Extremist Content Removal on Social Media: A Process of Cutting Corners

Connor Rees
872245

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Master of Science

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

November 13, 2020

Declaration

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

Signed ..... (candidate)

Date .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give my 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 ..... (candidate)

Date .....

I would like to dedicate this work to Anthony, my dad and Caroline, my mum.
Among many other things you believed in my potential and pushed me to set my sights always a little higher.

Abstract

The adoption of social media by online extremists continues to leave users, governments and social media companies on the back foot. The primary mode of regulating extremist content comes in the form of extremism content removal processes. In turn, this leaves users, academics, stakeholders and governments alike asking the right questions in the wrong order for example, how do we make these processes more accurate and effective? However, throughout the course of this dissertation attention is drawn to the right questions i.e., how are these processes conducted and what are the human impacts of these primarily computational solutions. Through considering the regulatory factors that govern these content removal processes, analysing the strengths and limitations of the modes of content removal and consulting social media users who are on the wrong end of the counter-extremism stick.

Drawing from these notions, the research uncovers the understanding that social media companies are provided with little support and are subject to significant sanctions by governments for not removing extremist content in a timely manner. As a result it becomes increasingly clear that respect of human rights, due process and ethical processes are to an extent being neglected to meet these increasing demands. Consequently, the users that social media extremist content removal works to protect is the very things that is causing harm.

Acknowledgements

I have no doubt that this dissertation would not be what it is without my friends in the CDT and the staff that have supported me, not to mention my supervisor Dr Muller. And to the members of CYTREC who helped an encouraged me to push myself in an unfamiliar setting.

Contents

Glossary

Black Box Model The black box metaphor dates back to the early days of cybernetics and behaviourism, and typically refers to a system for which we can only observe the inputs and outputs, but not the internal workings. 34

Censorship The suppression of words, images, or ideas that are "offensive," which happens whenever some people succeed in imposing their personal political or moral values on others. 6

Data Privacy Or information privacy describes the use and governance of personal data—things like putting policies in place to ensure that consumers' personal information is being collected, shared and used in appropriate ways. 30

Displacement Is the forced closure of a platform that makes it impossible for the community to continue their social and informational activities as well as to capture the strengths of reactions. 15

Extremism Is the belief that an in-group's success or survival can never be separated from the need for hostile action against an out-group. 6

Grey-zone Content Potentially problematic social media content commonly detected by automated systems to be reviewed by human content moderators. 6

Moral Panic Is a feeling of fear due to an evil threatens the well-being of society usually as a result of the mass media. 7

Privacy Is a state in which one is not observed or disturbed by other people. 6

Radical Right Adopts a similar definition as right-wing extremism, but it also incorporates some aspects of mainstream conservatism in addition. 10

Radicalisation Is a phased and complex process in which an individual or a group embraces a radical ideology or belief that accepts, uses or condones violence, including acts of terrorism within the meaning of the Directive on combating terrorism, to reach a specific political or ideological purpose. 7

Right-Wing Extremism Is a phrase used to describe right-wing political, social and religious movements that exist outside of and are more radical than mainstream conservatism. 10

Terrorism Is an act/s intended or calculated to provoke a state of terror in the general public, a group of persons or particular persons for political purposes are in any circumstance unjustifiable, whatever the considerations of a political, philosophical, ideological, racial, ethnic, religious or any other nature that may be invoked to justify them. 7

Web 2.0 Is a term that was introduced in 2004 and refers to the second generation of the World Wide Web, following Web 1.0 and before a theoretical Web 3.0 which involves additional advanced technologies in the future. 8

List of Figures

Chapter 1

Introduction

As society delves deeper into the digital age, long existing societal problems follow the same path to an online existence, and extremism is no exception. What was once the responsibility of governments and local communities has quickly become the burden for social media companies who are tasked with removing extremist content from their platforms. The present paper sheds some light on the multi-dimensional factors that contribute to extremist content removal. The difficulty in addressing this topic is that akin to all research regarding extremism absence of a unifying understanding of the term extremism, research in this area lacks coherency and specificity. Therefore, in referring to this term the constructive definition [102] provided by J.M. Berger will be what this research understands by its meaning.

> "the belief that an in-group's success or survival can never be separated from the need for hostile action against an out-group." [9]

To expand further, the notion of a "hostile action" is inherently broad with attributes ranging from insults all of the way to the elimination of the out-group.

