unlikely to work with the regulator to resolve the issue. This is the compliance issue that is likely to require the regulator to utilise the tools towards the most draconian levels of the enforcement pyramid outlined below. The threat of such measures aims to incentivise the platforms to engage with the regulatory framework.

A tech platform may have just one of the above compliance issues or it may have two or all three. It is also possible that a platform may not have any compliance issues. In situations where a tech platform has more than one compliance issue, figure 1 demonstrates the order in which the compliance issues should be addressed.

Figure 1. Order in which compliance issues should be addressed



The issue of awareness of the regulatory framework must be addressed first because you cannot gauge the other issues until the platform is aware of the regulatory demands. Willingness must be addressed next because the following compliance issues require engagement with the regulator which cannot be done without willingness. Awareness of how to comply must be addressed next to ensure that a platform fully understands the issues regarding terrorist use of their platform and how to comply with the mandatory regulatory standards. Finally, capacity would be addressed last to increase resources/make existing resources go further (now that the tech platform has all the necessary knowledge and expertise to utilise them) in order to carry out compliance with the mandatory regulatory standards. Once all compliance issues are addressed, the tech platform will move to a system of (amended) enforced self-regulation whereby compliance is monitored by the regulator. However, as already noted, terrorist groups are constantly evolving and adapting and so there may be times at which a tech platform needs to revisit a regulatory track to receive help and guidance with compliance.

Educative Approach

This thesis recognizes that, whilst a standardized industry-wide approach is necessary, without these distinct regulatory tracks, many tech platforms will be penalized, not necessarily because of a refusal to comply but because of an inability to do so. As seen in chapter 4, many existing regulatory frameworks focus on punishing non-compliance and have often been criticised as creating challenges for smaller platforms or platforms that lack the required awareness and resources, subsequently restricting competition (Hadley and Berntsson, 2020; Article19, 2020). Research has shown that smaller organisations across other industries face similar regulatory challenges (Gunningham, 2002). Punitive approaches are further argued to result in a regulatory game of cat and mouse in which companies could be tempted to engage in creative compliance. Companies may make it seem as though they are complying when in fact they are not or platforms may become unable to survive due to the penalties and lack of ability to overcome their compliance issues (Black, 2001b). Since the imposition of unnecessary burdens, punishments or other actions that could reduce market competitiveness for tech platforms is unlikely to lead to effective change, the proposed regulatory framework adopts a more holistic approach. It seeks to educate and provide assistance where possible, and exhaust these options before resorting to enforcement action. However, enforcement action is available

where tech platforms are unwilling to comply. This may encourage companies to take the carrot because the use of a stick is a possibility.

The creation of regulatory tracks aims to create a synergy between educative and punitive approaches in responding to compliance issues. An educative approach will be taken where possible, with a punitive approach utilised only when platforms are unwilling to engage. An educative approach is based on advice-giving and training with the enforcer playing the role of consultant (Fairman and Yapp, 2005). This approach is thought to help companies to make sense of what is required of them in order to be able to comply (Weick, 2001), help companies to internalise the rules and principles that they face (Honneland, 2000), and to reinforce good practice (Fairman and Yapp, 2005). Parker (1997b) has argued that regulators need to spend more time focused on improving education, in addition to the role of punisher and persuader, with a specific focus on building regulatory capacity within companies.

Research by Fairman and Yapp (2005) which interviewed small and medium-sized enterprises (SMEs) found that most of the SMEs interviewed (n=81) were unable to accurately judge whether they were complying or not because of a lack of knowledge and understanding of what was required of them. The research also found that many companies believed they had complied when in fact they had they had not. Further questioning in the interviews regarding this revealed that when a company did not understand what was required of them, the most common response from the company was that they chose to ignore the requirements. Oftentimes the company reported not understanding how the requirements related to their organisation/business activities. When investigating why companies might be happy to ignore the requirements asked of them, the response was that the company was aware that severely punitive actions were only taken in extreme circumstances. Companies were therefore confident that they were not going to be severely punished and that the effort of trying to figure out how to comply was not worth it because they were not scared of being punished. Some companies were even found to be unaware that they had been sent warning letters. For example, one company that was interviewed that had been sent an improvement notice said “Did they?...I mean I get loads of things through the post…” (Fairman and Yapp, 2005, p. 506). Many SMEs did not understand the risks created by their services and could not afford the necessary safety specialists. As a result, many companies were found to rely on the inspector to tell them exactly what to do to meet what was required of them.

This research into SMEs is particularly important and relevant because many of the platforms exploited by terrorist organisations are SMEs (Tech Against Terrorism, 2019a). Although an educative approach may be described as resource and time-intensive, it is important for several reasons that arguably outweigh this limitation. An on-going educational approach is important in order to prevent the problems identified in Fairman and Yapp's (2005) research from happening with the platforms that fall under the scope of this framework. Further, compliance is a continual process (Hawkins, 1984; Hutter, 1997) and education is required first to ensure that platforms realise this, and second, to ensure that platforms have access to on-going education in order to allow them to maintain compliance (Reiss, 1984). This is particularly so given that both the tech industry and terrorist organisations evolve so quickly. This is also important given that Fairman and Yapp's (2005) research into SMEs showed that many SMEs tend to view compliance as an outcome and as static. It is important that through the proposed regulatory tracks and engagement with the regulator that platforms learn that compliance is a continual process and that they have access to the relevant educational information that will allow them to fulfil the mandatory regulatory standards on a continual basis. Fairman and Yapp (2005) found that where an educative approach is taken, SMEs are more likely to comply with and understand what is required of them. This research is supported by Braithwaite and Makkai (1991) who found that non-coercive and informal approaches are more likely to result in compliance than enforcement approaches that are considered punitive and coercive.

Enforcement Pyramid

Enforcement pyramids are used as a tool in a responsive approach to regulation. Responsive regulation focuses on what triggers a regulatory response as well as what the regulatory response will be (Ayres and Braithwaite, 1992). Ayres and Braithwaite (1992) argue,

> "that regulation be responsive to industry structure in that different structures will be conducive to different degrees and forms of regulation. Government should also be attuned to the differing motivations of regulated actors. Efficacious regulation should speak to the diverse objectives of regulated firms…regulations themselves can affect structure…and can affect motivations…" (p. 4).

On this approach, in order for regulation to be effective, efficient and viewed legitimately it should take neither a solely deterrent nor solely cooperative approach (Nielsen and Parker, 2009). Responsive regulation proposes an approach that combines several theories of compliance and enforcement (Nielsen and Parker, 2009). Arguably, the main contribution of

responsive regulation theory is the understanding that enforcement has on compliance and the argument that different companies have different motivations for complying or failing to comply, and that a company can have multiple, potentially conflicting, motivations regarding compliance (Ayres and Braithwaite, 1992; Braithwaite, 2002b).

How does the enforcement pyramid work?

The idea of an enforcement pyramid is that the lower levels take a persuasive approach that does not involve sanctions. The levels of the pyramid progress upwards, involving sanctions that increase in severity (Black, 2001b). The idea is that, typically, the regulator will begin at the lower levels of the pyramid and only when a company displays unwillingness to comply will the regulator move up the levels, taking a sanctioning approach. The regulator should, however, be ready to de-escalate the sanctioning at the first sign of goodwill from a company (Mascini, 2013). When a company chooses to display behaviour that shows cooperation, the regulator can therefore begin to move back down the levels (Ayres and Braithwaite, 1992). This approach whereby the regulator responds to a company based on its behaviour is sometimes referred to as a "tit-for-tat" strategy" (Nielsen and Parker, 2009). Johnstone (2003) argued, however, that for a tit-for-tat strategy to work, the regulator must be able to identify what kind of company it is regulating. This is why this thesis has already identified four categories that the regulator must assign platforms to, as the first task, upon implementation of the proposed framework.

Overall, the aim of the pyramid is to encourage cooperation of organisations at the bottom levels of the pyramid, providing encouragement, advice and guidance, as typical of an educative approach, and become punitive only when the more persuasive, good faith methods fail thereby only affecting recalcitrant companies (Ayres and Braithwaite, 1992; Fineman and Sturdy, 1999). The lower levels of the pyramid are more concerned with repair and results than with retribution (Hawkins, 1984). The lower levels of the pyramid also ensure that there is a level playing field and that punishment is dependent only on willingness to comply, not ability to comply (Gunningham, 2010). Companies that are willing to comply but face issues receive help instead of sanctions. The upper levels of the pyramid are to deter companies from non-compliance. Punishment is thought to work when penalties are severe, assuming companies are rational, future-oriented and concerned with their reputation (Braithwaite, 1985; Gunningham, 2010). Regulators should aim to punish in a way that maintains dignity and respect with the regulatee, avoiding labels such as "irresponsible" and "untrustworthy"

(Braithwaite, 1985). Braithwaite (1985) has argued that it should not be a case of asking whether to punish or persuade but when to punish and when to persuade. The chosen strategy should match the company's ability and motivation to comply (Black, 2001b). The enforcement pyramid offers answers to this question (Braithwaite, 2002b).

Advantages of an enforcement pyramid

The platforms that are being regulated are part of a globalised industry, with many of the companies set up across several jurisdictions. Regulating and carrying out enforcement strategies on such companies is a highly complex task. Due to the diversity of platforms, if the regulator only had one enforcement tool at its disposal, the regulator would likely face many problems enforcing it across the board. The enforcement pyramid provides the regulator with a variety of options if it faces trouble implementing or enforcing any of the enforcement strategies and in case the lower levels of the pyramid have no effect on the platforms non-compliance. Braithwaite (1990) argues that failing to comply is a less attractive option for a company when faced with a regulator armed with a range of enforcement strategies in an enforcement pyramid than a regulator who only has one option. Research undertaken by Clinard and Yeager (1980); Orland, (1979); Stone (1976); Fisse and Braithwaite (1984) all argue that some penalties work better for some companies than others, therefore, it makes sense to have a range of penalties to cover all bases. This is the case even when the one enforcement option is very severe. In such cases companies are usually aware that the regulator cannot use the one enforcement option lightly and that they will have to do something terrible for it to be sanctioned against them. As Braithwaite (1990) points out, regulators have more power and credibility "when they can escalate deterrence in a way that is responsive to the degree of uncooperativeness of the firm, and the moral and political acceptability of the response" (p. 63). Braithwaite (1990) refers to the use of an enforcement pyramid and tit-for-tat strategy as "benign big guns" where the regulator walks softly whilst carrying very big sticks.

Target-analytic approach

One question that appears in the responsive regulation literature is where the regulator should begin. Should they always begin at the lower persuasive levels of the pyramid or can cases be made for starting further up the pyramid? (Black, 2001b). One benefit of having the four tracks proposed in this chapter is that they can be used to answer this question. If a platform is unwilling to comply (track 4) then it is unlikely that the lower levels of the pyramid will be effective (Macdonald and Bishop, 2019). To start low could waste time and resources that are

better spent elsewhere. Therefore, regulators can justify starting higher up the pyramid when dealing with non-compliance from track 4 platforms. However, if the platform is assigned to any of the other three tracks, then this means that they are willing to comply (despite those in tracks 2 and 3 having other compliance issues). It is therefore arguably unnecessary and unfair for the regulator to begin anywhere on the pyramid other than the lowest level. "When there is willingness to do the right thing, across-the-board punishment is simply not the best strategy for maximizing compliance" (Braithwaite, 1985, p. 99). Starting any higher risks creating unnecessary conflict with a platform, alienating them and ruining their relationship with the regulator (Braithwaite, 1985). This is supported by Ayres and Braithwaite (1992) who argue that the regulator should begin with the most minimal sanction necessary to ensure compliance. Kagan and Scholz (1984) suggest that such a "target-analytic approach" should be used to decide where best to begin on the pyramid. They argue that if knowledge and experience suggest that a company is unlikely to be persuaded then it makes sense to begin higher up the pyramid. It is therefore preferable for the regulator to start low on the pyramid for platforms from tracks 1-3 and take the earlier mentioned tit-for-tat approach. Starting higher for platforms in track 4 supports the argument that companies only comply when they know that companies who have failed to comply are punished (Ayres and Braithwaite, 1992; Bardach and Kagan, 1984). Bruhn (2006) argues that it is best not to view such an approach as "linear models of enforcement escalation, but rather as different ideal types of action, some more important than others, to be practised in a flexible way depending on a given, unique, situation" (p.9). There needs to be an understanding from both the regulator and regulatee about what is necessary to avoid escalation up the pyramid and what is required to be moved back down (Parker, 1997a). Further, when a company does then decide to comply, it can be moved back down the pyramid which can be seen as a 'reward' from the regulator (Black, 2001b). It could be argued that having the regulator start at different levels increases the complexity of its task, however, this is arguably less of a limitation than wasting time and resources by always starting at the lowest level. It is important, however, that the regulator provides clear transparency around these decisions and that there is an auditing process in place (Bruhn, 2006).

Enforcement pyramid levels

Some of the existing regulatory frameworks mentioned in chapter 4 focused primarily on imposing fines (for example, the NetzDG law), whilst others explored other options (for example, the UK Online Harms White Paper). The enforcement pyramid proposed in this framework consists of an array of enforcement tools. Although fines may be able to work as a

financial deterrent, there are some problems that can occur when dealing with large, profitable companies. The fine has to be very large to be able to have any impact on the company or make the company doubt that it is worth risking non-compliance (Braithwaite, 1985; Kennedy, 1985). However, if it is too large then there is a risk that the fine will result in the company increasing costs, this could work as a deterrent or subsequently negatively affecting consumers and/or in this case the companies that advertise on the platforms (Braithwaite, 1985; Macdonald and Bishop, 2019; Kennedy, 1985). Where jurisdictional issues exist, enforcing a fine may not be feasible (Macdonald and Bishop, 2019). There may be difficulties enforcing the fine across jurisdictions and the cost of the regulator trying to collect the fine may end up being greater than the value of the fine. Companies are aware of this and may intentionally fail to pay (Braithwaite, 1985). On a different note, the jurisdictional problems may end up disproportionately affecting a certain type of platform. For example, the major platforms tend to have offices based across several continents, whilst other platforms have only one base and may be able to evade enforcement more easily, creating consequences for some platforms but not others. Some unwilling platforms may be more likely to intentionally locate themselves in jurisdictions that are uncooperative. Therefore, fines alone, without the threat of more severe sanctions, may end up not being very effective at all or may only be applicable to the more willing platforms, subsequently creating issues around fairness and competitive practices.

Figure 2 displays the proposed enforcement pyramid for this framework. The bottom level of the pyramid starts with persuasion (Braithwaite, 1985). To undertake persuasion, the regulator will adopt the earlier mentioned educative approach that is largely based on providing advice, guidance and encouragement. This will primarily be for platforms that show a willingness to comply and engage with an educative approach, however, face some kind of compliance issue that requires awareness-raising and/or capacity-building. This level is based on both the regulator and regulatee working in good faith with open lines of communication. If the platform does not appear to be engaging with the educative approach, then the regulator will move up one level and provide the platform with a warning. This warning will detail the area in which the regulator believes the platform is not engaging or complying and will provide the platform with a stipulated timeframe in which to demonstrate to the regulator that it has decided it will make the effort to engage/comply in the problem area. If the platform makes these efforts then they can be moved back down to the bottom level. However, if the platform does not make any effort to take appropriate actions after the warning, then the regulator will move up a level again. The regulator must be mindful that a platform could end up in a loop of warning – display

of willingness- inaction – warning – display of willingness – inaction, that appears to be a way of evading punishment for non-compliance. If such a situation arises, then the maximum number of warnings should be three, at which point the regulator escalates the platform up the pyramid. Three warnings are chosen because it allows for misunderstandings or mistakes to occur without unfair punishment. It is, however, up to the discretion of the regulator to escalate before three warnings if the regulator has reason to believe at the point of the second warning that the platform is being disingenuous. It should be noted that platforms may need flexibility with the timeframes, for example, a specific action might take a major platform with a large number of employees only a few days to fix, but other platforms with a much smaller number of employees longer (Bishop and Macdonald, 2019). This is something the regulator must keep in mind when setting timeframes.

The middle levels of the pyramid are where the sanctions begin. The two levels are comprised of two different types of fines. The first type of fine is imposed on the company. This is a common sanction used in other frameworks across this industry (for example, German NetzDG law, UK Online Harms White Paper, European Commission Directive). As seen in the NetzDG law, fines can be applied to a number of different situations of non-compliance, for example failure to comply with flagging mechanisms or transparency reporting (Bishop and Macdonald, 2019). The intention of the fine is creating a direct economic impact that deters companies from continuing the path of non-compliance. In addition to the economic impact, there is also a chance of reputational damage. In the case of tech platforms this could affect the way both advertisers and users perceive the platform and whether or not they want to be associated with/continue to use the platform, which could lead to even further economic damage (Bishop and Macdonald, 2019). It is important that the fines are proportionate so as not to unfairly reduce market competitiveness and place smaller platforms at a greater disadvantage. Bishop and Macdonald (2019) have suggested basing the amount of the fine on the financial strength of the platform as a company (as is done in the GDPR legislation). This could be up to a certain amount or it could be a percentage of the company's total turnover in the preceding financial year. These figures or percentages would be for the regulator to decide and then should remain consistent in its application. If the company-level fine does not result in the platform deciding to make efforts to comply then the next level up the pyramid is the imposition of fines on individual members of the platform's senior management team. This is a sanction in both the UK Online Harms White Paper and Australia's Abhorrent Violent Material Act.

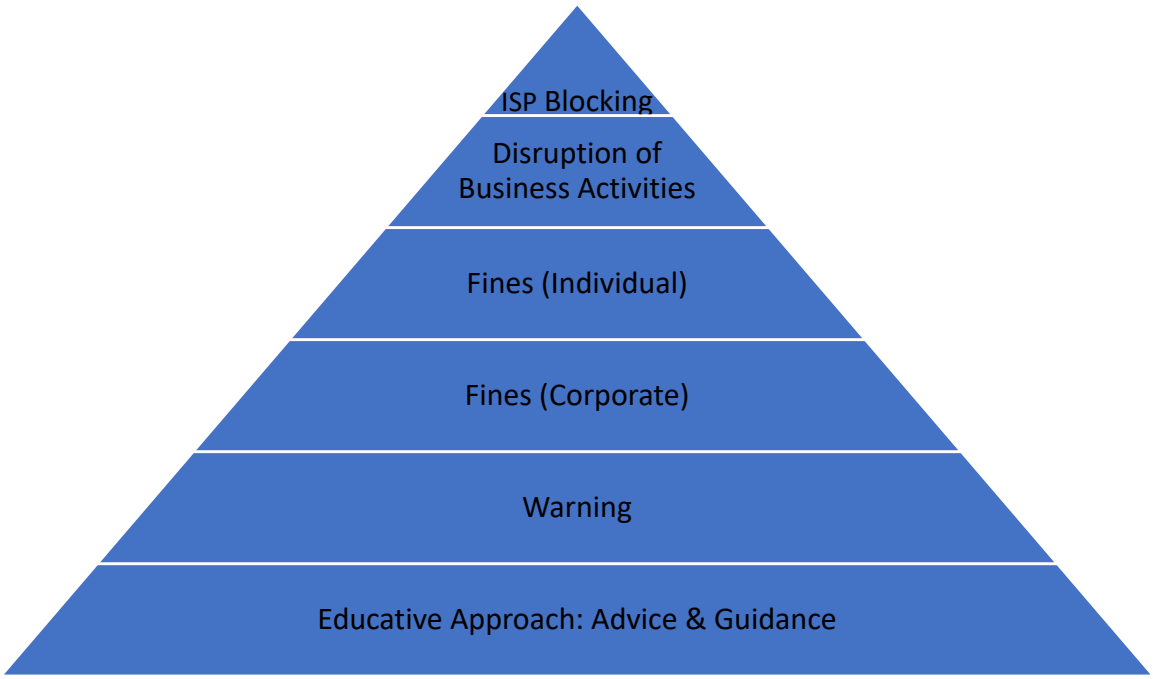
In the next escalation, the regulator will move up a level to fining senior individual employees of the platform. In order for this to take place, specific members of the senior management of each platform must be identified as responsible for taking actions that ensure compliance. If the identified member of senior management fails in this duty then the regulator can impose a fine on them as an individual. It is thought that the threat such a fine creates will really get the attention of the person responsible (Kennedy, 1985) and be a large incentive on senior management to ensure compliance takes place because the consequences will impact them on a personal financial level. It may also, as with the company-level fines, reflect badly on the individual's reputation, potentially hindering their future career opportunities elsewhere. As with most sanctions, there are potential limitations and issues. It may be difficult to pinpoint which member of senior management should be responsible for what actions. There are, similar to the company-level fines, jurisdiction issues here also. Finally, it can be difficult to demonstrate requirements such as neglect, especially in an industry with such complex management structures (Bishop and Macdonald, 2019).

The top two levels of the pyramid carry the most severe sanctions and should not be used lightly. Second from the top is disruption to business activities. This sanction involves penalising companies that provide supporting services, in order to persuade them into withdrawing their services. An example of the consequences of this for a tech platform would be removal from a search engine or app store. It is thought that platforms would not want to risk the impact this could have on their growth. This may, however, affect smaller or newer companies more than the already well-established major platforms (Bishop and Macdonald, 2019). There is also the example of Gab who, with effort, have overcome these issues to create their own third-party services, for example, their own browser (Gilbert, 2019b; Bishop, Looney, Macdonald, Pearson, and Whittaker, 2019). Failing this, the top level and last resort is for the regulator to enforce ISP blocking. This entails the blocking of access to a platform in a country. This is likely to have a big impact on platforms of all sizes and will affect their ability to grow. Given that this affects the lives of many people, for example, many people rely on certain platforms to run their business and document human rights violations, this is a very serious action to take with enormous consequences for not just the platform but the users of the platform, many of whom only post lawful, non-violating content. It could have an extremely significant socio-economic impact and be viewed as a prior restraint on free speech which requires special justification in a liberal democracy (Bishop and Macdonald, 2019). From a human rights perspective, this action may not be deemed proportionate depending on the ratio

of lawful to unlawful content on the platform (Bishop and Macdonald, 2019; UK Home Office 2020). Broughton and Jacques (2019) argue that before these top levels are implemented, the decision should be subject to checks around appeal, judicial review and public reporting requirements. There is also the limitation that users have learned technical ways to circumvent ISP blocking (Bishop and Macdonald, 2019). It is, therefore, very important to restate the importance of the regulator exhausting the middle levels of the pyramid before enforcing the highest levels.

It is unlikely that even the most carefully designed regulatory framework will put forth a perfect solution. It is in part due to the problem of most sanctioning tools containing limitations or issues regarding enforcement (against at least one category of platform). This highlights the importance of the regulator having a diverse range of sanctions at its disposal. If the regulator uses this pyramid effectively by, (1) not moving up the levels too readily but; (2) following through with increasing the severity of the sanction when required, then it may prove to be more successful than other enforcement approaches such as only using one sanctioning tool.

Figure 2. Proposed Enforcement Pyramid



Potential limitations with the enforcement pyramid

The levels of the pyramid will differ from pyramid to pyramid based on the industry and companies that are being regulated. The sanctions involved are chosen based on what is considered appropriate and most effective for compliance. The sanctions must be able to be carried out in order to be perceived as credible (for example, issuing fines across jurisdictions may be difficult to carry out in some cases (Macdonald and Bishop, 2019), if the companies know it cannot be enforced it will not hold any credibility). The sanctions involved will differ across industries. What is consistent, however, is that the severity of the sanctions increases (Black, 2001b). If the top levels of the pyramid are not seen as severe, then they will not be effective in incentivising platforms to comply. Ayres and Braithwaite (1992, p. 19) argue that regulators "will be more able to speak softly when they carry big sticks". Braithwaite (2011) insists that the harsher enforcement tools tend to be seen as more legitimate after a more persuasive style has already been attempted, and when regulation is perceived as more legitimate, the likelihood of compliance increases. Gunningham, Kagan and Thornton (2005) found that hearing about sanctions taken against other firms, led to companies reviewing and sometimes taking further action to ensure their own compliance. Braithwaite (1985) has said that companies must "fear the possibility that they will be among the few who will have the book thrown at them" (p. 142), whilst being fully aware that if they quickly express efforts to reform, the regulator may take more lenient actions. However, the top levels should not be

escalated too lightly due to their severity. This can bring with it challenges. Companies who may well be predisposed to comply will feel at a competitive disadvantage if they spend time, money and effort complying while other companies that do not bother go unpunished, resulting in the companies being disincentivised to comply (Shapiro and Rabinowitz, 1997). The enforcement pyramid is also argued to discourage game-playing and resisting compliance (Nielson and Parker, 2009). Gunningham (2011) believes that the rules of the enforcement pyramid are easily understandable for those using it. This is all dependent, however, on the earlier mentioned issue of the regulator being able to enforce the enforcement strategies and also on the regulator and regulatee's communicating with one another effectively (Mascini, 2013). It is also dependent on Braithwaite's (1985) assumption that,

> "Punishment presumes man to be a rational actor who weighs the benefits of non-compliance against the probability and costs of punishment. Persuasion presumes man to be reasonable, of good faith, and motivated to heed advice…" (p.100).

There can be problems with failing to recognize that, in terms of persuasion, there are some who are not reasonable, of good faith and motivated to heed advice, and regarding punishment, actors are not always rational (Braithwaite, 1985).

Enforcement Pyramid Summary

In summary, this framework proposes the enforcement approach used in responsive regulation theory: the enforcement pyramid. It has been argued throughout regulatory literature that a mix of enforcement strategies and sanctions is most effective (Braithwaite, 1985). The enforcement pyramid uses a combination of persuasion and punishment. For willing platforms, the regulator will begin at the lower levels of the pyramid applying a tit-for-tat strategy. For unwilling platforms, regulators may start at a higher level of the pyramid. The aim of this approach is for the regulator to impose punishment when required, however, without undermining the use of persuasion (Braithwaite, 1985). When punishment is imposed, it should be done so without negative labels that could result in permanent damage to a respectful and trusting relationship between the regulator and regulatee (Braithwaite, 1985). Although many of the sanctions proposed in the enforcement pyramid contain limitations or challenges, even the most carefully thought-out regulatory framework would struggle to propose a perfect solution. The best way to minimize such limitations and challenges is to ensure that the regulator has a diverse range of tools and sanctions at its disposal. Finally, it is important that the regulator exhausts all of the low and -middle levels of the pyramid, not moving too readily to the highest levels due to

the enormous societal consequences, particularly of ISP blocking. However, the regulator must move up the levels when required in order to be seen as a legitimate and credible threat and incentivise platforms to comply.

Regulatory Tracks

The proposed regulatory tracks are designed as a response to the three key compliance issues that have been identified and a fourth track is proposed to outline the procedure that should take place when a tech platform does not face any of the identified compliance issues. The regulatory tracks are designed to ensure that the regulator is responsive to the tech platform's attitude to engagement (willingness), degree of knowledge/expertise (awareness), and its level of resource (capacity). The regulator will, over the course of the four tracks, play a number of different roles (e.g., enforcer, punisher, advisor etc.). The four different tracks provide a clear role and path for the regulator depending on the compliance issue. The overarching goal is to provide the help, guidance and incentive that each tech platform requires in order to achieve compliance with the mandatory regulatory standards proposed in the previous chapter.

Table 5. Regulatory tracks

Regulatory Track	**Description**
4. Unwillingness	The regulator will identify tech platforms that have shown no willingness to comply with either the whole framework or specific standards. The regulator should investigate why the tech platform is unwilling to comply in order to assess whether this can be resolved without the need to take enforcement action. However, if the issue cannot be resolved then the regulator will utilise the tools available and escalate enforcement action up the enforcement pyramid until the tech platform engages.
3. Lack of Awareness	The regulator will identify tech platforms that face compliance issues with either the whole framework or specific standards because the platform lacks the necessary

	awareness, knowledge and expertise to comply. The regulator will assess the areas in which the tech platform requires assistance with knowledge and expertise and draw on the relevant areas of the regulatory track to provide help and assistance with this. This track proposes different ideas and strategies that can be drawn on to increase knowledge and expertise across the scope of the regulatory standards.
2. Limited Capacity	The regulator will identify tech platforms that face compliance issues with either the whole framework or specific standards because the platform lacks the necessary resources to comply. The regulator will assess the areas in which the tech platform requires assistance with capacity building. The regulator will draw on the relevant areas of the regulatory track to provide help and assistance with this. This track proposes different ideas and strategies that can be drawn on to increase access to the resources necessary/make existing resources go further to comply with each of the regulatory standards.
1. Enforced self-regulation	The regulator will identify tech platforms that do not have any of the above compliance issues. These tech platforms will undertake enforced self-regulation that will be monitored by the regulator. These tech platforms, as mentioned previously, may require assistance from any of the above

	tracks at any time if there are changes in how terrorists exploit their platform.

The regulator will create an assessment that will be used in regards to assessing the willingness, awareness and capacity of a platform to engage and comply with the framework. The regulator will contact the selected designated representative of each platform (this is the first mandatory regulatory standard, see Chapter 7) who must engage with the regulator and assessment process. Tech platforms must supply to the regulator the information necessary to assess whether or not the platform has the awareness and capacity to engage with the framework. Failure to do so will be classed as unwilling and automatically assigned to Track 4.

There are eight different combinations of the three compliance issues that a tech platform may face. Table 6 demonstrates these combinations and assigns the appropriate regulatory track(s) that must be utilised and the order in which they should be utilised in order to work towards full compliance with the mandatory regulatory standards.

Table 6. Assortment of compliance issues into the relevant regulatory track

Willingness	**Capacity**	**Awareness**	**Regulatory Track**
Willing	Capacity	Awareness	1
Willing	Capacity	No Awareness	3
Willing	No capacity	Awareness	2
Willing	No capacity	No Awareness	3,2
Not willing	Capacity	Awareness	4
Not willing	No capacity	Awareness	4,2
Not willing	Capacity	No Awareness	4,3
Not willing	No Capacity	No Awareness	4,3,2

An overall aim in proposing the framework and four regulatory tracks is to increase tech platforms abilities and capacity to comply with the framework. The aim is to use an educative approach as widely as is possible, only implementing the punitive sanctions of the enforcement pyramid when all other options have been exhausted. An educative approach is more likely to allow the regulator to uphold the four objectives of the framework than a punitive approach.

Further, being able to use an educative approach as much as is possible eliminates the limitations that were highlighted with the more punitive sanctions in the enforcement pyramid.

Regulatory Track 4: Unwillingness

This regulatory track is aimed at tech platforms that are failing to comply with the mandatory regulatory standards and unwilling to engage with the regulator. This is the track that will most likely require the regulator to apply sanctions from the enforcement pyramid. The regulator should begin by investigating why the tech platform is unwilling to engage. There are many different reasons as to why a platform may be unwilling, some of which may be easier to overcome than others. For example, a platform might be unwilling for financial reasons, however, this could be overcome by the threat of a fine, in which case, it may actually make better economic sense to engage with the regulator and seek to comply with the regulatory standards than pay the fine and risk even further escalation up the enforcement pyramid. Another reason for unwillingness could be that the platform is ignorant or misinformed. The platform may not understand what is required of them or the importance behind what is required of them. This could be fixed through engagement with the regulator and a move to track 3 (awareness). The platform may also have an organisational structure or corporate culture which mitigates against a cooperative relationship with the regulator. This could be fixed by the threat of a number of the sanctions in the pyramid. A reason that is more difficult to solve is if the platform is unwilling to comply because they have strong values that are in opposition to the objectives of the regulatory framework. These tech platforms may believe that compliance with the framework would force them to change their values, brand identity or unique selling point and consequently drive away their userbase. Such platforms are only likely to engage if the regulator applies severe sanctions. Even with the enforcement of the most severe sanctions, it is likely that some platforms, will try to find creative ways around compliance (as in the example of Gab mentioned earlier). Sanctions will still, however, create difficulties for the platform (e.g., it can be very expensive for platforms to create their own third-party services), and therefore the regulator should, where necessary, apply them. The regulator must be cautious of platforms that are unwilling but try to evade demonstrating this in order to circumvent being assigned to this track. As mentioned in the earlier enforcement pyramid section, platforms should be given a maximum of three warnings from a regulator before escalation can take place. The regulator does have discretion to escalate before the maximum use of warnings if there is reason to believe a platform is being disingenuous in their claims of engagement.

This track differs from tracks 2 and 3 in that, tracks 2 and 3 set out tasks and actions to be taken by both the regulator and regulatee, involving high levels of communication and engagement. Due to the nature of this track (platforms demonstrating unwillingness to engage and communicate), the focus of this track is centred on how the regulator should deal with these defiant platforms. The regulator ought to respond to these tech platforms with the proposed enforcement pyramid. As already mentioned above, when a company demonstrates an unwillingness to comply, it is unlikely that a persuasive approach will have any effect. It is, however, upon investigating the reasons why a platform is unwilling to comply, at the discretion of the regulator as to whether or not they think there is a chance that starting at the bottom of the pyramid could be effective in achieving compliance with an unwilling platform. If the regulator does not believe this to be the case then the regulator can decide to start at a level higher up the pyramid. If a platform makes the decision to begin demonstrating efforts towards compliance, then the regulator must reward the platform with a de-escalation of levels down the pyramid.

If the regulator is not strict with the tech platforms in this track and the use of the higher levels of the enforcement pyramid then the effectiveness of the lower levels of the pyramid are likely to decline, potentially creating problems with willingness to comply across all of the tracks. Tech platforms that are putting in a lot of time, effort and resources into complying may become disgruntled or less willing to continue with these efforts if they think that other platforms are getting away with non-compliance. The platforms in this track will put little effort into considering changing their ways and complying if they do not believe the higher levels of the pyramid to be a credible threat.

This track is likely to end in one of three ways. The platform may continue with their defiance to comply until the regulator escalates and enforces the top-level sanction of ISP blocking. Alternatively, the platform may decide at some stage of the escalation of the sanctions in the pyramid to change its ways and make efforts to comply. Regarding the latter, upon de-escalation, the regulator must then examine the next appropriate track the platform should be allocated to. Finally, the platform will appreciate the severity of the harshest levels of the pyramid once imposed and then adopt a more cooperative approach.

Regulatory Track 3: Lack of Awareness

This regulatory track is aimed at tech platforms that face compliance issues because of a lack of necessary knowledge or expertise. First, the regulator must assess that a tech platform is

fully aware of the regulatory framework and what is required of them. Then, as mentioned previously, some platforms may not fully understand how terrorists exploit online platforms or the extent of exploitation/volume of terrorist content that exists on their own platform (this may be for many reasons such as the platform only has a small handful of employees). Once this has been assessed and addressed, the regulator will have to assess which specific standards the companies require help with. The effectiveness of this track relies on the regulator and tech platform working together in order to achieve the overarching goal of compliance with all of the regulatory standards. The role of the regulator in this track is to take an educative approach and provide guidance, advice and encouragement (see Bruhn, 2006). The role of the tech platform is to engage with the regulator's efforts to provide knowledge and expertise, and use this knowledge and expertise to make the necessary changes required to comply with the regulatory standards. Overall, this track revolves around education in order to raise awareness to assist with compliance, which as mentioned earlier, has been shown to be a more beneficial strategy than other strategies, particularly punitive strategies (Braithwaite and Makkai, 2001; Fairman and Yapp, 2005). Only if the tech platform begins to demonstrate a lack of engagement or willingness to participate in the track should the regulator begin to move up the levels of enforcement pyramid.

If necessary, the first action that the regulator can take is to assist/guide the tech platform as to how to undertake an investigation into terrorist use of their platform if they are not aware of the extent to which their platform is exploited. This will help identify the specific areas in which the regulator can provide tailored help and advice.

Regarding how to comply with the mandatory regulatory standards, one method that will be utilised by the regulator is providing educative training materials for the employees of the tech platforms. The regulator will use its own expertise, as well as the expertise of others in the industry, relevant NGOs, CSOs and academia, to create such materials that provide knowledge and expertise on the following areas:

1) Terrorist use of tech platforms
2) How to create and maintain policies regarding terrorist content
3) Best practice on content removal across the industry
4) The importance of an appeals mechanism and how an appeals mechanism should work
5) The importance of a user flagging mechanism and how a flagging mechanism should work

6) The importance and potential limitations of granting user powers and how to do so
7) Key issues concerning employee well-being and how to improve employee well-being
8) What is a transparency report, the importance of transparency reports, how to collect the information required for the report, and how to write a transparency report
9) Existing collaborative ventures and how to join and engage with them
10) The importance of digital literacy and how to contribute to digital literacy programmes
11) User safety and undertaking risk assessments

Such training materials should be accessible at all times to the tech platforms. The materials should give an overview of the topic, explain the issue and offer solutions, and outline what is required of the platform to achieve compliance with the relevant standard. The regulator should also make any existing relevant and useful educational resources that have already been written (by various experts, NGOs, CSOs and academia) readily available alongside these materials.

Tech platforms may face unique issues and challenges with knowledge and expertise depending on the infrastructure of their platform and/or the ways in which terrorists exploit their platform. Some of these issues and challenges may not be covered in any of the above training materials. The regulator must, therefore, be accessible and available to offer more personalised and specific training or guidance or to connect the platform to areas of best practice, where possible/necessary. The regulator should also identify where it would be beneficial for tech platforms to work with NGOs, CSOs or academia to gain knowledge and expertise on a specific issue. The regulator is responsible for connecting tech platforms with the relevant expertise and the tech platforms are responsible for engaging with the organisations that the regulator connects them with.

Overall, this track is about raising awareness and ultimately revolves around education in order to provide tech platforms with the relevant knowledge and expertise to comply with the regulatory standards. The regulator must provide tech platforms with educational materials, be readily accessible and available to offer tailored help, advice and guidance, and where possible, connect platforms with other relevant expertise that can provide educational materials, advice and guidance. The tech platforms in this track have a responsibility to engage with these materials and to communicate with the regulator about which topics and challenges require tailored advice and guidance. It is argued that this educative approach will be more effective than a punitive approach based on the research mentioned earlier (e.g., Braithwaite and Makkai, 2001; Fairman and Yapp, 2005).

If the tech platform, upon receiving educational training materials, help and guidance from the regulator, does not implement the changes necessary to comply with the regulatory standards then the regulator will have to investigate whether this is down to requiring further help/education or a new found unwillingness from the platform. If it is the latter then the regulator should consider looking to escalate the platform up the enforcement pyramid and a move to track 4.

Regulatory Track 2: Capacity

This regulatory track is aimed at tech platforms that face compliance issues because they lack the resources necessary for compliance. As mentioned, the amount/type of resources needed will differ from platform to platform depending on, for example, the number of users and volume of content a platform has, and importantly, how the platform is exploited by terrorist groups. Therefore, the regulator must assess the areas in which the tech platform requires assistance with increasing their resources and making existing resources stretch further. The regulator will focus on helping the tech platforms to build capacity as much as is possible in ways that do not place a financial burden on the platform. This is an ambitious aim; however, the focus will be on other means such as sharing tools and best-practice across the industry. The effectiveness of this track relies on the regulator and tech platform working together to increase access to the required resources in order to achieve the overarching goal of compliance with all of the regulatory standards. This track also takes an educative approach. The role of the regulator in this track is one of guidance, advice and encouragement (see Bruhn, 2006). The role of the tech platform is to engage with the regulator's efforts to increase access to resources and implement these resources in order to comply with the regulatory standards. This track requires flexibility and innovation from the regulator and platforms across the industry.

The first method that can be employed by the regulator is the designing and sharing of resources across the industry. The regulator will design and collate a range of useful templates that can be accessed by the tech platforms. There will be a transparency report template and a model terms of service/policy template. Platforms can communicate to the regulator any other templates that they would find useful to assist with compliance and the regulator should make efforts to help where possible or connect platforms with other relevant experts who can help. During their consultation for the Terrorist Content Analytics Platform (TCAP), Tech Against Terrorism (2020) found that tech companies find producing transparency reports arduous, however, expressed that it would be helpful to receive support with the process. In the proposed

framework, platforms are permitted to diverge from the template where necessary because the framework and regulator acknowledge that platforms are diverse. Some parts of the template may not be relevant for a platform. Alternatively, there may be information that the platform acknowledges in the report that was not listed on the template. This thesis recognizes the need for such flexibility which is one of the main reasons why a responsive approach was chosen. The regulator should be available to provide assistance and guidance. The regulator will also make available the collated information that was undertaken in the regulator's multi-stakeholder consultation that involved input from range of stakeholders (users, employees, NGOs, CSOs, and IRUs) regarding policy and content removal decision-making processes.

The second method that the regulator can draw from is existing industry collaborations. Through this there can be 1) the sharing of best practice; and 2) the sharing of tools and technology. The tech platforms that have designed tools, technology and other relevant resources and undertake what is considered best practice have a responsibility to participate in and support collaborative initiatives to increase access to resources for platforms that do not have readily available access to such resources. There would be careful consideration regarding concerns platforms may have around sharing proprietary software and how this may affect competition between the platforms.

It must be acknowledged that the involvement of collaborative ventures that this track puts forth, such as the GIFCT, have come under recent criticism. One of many criticisms by Douek (2020) is the lack of oversight. Douek argues that without such oversight, it is unclear whether the GIFCT does in fact respect the rights that it claims to. Further, there is a lack of transparency regarding tools such as the hash sharing database (Llanso, 2019). It is unclear exactly what content is or is not stored on this database or how accurate the tool is in identifying terrorist content. There is also a lack of opportunity to challenge or appeal content in the database. Another criticism has been that any bias in one platform's tools and technology could then seep into the moderation on other platforms (Douek, 2020). Douek argues that such collaborations could result in allowing the major platforms "to decide standards for smaller players" (2020, p.28). This removes the ability for contestation and debate that normally arises from the market-place of ideas.

Although there are these limitations, the sharing of best practice, tools and technology is particularly useful in situations of cross-platform abuse (Llanso, 2020a). Such a standardised approach may help to minimize the current whack-a-mole problem whereby terrorist groups

flock to platforms that do not have the capacity to remove them. Without this sharing of tools and technology, platforms that lack capacity may believe that they must err on the side of caution and take a remove-everything approach to avoid punitive measures, infringing on freedom of speech (Douek, 2020), which is one of the objectives that this framework seeks to protect. Moreover, it allows smaller platforms to avoid the failures and pitfalls that other tech platforms have experienced before them (Llanso, 2020a). Overall, without such sharing, platforms that lack capacity will be vulnerable to punitive measures and face challenges that will burden them in ways that could disadvantage them and reduce market competitiveness. This thesis calls on collaborative ventures such as the GIFCT to do more regarding oversight and transparency. This thesis also acknowledges the limitations of such technology (e.g., bias and errors), however, due to the scale of the issue of terrorist content on tech platforms, it will be impossible to propose a framework that does not include the use of such technology.

The regulator should assess what resources a tech platform requires and investigate the options that are available for tool-sharing from other platforms. Some examples where best practice and/or tools/technology could be shared are how to implement appeal mechanisms and flagging mechanisms. Another example would be providing the platforms with access to the GIFCT's hash sharing database. A final example are the tools and methods used to track and collect data for transparency reports. This example has been noted by Tech Against Terrorism (2020) with the point that,

> "Many of the smaller platforms that are exploited by terrorist groups struggle to publish transparency reports due to lack of capacity: without automated data capturing process in place, compiling and publishing a transparency report can be time and labour intensive".

This is not an exhaustive list of examples where the regulator could ask tech platforms to share best practice, tools and technology and various other relevant resources that they use. It is the responsibility of the regulator and platforms to investigate and communicate which areas of the regulatory standards they have issues complying with and therefore what areas they could benefit from shared best practice, tools and technology.

The third method that the regulator can draw on is collaboration with academia, NGOs and CSOs. Through this there may also be opportunities to share tools and technology.

Finally, the regulator will make available industry-wide resources that guide tech platforms with how to implement programmes and practices. One example is implementing a Trusted

Flagger programme. This programme provides tools for those who are exceptionally interested in, and highly accurate at, reporting content to a platform for review (YouTube, 2016). Such a programme takes some pressure off a platform's technology and content moderator workloads, as well as reducing financial burdens. Trusted flaggers have been reported to accurately flag content in over 90 percent of cases which is three times more accurate than the average flagger (YouTube, 2016).

Overall, this track is also based on an educative approach. The purpose of this track is to help and assist platforms build their resources and thus their capacity to comply with the mandatory regulatory standards. However, this is done in a way that does not place a financial burden on any platform. The aim, where possible, is to help make existing resources go further, to learn from each other and share resources. The approach is based on seeking resources created by the regulator or other relevant experts, sharing best practice, tools and technology across the industry and collaborating with experts and other platforms where possible.

If the tech platform, upon receiving assistance with building capacity refrains from implementing the necessary changes to comply with the regulatory standards then the regulator will have to investigate whether this is down to requiring further help/education or down to a new found unwillingness from the platform. If it is the latter, then the regulator should consider looking to escalate the platform up the enforcement pyramid and a move to track 4.

Regulatory Track 1: Enforced Self-Regulation

This regulatory track is considered the "Gold Standard". Tech platforms in this track will (at the time of assignment to this track) be willing to comply with the framework in full and have both the awareness and capacity to fully comply. This track proposes a move away from the self-regulated approach that tech platforms have experienced over the last few years to an enforced self-regulation approach (also sometimes referred to as regulated self-regulation) Bruhn (2006). Self-regulation is defined by Ogus and Carbonara (2011, p.587) as "law formulated by private agencies to govern professional and trading activities". Self-regulation, therefore, places the responsibility of ensuring compliance on the organisation itself. Ogus and Carbonara (2011) criticize that a self-regulatory approach has much potential for abuse, and this is supported by the research mentioned in chapter 4; the Jugendschutz.net study found widespread inconsistency across tech platforms removal of content under a self-regulatory approach (2017).

An enforced self-regulation approach has been increasingly used in environmental protection and health and safety in order to accommodate both rule and principle-based regulatory standards (Fairman and Yapp, 2005). Such a traditional enforced self-regulation approach is the inspiration for the approach taken in this track.

According to Braithwaite, who first coined the term enforced self-regulation, (1982, p.1470),

> "Under enforced self-regulation, the government would compel each company to write a set of rules tailored to the unique set of contingencies facing that firm. A regulatory agency would either approve these rules or send them back for revision if they were insufficiently stringent. At this stage in the process, citizens' groups and other interested parties would be encouraged to comment on the proposed rules".

Hutter (2001, p.380) describes enforced self-regulation as,

> "a mix of state and corporate regulatory efforts. The government lays down broad standards which companies are then expected to meet. This involves companies in developing risk management systems and rules…regulatory officials oversee this process. They undertake monitoring themselves and can impose public sanctions for non-compliance".

This regulatory track is based on this enforced self-regulation approach, however, has some minor differences that must be noted. The framework has proposed a set of twelve mandatory regulatory standards. The tech platforms that fall under this regulatory track must comply with these twelve standards. Therefore, unlike enforced self-regulation, the development of regulatory standards is not a joint enterprise. However, in line with enforced self-regulation, tech platforms have flexibility in how the regulatory standards are achieved. Two platforms may adopt completely different methods to one another in their efforts to comply with the same regulatory standard, however, still achieve the same outcome. If the regulator assesses that the decisions/methods tech platforms have implemented to comply with the standards are not sufficient then, in an enforced self-regulation style, the regulator will inform the platform that they must revise these methods. If the tech platform fails to revise the methods then the regulator will take enforcement action, beginning at the bottom of the enforcement pyramid and working up the levels until the tech platform complies.

Therefore, unlike traditional enforced self-regulation, the tech platforms are not free to write their own standards because the framework proposed in this thesis puts forth twelve mandatory regulatory standards. However, the tech platforms do have some freedom and flexibility in deciding how they undertake compliance with the standards and this will be overseen by a regulator and either result in approval with on-going monitoring, or revision and monitoring.

This approach distinctly defines the separate roles of the tech platforms and the regulator. The framework proposed twelve mandatory regulatory standards. The role of the tech platforms is to comply with these standards, however, as mentioned there is flexibility with the methods a platform chooses, as long as these methods achieve the desired outcome. Regarding this flexibility, some standards may allow greater discretion than others. For example, under mandatory regulatory standard 3. "Implement a multi-stakeholder approach to policy-making", platforms can decide whether to engage with the regulator's multi-stakeholder consultation process or whether to design and undertake their own multi-stakeholder consultation process. If a platform decides to undertake their own consultation process then the platform will have to submit a plan detailing this to the regulator which may or may not be approved. Other mandatory regulatory standards, such as 11. "Support and engage with digital literacy programmes" provide some freedom and flexibility with how tech platforms choose to comply with the outcome of supporting and engaging with digital literacy programmes. The role of the regulator is to monitor that the tech platforms are implementing and designing methods that are assessed as complying with the standard and fulfilling the outcome the standard aims to achieve. The regulator should be available to answer any questions that platforms may have about whether they are fulfilling compliance. The regulator is responsible for on-going monitoring of compliance with the standards and implementing enforcement action if required. Finally, the regulator must also investigate any user-complaints regarding systematic failings with any of the standards.

It is important to note that no regulatory approach is perfect, and therefore, there are some potential drawbacks and limitations to this approach. Enforced self-regulation may lead to tensions between the companies being regulated and the regulator, as well as the regulator and the government (Kaye, 2006). There is also a risk of regulatory overkill and making the regulatory landscape more complex (Kaye, 2006). Critics of the approach may even try to argue that it is simply indirect government regulation (Kaye, 2006). Bruhn (2006) raises the problem of the "inspectors dilemma" whereby the regulator will face challenges when monitoring compliance. For example, the regulator will face dealing with competing values between the

different parties and the expectations that contradictory demands and objectives are solved or fulfilled. The framework has already outlined potential conflicts between the four objectives in the previous two chapters and how they should be resolved, this does not ignore, however, the difficulty of such conflicts for the regulator.

Although there are limitations, an enforced self-regulation approach aims to overcome the problems identified with self-regulation and has many advantages. In response to the limitation that it could lead to tensions between the regulator and platforms, it also creates opportunities for strong relationships to form between the two. It opens a channel of communication that can minimize misunderstandings and open pathways to clarifications and help if necessary. In response to the criticism that it could make the regulatory landscape more complex, under this framework where the 12 mandatory regulatory standards are proposed in an industry-wide manner, it is more likely that this approach will help to make the regulatory landscape clearer.

In response to criticism regarding indirect-government regulation, this approach provides the platforms with some freedom and responsibility to choose the methods they are going to use to comply with regulation. Braithwaite (1982) and Hutter (2001) argue that such an approach is likely to make engagement and compliance more appealing to companies. A subsequent benefit of this is that the platforms will not be able to plead ignorance in situations of failed compliance (Braithwaite, 1982). Moreover, enforced self-regulation encourages the inclusion of other parties/stakeholders in the decision-making processes (Braithwaite, 1982) which is an idea that has already been argued as important in the framework. Other advantages include that it may instil a new sense of trust from the public (Kaye, 2006). This is particularly so in this context because the standards have been set by an independent regulator and the monitoring undertaken by the regulator ensures the accountability that self-regulation is missing. The on-going monitoring and threat of enforcement action creates a compliance incentive (Braithwaite, 1982). Enforced self-regulation allows platforms to retain a level of flexibility and freedom regarding their own specific operational requirements when complying (Hutter, 2001). This encourages and ensures innovation (Hutter, 2001) which is one of the key objectives of the framework. It also brings together the expertise of both the regulator and the tech platforms which Bruhn (2006) argues is important because no party has the complete knowledge and comprehensive picture required to regulate.

Finally, enforced self-regulation is less time-consuming than other models of regulation, thus allowing the regulator to put some of the effort on the tech platforms (Hutter, 2001). An

example of existing enforced self-regulation with these benefits is in civil aviation (Braithwaite, 1982). In civil aviation, companies must implement methods (e.g., concerning survival equipment and operating procedures) with their own aircraft and flight paths in mind, allowing innovation. These methods must be approved by the relevant civil aviation authority. Such an authority can punish non-compliance thus creating accountability.

Further to this enforced self-regulation approach and on-going compliance with the mandatory regulatory standards, tech platforms in this track are expected to become more critical of themselves and develop a culture of self-evaluation (even though they will be evaluated and monitored by the regulator). The aim of this is to develop a culture of continuous improvement and development (Carnino, 2000) in order to meet the framework's four objectives (promoting innovation, free speech and user autonomy, and preventing harm). This is an approach that has been heavily encouraged in the Occupational Health and Safety industry (under social regulation theory) for many years (Parker, 2002). It is argued that the tech platforms in this track are capable of this because they have been assessed as willing and have the knowledge, expertise and resources required to do so. Under such requirements, a tech platform should direct efforts to improving "communications, training, management style, and improving efficiency and effectiveness", with long-term goals in mind, where possible (Carnino, 2000, p4). The platforms would be encouraged to set themselves targets and goals that will help them meet the framework's objectives. The platforms could create indicators to self-assess themselves, for example, "the percentage of employees who received training on X regarding safety", "percentage of employee communications briefs that included safety information on X" or the "percentage of employee suggestions that were implemented to improve X" (Carnino, 2000). With regular review of such indicators, the platform will be able to assess whether or not they are meeting their goals and targets for continuous improvement. This will take commitment from employees at all levels of the platform (Carnino, 2000).

As this track is the regulatory track with the most hands-off approach from the regulator, the regulator will need to be cautious of attempts by tech platforms to evade a more hands-on regulatory track and claims of compliance when in fact only superficial or cosmetic efforts have been made (Krawiec, 2004). It is important that the tech platforms are aware that the regulator can remove them from this track at any time and place them on any of the other three tracks. It is important that the tech platforms are also aware of the tools the regulator has at its disposal in responding to non-compliance and how the enforcement pyramid works. If the regulator does not remove platforms from this track during times of non-compliance, either

because the platform requires assistance with awareness and/or capacity, or because the platform has become unwilling to comply in one way or another, then the regulator will lose credibility in terms of both its approachability and supportive role, as well as in its more punitive enforcement role.

Overall, this track is the "Gold Standard" of tracks and the end goal for the platforms that fall under the scope of this framework. It has been established that self-regulation was not working as intended for tech platforms. As a result, this approach is based on an enforced self-regulation approach. Although the platforms do not create their own standards, they do have some flexibility in some of the decisions they make regarding how they will fulfil compliance and this will be monitored and approved or disapproved by the regulator. This approach removes many of the disadvantages of the self-regulation approach that was previously taken. It creates the accountability that a self-regulatory approach lacks whilst still allowing platforms to be innovative and retain some freedom and flexibility in how they achieve compliance. This track also aims to develop a culture of continuous improvement and development to aid in meeting the framework's four objectives. If at any time, the platform ceases to meet the required criteria for this track, the regulator will demote the platform to one or more of the other tracks.

Conclusion

In conclusion, this chapter discussed research that has been undertaken across other industries that revealed that there are benefits to categorising types of companies. The main benefit of doing so is that it helps to create a profile of each different category of company and this can be used to create tailored responses to their compliance or non-compliance. This thesis categorises tech platforms into four categories (lacking willingness, lacking awareness, lacking capacity, or having all three) and as such proposes four regulatory tracks as an approach that the regulator can take to try and help each platform achieve and maintain compliance with the twelve mandatory regulatory standards put forth in the previous chapter. The response to non-compliance that is argued to be most appropriate is the enforcement pyramid that is put forth under a responsive regulation approach. This is because it offers the regulator two choices: persuasion or punishment. The lower levels of the pyramid start with persuasion, taking an educative approach, and work up to sanctions that get more severe the higher the level of the pyramid.

Platforms that lack willingness (track 4) are most likely going to require the regulator to enforce sanctions from the enforcement pyramid in order to incentivise compliance. Platforms in tracks

2 and 3 are going to require the regulator to take an educative approach. Platforms in track 1 will be assessed as having the necessary willingness, awareness and capacity to comply with the mandatory regulatory standards. The approach taken under this track will therefore be an enforced self-regulatory approach. It has been argued that a self-regulatory approach is not effective enough and as such another approach is required. Track 1 is where the regulator aspires to get the platforms that fall under the scope of this framework to be. However, it is important to acknowledge the range of compliance issues and difficulties that must be overcome. It is also important to note that compliance is not static or permanent, especially not in such a complex, rapidly evolving industry such as the tech industry. Therefore, platforms may move back and forth between tracks depending on changes to their site's infrastructure, the services they offer, and also the way that terrorist organisations evolve in the ways that they exploit such platforms.

These four regulatory tracks are proposed as a solution to the issue that a punitive approach to non-compliance with existing regulatory frameworks trying to counter terrorist content on tech platforms have not worked well enough. Regulatory frameworks cannot overlook compliance issues, whether these issues are intentional or unintentional. These four regulatory tracks aim to offer solutions and strategies that can be used to overcome some of the key compliance issues. Education is offered as the main solution with a punitive approach only implemented as a last resort. If this fails, the regulator then has a diverse array of sanctioning tools that can be drawn on.

Chapter 9: Conclusion

Prior to around 2016, terrorist organisations were free to use tech platforms to fulfil a variety of their needs because tech platforms were going largely unregulated. However, around this time, governments around the world increased pressure on tech platforms to counter terrorist use of their sites. At first, tech platforms were trusted with a self-regulatory approach, however, research, such as the Jugendschutz.net study (Cited in Schmitz and Berndt, 2018), revealed that this approach was not working consistently across the industry. As a result, many governments have since either proposed or implemented an array of regulatory frameworks that seek to counter terrorist content on tech platforms. These frameworks have been widely criticised for a number of reasons, most importantly the effect on free speech and placing unfair burdens on smaller platforms. Therefore, the aim of this thesis was to undertake a thorough examination of the counter-online terrorist content regulatory landscape in order to develop and propose a new regulatory framework.

For this purpose, this thesis first, in chapter 2, undertook a literature review investigating terrorist exploitation of tech platforms. Then, in chapter 3, an analysis was undertaken on tech platform blogposts (because they are the main way that platforms communicate with their users and are understudied in this context) in order to investigate tech platform responses to terrorist use of their sites. This included examining differences between platforms and the challenges that platforms face. In chapter 4, a sample of regulatory frameworks were examined to investigate what lessons could be learned for future regulation. Next, social regulation theory was identified as a theory that could be applied in this context. Chapter 5, therefore, examined social regulation theory in three different regulatory contexts (environmental protection, consumer protection, and occupational health and safety). In this chapter, regulatory strategies that have been used in these contexts were examined in order to understand what can be learned from them and whether they could be adopted in the regulatory framework that is proposed in this thesis. The findings from chapter 2 through to chapter 5 were then used to influence the development of the regulatory framework that is proposed in chapters 6-8 of this thesis. Chapter 6 proposes four objectives and the ethos that underpin the framework and why. Chapter 7 proposes twelve mandatory regulatory standards that tech platforms must comply with and in doing so, identifies three main compliance issues that tech platforms may face. Chapter 8 proposes four regulatory tracks as solutions to address these three compliance issues.

This framework proposed in thesis is intended to be applied in the UK. This is because it could be argued that a good approach to new regulation would be to test it first on a smaller scale (e.g., one country), in this case, in the UK, in order to examine any challenges or limitations that may require amending. After this testing phase, amendments could be made and consideration could be given to rolling out the regulation more widely, for example, regionally, across Europe, and perhaps even internationally. However, an issue with this, and the second reason as to why it is aimed solely at the UK at this stage is that such a framework will face challenges regarding geographic reach because of the transnational nature of cyberspace. There is the possibility that people will exploit the features of cyberspace to try and circumvent the regulatory framework and that the regulator will face difficulties carrying out enforcement actions across jurisdictions. It is also difficult to achieve international adoption due to the significant differences that exist between countries and their varying levels of protection of free speech (e.g., America with the First Amendment, and also authoritarian countries such as China) (Aziz, 2015; Bychawska-Siniarska, 2017; Barednt, 2007). This is discussed further in the section titled 'Regulatory challenges and future research' below.

In summary, this thesis focused on the creation of regulation, drawing on social regulation theory, to propose a new regulatory framework to counter online terrorist content on tech platforms. The decision to focus on terrorist content was partly due to the online terrorist content landscape that was in place in 2016 when this PhD began. However, there was also a decision to focus on terrorist content only because expanding the regulatory scope to other types of content, such as extremist content and hate speech would create a much wider and more complex regulatory scope that would require different free speech considerations. This thesis had a wide scope of researching a number of key areas of expertise that are argued as necessary for developing a well-researched regulatory framework, including the ecosystem of platforms used by terrorist organisations, what tech platforms already report doing to counter terrorist content and the challenges they face, criticisms of existing regulatory frameworks in this area, protecting human rights, protecting platforms business interests, and how social regulation has been applied elsewhere. Although an important and still misunderstood issue, this thesis did not set out to discuss who decides, and how they decide, what is terrorist content or not. However, this thesis acknowledges that is these questions are crucial to the success of regulation and should be addressed in future research.

Research Findings

This research investigated nine research questions that are argued as key to the development of the new regulatory framework. The first three research questions asked:

1) Which tech platforms are exploited by terrorist organisations?
2) How do terrorist organisations exploit tech platforms?
3) What considerations do terrorist organisations have when choosing which platform to use?

These questions were addressed by the literature review that was undertaken in chapter 2 into the ways in which terrorists exploit tech platforms. The findings supported the existing argument that terrorist organisations utilize a whole ecosystem of platforms: social media platforms, alternative platforms, file-sharing sites, instant messaging sites, and archive sites (Frampton et al. 2017; Fisher et al. 2019, Conway, Khawaja, Lakhani, Reffin, Robertson and Weir, 2019; Macdonald, Grinnell, Kinzel and Lorenzo-Dus, 2019). The literature review found that terrorist organisations are thought to use this whole ecosystem for two main reasons. The first is to be able to remain online. It is too risky for a terrorist organisation to rely on a small handful of platforms to maintain their online presence. Any disruption could result in the organisation losing their content, followers and community. Terrorist organisations have realised that it is easier to circumvent disruption strategies if they utilize a wide array of platforms and implement a strategy of continuous signposting to less censored platforms. The second reason is that is appears that no single platform can provide all of the services, protections and audiences that a terrorist organisation requires to undertake all of the activities that they require. Each platform offers a limited but unique set of services, protections and audiences. For example, platforms, such as Twitter, provide access to a mass audience and opportunities to signpost followers to less-censored platforms. However, platforms such as Telegram provide privacy and security. Platforms such as YouTube, as well as a range of file-sharing sites, provide content repositories, and platforms such as Gab provide an unwillingness to remove speech that would be removed elsewhere. A tech platform's capacity and willingness to identify and remove terrorist content was also found as a likely factor in terrorists' choice of platforms or in how they use the platform. Therefore, the more platforms a terrorist organisation utilizes, the more resiliency it builds against disruption, and the more functions, protections and audiences it will have access to. The main considerations that arose from this chapter for the development of the regulatory framework that was proposed in this thesis was that the framework must include the whole ecosystem of platforms that are used by terrorist organisations as they all play a vital role in terrorist organisations online operations. Another

finding was that platforms do not all contain the same levels of capacity and expertise to counter terrorist content on their sites. They also differ on how willing they are to counter terrorist content on their platforms.

The next questions were addressed in chapter 3:

4) In their blogposts, what efforts do tech platforms report taking to counter terrorist content on their services?
5) What challenges do tech platforms face in their efforts to counter terrorist content on their services that could affect their compliance with regulation?

The findings of chapter 3 that undertook a content analysis of tech platform blogposts to investigate tech platforms' reported efforts to counter terrorist content on their platforms found that platforms differ from one another in several ways. First, the platforms differ in terms of size which includes both the size of the userbase and number of employees. They also differ in how they earn revenue. The findings of the thematic analysis of blogposts include differences in their policymaking; use of terms; response to ideologies, groups and movements; use of technology; use of human review; collaborations; CVE efforts; and attitudes towards regulation. Therefore, tech platforms do not report equal levels of effort in countering terrorist content on their services. Although there are limitations to what can be ascertained from analysing the blogposts, this is one of the only ways that the tech platforms communicate the efforts that they claim to be making. Some platforms report implementing a wide array of strategies and efforts, while other platforms, report very little in the way of efforts and strategies. Platforms therefore differ regarding how transparent they are around their efforts. Platforms also differ in the challenges they face (for example, Twitter and former US President Trump). Reasons for these differences could be because of issues around platform capacity and expertise to counter terrorist content. For example, platforms may not have the necessary financial resources or number of staff required to manage the volume of content and size of their userbase. Platforms may also be unwilling to counter terrorist content if they have missions and values that are anti-censorship. If there are this many differences in a small sample of six platforms, then it is likely that there are many more in the whole ecosystem of platforms that are utilized by terrorist organisations. These differences and challenges are likely to affect platform abilities to comply with regulation. Future regulation must therefore acknowledge these issues during the development process and consider how they could be minimized or

overcome if possible. These findings were used to inform the development of the regulatory framework proposed in this thesis.

Question 6 asked:

6) What has and has not been effective in existing regulatory frameworks that seek to counter online terrorist content?

This was answered in chapter 4 via an examination of a sample of government responses that sought to counter terrorist content on tech platforms. There were several findings regarding what these frameworks did well and where there are gaps and limitations. The findings revealed the importance of mandatory inclusion of complaints mechanisms, appeals mechanisms and transparency reports in order to increase platform transparency and accountability, and user participation, as well as to ensure the protection of free speech. Another finding was the importance of the regulator having a range of enforcement tools at their disposal. It was found that limiting the scope of the regulatory framework to a certain portion of platforms meant that terrorist operations online were only going to be partially disrupted. However, if all platforms in the ecosystem are to be included under the regulatory scope, then there must be consideration as to how to ensure that smaller platforms are not unfairly burdened. The findings revealed that consideration of tech platform employee well-being was missing in many of the existing frameworks despite these employees playing a vital role in the content removal process and reports of mental health conditions arising from working in content removal without satisfactory training and support from the platforms. Another gap in the frameworks was the need to safeguard at the design stage of new services and features. Further, some frameworks neglected the importance of digital literacy, collaborations and CVE efforts. One of the greatest challenges appeared to be the implementation of a timeframe for content removal. Some were too strict, and others lacked clarity. Finally, it was found to be important that future regulation is informed by research findings, existing regulatory strategies and consultations with a range of stakeholders. All of these findings informed the development of the regulatory framework that is proposed in this thesis.

The final three research questions asked:

7) Is social regulation theory applicable to this regulatory context?
8) What is there to be learned from examining social regulation in other regulatory contexts?
9) Could these strategies be applied in this regulatory context?

Chapter 5 introduced social regulation theory because it was identified as being applicable to this regulatory context and having an abundance of relevant research and regulatory lessons that could be learned from and applied to the creation of new regulation. Social regulation theory had not yet been applied in this context, however, has been used for decades in three other regulatory contexts: environmental protection; consumer protection; and occupational health and safety. Social regulation theory is concerned with a broad array of non-economic issues. These include regulating issues that concern the public interest (Ogus, 1994), the promotion of human rights, social solidarity, social inclusion and general societal good (Prosser, 2010; Wilson,1984). It is appropriate for this context because the issues that it is concerned with are directly endangered by terrorist exploitation of tech platforms. It also provides an approach that overcomes the difficulty of proving a causal effect between the activity to be regulated and potential harms it is thought to cause (the Precautionary Principle). Further, given the findings that the internet is a facilitative tool for radicalisation and terrorist activities, the failure to regulate and consider social issues could result in a failure to protect individuals who are vulnerable to radicalisation and those who are the target of terrorist content. The scale of the consequences of these failures is unknown but has the potential to be catastrophic. Social regulation can help overcome a number of market failures that have been identified as occurring during the period of self-regulation that has taken place in recent years.

Finally, social regulation has the advantage of providing a plethora of previous research to draw on and learn from. It cannot be expected that regulation will be effective without having researched the approach in other industries. Several social regulatory strategies were identified across the three regulatory contexts that informed the development of the regulatory framework that is proposed in this thesis. In some cases, only specific aspects of the strategy were identified as applicable. These strategies were the precautionary principle, prior approval, information regulation, and the implementation of a health and safety culture. These strategies bring a new approach and arguably overcome challenges in this regulatory context (for example, the precautionary principle), and assist with the objective of preventing harm (prior approval, and health and safety culture), whilst promoting user autonomy (information regulation). These strategies would not have been identified without having investigated this (social regulatory) body of work. It is argued elsewhere (see Windholz, 2010) that undertaking research into existing regulatory strategies (as seen in chapters 4 and 5) prior to the development of new regulatory proposals can only strengthen the regulatory framework.

The findings of the nine research questions were addressed in chapters 2-5. These findings informed the development of the regulatory framework that was proposed in chapters 6-8. Chapter 6 began introducing the proposed regulatory framework by introducing the four objectives and ethos that have been developed to underpin the framework. The objectives and ethos aimed to ensure that the many benefits of tech platforms are not lost as a result of the proposed regulation to counter online terrorist content on tech platforms. The four objectives were: 1) to preserve and promote innovation; 2) to promote freedom of speech; 3) to promote user autonomy; and 4) to prevent harm. Upon discussion of these objectives, there were three main findings. The first was that the objectives can have synergies with one another, the second, was that the objectives can having differing interpretations, and finally, the objectives will sometimes conflict with one another. This demonstrates the complexity of regulating terrorist content on tech platforms and the importance of a framework that acknowledges this. The promotion of innovation is necessary for two reasons. Tech platforms require innovation to prevent the reduction of market competitiveness and the burdening of certain platforms, but also to ensure that they can keep up with the rapid pace in which terrorist organisations adapt. The other objectives provide clarity surrounding the definitions of autonomy, free speech, and harm. This is crucial because it aids the regulator in cases where objectives conflict. The underpinning ethos of the proposed framework is collaboration, transparency and accountability, and finally, drawing on research in social regulation, the adoption of a health and safety culture.

In order to achieve the objectives and nurture this ethos, chapter 7 introduced the proposal of twelve mandatory regulatory standards that tech platforms must comply with. This chapter explains that the regulatory framework should have a wide regulatory scope, encompassing all tech platforms that have been identified as forming part of the whole ecosystem that is used by terrorist organisations in order to fully disrupt their operations. There is recognition, however, of the need to ensure that smaller tech platforms are not unnecessarily burdened unlike in other frameworks (Tech Against Terrorism, 2020f). A new regulator is proposed that will be funded by a fee that is applied to platforms that earn over a certain amount of annual revenue. All tech platforms must:

1. Designate representatives in their organisation as the point of contact for the regulator
2. Maintain clear up-to-date policies regarding online terrorist content
3. Implement a multi-stakeholder approach to policy-making processes

4. Make all reasonable efforts to remove online terrorist content
5. Implement an appeals mechanism
6. Implement a user flagging mechanism
7. Implement user powers
8. Implement processes to ensure employee well-being
9. Publish bi-annual transparency reports
10. Engage in collaborative ventures
11. Support and engage with digital literacy programmes
12. Create appropriate mechanisms to ensure that user safety is considered in the design and development of new features

The proposed mandatory regulatory standards are comprised of a variety of rules, rules with exceptions, and principles in order to provide flexibility to innovate and minimize unnecessary burdens on platforms. These standards are proposed to support the four objectives and overall ethos of the framework. These twelve mandatory standards draw on the social regulation strategies examined under the three contexts in chapter 5 (environmental protection; consumer protection; and occupational health and safety). The aim for this framework was to incorporate identified best practice from existing frameworks and identify lessons to be learned from social regulation. The framework proposed takes a standardized approach that aims to promote innovation, free speech, user autonomy, and to prevent harm, whilst being underpinned by an ethos of collaboration, transparency and accountability, and a health and safety culture. However, during the development of this chapter, three main compliance issues were identified. The first was ‘awareness’ which is when the platform does not have the knowledge or expertise to comply. The second was ‘capacity’ which is when the platform does not have resources to comply. The final compliance issue was ‘willingness’ which is when platforms are unwilling to comply. Many existing frameworks have failed to identify and address these compliance issues. Overall, this chapter argues that mandatory regulatory standards are crucial to achieve the standardization that regulation for this industry requires. However, are unlikely to be effective unless the three compliance issues are addressed.

Chapter 8, which is the final chapter that proposes the regulatory framework, addresses the final challenge which is the three identified compliance issues. This chapter proposed four regulatory tracks that aim to address these issues. These four regulatory tracks require the regulator and tech platforms to work together to assist with the overarching goal of achieving

compliance with the mandatory regulatory standards. These tracks are a response to identifying that a standardized approach is required, however, platforms are not homogeneous, they vary in terms of capacity, expertise and willingness to comply and each face unique challenges. This chapter was developed on the basis of research that has been undertaken in other industries.

First, this research revealed that there are benefits to categorizing types of companies (Gunningham, 2007; Gunningham, 2010; Black, 2001b; Baldwin, 1997). This helps to create a profile of each different category of company and this can be used to create tailored responses to compliance issues. The four categories of tracks created in this framework are lacking willingness; lacking awareness; lacking capacity; or containing the awareness, capacity and willingness to comply. The response to non-compliance that is argued to be most appropriate is the enforcement pyramid in responsive regulation. This is because it offers the regulator two choices: persuasion or punishment (Braithwaite, 1985). It was identified that many of the existing regulatory frameworks were based on punishment despite issues with this, therefore, this framework chose an alternative approach. The lower levels of the pyramid start with persuasion, taking an educative approach, and work up to sanctions that get more severe. Platforms that lack willingness (track 4) are most likely going to require the regulator to enforce sanctions from the enforcement pyramid in order to incentivise compliance. Platforms in tracks 2 (lacking capacity) and track 3 (lacking awareness) require the regulator to take an educative approach, providing them with resources and knowledge to boost their capacity and awareness. Platforms in track 1 are assessed as having the necessary willingness, awareness, and capacity to comply with the mandatory regulatory standards. The approach taken under this track will therefore be an (amended) enforced self-regulatory approach. Track 1 is where the regulator will aspire to get platforms to. These four tracks are proposed as a solution to the issue that a punitive approach to non-compliance has not worked well enough for existing frameworks. Regulatory frameworks cannot overlook compliance issues, whether they are intentional or unintentional. These four regulatory tracks aim to offer solutions and strategies that can be used to overcome these compliance issues. Education is offered as the main solution with a punitive approach only implemented as a last resort. If this fails, the regulator then has a diverse array of sanctioning tools that can be enforced.

Overall, this thesis identified a problem: that existing regulatory frameworks are not working well enough to counter online terrorist content. This thesis therefore designed 9 research questions to investigate a number of research areas that are key to developing a new well-researched social regulatory approach to countering online terrorist content on tech platforms.

Further to this, this regulatory framework addressed criticisms in existing frameworks and proposed solutions to identified compliance issues.

Regulatory challenges and future research

Whilst the regulatory framework proposed in this thesis addresses many of the key criticisms of existing frameworks in this area, the implementation of this framework is not without its challenges. Perhaps the greatest challenge that this (or any other) framework faces is the issue of its geographic reach when set against the transnational nature of cyberspace. There is the possibility that people will exploit the features of cyberspace to try and circumvent the regulatory framework and that the regulator will face difficulties carrying out enforcement actions across jurisdictions.

It is important, therefore, that wider adoption of the regulatory framework, beyond the UK is encouraged. Regional adoption of the framework (e.g., Europe) is a more realistic ambition than international adoption given the significant differences that exist between countries, particularly in respect of the right to free speech (e.g., America with the First Amendment, and also authoritarian countries such as China) (Aziz, 2015; Bychawska-Siniarska, 2017; Barednt, 2007). Regional adoption would seek to minimise the issue of fragmentation that currently exists that tech platforms have reported struggling with (Echikson and Knodt, 2018; European Commission, 2018). However, upon regional adoption, international adoption could be encouraged. Such an approach has been seen with the Budapest Convention on Cybercrime.[55] The Budapest Convention came into force in 2004 and was the world's first cybercrime treaty (Daskal and Kennedy-Mayo, 2020). It aimed to harmonize laws and make it easier to cooperate across borders and internationally. As of 2020, sixty-four countries have signed the convention, including a number of non-Council of Europe countries (Daskal and Kennedy-Mayo, 2020). A possible supplement or alternative approach could be that countries make an agreement on enforcement action. Under such an agreement, a signatory state would agree to enforce any sanctions imposed by another signatory state.

Although international adoption would be difficult and some countries may never agree to the framework, this does not mean that the regulation will not be effective. The regulatory framework proposed addresses many of the key criticisms of existing frameworks and puts forth new (social regulation) approaches that have been effective in other regulatory contexts.

[55] https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/185

This framework ensures input from a range of stakeholders and protects the well-being of employees, both of which were missing from existing frameworks. This framework aims to overcome the concern around over-blocking through mandatory implementation of an appeal mechanism and proposed a solution to the issue of an overly strict or arbitrary removal timeframe. This framework will minimize unnecessary and unfair burdens on smaller platforms, unlike existing frameworks which risked reducing market competitiveness. Finally, this framework increases the number of tech platforms that can counter online terrorist content due to the four regulatory tracks that will aid smaller platforms efforts in ways that existing frameworks do not.

Future research projects could focus on this question of implementation of the proposed framework. There would be value in a project involving stakeholder interviews in order to investigate further the practicalities of implementing the proposed framework. This would also provide an opportunity to find out what aspects of the framework stakeholders believe could be effective, whether additional regulatory standards are required, and whether additional regulatory tracks are needed, as well as identifying areas of disagreement that need to be resolved in order for the framework to receive wide acceptance. Another research project could investigate other attempts at regulating activities in cyberspace in order to understand how challenges such as geographic reach and jurisdiction have been addressed in these other contexts and assess the effectiveness of these efforts.

Contributions to the field

This research put forth a number of contributions that are argued to be insightful and missing from existing regulatory frameworks that seek to counter online terrorist content on tech platforms. Although the findings of the literature review in chapter 2 are not new, for example, that terrorist organisations utilize a whole ecosystem of platforms, many existing frameworks failed to consider this in their development. As a result, some of the existing frameworks (for example, NetzDG), only targeted a small percentage of platforms in the ecosystem, thereby only partially disrupting the networks and operations of these organisations. This thesis used this literature to inform which tech platforms should fall under the scope of the proposed regulatory framework.

The next contribution is the analysis of the tech platform blogposts which are understudied in the counter-terrorism context. The findings revealed stark differences in what tech platforms report doing to counter terrorist content on their services across a small sample of tech

platforms, particularly between the major platforms and other platforms. These differences include the reported use of technology, human review, collaborations, CVE efforts, transparency efforts, the platforms missions and values, what ideologies and attacks the platforms focused on and addressed, and attitudes towards regulation. These findings suggest that these differences, particularly regarding missions and values, and attitudes towards regulation, could affect the willingness of a platform to comply with regulation. This chapter also compared the size of userbases, the number of employees, and the methods of earning revenue across the platforms. Some platforms have stable methods of earning revenue and a large number of employees, whereas other platforms do not. These findings, combined with the findings of the content analysis of the blogpost suggest that platforms differ regarding the capacity (resources) and expertise that they have to counter terrorist content on their platforms. However, it is important to consider these findings with caution because the analysis only examined what the platforms reported which is not necessarily accurate regarding what they actually do. Platforms may undertake more efforts than they report or vice versa. There is also likely to be an element of PR in the blogposts. However, it is argued that the findings do provide a level of insight, particularly into how the platforms wish to be perceived in their efforts in this area and into the capacity and willingness of the platforms. Therefore, the findings of this chapter aided in the development of identifying potential compliance issues and was therefore able to address them in the framework proposed. This was arguably missing from the existing frameworks examined in this thesis.

The two main contributions of this research are the novel application of social regulation theory and the regulatory framework itself. Social regulation theory had yet to be applied to the countering online terrorist content context. However, this research identified that it is highly applicable. Social regulation theory is concerned with issues that affect the public interest, the promotion of human rights, social solidarity, social inclusion and general societal good (Ogus, 1994; Prosser, 2010; Wilson, 1984), all of which are affected by online terrorist content. Further, tech platforms themselves have described their work as a social mission on many occasions (Miller, 2019). Unlike the previous self-regulatory approach, social regulation ensures that the tech platforms will be held accountable if they fail to fulfil regulatory demands. Social regulation has been applied in three other regulatory contexts for decades (environmental protection, consumer protection, and occupational health and safety), and therefore, provides a plethora of research and experience to learn from.

Many of the key insights and strategies that informed the regulatory framework proposed in this thesis would not have been identified without the body of social regulation literature in other regulatory contexts. One example is the precautionary principle (in environmental protection) which states that if an activity creates a threat to public health or the environment, precautionary measures should be put in place, even if cause and effect relationships have not been reliably or scientifically confirmed (Baldwin et al., 2010). The principle allows the regulatory framework to fulfil the objective of preventing harm, despite the difficulty in assessing a causal link between viewing terrorist content and undertaking terrorist activities. The principle considers important non-scientific factors such as public opinion and social values. The precautionary principle literature also provides examples of conflicting objectives (as seen in the Clean Air example in chapter 5). These examples may be helpful for regulators to examine when faced with conflicting objectives in this context.

The consumer protection literature revealed two approaches. These were prior approval and information regulation. While neither approach is an ideal fit for this regulatory context, elements of both approaches were identified as relevant for the current regulatory context. One element of the prior approval approach that was identified as relevant was the implementation of a thorough risk assessment at the design and development stage of new features or services. This was missing from the existing frameworks that were identified, however, after attacks such as the Christchurch attack, it is argued that tech platforms may not be doing enough in this area. This therefore informed the twelfth mandatory regulatory standard to "create appropriate mechanisms to ensure that user safety is considered in the design and development of new features". The information regulation approach highlights that there are certain groups in society that would particularly be affected by a lack of information or would benefit greatly from more information. In this context, users who are trying to document human rights violations (which have been erroneously removed by tech platforms in the past, see Kayyali and Althaibani (2017)) by posting them on tech platforms, require clear information about what content will be permanently removed. Information regulation also promotes consistency and standardisation across an industry which has benefits for fair and competitive practices. Information regulation therefore influenced the formation of several of the mandatory regulatory standards including the second, "maintain clear up-to-date policies regarding online terrorist content", and the ninth, "publish biannual transparency reports".

Finally, the health and safety culture approach in the occupational health and safety literature influenced the development of several mandatory regulatory standards including number four,

“make all reasonable efforts to remove online terrorist content”, number eight, “implement processes to ensure employee well-being”, and number twelve “create appropriate mechanisms to ensure that user safety is considered in the design and development of new features”. It also highlighted the issue that regulation also neglects psycho-social risks which is a problem in the current context that the existing frameworks failed to address (Article 19, 2017; Boran, 2020; The Guardian, 2018; Gilbert, 2019a).

The final contribution is the proposed regulatory framework. First, the objectives and ethos provide the regulator and tech platforms with the clarity that is required for the inevitable situations where objectives will conflict. In addition to overcoming the limitation of other existing frameworks that only included a small number of tech platforms under its scope, and the limitations that were overcome by the social regulation strategies mentioned above, a number of other criticisms were addressed in the proposed mandatory regulatory standards. Some of the existing frameworks (e.g., NetzDG) failed to implement an appeals mechanism which was highly criticised due to the infringements that this can have on free speech. This regulatory framework included the standard that tech platforms must implement an appeals mechanism. Many of the existing frameworks focused solely on content removal. This framework proposes approaches that will compliment content removal, for example, a marginalisation strategy, collaborations with NGOs and CSOs that promote a range of CVE approaches, and digital literacy programmes. Another criticism was that existing frameworks lacked user participation in policy and regulatory decision-making. One of the ways in which this framework tried to overcome this was stating that tech platforms must implement user powers. A further criticism in existing frameworks was that the timeframe to remove content was either too short and strict or lacked clarity. This regulatory framework acknowledged the complexity and the need for the regulator to consider many factors, such as, the size of the company, the number of users that have viewed the content, the nature of the service it provides, and the volume of content that is posted, amongst other factors. This framework proposes that there should be an element of discretion and flexibility with timeframes. The regulator will assess and consider the ratio of content to staff for tech platforms to ensure that platforms are not working under unreasonable timeframes that would create burdens for some platforms or encourage and incentivise an overly cautious, rights infringing strategy that could lead to errors and a chilling effect on free speech.

Finally, the four regulatory tracks identify that there are three main compliance issues that are likely to affect the effectiveness of the regulatory framework. The existing frameworks did not

appear to have identified, nor addressed, potential compliance issues, and took punitive enforcement actions. A consequence of which is the criticism that many existing frameworks would result in unnecessary and unfair burdens on smaller platforms, thereby reducing market competitiveness. The educative approach used in tracks 2 and 3 to counter a lack of capacity and awareness aim to address this criticism. These tracks provide the platforms with the resources and expertise that is required to fully comply with the regulatory framework without unnecessary enforcement actions that could cause burdens and reduce market competitiveness.

Glossary

Availability heuristic - occurs when people lack statistical knowledge about a particular situation, so they consider risks to be significant only if they can easily think of instances when those risks actually materialised

Behavioural control - when individuals' decisions are overridden with the justification not of protecting their own welfare, safety, happiness, needs and values but that of others

Benefit-cost analysis – balancing the risks and benefits of the regulation or a decision in order to decide if the benefits outweigh the costs

Bounded rationality - the idea that the capacity of individuals to receive, store, and process information is limited and therefore may not necessarily lead to good decision-making

Brandenburg Test – a legal framework that is used to determine whether free speech can be limited in cases where the speech incites violence or crime

Disrupting business activities – an enforcement tool in which a third-party company must withdraw their services from the company that is being regulated

Due process – a legal requirement that the state must respect all legal rights that are owed to a person

Duty of care – a legal obligation imposed on a person that requires a standard of reasonable care to be taken

eSafety Commissioner (Australia) – leader of a government agency that aims to keep citizens safe online

Externalities – in economics, externalities are costs or benefits experienced by third party's that did not consent to the cost or benefit

Fait accompli - a legal concept that once something has already occurred, it cannot be undone, the world has already changed

Individual liability – an enforcement tool (e.g., a fine or jail time) that can be enforced on an individual if they do not comply with aspects of regulation in which they were responsible for

Information asymmetries – in economics, whereby one party has more or better information than another party

International Government Organisations – an organisation composed of sovereign states or other intergovernmental organisations

ISP blocking – when internet service providers impose censorship or limits on what their users can do and access

Market failure – in economics, this is when there is an inefficient distribution of goods and services in the free market

Nemo judex in sua causa - legal principle that means no man should be the judge in his own case because the judge's impartiality may reasonably be questioned

Ofcom – also known as the Office of Communications and is the government-approved regulatory authority for the broadcasting, telecommunications and postal industries in the UK

Penalty Unit - is a standard amount of money that is used to calculate fines when the law has been breached

Probability neglect – during decision-making, people ignore probability and focus on the worst-case scenario

Referral Action Days (RADs) - coordinated referral campaigns (Europol) that focus on flagging as much online terrorist content on certain platforms, usually within a one- or two-day period

Sine qua non causation – legal principle that asks would Y harm/outcome have occurred if X actions had not occurred?

The Aarhus Convention – convention by the United Nations Economic Commission for Europe that grants rights to the public regarding the environment

The Robens Report – put forth the idea of self-regulation regarding health and safety at work

The Stern Review – a report released by economist Nicholas Stern for the UK government regarding the economics of climate change

Transparency Reports – reports that are published on a regular basis by an organisation that discloses specific information

Virtual Private Network (VPN) – a secure connection between an individual and the internet whereby data traffic is routed through an encrypted virtual tunnel and disguises the individual's IP address

1998 Wingspread Declaration – A statement that was released after the Wingspread conference where the precautionary principle was defined

Bibliography

Abbott, K. W., & Snidal, D. (2000). Hard and soft law in international governance. *International organization*, *54*(3), 421-456.

Al Darwish, M. (2019) From Telegram to Twitter: The lifecycle of Daesh propaganda material. *VOX-Pol Blog*. Accessed 18 October 2020 via https://www.voxpol.eu/from-telegram-to-twitter-the-lifecycle-of-daesh-propaganda-material/

Alexander, A., and Braniff, W. (2018) Marginalizing Violent Extremism Online. *Lawfare Blog*. Accessed 9 March 2020 via https://www.lawfareblog.com/marginalizing-violent-extremism-online

Alkiviadou, N. (2019). Hate speech on social media networks: towards a regulatory framework?. *Information & Communications Technology Law*, *28*(1), 19-35.

Allen, R. (2017) Hard Questions: Who should decide what is hate speech in an online global community? *Facebook Newsroom*. Accessed June 2019 via https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/

Archer, J. (2018) 'Far-right social network' Gab goes offline after tech firms pull support. *The Telegraph.* Accessed July 2019 via https://www.telegraph.co.uk/technology/2018/10/29/far-right-social-network-gab-goes-offline-tech-firms-pull-support/

Ardia, D. S. (2009). Free speech savior or shield for scoundrels: An empirical study of intermediary immunity under Section 230 of the Communications Decency Act. *Loy. LAL Rev.*, *43*, 373.

Arjoon, S. (2006). Striking a balance between rules and principles-based approaches for effective governance: A risks-based approach. *Journal of Business Ethics*, *68*(1), 53-82

Arquilla, J. (1996). *The advent of netwar*. Rand Corporation.

Article 19 (2017) Germany: The act to improve enforcement of the law in social networks. Accessed 12 November 2019 via https://www.article19.org/wp-content/uploads/2017/09/170901-Legal-Analysis-German-NetzDG-Act.pdf

Article 19 (2020) EU: Terrorist content regulation must protect freedom of expression rights. *Article 19*. Accessed 10 June 2020 via https://www.article19.org/resources/eu-terrorist-content-regulation-must-protect-freedom-of-expression-rights/

Asch, P. (1988) *Consumer safety regulation: Putting a price on life and limb*. Oxford University Press on Demand

Ashworth, A., & Zedner, L. (2012). Prevention and criminalization: justifications and limits. *New Criminal Law Review: An International and Interdisciplinary Journal*, *15*(4), 542-571.A

Australian Government (2020) 89-2020-Increase to Commonwealth penalty unit value. Accessed 17 December 2020 via https://www.agriculture.gov.au/import/industry-advice/2020/89-2020

Ayres, I., and Braithwaite, J. (1992) *Responsive Regulation*. Oxford: OUP

Aziz, M. H. (2015). Counter Terrorism Measures via Internet Intermediaries: A First Amendment & National Security Dilemma. *JL & Cyber Warfare*, *4*, 1.
Baele, S. J., Bettiza, G., Boyd, K. A., & Coan, T. G. (2019). ISIS's Clash of Civilizations: constructing the "West" in terrorist propaganda. *Studies in Conflict & Terrorism*, 1-33

Baele, S. J., Boyd, K. A., & Coan, T. G. (2019). Lethal images: Analysing extremist visual propaganda from ISIS and beyond. *Journal of Global Security Studies*.

Baldwin, R. (1995) *Rules and government*. Oxford: OUP

Baldwin, R. (1997). *Rules and government*. Clarendon Press

Baldwin, R., Cave, M., & Lodge, M. (Eds.). (2010). *The Oxford handbook of regulation*. Oxford University Press.

Balkin, J. (2014) Old-school/new-school speech regulation. Harvard Law Review 127: 2296
Bantam Books v. Sullivan (1963) 372 U.S. 58.

Bardach, E. and Kagan, R.A. (1982) *Going by the book: The problem of regulatory unreasonableness.* Routledge

Bardach, E., and Kagan, R. (1984) *Going by the book: The problem of regulatory unreasonableness.* Philadelphia, Penn: Temple Press

Barendt, E. (2005). *Freedom of speech*. OUP Oxford.

Barendt, E. (2007) *Freedom of Speech*. Oxford: Oxford University Press

Barlow, J.P. (1996) A declaration of the independence of cyberspace. *Electronic Frontier Foundation*. Accessed 23 March 2021 via https://www.eff.org/cyberspace-independence

BBC (2017) Telegram to block terror channels after Indonesian ban. *BBC*. Accessed November 2017 via http://www.bbc.co.uk/news/business-40627739

BBC (2019a) Facebook: New Zealand attack video viewed 4,000 times. *BBC*. Accessed 26 May 2020 via https://www.bbc.co.uk/news/business-47620519

BBC (2019b) US says it will not join Christchurch Call against online terror. Accessed 18 December 2020 via https://www.bbc.co.uk/news/technology-48288353

BBC (2019c) Christchurch attacks: Facebook curbs Live feature. Accessed 22 March 2021 via https://www.bbc.co.uk/news/technology-48276802

BBC (2020) Social media: How do other governments regulate it. Accessed 26 March 2021 via https://www.bbc.co.uk/news/technology-47135058

Beauchere, J. (2017) The importance of reporting concerns about online content and conduct to tech companies. *Microsoft Blog.* Accessed 23 February 2021 via https://blogs.microsoft.com/on-the-issues/2017/11/06/importance-reporting-concerns-online-content-conduct-tech-companies/

Beattie, A. (2020) How YouTube makes money off videos. *Investopedia*. Accessed 18 February 2021 via https://www.investopedia.com/articles/personal-finance/053015/how-

youtube-makes-money-videos.asp#:~:text=Key%20Takeaways,well%20as%20promoting%20featured%20content.

Beck, U. (1992) *Risk Society*. London: Sage

Beckett, L. (2018) Pittsburgh shooter was fringe figure in online world of white supremacist rage. *The Guardian*. Accessed 26 May 2020 via https://www.theguardian.com/us-news/2018/oct/30/pittsburgh-synagogue-shooter-was-fringe-figure-in-online-world-of-white-supremacist-rage

Bennett, T. (2018) Gab is the alt-right social network racists are moving to. *Vice*. Accessed 16 January 2021 via https://www.vice.com/en/article/ywxb95/gab-is-the-alt-right-social-network-racists-are-moving-to

Benson, D. C. (2014). Why the internet is not increasing terrorism. *Security Studies*, *23*(2), 293-328.

Berger, J. M. (2018). The alt-right Twitter census: Defining and describing the audience for alt-right content on Twitter. *VOX-Pol (October 15), http://bit. ly/2JQrUNh*.

Berger, J. M., & Morgan, J. (2015). The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings project on US relations with the Islamic world*, *3*(20), 4-1.

Berger, J. M., & Perez, H. (2016). *The Islamic State's Diminishing Returns on Twitter: How Suspensions are Limiting the Social Networks of English-speaking ISIS Supporters*. George Washington University.

Bernstein, M. H. (2015). *Regulating business by independent commission*. Princeton University Press.

Berthélémy, E. and Naranjo, D. (2020) Blind faith in technology diverts EU efforts to fight terrorism. *VOX-Pol Blog.* Accessed 29 June 2020 via https://www.voxpol.eu/blind-faith-in-technology-diverts-eu-efforts-to-fight-terrorism/

Bertram, L. (2016). Terrorism, the Internet and the Social Media Advantage: Exploring how terrorist organisations exploit aspects of the internet, social media and how these same platforms could be used to counter-violent extremism. *Journal for Deradicalisation*, (7), 225-252.

Bickert, M. (2017a) Facebook's community standards: How and where we draw the line. *Facebook Newsroom.* Accessed 8 December 2019 via https://about.fb.com/news/2017/05/facebooks-community-standards-how-and-where-we-draw-the-line/

Bickert, M. (2017b) Hard Questions: How We Counter Terrorism. *Facebook*. Accessed September 2017 via https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/

Bickert, M. (2017c) Hard Questions: Are we winning the war on terrorism online? *Facebook News Room*. Accessed December 2017 via https://newsroom.fb.com/news/2017/11/hard-questions-are-we-winning-the-war-on-terrorism-online/

Bickert, M. (2018a) Hard questions: What are we doing to stay ahead of terrorists? *Facebook Newsroom.* Accessed 16 June 2020 via https://about.fb.com/news/2018/11/staying-ahead-of-terrorists/

Bickert, M. (2018b) Hard Questions: How effective is technology in keeping terrorists off Facebook? Accessed June 2019 via https://newsroom.fb.com/news/2018/04/keeping-terrorists-off-facebook/

Bickert, M. (2019) Updating the values that inform our community standards. Facebook Newsroom. Accessed September 2019 via https://newsroom.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/

Bickert, M. (2020) Charting a way forward: Online content regulation. *Facebook report.* Accessed 21 February via https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf

Bickert, M., and Fishman, B. (2018) Hard questions: What are we doing to stay ahead of terrorists? *Facebook Newsroom.* Accessed November 2018 via https://newsroom.fb.com/news/2018/11/staying-ahead-of-terrorists/

Bishop, P. & Macdonald, S. (2019). Terrorist Content and the Social Media Ecosystem: The Role of Regulation. In Francesco Marone (Ed.), *Digital Jihad: Online Communication and Violent Extremism* (pp. 135-152). ISPI.

Bishop, P., Looney, S., Macdonald, S., Pearson, E., and Whittaker, J. (2019) Response to the Online Harms White Paper. *CYTREC, Swansea University*. Accessed 7 December 2019 via https://www.swansea.ac.uk/media/Response-to-the-Online-Harms-White-Paper.pdf

Black, J. (2001a) Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a "Post-Regulatory" World. *Current Legal Problems, 54*: 103-147

Black, J. (2001b). Managing discretion. *Unpublished manuscript, London School of Economics, UK*.

Black, J. (2002). Critical reflections on regulation. *Austl. J. Leg. Phil., 27*, 1.

Blasi, V. (1981). Toward a Theory of Prior Restraint: The Central Linkage. *Minn. L. Rev., 66*, 11.

Bloom, M., Tiflati, H., & Horgan, J. (2019). Navigating ISIS's preferred platform: Telegram1. *Terrorism and Political Violence, 31*(6), 1242-1254.

Bluff, L., & Gunningham, N. (2003). *Principle, process, performance or what? New approaches to OHS standards setting*. The Federation Press.

Bogle, A. (2019) Laws targeting terror videos on Facebook and YouTube 'rushed' and 'knee-jerk', lawyers and tech industry say. *ABC.* Accessed 18 December 2019 via https://www.abc.net.au/news/science/2019-04-04/facebook-youtube-social-media-laws-rushed-and-flawed-critics-say/10965812

Boran, M. (2020) Life as a Facebook moderator: 'People are awful. This is what job has taught me. *Irish Times.* Accessed 10 March 2020 via https://www.irishtimes.com/business/technology/life-as-a-facebook-moderator-people-are-awful-this-is-what-my-job-has-taught-me-1.4184711

Braithwaite, J. (1982). Enforced self-regulation: A new strategy for corporate crime control. *Michigan law review*, *80*(7), 1466-1507

Braithwaite, J. (1984). *Corporate crime in the pharmaceutical industry (Routledge Revivals)*. Routledge.

Braithwaite, J. (1985). *To punish or persuade: Enforcement of coal mine safety*. SUNY Press.

Braithwaite, J. (1990). Convergence in models of regulatory strategy. *Current Issues in Criminal Justice*, *2*(1), 59-65.

Braithwaite, J. (2002a). Rules and principles: A theory of legal certainty. *Austl. J. Leg. Phil.*, *27*, 47.

Braithwaite, J. (2002b). *Restorative justice & responsive regulation*. Oxford University press on demand.

Braithwaite, J. (2011). The essence of responsive regulation. *UBCL Rev.*, *44*, 475.

Braithwaite, J., & Makkai, T. (1991). Testing an expected utility model of corporate deterrence. *Law & Soc'y Rev.*, *25*, 7.

Brandt, L., and Dean, G. (2021) Gab, a social-networking site popular among the far right, seems to be capitalizing on Twitter bans and Parler being forced offline. It says it's gaining 10,000 new users an hour. *Business Insider.* Accessed 31 January 2021 via https://www.businessinsider.com/gab-reports-growth-in-the-midst-of-twitter-bans-2021-1?r=US&IR=T

Brems, E., & Lavrysen, L. (2013). Procedural justice in human rights adjudication: The European Court of Human Rights. *Human Rights Quarterly*, 176-200.

Bright, J., Marchal, N., Ganesh, B., & Rudinac, S. (2020). Echo Chambers Exist!(But They're Full of Opposing Views). *arXiv preprint arXiv:2001.11461*.

Brink, D.O. (2008) "Mill's Liberal Principles and Freedom of Expression", In C.L. Ten (2008) *Mill's "On Liberty": A Critical Guide*. New York: Cambridge University Press.

Brock, D. W. (1988). Paternalism and autonomy.

Broughton, S.M, and Jacques, S. (2019) HM Government's Online Harms White Paper Consultation response from the Centre for Competition Policy University of East Anglia.

Brown, D. (2020) Draft EU Regulation on 'Terrorist Content' Online Threatens Rights: Proposals Risk Undermining Free Speech, Judicial Authority. *Human Rights Watch.* Accessed 20 November 2020 via https://www.hrw.org/news/2020/11/16/draft-eu-regulation-terrorist-content-online-threatens-rights?utm_source=Tech+Against+Terrorism&utm_campaign=0cef7d7c5d-EMAIL_CAMPAIGN_2019_03_24_07_51_COPY_01&utm_medium=email&utm_term=0_cb464fdb7d-0cef7d7c5d-162586859

Bruhn, A. (2006). The inspector's dilemma under regulated self-regulation. *Policy and Practice in Health and Safety*, *4*(2), 3-23.

Bruns, A. (2017). Echo chamber? What echo chamber? Reviewing the evidence.

Burgess, A. (2001). Flattering consumption: Creating a Europe of the consumer. *Journal of Consumer Culture*, *1*(1), 93-117.

Bychawska-Siniarska, D. (2017). *Protecting the right to freedom of expression under the European convention on human rights: A handbook for legal practitioners*. Council of Europe.

Carnino, A. (2000, September). Management of safety, safety culture and self assessment. In *International Conference Nuclear Energy in Central Europe* (pp. 11-14).

Carter, J. (2016) Growing our Trusted Flagger program into YouTube Heroes. *YouTube Blog*. Accessed November 2020 via https://blog.youtube/news-and-events/growing-our-trusted-flagger-program/#:~:text=Our%20Trusted%20Flaggers'%20results%20around,accurate%20than%20the%20average%20flagger.

Cartes, P. (2016) Announcing the Twitter Trust & Safety Council. *Twitter Blog*. Accessed 23 February 2021 via https://blog.twitter.com/en_us/a/2016/announcing-the-twitter-trust-safety-council.html

CCPR/C/GC/34. Available at: https://www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf
Chalmers, J., & Leverick, F. (2018). Criminal law in the shadows: creating offences in delegated legislation. *Legal Studies*, *38*(2), 221-241.

Chemerinsky, E. (2018) *The First Amendment*. Wolters Kluwer: New York

Christchurch Call (2019) The Call. Accessed 18 December 2020 via https://www.christchurchcall.com/call.html

Christoffel, T., and Christoffel, K.K. (1989) The Consumer Product Safety Commission's Opposition to Consumer Product Safety: Lessons for Public Health Advocates. *Public Health and the Law.*

Clarke, J., Smith, N., & Vidler, E. (2005). Consumerism and the reform of public services: inequalities and instabilities. *Social policy review*, *17*, 167-182.

Clegg, N. (2019) Charting a course for an oversight board for content decisions. *Facebook Newsroom.* Accessed June 2019 via https://newsroom.fb.com/news/2019/01/oversight-board/

Clifford, B., and Powell, H. (2019a) De-platforming and the online extremist's dilemma. *Lawfare Blog.* Accessed 8 December 2019 via https://www.lawfareblog.com/de-platforming-and-online-extremists-dilemma

Clifford, B., and Powell, H. (2019b) Encrypted Extremism: Inside the English-speaking Islamic State ecosystem on Telegram. *Program on Extremism, The George Washington University.*

Clinard, M., & Yeager, P. (2011). *Corporate crime* (Vol. 1). Transaction Publishers.

Cohen, F. (2002). Terrorism and cyberspace. *Network Security*, *2002*(5), 17-19.

Cohenalmagor, R. (1993). Harm principle, offence principle, and the Skokie affair. *Political Studies*, *41*(3), 453-470.

Cohenalmagor, R. (2001). Harm Principle, Offence Principle, and Hate Speech. In *Speech, Media and Ethics* (pp. 3-23). Palgrave Macmillan, London.

Colley, T., & Moore, M. (2020). The challenges of studying 4chan and the Alt-Right:'Come on in the water's fine'. *New Media & Society*, 1461444820948803.

Conway, M. (2006). Terrorism and the Internet: New media—New threat?. *Parliamentary Affairs*, *59*(2), 283-298.

Conway, M. (2019) Cited in Sahinkaya, E. (2019) Europol goes after IS propaganda online. *VOA News*. Accessed 28 September 2020 via https://www.voanews.com/extremism-watch/europol-goes-after-propaganda-online

Conway, M. (2020). Routing the Extreme Right: Challenges for Social Media Platforms. *The RUSI Journal*, 1-6.

Conway, M., & McInerney, L. (2008, December). Jihadi video and auto-radicalisation: Evidence from an exploratory YouTube study. In *European Conference on Intelligence and Security Informatics* (pp. 108-118). Springer, Berlin, Heidelberg.

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2017) Disrupting Daesh: measuring takedown of online terrorist material and its impacts. VoxPol Network of Excellence for Research in Violent Online Political Extremism.

Cosh, A., & Hughes, A. (1996). The Changing State of British Enterprise. *ESRC Centre for Business Research, Cambridge*.

Corera, G. (2017) Facebook reveals measures to remove terrorist content. *BBC.* Accessed 17 February 2021 via https://www.bbc.co.uk/news/technology-40290258

Costello, R. (2019) Twitter builds partnership with UNESCO on media and information literacy. Twitter Blog. Accessed 27 December 2019 via https://blog.twitter.com/en_us/topics/company/2019/twitter-launches-new-media-literacy-handbook-for-schools.html

Council of Europe/European Court of Human Rights. (2018). Guide on Article 8 of the European convention on human rights: Right to respect for private and family life, home and correspondence.

Council of Europe/European Court of Human Rights. (2019a). Guide on Article 9 of the European convention on human rights: Freedom of thought, conscience and religion.

Council of Europe/European Court of Human Rights. (2019b). Guide on Article 17 of the European convention on human rights: Prohibition of abuse of rights.

Council of Europe/European Court of Human Rights. (2020). Factsheet – Hate speech.

Counter Extremism Project (2017) Terrorists on Telegram. *Counter Extremism Project*. Accessed November 2017 via https://www.counterextremism.com/terrorists-on-telegram

Counter Terrorism Policing (2018) Together, we're tackling online terrorism. Accessed 17 December 2020 via https://www.counterterrorism.police.uk/together-were-tackling-online-terrorism/

Crandall, R. W., & Graham, J. D. (1984). Automobile safety regulation and offsetting behaviour: Some new empirical estimates. *The American Economic Review*, *74*(2), 328-331.

Cuthbertson, A. (2021) WhatsApp rival Telegram sees new users rise 500% amid exodus from Facebook-owned app. *The Independent.* Accessed 31 January 2021 via https://www.independent.co.uk/life-style/gadgets-and-tech/telegram-whatsapp-delete-facebook-privacy-b1786161.html

Dardis, R. (1988). Risk regulation and consumer welfare. *Journal of consumer affairs*, *22*(2), 303-318.

Darme, Z.M., Miller, M., and Steeves, K. (2019) Global Feedback & Input on the Facebook Oversight Board for Content Decisions. Accessed July 2019 via https://fbnewsroomus.files.wordpress.com/2019/06/oversight-board-consultation-report-2.pdf

Daskal, J., and Kennedy-Mayo, D. (2020) Budapest Convention: What is it and how is it being updated? *Cross Border Data Forum*. Accessed 13 April 2021 via https://www.crossborderdataforum.org/budapest-convention-what-is-it-and-how-is-it-being-updated/?cn-reloaded=1

De Sadeleer, N. (2002). *Environmental principles: from political slogans to legal rules*. Oxford University Press on Demand.

Defence, A. (2020) Mapping extremist communities: A social network analysis approach. Accessed 18 October 2020 via https://www.voxpol.eu/download/report/web_stratcom_coe_mapping_extremist_strategies_31.03.2020_v2.pdf

Den Hertog, J. A. (2010). Review of economic theories of regulation. *Discussion Paper Series/Tjalling C. Koopmans Research Institute*, *10*(18).

Denes, B. (2020) Open letter on behalf of civil society groups regarding the proposal for a Regulation on Terrorist Content Online. Accessed 20 November 2020 via file:///D:/PhD/Lit%20review%20research/Reading%20my%20lit%20searches/Second%20lit%20review%20search/Law%20chapter/European%20Commission/Letters%20against%20dissemination%20of%20terrorist%20content%20online%20proposal/TERREG_Openletter_Liberties.pdf

Department of Homeland Security (2010) DHS terrorist use of social networking Facebook case study. *Public Intelligence*. Accessed 27 September 2020 via https://publicintelligence.net/ufouoles-dhs-terrorist-use-of-social-networking-facebook-case-study/

Dorsey, G. L. (1953). The necessity of authority to freedom. *Charles W. Hendel," Freedom and Authority as Functions of Civilization*.

Douek, E. (2019). Australia's' Abhorrent Violent Material. *Australian Law Journal*, *2020*.

Douek, E. (2019). Facebook's Oversight Board: Move Fast with Stable Infrastructure and Humility. *NCJL & Tech.*, *21*, 1.

Douek, E. (2020). The Rise of Content Cartels. *Knight First Amendment Institute at Columbia*.

Driesen, D. M. (1997). The Societal Cost of Environmental Regulation: Beyond Administrative Cost-Benefit Analysis. *Ecology LQ*, *24*, 545.

Duarte, N., Llanso, E., & Loup, A. (2017). Mixed Messages?. *Center for Democracy and Technology*.

Dworkin, G. (2015) The nature of autonomy. *Nordic Journal of Studies in Educational Policy*, pp.7-14

Dyzenhaus, D. (1992). John Stuart Mill and the harm of pornography. *Ethics*, *102*(3), 534-551.

Echikson, W., & Knodt, O. (2018). Germany's NetzDG: A key test for combatting online hate. CEPS Research Reports No. 2018/09, November 2018

Emerson, T (1970) The System of Free Expression. New York: Vintage Books
Endicott, T. Due Process. In T, Endicott (2018) *Administrative Law*. Oxford University Press

Engstrom, E., and Feamster, N. (2017) The limits of filtering: A look at the functionality and shortcomings of content detection tools. Accessed 19 November 2019 via https://perma.cc/UV5H-89SK

European Agency for Safety and Health at Work (2020) Psychosocial risks and stress at work. Accessed 28 August 2020 via https://osha.europa.eu/en/themes/psychosocial-risks-and-stress#:~:text=Psychosocial%20risks%20and%20work%2Drelated,in%20occupational%20safety%20and%20health.&text=Around%20half%20of%20European%20workers,of%20all%20lost%20working%20days.

European Commission (2016) Remarks by Commissioners Avramopoulos and King at the press conference ahead of the EU Internet Forum. Accessed 2 December 2019 via https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_16_4329

European Commission (2017) Fighting Terrorism Online: Internet Forum pushes for automatic detection of terrorist propaganda. Accessed 2 December 2019 via https://ec.europa.eu/commission/presscorner/detail/en/IP_17_5105

European Commission (2018) Code of Conduct on countering illegal hate speech online: Questions and answers on the fourth evaluation. Accessed 19 November 2019 via https://ec.europa.eu/commission/presscorner/detail/en/MEMO_19_806

European Commission (2019a) Security Union: After Christchurch, EU Internet Forum discusses operational measures to tackle terrorist content online. Accessed 2 December 2019 via https://ec.europa.eu/home-affairs/news/20190506_security-union-after-christchurch-eu-internet-forum-discusses-operational-measures-tackle-terrorist-content-online_en

European Commission (2019b) Fighting terrorism online: EU Internet Forum committed to an EU-wide crisis protocol. Accessed 4 March 2021 via https://ec.europa.eu/commission/presscorner/detail/en/ip_19_6009

European Parliament legislative resolution of 17 April 2019 on the proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (COM(2018)0640 –C8-0405/2018 –2018/0331(COD)) ("Terrorist Content Regulation"), Recitals 9 and 13, https://perma.cc/2MKH-WG8V

European Union (2016) Code of Conduct on Countering Illegal Hate Speech Online, https://perma.cc/74LC-3CJB.

Europol (2015) Europol's internet referral unit to combat terrorist and violent extremist propaganda. Accessed 3 December 2019 via https://www.europol.europa.eu/newsroom/news/europol%E2%80%99s-internet-referral-unit-to-combat-terrorist-and-violent-extremist-propaganda

Europol (2016) 211 Terrorist attacks caried out in EU Member States in 2015, new Europol report reveals. Accessed 23 March 2021 via https://www.europol.europa.eu/newsroom/news/211-terrorist-attacks-carried-out-in-eu-member-states-in-2015-new-europol-report-reveals

Europol (2017) EU Internet Referral Unit Transparency Report 2017. Accessed 3 December 2019 via file:///C:/Users/Amy-Louise.Watkin/Downloads/eu_iru_transparency_report.pdf

Europol (2019a) Europol and Telegram Take on Terrorist Propaganda Online. *Europol Newsroom.* Accessed 11 October 2020 via https://www.europol.europa.eu/newsroom/news/europol-and-telegram-take-terrorist-propaganda-online

Europol (2019b) EU Internet Referral Unit – EU IRU. Accessed 3 December 2019 via https://www.europol.europa.eu/about-europol/eu-internet-referal-unit-eu-iru

Europol (2019c) Europol and Telegram take on terrorist propaganda online. *Europol Press Release.* Accessed 02/02/2021 via https://www.europol.europa.eu/newsroom/news/europol-and-telegram-take-terrorist-propaganda-online

Eyre, S., Heims, E., Koop, C., Lodge, M., Stirton L., and Vibert, F. (2016) Regulatory agencies under challenge. *Centre for analysis of risk and regulation*. Accessed 30 October 2020 via https://www.lse.ac.uk/accounting/assets/CARR/documents/D-P/Disspaper81.pdf

Facebook (2016). Partnering to Help Curb Spread of Online Terrorist Content. *Facebook*. Accessed September 2017 via https://newsroom.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/

Facebook (2017) Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism. *Facebook*. Accessed September 2017 via https://newsroom.fb.com/news/2017/06/global-internet-forum-to-counter-terrorism/

Facebook (2018a) NetzDG-Transparenzbericht, https://fbnewsroomus.files.wordpress.com/2018/07/face-book_netzdg_juli_2018_deutsch-

1.pdf, Cited in Schmitz, S., & Berndt, C. M. (2018). The German Act on Improving Law Enforcement on Social Networks (NetzDG): A Blunt Sword?. *Available at SSRN 3306964*

Facebook (2018b) Hard Questions: The Line Between Hate and Debate. *Facebook Newsroom.* Accessed August 2018 via https://newsroom.fb.com/news/2018/08/the-line-between-hate-and-debate/

Facebook (2019a) Combating hate and extremism. *Facebook Newsroom.* Accessed September 2019 via https://newsroom.fb.com/news/2019/09/combating-hate-and-extremism/

Facebook (2019b) Standing Against Hate, *Facebook Newsroom.* Accessed June 2019 via https://newsroom.fb.com/news/2019/03/standing-against-hate/

Facebook (2019c) Hard Questions: What is Facebook doing to address the challenges it faces? Facebook Newsroom. Accessed June 2019 via https://newsroom.fb.com/news/2019/02/addressing-challenges/

Facebook Investor Relations (2019) FAQs. Accessed 17 February 2021 via https://investor.fb.com/resources/default.aspx#:~:text=Founded%20in%202004%2C%20Facebook's%20mission,express%20what%20matters%20to%20them.

Fairman, R., & Yapp, C. (2005). Enforced self-regulation, prescription, and conceptions of compliance within small businesses: The impact of enforcement. *Law & Policy*, *27*(4), 491-519.

Feinberg, J. (1982). Autonomy, sovereignty, and privacy: Moral ideals in the constitution. *Notre Dame L. Rev.*, *58*, 445.

Feinberg, J. (1984). *Harm to others* (Vol. 1). Oxford University Press on Demand.

Feinberg, J. (1988). *Offense to others* (Vol. 2). Oxford University Press on Demand.

Feinberg, J. (1989). *The moral limits of the criminal law: volume 3: harm to self*. Oxford University Press on Demand.

Feintuck, M. (2010) Regulatory Rationales Beyond the Economic: In Search of the Public Interest. In Baldwin, R., Cave, M., & Lodge, M. (Eds.). (2010). *The Oxford handbook of regulation*. Oxford University Press.

Fineman, S., & Sturdy, A. (1999). The emotions of control: A qualitative exploration of environmental regulation. *Human Relations*, *52*(5), 631-663.

Fisher, A., Prucha, N., and Winterbotham, E. (2019). Mapping the Jihadist Information Ecosystem: Towards the next generation of disruption capability. *Global Research Network on Terrorism and Technology: Paper No. 6*

Fishman, B. (2019) Crossroads: Counter-Terrorism and the Internet. *Texas National Security Review* 2(2)

Fisse, B., & Braithwaite, J. (1984). Sanctions against corporations: dissolving the monopoly of fines. *Business Regulation in Australia*, *129*, 146.

Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in practice*, *17*(4-5), 663-671.

Frampton, M., Fisher, A., Prucha, N., & Petraeus, D. H. (2017). *The new Netwar: Countering extremism online*. Policy Exchange.

Freedman, D. (2012). Outsourcing internet regulation. *Misunderstanding the internet*, 95-120.

Furnell, S. M., & Warren, M. J. (1999). Computer hacking and cyber terrorism: The real threats in the new millennium?. *Computers & Security*, *18*(1), 28-34.

Fussell, S. (2019) Why the New Zealand shooting video keeps circulating. *The Atlantic*. Accessed 19 November 2019 via https://www.theatlantic.com/technology/archive/2019/03/facebook-youtube-new-zealand-tragedy-video/585418/

Gab (2019) Gab About. Accessed July 2019 via https://gab.com/about

Gab Homepage (2019) Gab Homepage. Accessed July 2019 via https://gab.com/

Gab (2021a) Gab AI Inc Terms of Service. Accessed 31 January 2021 via https://gab.com/about/tos

Gab (2021) Gab.com. Accessed 19 February 2021 via https://gab.com/#:~:text=A%20social%20network%20that%20champions,free%20flow%20of%20information%20online.

Gadde, V., and Harvey, D. (2018) Creating new policies together. *Twitter Blog*. Accessed 9 March 2020 via https://blog.twitter.com/en_us/topics/company/2018/Creating-new-policies-together.html

Gallup and Knight Foundation (2020) Free expression, harmful speech and censorship in a digital world. *Knight Foundation.* Accessed 21 June 2020 via https://knightfoundation.org/wp-content/uploads/2020/06/KnightFoundation_Panel6-Techlash2_rprt_061220-v2_es-1.pdf

Gander, P., Hartley, L., Powell, D., Cabon, P., Hitchcock, E., Mills, A., & Popkin, S. (2011). Fatigue risk management: Organisational factors at the regulatory and industry/company level. *Accident Analysis & Prevention*, *43*(2), 573-590.

Gardbaum, S. (1996). Liberalism, autonomy, and moral conflict. *Stanford Law Review*, 385-417.

Garrett, R. K. (2017). The “echo chamber” distraction: Disinformation campaigns are the problem, not audience fragmentation.

Gaudette, T., Scrivens, R., Davies, G., & Frank, R. (2020). Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society*, 1461444820958123.

Geroski, P., & Machin, S. (1992). Do innovating firms outperform non-innovators?. *Business Strategy Review*, *3*(2), 79-90.

GIFCT (2019) Global Internet Forum to Counter Terrorism: An update on our progress. *YouTube Blog*. Accessed July 2019 via https://youtube.googleblog.com/2019/07/global-internet-forum-to-counter.html

GIFCT (2020) Global Internet Forum to Counter Terrorism. Accessed 27 May 2020 via https://www.gifct.org/

GIFCT (2021) Global Internet Forum to Counter Terrorism. Accessed 15 January 2021 via https://www.gifct.org/

Gil, N., Miozzo, M., & Massini, S. (2012). The innovation potential of new infrastructure development: An empirical study of Heathrow airport's T5 project. Research Policy, 41, 452–466

Gilbert, D. (2019a) Bestiality, stabbings and child porn: Why Facebook moderators are suing the company for trauma. *Vice*. Accessed 10 March 2020 via https://www.vice.com/en_uk/article/a35xk5/bestiality-stabbings-and-child-porn-why-facebook-moderators-are-suing-the-company-for-trauma

Gilbert, D. (2019b) Here's how big far right social network Gab has actually become. *Vice*. Accessed 18 August 2020 via https://www.vice.com/en_uk/article/pa7dwg/heres-how-big-far-right-social-network-gab-has-actually-become

Gilbert, D.T. and Wilson, T.D. (2000) Miswanting: Some Problems in the Forecasting of Future Affective States. In Forgas, J.P. (2000) *Feeling and Thinking: The Role of Affect in Social Cognition.* Cambridge: Cambridge University Press

Gill, P., & Corner, E. (2015). Lone actor terrorist use of the Internet and behavioural correlates. In *Terrorism Online* (pp. 47-65). Routledge.

Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., & Horgan, J. (2017). Terrorist use of the Internet by the numbers: Quantifying behaviors, patterns, and processes. *Criminology & Public Policy*, *16*(1), 99-117.

Gill, P., Horgan, J., & Deckert, P. (2014). Bombing alone: Tracing the motivations and antecedent behaviors of lone-actor terrorists. *Journal of forensic sciences*, *59*(2), 425-435.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., ... & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4), Article-number.

Goh, A. L. (2005). Promoting innovation in aid of industrial development: The Singaporean experience. International Journal of Public Sector Management, 18(3), 216–240

Gray, J. (2013). *Mill on liberty: a defence*. Routledge.

Gunningham, N. (2002). Regulating small and medium sized enterprises. *Journal of Environmental Law*, 3-32.

Gunningham, N. (2007). Designing OHS standards: process, safety case and best practice. *Policy and practice in health and safety*, *5*(2), 3-24.

Gunningham, N. (2010). Enforcement and compliance strategies. *The Oxford handbook of regulation*, *120*, 131-35.

Gunningham, N. (2011). Strategizing compliance and enforcement: responsive regulation and beyond. *Explaining compliance: business responses to regulation*, 199-221.

Gunningham, N., & Sinclair, D. (2017). Smart regulation. *Regulatory theory: Foundations and applications*, 133-148.

Gunningham, N., Kagan, R. A., & Thornton, D. (2003). *Shades of green: business, regulation, and environment*. Stanford University Press.

Guynn, J (2019) Facebook while black: Users call it getting'Zucked,' say talking about racism is censored as hate speech. *USA Today.* Accessed 26 June 2020 via https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

Hadley, A. and Berntsson, J. (2020) Regulation: Concerns about effectiveness and impact on smaller tech platforms. *VOX-Pol Blog*. Accessed 3 July 2020 via https://www.voxpol.eu/the-eus-terrorist-content-regulation-concerns-about-effectiveness-and-impact-on-smaller-tech-platforms/

Hale, A.R., and Hovden, J. Management and culture: the third age of safety. In A-M. Feyer, A, Williamson, (1998) *Occupational injury: risk prevention and intervention*. London: Taylor & Francis, p.129-66

Haq, M., and Forks, B. (2019) Giving you more control over your conversations. *Twitter Blog*. Accessed 13 February 2020 via https://blog.twitter.com/en_us/topics/product/2019/morecontrolofconversation.html

Hardinghaus, A., Kimmich, R., and Schonhofen, S. (2020) German government introduces new bill to amend Germany's Hate Speech Act, establishing new requirements for social networks and video-sharing platforms. *Technology Law Dispatch.* Accessed 30 October 2020 via https://www.technologylawdispatch.com/2020/04/regulatory/german-government-introduces-new-bill-to-amend-germanys-hate-speech-act-establishing-new-requirements-for-social-networks-and-video-sharing-platforms/

Hardy, K. (2020). Navigating radicalization concepts: A role for the harm principle. *Radicalization and Counter-Radicalization*.

Harris, B. (2019) Getting input on an oversight board. *Facebook Newsroom.* Accessed 15 January 2021 via https://about.fb.com/news/2019/04/input-on-an-oversight-board/

Hart, H.L.A. Causation and Sine Qua Non. In Hart, H.L.A., and Honore, T. (1985) *Causation in the Law*. OUP Oxford

Hawkins, K. (1984) *Environment and Enforcement*. Oxford: OUP

Hayden, M. (2019) A problem of epic proportions. *Southern Law Poverty Center*. Accessed 18 October 2020 via https://www.splcenter.org/hatewatch/2019/01/11/problem-epik-proportions

Hayek, F.A. (1960) *The Constitution of Liberty*. Chicago: University of Chicago Press

Health and Safety Commission (1993) ACSNI Study Group on Human Factors. 3rd Report: Organizing for Safety. London (UK): HS

Heldt, A. (2020) Germany is amending its online speech act NetzDG…but not only that. *Policy Review*. Accessed 6 November 2020 via

https://policyreview.info/articles/news/germany-amending-its-online-speech-act-netzdg-not-only/1464

Hidvegi, F. (2019) Open letter to EU Parliament on the Terrorism Database. *Access Now.* Accessed 24 November 2020 via https://www.accessnow.org/open-letter-to-eu-parliament-on-the-terrorism-database/

Hilf, M. (2001). Power, rules and principles-which orientation for WTO/GATT law?. *Journal of International Economic Law*, *4*(1), 111-130

HM Government (2019) UK Online Harms White Paper. Accessed 30 October 2020 via https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

HM Government (2020) Online harms white paper: Full government response to the consultation. Updated 15 December 2020. Accessed 21 December 2020 via https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response

Home Office (2020) Online Harms White Paper – Initial consultation response (2020) Accessed 13 August 2020 via https://www.gov.uk/government/consultations/online-harms-white-paper/public-feedback/online-harms-white-paper-initial-consultation-response

Honneland, G. (2000). *Coercive and Discursive Compliance Mechanisms in the Management of Natural Resources: A Case Study from the Barents Sea Fisheries* (Vol. 23). Springer Science & Business Media.

Hope, C. (2017) Google, Facebook and Twitter told to take down terror content within two hours or face fines. *The Telegraph*. Accessed 17 February 2021 via https://www.telegraph.co.uk/news/2017/09/19/google-facebook-twitter-told-take-terror-content-within-two/

House of Lords (2019) Regulating in a digital World. *Select Committee on Communications: 2nd Report Session of 2017-2019*.

Husak, D. N. (1981). Paternalism and autonomy. *Philosophy & Public Affairs*, 27-46.

Hutter, B. M. (2001). Is enforced self-regulation a form of risk taking?: The case of railway health and safety. *International Journal of the Sociology of Law*, *29*(4), 379-400.

Hutter, B.M. (1997) *Compliance: Regulation and Environment*. Oxford: OUP

Information Commissioner's Office (2019) Guide to the General Data Protection Regulation (GDPR). Accessed 22 June 2020 via https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf

International Atomic Energy Agency (2006) Application of the management system for facilities and activities. Safety Guide Series No. GS-G-3.1. Accessed 27 May 2020 via https://www-pub.iaea.org/mtcd/publications/pdf/pub1253_web.pdf

International Labour Organisation (2006) Promotional Framework for Occupational Health and Safety Convention (No. 187)

International Labour Organisation (ILO). (2009) Information on decent work and a health and safety culture [Internet]. Geneva (Switzerland): Office

Isler, C. (2001). *The Right to Free Speech*. The Rosen Publishing Group.

Iqbal, M. (2021) Telegram revenue and usage statistics. *Business of Apps.* Accessed 18 February 2021 via https://www.businessofapps.com/data/telegram-statistics/#:~:text=Telegram%20does%20not%20generate%20revenue,privately%20generated%20by%20Pavel%20Durov.

Jackson, S. (2019) The double-edged sword of banning extremists from social media.

Jee, C. (2020) Facebook needs 30,000 of its own content moderators, says a new report. Technology Review. Accessed 22 November 2020 via https://www.technologyreview.com/2020/06/08/1002894/facebook-needs-30000-of-its-own-content-moderators-says-a-new-report/

Johnston, M. (2021) Facebook sells most of its revenue from selling advertising space. *Investopedia*. Accessed 18 February 2021 via https://www.investopedia.com/ask/answers/120114/how-does-facebook-fb-make-money.asp#:~:text=Facebook%20sells%20ads%20on%20social,by%20the%20COVID%2D19%20pandemic.

Johnstone, R. (2003). *From Fiction to Fact-Rethinking OHS Enforcement', National Research Centre for Occupational Health and Safety Regulation* (Vol. 11). Working Paper 11.

Johnstone, R., Quinlan, M., & McNamara, M. (2011). OHS inspectors and psychosocial risk factors: Evidence from Australia. *Safety Science*, *49*(4), 547-557.

Jordan, A., & O'Riordan, T. (1995). The precautionary principle in UK environmental law and policy. In *UK Environmental Policy in the 1990s* (pp. 57-84). Palgrave Macmillan, London.

Jourová, V. (2016) Code of Conduct on countering illegal hate speech online: First results on implementation. European Commission.

Jugendschutz.net (2017) Löschung rechtswidriger Hassbeiträge bei Facebook, YouTube und Twitter, Cited in Schmitz, S., & Berndt, C. M. (2018). The German Act on Improving Law Enforcement on Social Networks (NetzDG): A Blunt Sword?. *Available at SSRN 3306964*.

Kant, I. (1981). Groundwork for Metaphysics of Morals (J. W Ellington, trans.). *Hackett, Indiana (original work published 1785)*.

Kantrowitz, A. (2018) YouTube is assembling new teams to spot inappropriate content early. *Buzzfeed News*. Accessed 10 March 2020 via https://www.buzzfeednews.com/article/alexkantrowitz/youtube-intelligence-desk-will-spot-inappropriate-content

Kaye, D. (2018) "Report of the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression," April 6, 2018, https://perma.cc/9XWD-7JQU.

Kaye, R. P. (2006). Regulated (self-) regulation: A new paradigm for controlling the professions?. *Public Policy and Administration*, *21*(3), 105-119.

Kayode-Adedeji, T., Oyero, O., & Aririguzoh, S. (2019). Dataset on Online mass media engagements on YouTube for terrorism related discussions. *Data in brief*, *23*, 103581.

Kayyali, D., and Althaibani, R. (2017) Vital human rights evidence in Syria is disappearing from YouTube. *VOX-Pol Blog.* Accessed 4 October 2020 via https://www.voxpol.eu/vital-human-rights-evidence-syria-disappearing-youtube/

Keller, D. (2018) Observations on Speech, Danger and Money. *Aegis Series Paper No. 1807*

Kennedy, C. (1985). Criminal sentences for corporations: Alternative fining mechanisms. *Calif. L. Rev.*, *73*, 443.

Khan, I., and Ní Aoláin, F. (2020) Mandates of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression and the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism. Accessed 20 November 2020 via file:///D:/PhD/Lit%20review%20research/Reading%20my%20lit%20searches/Second%20lit%20review%20search/Law%20chapter/European%20Commission/Letters%20against%20dissemination%20of%20terrorist%20content%20online%20proposal/DownLoadPublicCommunicationFile.pdf

Kidron, B., Evans, A. and Afia, J. 2018. 'Disrupted Childhood: The Cost of Persuasive Design'. Retrieved from: www.5rightsframework. com/static/5Rights-Disrupted-Childhood.pdf

Kim, Y., Park, J., & Park, M. (2016). Creating a culture of prevention in occupational safety and health practice. *Safety and health at work*, *7*(2), 89-96.

Klausen, J., Barbieri, E. T., Reichlin-Melnick, A., & Zelin, A. Y. (2012). The YouTube Jihadists: A social network analysis of Al-Muhajiroun's propaganda campaign. *Perspectives on Terrorism*, *6*(1), 36-53.

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, *131*, 1598.

Klonick, K. (2019). The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression. *Yale LJ*, *129*, 2418.

Klonick, K. (2020) The Facebook oversight board: creating an independent institution to adjudicate online free expression. *The Yale Law Journal*.

Koebler, J. (2020) How is Twitter going to moderate these voice recordings? *Vice*. Accessed 2 July 2020 via https://www.vice.com/en_us/article/xg8w8j/how-is-twitter-going-to-moderate-these-voice-recordings

Kosti, N., Levi-Faur, D., & Mor, G. (2019). Legislation and regulation: three analytical distinctions.

Krasenberg, J. (2019) EU Policy: Preventing the dissemination of terrorist content online. *George Washington Program on Extremism: Legal Perspectives on Tech Series.* Accessed 26

November 2019 via https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EU%20Policy%20-%20Preventing%20the%20Dissemination%20of%20Terrorist%20Content%20Online.pdf

Krawiec, K. D. (2004). Organisational misconduct: Beyond the principal-agent model. *Fla. St. UL Rev.*, *32*, 571.

Kreimer, S. (2006) *supra* note 6, at 28. Cited in Ardia, D. S. (2009). Free speech savior or shield for scoundrels: An empirical study of intermediary immunity under Section 230 of the Communications Decency Act. *Loy. LAL Rev.*, *43*, 373.

Kriebel, D., Tickner, J., Epstein, P., Lemons, J., Levins, R., Loechler, E. L., ... & Stoto, M. (2001). The precautionary principle in environmental science. *Environmental health perspectives*, *109*(9), 871-876.

LaFree, G. (2017). Terrorism and the Internet. *Criminology & Pub. Pol'y*, *16*, 93.

Lakoff, S. (1990). Autonomy and liberal democracy. *The Review of politics*, *52*(3), 378-396.

Lave, L. (1981). The strategy of social regulation: Decision frameworks for policy. *Washington, DC*.

Leka, S., Jain, A., Iavicoli, S., & Di Tecco, C. (2015). An evaluation of the policy context on psychosocial risks and mental health in the workplace in the European Union: achievements, challenges, and the future. *BioMed research international*, *2015*.

Levi-Faur, D. (2011). Regulation and regulatory governance. *Handbook on the Politics of Regulation*, *1*(1), 1-25.

Lieberman, A. V. (2017). Terrorism, the internet, and propaganda: A deadly combination. *J. Nat'l Sec. L. & Pol'y*, *9*, 95.

Lin, Y. (2020) 10 Twitter statistics every marketer should know in 2021. *Oberlo*. Accessed 18 January 2021 via https://www.oberlo.co.uk/blog/twitter-statistics#:~:text=There%20are%20330%20million%20monthly,are%20between%2035%20and%2065.

Llanso, E. (2019) Platforms want centralized censorship. That should scare you. *Wired*. Accessed 14 September 2020 via https://www.wired.com/story/platforms-centralized-censorship/

Llanso, E. (2020a) Content moderation knowledge sharing shouldn't be a backdoor to cross-platform censorship. *Techdirt.* Accessed 30 August 2020 via https://www.techdirt.com/articles/20200820/08564545152/content-moderation-knowledge-sharing-shouldnt-be-backdoor-to-cross-platform-censorship.shtml:?utm_source=Tech+Against+Terrorism&utm_campaign=6ebef31ab9-EMAIL_CAMPAIGN_2019_03_24_07_51_COPY_01&utm_medium=email&utm_term=0_cb464fdb7d-6ebef31ab9-140969075

Llansó, E. J. (2020b). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, *7*(1), 2053951720920686.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, *92*(368), 805-824.Ogus, A. I. (1994). *Regulation: Legal form and economic theory*. Oxford: Clarendon Press

Lord Robens, A. (1972) Safety and Health at Work, vol 1. Report of the Committee. HMSO, London.

Lunt, P., Livingstone, S. and Kelay, T. (2005) 'Risk and Regulation in Financial Services and Communications', SCARR Working Paper Series

Mac Síthigh, D. (2019). The road to responsibilities: new attitudes towards Internet intermediaries. *Information & Communications Technology Law*, 1-21.

Macdonald, S. & Lorenzo-Dus, N. (2020). Intentional and Performative Persuasion: The Linguistic Basis for Criminalizing the (Direct and Indirect) Encouragement of Terrorism. *Criminal Law Forum*, *31*(4), 473-512. https://doi.org/10.1007/s10609-020-09405-x,

Macdonald, S. (2018). *Text, Cases and Materials on Criminal Law*. Pearson Higher Ed.

Macdonald, S., & Mair, D. (2015). Terrorism online: a new strategic environment. In *Terrorism Online* (pp. 22-46). Routledge.

Macdonald, S., Correia, S. G., & Watkin, A. L. (2019). Regulating terrorist content on social media: automation and the rule of law. *International Journal of Law in Context*, *15*(2), 183-197.

Macdonald, S., Grinnell, D., Kinzel, A., and Lorenzo-Dus, N. (2019) A study of outlinks contained in tweets mentioning 'Rumiyah'. *The Global Research Network on Terrorism and Technology*

Majone, G. (1997). The new European agencies: regulation by information. *Journal of European Public Policy*, *4*(2), 262-275.

März (2018) Leitlinien zur Festsetzung von Geldbußen im Bereich des Netzwerkdurchsetzungsgesetzes (NetzDG) vom 22., p. 3., Cited in Schmitz, S., & Berndt, C. M. (2018). The German Act on Improving Law Enforcement on Social Networks (NetzDG): A Blunt Sword?. *Available at SSRN 3306964*.

Mascini, P. (2013). Why was the enforcement pyramid so influential? And what price was paid?. *Regulation & Governance*, *7*(1), 48-60.

Matsakis, L. (2019) Twitter Trust and Safety Advisers Say They're Being Ignored. *Wired.* Accessed 17 February 2020 via https://www.wired.com/story/twitter-trust-and-safety-council-letter/

May, T (2018) Theresa May's Davos address in full, World Economic Forum, 25 January. Available at https://www.weforum. org/agenda/2018/01/theresa-may-davos-address/ (accessed 22 February 2019).

May, T. (1994). The concept of autonomy. *American Philosophical Quarterly*, *31*(2), 133-144.

Mayton, W. T. (1981). Toward a Theory of First Amendment Process: Injunctions of Speech Subsequent Punishment and the Costs of the Prior Restraint Doctrine. *Cornell L. Rev.*, *67*, 245.

McCulloch, J., & Pickering, S. (2011). Counter-terrorism law: preventing terrorism or pre-empting the future?. *Precedent*, (102), 4-9.

Meadowcroft, J. (2005). Environmental political economy, technological transitions and the state. *New Political Economy*, *10*(4), 479-498.

Meadowcroft, J. (2012). Greening the state. *Comparative environmental politics: Theory, practice, and prospects*, 63-87.

Meichtry, S., and Schechner, S. (2016) How Islamic State weaponized the chat app to direct attacks on the West. *The Wall Street Journal*. Accessed 2 October 2020 via https://www.wsj.com/articles/how-islamic-state-weaponized-the-chat-app-to-direct-attacks-on-the-west-1476955802

Meiklejohn, A. (2000). *Free speech and its relation to self-government*. The Lawbook Exchange, Ltd.

Mendel, T. (2017). A Guide to the Interpretation and Meaning of Article 10 of the European Convention on Human Rights. *Council of Europe Publications*, 1-91.

Meserole, C., and Byman, D. (2019) Terrorist definitions and designation lists: What technology companies need to know. *Global Research Network on Terrorism and Technology:* Paper No. 7

Microsoft (2016) Microsoft's approach to terrorist content online. Accessed 31 January 2021 via https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/

Microsoft (2017) Bing ads pilot program to counter radicalisation shows early progress, will be renewed for 2018. Accessed 23 February 2021 via https://blogs.microsoft.com/on-the-issues/2017/12/19/bing-ads-pilot-program-counter-radicalisation-shows-early-progress-will-renewed-2018/

Microsoft (2021a) Report terrorist content posted to a Microsoft consumer service. Accessed 31 January 2021 via https://www.microsoft.com/en-us/concern/terroristcontent

Microsoft (2021b) Report hate speech content posted to a Microsoft hosted consumer service. Accessed 31 January 2021 via https://www.microsoft.com/en-us/concern/hatespeech

Microsoft (2021c) About. Accessed 18 February 2021 via https://www.microsoft.com/en-gb/about/

Mill, J. S. (1863). On Liberty, (Boston: Ticknor and Fields).

Mill, J. S. (1998). *John Stuart Mill's Social and Political Thought: Freedom* (Vol. 2). Taylor & Francis.

Miller, C (2019) It's time to forcibly reform big tech. *Wired.* Accessed 26 August 2020 via https://www.wired.co.uk/article/tech-reform-regulation

Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2020, May). "And We Will Fight for Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 452-463).

Mohsin, M, (2020) 10 YouTube stats every marketer should know in 2021. *Oberlo.* Accessed 18 January 2021 via https://www.oberlo.co.uk/blog/youtube-statistics#:~:text=1.-,Monthly%20Active%20YouTube%20Users,have%20on%20a%20monthly%20basis.

Moore, A. D. (2010). *Privacy rights: Moral and legal foundations*. Penn State Press.

Morris, J. (2010) Free speech and online intermediaries in an age of terror recruitment. *Centre for Democracy and Technology – Statement of John B. Morris Jr. General Counsel.*

Murphy, L. (2019) Facebook's Civil Rights Audit – Progress Report. Accessed July 2019 via https://fbnewsroomus.files.wordpress.com/2019/06/civilrightaudit_final.pdf

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366-4383.

Nash, V. (2019). Revise and resubmit? Reviewing the 2019 Online Harms White Paper. *Journal of Media Law*, 1-10.

Needham, C. (2003) *Citizen-Consumers: New Labour's Marketplace Democracy*. London: Catalyst Forum

Neely, A., & Hii, J. (1998). Innovation and business performance: a literature review. *The Judge Institute of Management Studies, University of Cambridge*, 0-65.

Nelson, R. R., & Rosenberg, N. (1993). National innovation systems: A comparative analysis. New York: Oxford University Press

Network Enforcement Act (2017) Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act). Accessed 12 November 2019 via https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2

Ní Aoláin, F. (2018) UN human rights experts concerned about EU's online counter-terrorism proposal. *United Nations Human Rights Office of the High Commissioner.* Accessed 26 November 2019 via https://www.ohchr.org/en/NewsEvents/Pages/DisplayNews.aspx?NewsID=24013&LangID=E

Nouri, L., Lorenzo-Dus, N., and Watkin, A. (2019) Following the whack-a-mole: Britain First's Visual Strategy from Facebook to Gab. Global Research Network on Terrorism and Technology: Paper No. 4

Nouri, L., Lorenzo-Dus, N., and Watkin, A. (2021) Impacts of radical right groups' movements across social media platforms – a case study of changes to Britain First's visual strategy in its removal from Facebook to Gab. *Studies in Conflict and Terrorism*

NPCC (2018) Counter terrorism policing urging public to ACT against online extremism. Accessed 17 December 2020 via https://news.npcc.police.uk/releases/counter-terrorism-police-urge-public-to-act-against-online-extremism

Nyirenda, V., Chinniah, Y., Agard, B. (2015) Identifying Key Factors for an Occupational Health and Safety Risk estimation Tool in Small and Medium-size Enterprises. *IFAC-PapersOnLine*, 48, 541-546

Oboler, A. (2019) In Australia, a new law on livestreaming terror attacks doesn't take into account how the internet actually works. *Nieman Lab*. Accessed 18 December 2019 via https://www.niemanlab.org/2019/04/in-australia-a-new-law-on-livestreaming-terror-attacks-doesnt-take-into-account-how-the-internet-actually-works/

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, *33*(4), 459-478.

OECD, F. M. (1981). The measurement of scientific and technical activities.

Ogus, A. I. (1994). *Regulation: Legal form and economic theory*. Oxford: Clarendon Press

Ogus, A., & Carbonara, E. (2011). Self-regulation. In *Production of Legal Rules*. Edward Elgar Publishing.

Orland, L. (1979). Reflections on corporate crime: Law in search of theory and scholarship. *Am. Crim. L. Rev.*, *17*, 501.

Ottoni, R., Cunha, E., Magno, G., Bernardina, P., Meira Jr, W., & Almeida, V. (2018, May). Analysing right-wing youtube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 323-332).

Owler (2021a) YouTube. Accessed 18 January 2021 via https://www.owler.com/company/youtube

Owler (2021b) Telegram. Accessed 17 February via https://www.owler.com/company/telegram

Park, M. (2013). Hearing conservation culture change and noise control program implementation; a reflection on 40 years of hearing conservation history at a multinational company. *Graduate School of Public Health, Seoul National University: Doctoral thesis*.

Parker, C. (1997a) Converting the lawyers: The dynamics of competition and accountability reform. *The Australian and New Zealand journal of sociology*, *33*(1), 39-55

Parker, C. (1997b). The open corporation: Self-regulation and corporate citizenship. *Faculty of Law, University of Toronto.*

Parker, C. (2002). *The open corporation: Effective self-regulation and democracy*. Cambridge University Press.

Pasquale, F. (2015). *The black box society*. Harvard University Press.

Patanakul, P., & Pinto, J. K. (2014). Examining the roles of government policy on innovation. *The Journal of High Technology Management Research*, *25*(2), 97-107.

Payne, K. (2009). Winning the battle of ideas: Propaganda, ideology, and terror. *Studies in Conflict & Terrorism*, *32*(2), 109-128.

Pearson, E. (2018). Online as the new frontline: Affect, gender, and ISIS-take-down on social media. *Studies in Conflict & Terrorism*, *41*(11), 850-874.

Peltzman, S. (1975). The effects of automobile safety regulation. *Journal of political Economy*, *83*(4), 677-725.

Pendlebury, M. (2004). Individual Autonomy and Global Democracy. *Theoria*, *51*(103), 43-58.

Phadke, S., & Mitra, T. (2020, April). Many Faced Hate: A Cross Platform Study of Content Framing and Information Sharing by Online Hate Groups. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Pickles, N. (2019) Strengthening our trust and safety council. *Twitter Blog*. Accessed 27 December 2019 via https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-trust-and-safety-council.html

Plumb, R. (2019) An independent report on how we measure content moderation. *Facebook Newsroom.* Accessed June 2019 via https://newsroom.fb.com/news/2019/05/dtag-report/

Portaru, A. (2017). Freedom of expression online: The code of conduct on countering illegal hate speech online. *RRDE*, 77.

Pötzsch, S. (2010). Influence of perceived privacy and anonymity on forum users. *Mensch & Computer 2010: Interactive Cultures*

Prosser, T. (2010) Models of Economic and Social Regulation. In Oliver, D., Prosser, T., & Rawlings, R. (Eds.). (2010). *The regulatory state: constitutional implications*. Oxford University Press, USA.

Putra, M. D. (2016). New Media and Terrorism: Role of the Social Media to Countering Cyber Terrorism and Cyber Extremism for Effective Response. *Available at SSRN 2754370*.

Raffensperger, C., & Tickner, J. (1999). Introduction: to foresee and forestall. *Protecting public health and the environment: Implementing the Precautionary Principle*, 1-11.

Ramraj, V. V. (2004). Four models of due process. *International Journal of Constitutional Law*, *2*(3), 492-524.

Rauchfleisch, A., & Kaiser, J. (2020). The German Far-right on YouTube: An Analysis of User Overlap and User Comments. *Journal of Broadcasting & Electronic Media*, *64*(3), 373-396.

Raz, J. (1986) *The Morality of Freedom*. New York: Oxford University Press

Rea, S. Regulating occupational health and safety, Cited in Dewees, D., (1983) *The Regulation of Quality: Products, services, workplaces, and the environment.* Butterworths.

Redish, M. H. (2013). *The Adversary First Amendment: Free Expression and the Foundations of American Democracy*. Stanford University Press.

Reed, A., Whittaker, J., Votta, F., and Looney, S. (2019) Radical filter bubbles: Social media personalisation algorithms and extremist content. *Global Research Network on Terrorism and Technology: Paper No. 8.*

Regional Comprehensive Economic Partnership (1988) 12th Report – Best Practicable Environmental Option. Cited in Ogus, A. I. (1994). *Regulation: Legal form and economic theory*. Oxford: Clarendon Press

Reiff, N. (2020) How Twitter makes money. *Investopedia.* Accessed 18 February 2021 via https://www.investopedia.com/ask/answers/120114/how-does-twitter-twtr-make-money.asp

Riley, J. (1998). *Mill on liberty*. Psychology Press.

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Robins-Early, N. (2019) Like ISIS before them, far-right extremists are migrating to Telegram. *Huffington Post.* Accessed 7 December 2019 via https://www.huffpost.com/entry/telegram-far-right-isis-extremists-infowars_n_5cd59888e4b0705e47db36ef

Rogers, E. M. (1983). Diffusion of innovation (3rd ed.). New York: Free Press

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 0267323120922066.

Rosen, G. (2019) Community Standards Enforcement Report, November 2019 Edition. *Facebook Newsroom*. Accessed 22 December 2019 via https://about.fb.com/news/2019/11/community-standards-enforcement-report-nov-2019/

Rushkoff, D. (1995) Cyberia: Life in the trenches of hyperspace. New York: Harpercollins Publishers

Scanlon, T. (1972). A theory of freedom of expression. *Philosophy & Public Affairs*, 204-226.

Schmitz, S., & Berndt, C. M. (2018). The German Act on Improving Law Enforcement on Social Networks (NetzDG): A Blunt Sword?. *Available at SSRN 3306964*.

Schulz, W. (2018). Regulating intermediaries to protect privacy online–the case of the German NetzDG. *HIIG Discussion Paper Series.*

Scott, C. (2010) Standard-Setting in Regulatory Regimes. In Baldwin, R., Cave, M., & Lodge, M. (Eds.). (2010). *The Oxford handbook of regulation*. Oxford University Press.

Shaffer, G. C., & Pollack, M. A. (2009). Hard vs. soft law: Alternatives, complements, and antagonists in international governance. *Minn. L. Rev.*, *94*, 706.

Shapiro, S. A., & Rabinowitz, R. S. (1997). Punishment versus cooperation in regulatory enforcement: A case study of OSHA. *Admin. L. Rev.*, *49*, 713.

Shearing, C. D., & Ericson, R. V. (1991). Towards a figurative conception of action. *British Journal of Sociology*, *42*, 481-506.

Shehabat, A., & Mitew, T. (2018) Black-boxing the black flag: anonymous sharing platforms and ISIS content distribution tactics. *Perspectives on Terrorism*, *12*(1).

Silke, A. (Ed.). (2018). *Routledge Handbook of Terrorism and Counterterrorism*. Routledge.

Simester, A. P., & Von Hirsch, A. (2011). *Crimes, harms, and wrongs: On the principles of criminalisation*. Bloomsbury Publishing.

Simon, H. (1975) Administrative Behaviour. *New York: Free Press*. Cited in Ogus, A. I. (1994). *Regulation: Legal form and economic theory*. Oxford: Clarendon Press

Simpson, A. (2000) Regulating Transition: Australian Telecommunications Industry Liberalization, 1989-1999 (unpublished PhD thesis, Faculty of Law, University of Sydney). Cited in Braithwaite, J. (2002a). Rules and principles: A theory of legal certainty. *Austl. J. Leg. Phil.*, *27*, 47.

Sinclair, D. (1997). Self-regulation versus command and control? Beyond false dichotomies. *Law & Policy*, *19*(4), 529-559.

Smet, S. (2010). Freedom of Expression and the Right to Reputation: Human rights in Conflict. *Am. U. Int'l L. Rev.*, *26*, 183.

Smith, G. (2020) The online harms edifice takes shape. *Cyberleagle.* Accessed 20 December 2020 via https://www.cyberleagle.com/2020/12/the-online-harms-edifice-takes-shape.html?utm_source=Tech+Against+Terrorism&utm_campaign=7aa1cb3e55-EMAIL_CAMPAIGN_2019_03_24_07_51_COPY_01&utm_medium=email&utm_term=0_cb464fdb7d-7aa1cb3e55-162586859

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333-339.

Sonnemaker, T. (2020) As Facebook prepares to outsource tough content decisions to its new 'Supreme Court', experts warn it still operates within a dictatorship and can't legislate a better government. *Business Insider*. Accessed 17 February 2021 via https://www.businessinsider.com/facebook-oversight-board-is-a-supreme-court-in-zuckerberg-dictatorship-2020-9?r=US&IR=T

Splittgerber, A., and Detmering, F. (2017) Germany's new hate speech act in force: what social network providers need to do now. *Technology Law Dispatch. Accessed October 2019 via* https://www.technologylawdispatch.com/2017/10/social-mobile-analytics-cloud-smac/germanys-new-hate-speech-act-in-force-what-social-network-providers-need-to-do-now/?utm_content=bufferd5f9a&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer#page=1

Starr, C. (1969). Social benefit versus technological risk. *Science*, 1232-1238.

Statista (2020a) Number of monthly active Facebook users worldwide as of 3rd quarter 2020. Accessed 18 January 2021 via https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/#:~:text=With%20over%202.7%20billion%20monthly,the%20biggest%20social%20network%20worldwide.

Statista (2020b) Number of full-time Facebook employees from 2004-2019. Accessed 18 January 2021 via https://www.statista.com/statistics/273563/number-of-facebook-employees/#:~:text=The%20social%20network%20had%2044%2C942,Sandberg%20and%20CFO%20David%20Wehner.

Statista (2020c) Number of employees at the Microsoft Corporation from 2005 to 2020. Accessed 17 February 2021 via https://www.statista.com/statistics/273475/number-of-employees-at-the-microsoft-corporation-since-2005/#:~:text=The%20American%20technology%20company%20Microsoft,in%20full%2Dtime%20positions%20worldwide.

Stern, N. (2006) Stern Review. Accessed 27 August 2020 via http://mudancasclimaticas.cptec.inpe.br/~rmclima/pdfs/destaques/sternreview_report_complete.pdf

Stewart, H., and Elgot, J. (2018) May calls on social media giants to do more to tackle terrorism. *The Guardian*. Accessed 7 December 2019 via https://www.theguardian.com/business/2018/jan/24/theresa-may-calls-on-social-media-giants-to-do-more-to-tackle-terrorism

Stewart, R. (1983) Regulation in a Liberal State: The Role of Non-Commodity Values. *The Yale Law Journal, 92(8),* 1537-1590

Stigler, G. J. (1971). The theory of economic regulation. *The Bell journal of economics and management science*, 3-21.

Stone, C. (1976). Where the law ends: The social control of corporate behavior. *Business Horizons*, *19*(3), 84-87.

Sullivan, G. R., & Dennis, I. (Eds.). (2012). *Seeking security: Pre-empting the commission of criminal harms*. Bloomsbury Publishing.

Sunshine, J., & Tyler, T. R. (2003). The role of procedural justice and legitimacy in shaping public support for policing. *Law & society review*, *37*(3), 513-548.

Sunstein, C. R. (2005). *Laws of fear: Beyond the precautionary principle* (Vol. 6). Cambridge University Press.

Tan, R. (2017) Terrorists' love for Telegram, explained. *Vox*. Accessed November 2017 via https://www.vox.com/world/2017/6/30/15886506/terrorism-isis-telegram-social-media-russia-pavel-durov-twitter

Tech Against Terrorism (2017a) Welcome to the knowledge sharing platform. Accessed 19 December 2019 via https://ksp.techagainstterrorism.org/

Tech Against Terrorism (2017b) Official Launch of the Knowledge Sharing Platform at the United Nations New York. Accessed 19 December 2019 via https://www.techagainstterrorism.org/2017/12/05/official-launch-of-the-knowledge-sharing-platform-at-the-united-nations-new-york/

Tech Against Terrorism (2017c) Tech Against Terrorism at Chatham House. *Tech Against Terrorism*. Accessed October 2017 via https://techagainstterrorism.org/2017/07/12/tat-at-chatham-house/

Tech Against Terrorism (2018) Tech Against Terrorism at TechHub London. Accessed November 2020 via https://www.techagainstterrorism.org/2018/03/09/tech-against-terrorism-at-techhub-london/

Tech Against Terrorism (2019a) ISIS use of smaller platforms and the DWeb to share terrorist content. *Tech Against Terrorism*. Accessed 30 August 2019 via https://www.techagainstterrorism.org/2019/04/29/analysis-isis-use-of-smaller-platforms-and-the-dweb-to-share-terrorist-content-april-2019/

Tech Against Terrorism (2019b) Analysis: The use of open-source software by terrorists and violent extremists. *Tech Against Terrorism.* Accessed 6 December 2019 via https://www.techagainstterrorism.org/2019/09/02/analysis-the-use-of-open-source-software-by-terrorists-and-violent-extremists/

Tech Against Terrorism (2019c) Press release: 10th April 2019 – launching an updated version of Jihadology to limit terrorist exploitation of the site. Accessed 19 December 2019 via https://www.techagainstterrorism.org/2019/04/10/press-release-10th-april-2019-launching-an-updated-version-of-jihadology-to-limit-terrorist-exploitation-of-the-site/

Tech Against Terrorism (2019d) Press release: Tech Against Terrorism awarded grant by the Government of Canada to build Terrorist Content Analytics Platform. Accessed 19 December 2019 via https://www.techagainstterrorism.org/2019/06/27/press-release-tech-against-terrorism-awarded-grant-by-the-government-of-canada-to-build-terrorist-content-analytics-platform/

Tech Against Terrorism (2020a) Tech Against Terrorism. Accessed 27 May 2020 via https://www.techagainstterrorism.org/

Tech Against Terrorism (2020b) Press release: Tech Against Terrorism awarded grant by the Government of Canada to build Terrorist Content Analytics Platform. Accessed 27 May 2020 via https://www.techagainstterrorism.org/2019/06/27/press-release-tech-against-terrorism-awarded-grant-by-the-government-of-canada-to-build-terrorist-content-analytics-platform/?utm_source=Tech+Against+Terrorism&utm_campaign=60533678bb-EMAIL_CAMPAIGN_2019_03_24_07_51_COPY_01&utm_medium=email&utm_term=0_cb464fdb7d-60533678bb-67710343

Tech Against Terrorism (2020c) The online regulation series | The European Union. Accessed 20 November 2020 via https://www.techagainstterrorism.org/2020/10/19/the-online-regulation-series-the-european-union/

Tech Against Terrorism (2020d) The online regulation series | The United Kingdom. Accessed 6 November 2020 via https://www.techagainstterrorism.org/2020/10/22/online-regulation-series-the-united-kingdom/?utm_source=Tech+Against+Terrorism&utm_campaign=d1b6edac7b-EMAIL_CAMPAIGN_2019_03_24_07_51_COPY_01&utm_medium=email&utm_term=0_cb464fdb7d-d1b6edac7b-140969075

Tech Against Terrorism (2020e) REPORT: Conclusions from the online consultation process for the Terrorist Content Analytics Platform (TCAP) – August 2020. *Tech Against Terrorism.* Accessed 14 August 2020 via https://static1.squarespace.com/static/5e95a8728adf0c46b944604d/t/5f33baa98546545a7c020543/1597225686130/Report+--+TCAP+Online+Consultation+Report.pdf

Tech Against Terrorism (2020f) Transparency reporting for smaller platforms. Accessed 3 April 2020 via https://www.techagainstterrorism.org/2020/03/02/transparency-reporting-for-smaller-platforms/?utm_source=Tech+Against+Terrorism&utm_campaign=d7804fffb6-EMAIL_CAMPAIGN_2019_03_24_07_51_COPY_01&utm_medium=email&utm_term=0_cb464fdb7d-d7804fffb6-140969075

Telegram (2021a) Telegram Homepage. Accessed 02/02/2021 via https://telegram.org/

Telegram (2021b) Telegram Terms of Service. Accessed 02/02/2021 via https://telegram.org/tos

Telegram (2021c) Telegram FAQ. Accessed 02/02/2021 via https://telegram.org/faq

The Guardian (2018) Facebook failing to protect moderators from mental trauma, lawsuit claims. *The Guardian*. Accessed 10 March 2020 via https://www.theguardian.com/technology/2018/sep/24/facebook-moderators-mental-trauma-lawsuit

The Parliament of the Commonwealth of Australia (2019) Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019

The Santa Clara Principles (2018) The Santa Clara Principles on Transparency and Accountability in Content Moderation. Accessed 9 April 2020 via https://www.santaclaraprinciples.org/

Theil, S. (2019). The Online Harms White Paper: comparing the UK and German approaches to regulation. *Journal of Media Law*, 1-11.

Thomas, M.W. (1948) The early factory legislation: A study in legislative and administrative Evolution

Thomas, T. L. (2003). *Al Qaeda and the Internet: The Danger of'Cyberplanning'*. Foreign Military Studies Office (ARMY) Fort Leavenworth Ks.

Timberg, C., Harwell, D., Dwoskin, E., and Brown, E. (2018) From Silicon Valley elite to social media hate: The radicalisation that led to Gab. The Washington Post. Accessed July 2019 via https://www.mercurynews.com/2018/11/01/from-silicon-valley-elite-to-social-media-hate-the-radicalisation-that-led-to-gab/

Torba, A. (2019) It's time to build a free speech internet of our own. *Gab News.* Accesed 31 January 221 via https://news.gab.com/2019/10/18/its-time-to-build-a-free-speech-internet-of-our-own/

Trujillo, M., Gruppi, M., Buntain, C., & Horne, B. D. (2020). What is BitChute? Characterizing the" Free Speech"Alternative to YouTube. *arXiv preprint arXiv:2004.01984*.

Tunick, M. (2014). *Balancing privacy and free speech: Unwanted attention in the age of social media*. Routledge.

Turner, P. N. (2014). "Harm" and Mill's Harm Principle. *Ethics, 124*(2), 299-326.

Twitter (2016). Combatting violent extremism. *Twitter Blog*. Accessed October 2017 via https://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html

Twitter (2018) Netzwerkdurchsetzungsgesetzbericht: Januar – Juni 2018, p. 6. Cited in Schmitz, S., & Berndt, C. M. (2018). The German Act on Improving Law Enforcement on Social Networks (NetzDG): A Blunt Sword?. *Available at SSRN 3306964*

Twitter Safety (2017) Clarifying the Twitter Rules. *Twitter Blog*. Accessed May 2018 via https://blog.twitter.com/official/en_us/topics/company/2017/Clarifying_The_Twitter_Rules.html

Twitter (2018) World leaders on Twitter. Accessed 24 January 2021 via https://blog.twitter.com/en_us/topics/company/2018/world-leaders-and-twitter.html

Twitter Investor Relations (2019) Investor Fact Sheet. Accessed 18 January 2021 via https://s22.q4cdn.com/826641620/files/doc_financials/2019/q3/Q3_19_InvestorFactSheet.pdf

Twitter Investor Relaitons (2021) FAQs. Accessed 18 February 2021 via https://investor.twitterinc.com/contact/faq/default.aspx#:~:text=The%20mission%20we%20serve%20as,a%20free%20and%20global%20conversation.

Twitter (2019) World Leaders on Twitter: Principles & Approach. *Twitter Blog.* Accessed 24 January 2021 via https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html

Twitter (2021a) Permanent suspension of @realDonaldTrump. Accessed 24 January 2021 via https://blog.twitter.com/en_us/topics/company/2020/suspension.html

Twitter (2021b) Insights from the 17th Twitter Transparency Report. Accessed 24 January 2021 via https://blog.twitter.com/en_us/topics/company/2020/ttr-17.html

Twitter (2021) Twitter Rules. Accessed 24 January 2021 via https://help.twitter.com/en/rules-and-policies#twitter-rules

Tworek, H., and Leerssen, P. (2019) An Analysis of Germany's NetzDG Law. *Translantic Working Group.* Accessed 21 November 2019 via https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

Tyler, T. R. (1990) *Why people obey the law*. New Haven: Yale University

Tyler, T. R. (2006). *Why people obey the law*. Princeton University Press.

Tyler, T. R., & Huo, Y. (2002). *Trust in the law: Encouraging public cooperation with the police and courts*. Russell Sage Foundation.

Tyler, T. R., & Lind, E. A. (1992). A relational model of authority in groups. In *Advances in experimental social psychology* (Vol. 25, pp. 115-191). Academic Press.

Tyler, T. R., & Mitchell, G. (1993). Legitimacy and the empowerment of discretionary legal authority: The United States Supreme Court and abortion rights. *Duke LJ*, *43*, 703.

Tyler, T.R., Boeckmann, R.J., Smith, H.J., and Huo, Y.J. (1997) *Social Justice in a diverse society*. Boulder, CO: Westview Press

United Nations Human Rights Committee (2011) General comment 34: Article 19: Freedoms of opinion and expression.
United States Consumer Product Safety Commission (2017) Injury Statistics. Accessed 28 August 2020 via https://www.cpsc.gov/Research--Statistics/Injury-Statistics

Urman, A., & Katz, S. (2020). What they do in the shadows: examining the far-right networks on Telegram. *Information, communication & society*, 1-20.

Van Snellenberg, T., & van de Peppel, R. (2002). Perspectives on compliance: non-compliance with environmental licences in the Netherlands. *European Environment*, *12*(3), 131-148.

VanDeVeer, D. (1986) *Paternalistic Intervention*. Princeton, N.J.: Princeton University Press

Van der Vegt, I., Gill, P., Macdonald, S., and Kleinberg, B. (2019) Shedding light on terrorist and extremist content removal. *Global Research Network on Terrorism and Technology:* Paper No. 3.

Veljanovski, C. (2010) Economic Approaches to Regulation. In Baldwin, R., Cave, M., & Lodge, M. (Eds.). (2010). *The Oxford handbook of regulation*. Oxford University Press.

Venkatesh, V., Podoshen, J. S., Wallin, J., Rabah, J., & Glass, D. (2018). Promoting extreme violence: visual and narrative analysis of select ultraviolent terror propaganda videos produced by the Islamic State of Iraq and Syria (ISIS) in 2015 and 2016. *Terrorism and Political Violence*, 1-23.

Vermeule, A. (2012). Contra Nemo Iudex in Sua Causa: the limits of impartiality. *Yale LJ*, *122*, 384.

Visnji, M. (2019) How Microsoft makes money? Understanding Microsoft Business Model. *Revenues and Profits*. Accessed 18 February 2021 via https://revenuesandprofits.com/how-microsoft-makes-money-understanding-microsoft-business-model/#:~:text=Microsoft%20generates%20revenues%20from%20individual,licensing%20of%20Windows%20operating%20system.&text=Microsoft%20also%20earns%20revenues%20through,third%2Dparty%20video%20game%20royalties.

Vogel, D. (1990). When consumers oppose consumer protection: The politics of regulatory backlash. *Journal of Public Policy*, *10*(4), 449-470.

Vogel, D. (2003). The hare and the tortoise revisited: the new politics of consumer and environmental regulation in Europe. *British Journal of Political Science*, *33*(4), 557-580.

Von Behr, I., Reding, A., Edwards, C., & Gribbon, L. (2013). Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism.

Walker, K. (2017) Working together to combat terrorists online. *Google*. Accessed September 2017 via https://www.blog.google/topics/public-policy/working-together-combat-terrorists-online/

Warner, B. (2019) Tech companies are deleting evidence of war crimes. *The Atlantic.* Accessed 19 November 2019 via https://www.theatlantic.com/ideas/archive/2019/05/facebook-algorithms-are-making-it-harder/588931/

Watkin and Conway (2021) Building social capital to counter polarization? An analysis of tech platforms' official blog posts. *First Monday*, Forthcoming

Waters and Postings (2018) Spiders of the Caliphate: Mapping the Islamic State's global support network on Facebook. *Counter Extremism Project.* Accessed June 2018 via https://www.counterextremism.com/sites/default/files/Spiders%20of%20the%20Caliphate%20%28May%202018%29.pdf

Weick, K. (2001) *Making sense of the organisation*. London: Blackwell

Weimann, G. (2004). *www. terror. net: how modern terrorism uses the Internet* (Vol. 31). United States Institute of Peace.

Weimann, G. (2010). Terror on facebook, twitter, and youtube. *The Brown Journal of World Affairs*, *16*(2), 45-54.

Weirman, S., & Alexander, A. (2018) Hyperlinked Sympathizers: URLs and the Islamic State. *Studies in Conflict & Terrorism*, 1-19.

Welch, C. (2020) Twitter starts rolling out audio tweets on iOS. *The Verge.* Accessed 2 July 2020 via https://www.theverge.com/2020/6/17/21294481/twitter-audio-tweets-now-available-iphone-ios

Welch, T. (2018). Theology, heroism, justice, and fear: an analysis of ISIS propaganda magazines Dabiq and Rumiyah. *Dynamics of Asymmetric Conflict*, *11*(3), 186-198.

Whittaker, J. (2019) How content removal might help terrorists. *Lawfare Blog.* Accessed 8 December 2019 via https://www.lawfareblog.com/how-content-removal-might-help-terrorists

Wilcke, G. (1971) Federal job plan goes into effect. *New York Times.* Cited in Wahl, A. M., & Gunkel, S. E. (1999). Due process, resource mobilization, and the Occupational Safety and Health Administration, 1971–1996: The politics of social regulation in historical perspective. *Social Problems*, *46*(4), 591-616.

Wilson, G. K. (1984). Social regulation and explanations of regulatory failure. *Political Studies*, *32*(2), 203-225.

Windwehr, S. and York, J. (2020) Facebook's most recent transparency report demonstrates the pitfalls of automated content moderation. *Electronic Frontier Foundation.* Accessed 14 January 2020 via https://www.eff.org/deeplinks/2020/10/facebooks-most-recent-transparency-report-demonstrates-pitfalls-automated-content

Windholz, E. (2010). Evaluating the Harmonisation of Australia's OHS Laws: Challenges and Opportunities. *Asia Pacific Journal of Public Administration*, *32*(2), 137-162.

Wingfield, R. (2019) Approaches to content regulation - #4: Transparency reporting. *Global Partners Digital.* Accessed 17 June 2020 via https://www.gp-digital.org/approaches-to-content-regulation-4-transparency-reporting/

Winter, D. (2018) AI errors vs human errors. *International Director*. Accessed 19 June 2020 via https://internationaldirector.com/technology/ai-errors-vs-human-errors/

Wojcikci, S. (2017) Expanding our work against abuse of our platform. *YouTube Blog.* Accessed 22 February 2021 via https://blog.youtube/news-and-events/expanding-our-work-against-abuse-of-our

Wong, J.C. (2019) Facebook finally responds to New Zealand on Christchurch attack. *The Guardian.* Accessed 4 October 2020 via https://www.theguardian.com/us-news/2019/mar/29/facebook-new-zealand-christchurch-attack-response

Woods, L. (2019). The duty of care in the Online Harms White Paper. *Journal of Media Law*, 1-12.

Wu, T. S. (1996). Cyberspace Sovereignty--The Internet and the International System. *Harv. JL & Tech.*, *10*, 647.

Wu, P. (2015). Impossible to regulate: Social media, terrorists, and the role for the UN. *Chi. J. Int'l L.*, *16*, 281.

Xie, S. (2019) More control over your conversations: now available globally. *Twitter Blog*. Accessed 27 December 2019 via https://blog.twitter.com/en_us/topics/product/2019/more-control-over-your-conversations-globally.html

Yapp, C., & Fairman, R. (2006). Factors affecting food safety compliance within small and medium-sized enterprises: implications for regulatory and enforcement strategies. *Food control*, *17*(1), 42-51.

Yar, M. (2018). A failure to regulate? The demands and dilemmas of tackling illegal content and behaviour on social media. *International Journal of Cybersecurity Intelligence & Cybercrime*, *1*(1), 5-20.

Yau, B. (2014). Occupational safety culture index e Measuring the community and employees awareness, attitude and knowledge towards workplace safety and health in Hong Kong [power point slides]. In *XX World Congress on Safety and Health at Work--Global Forum for Prevention Frankfurt (Germany)*.

York, J. and Schmon, C. (2021) The EU Online Terrorism Regulation: a Bad Deal. *Electronic Frontier Foundation*. Accessed 8 April 2021 via https://www.eff.org/deeplinks/2021/04/eu-online-terrorism-regulation-bad-deal

YouTube (2016). Growing our Trusted Flagger program into YouTube Heroes. *YouTube*. Accessed September 2017 via https://youtube.googleblog.com/2016/09/growing-our-trusted-flagger-program.html

YouTube, (2017a) An update on our commitment to fight violent extremist content online. *YouTube Blog*. Accessed December 2017 via https://youtube.googleblog.com/2017/10/an-update-on-our-commitment-to-fight.html

YouTube, (2017b). Bringing new Redirect Method features to YouTube. *YouTube*. Accessed September 2017 via https://youtube.googleblog.com/2017/07/bringing-new-redirect-method-features.html

YouTube (2018) Entfernungen von Inhalten nach dem Netzwerkdurchsetzungsgesetz 01. Januar 2018 – 20. Juni 2018, p. 20. Cited in Schmitz, S., & Berndt, C. M. (2018). The German Act on Improving Law Enforcement on Social Networks (NetzDG): A Blunt Sword?. *Available at SSRN 3306964*.

YouTube (2019) The four R's of responsibility, part 1: removing harmful content. *YouTube Blog*. Accessed September 2019 via https://youtube.googleblog.com/2019/09/the-four-rs-of-responsibility-remove.html

YouTube (2021a) YouTube Community Guidelines. Accessed 24 January 2021 via https://www.youtube.com/intl/ALL_uk/howyoutubeworks/policies/community-guidelines/#community-guidelines

YouTube (2021b) About. Accessed 18 February 2021 via https://www.youtube.com/intl/en-GB/about/#:~:text=Our%20mission%20is%20to%20give,a%20community%20through%20our%20stories.

Zannettou, S., Blackburn, J., De Cristofaro, E., Sirivianos, M., & Stringhini, G. (2018). Understanding web archiving services and their (mis) use on social media. *arXiv preprint arXiv:1801.10396*.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., & Blackburn, J. (2018). What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018* (pp. 1007-1014).

Zedner , L. (2007a), ' Pre-Crime and Post-Criminology? ' , Theoretical Criminology, 11 : 261 – 81 .

Zedner, L. (2007b). Preventive justice or pre-punishment? The case of control orders. *Current Legal Problems*, *60*(1), 174.

Zhi, C.Y. (1992) Micro-Economics of regulation. Cited in Liu, M. P. (2014). The Logic to Quantify Operation of Social Regulation. *AASRI Procedia*, *7*, 88-93.

Zoller, E. (2009). Foreword: Freedom of Expression: Precious Right in Europe, Sacred Right in the United States. *Ind. LJ*, *84*, 803.

Zuckerberg, M. (2018) A blueprint for content governance and enforcement. *Facebook post.* Accessed July 2019 via https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/

Zuckerberg, M. (2019) Four ideas to regulate the internet. *Facebook Newsroom*. Accessed 4 December 2019 via https://about.fb.com/news/2019/03/four-ideas-regulate-internet/