When it comes to extremist content removal on social media there is a lot of movement around the idea of efficiency and effectiveness. These discussions notably began to pick up speed after the New Zealand Christchurch shootings in 2019, the brief event of which can be summarised as Brenton Tarrant who murdered 51 Muslims and injured another 49 following a targeted shooting in a mosque [76]. Beyond the devastation of this attack, change was sparked following the significant and fast spread of content which showed this attack across several social media sites, including Facebook, Twitter and Youtube. In response to this mass spread of

extremist content Facebook for example, removed 1.5 million variations of the video showing the attack following the first 24 hours of its occurrence [98]. The response to which saw calls for improved responses to extremist content on social media platforms, and to a varying degree this improvement can be evidenced, as will become apparent later in this document. In turn, this document poses the question to what end accuracy ratings are chased and governments appeased? As a result, to what extent are corners cut, human right set aside, responsible innovation ignored and ethical procedures seen as secondary? In the wake of the methods posed in response of the Christchurch attacks, this paper offers the sentiments that in response to the cutting of significant corners, how is this to anyone's benefit? Especially when we still see the all too clear spread of hate, propaganda and misinformation and extremist content on social media. The significance of this issue becomes ever too apparent when it takes celebrities like Kim Kardashian West to point these issues when governments and third parties will not [7]. Thus, just as it is deemed important to push social media companies to remove extremist content, it may be deemed just as important to have a third party regulating body to enforce ethical procedures when doing so. Limitations on freedoms in the context of social media content may be deemed trivial by some, however, if left unchecked the accusation of such failures will all too quickly become pervasive.

1.1 Motivation

The principle notion of this research stems from the growing trend of concern regarding online content and discourse. With fake news, deep fakes paired with platforms that give any and everyone a voice, the lines between extremism, hate speech and offensive content have never been less clear and more difficult to differentiate. As a result, the way in which the aforementioned content is removed from social media platforms has never been more important and not just important to several people, but on a global scale that encompasses the billions of users across multiple platforms that play an increasing role in human existence. By extension any wrongdoings, shortcuts taken and corners cut that limit ethical processes and the preservation of human rights reach all corners of the globe. Thus, the motivations for this research is to analyse how human rights and ethical conduct is if at all taking place in the social media extremist content removal process.

1.1.1 Objective

The objective of this document is to analyse the several factors which contribute to how extremist content removal is conducted on social media, where corners are being cut and beginning to gauge how users understand and acknowledge how their rights and liberties are being affected as a result.

1.2 Structure

This section briefly outlines the structure of the dissertation. Following this introduction the subsequent Chapter 2 considers the substantial understandings drawn from the academic literature regarding the key concepts covered in this documents; including extremism, extremists use of the internet and extremist content removal. As a result this chapter aims to provide the reader with the necessary tools to develop their own judgement from the arguments presented in the remainder of the dissertation. Chapter 3 then identifies and reviews the techniques used in the mixed methodology approach adopted by this dissertation. This includes finding academic and non-academic resources online to analyse the factors involved with conducting an online survey. In Chapter 4 analysing the legal and ethical implementation of AI is represented through analysing UK/EU examples of legal and ethical regulations to showcase the difficulties and limitations put in place to theoretically protect human rights and encourage responsible innovation. Following this, Chapter 5 looks at analysing effective extremist content removal by analysing the strengths and limitation of the three primary content removal methodologies (human, automated and hybrid human-automated extremist content removal). The third and final research question found in Chapter 6 looks at social perceptions of extremist content removal on social media. In doing so, this considers an online survey to grasp people's views and opinions on the state of content removal and the various factors that effect it. And finally Chapter 7 consists of a summary of the dissertation, the acknowledgement of its limitations, the main contributions and any scope for future work.

1.3 Contributions

The main contributions of this work can be seen as follows:

- **Addressing AI regulation based on ethical principles**

This element of the dissertation addresses the recent trend in producing AI ethics principle documents and applying them to extremist content removal. Reviewing such documents is an angle that has seen recent growth in the academic literature, however, these documents have yet to be covered in the context of extremism. By shining a light on such documents, the integrity is to be brought into question as well as future recommendations. As a result this contribution leaves scope for future research.

- **Analysing regulatory impacts on the capabilities of extremist content removal**

  Reviewing social media content removal techniques is not a new angle in the literature. However, exploring how legal and ethical regulations effect to varying degrees how social media extremist content removal is conducted is an unexplored narrative within this context. How each of these factors shape and mold content removal creates questions regarding who is to be held accountable for such practices which are subject to more than its fair share of criticisms.

- **Exploring human perspective on extremist content removal**

  This component of the dissertation sheds some light on a significant angle that is yet to be explored in academic research. In that, the opinions and views of social media users and stakeholders are not necessarily reflected in the way that extremist content removal is conducted. In addition, this leaves significant scope for a wealth of future research on this subject matter.

Chapter 2

Literature Review

2.1 Introduction

The use of online platforms such as social media by extremist activists and extremist groups' has been widely documented and broadly understood in academic literature. Not to take this understanding for granted, this chapter will aim to unpick key findings and understandings that contribute to how effectively content removal can be conducted on social media, and the necessity for undertaking this process whilst maintaining ethical processes. In unpicking the initial statement, this chapter will develop a discussion surrounding the three primary modes of extremist content removal removal on social media [113] in addition to the necessity of effectively implementing social media content removal. To effectively discuss the areas which fall within the potentially broad scope of this topic, the following two-tiered structure will be adopted. The topics of the two definitive sub-chapters are listed as follows: exploring extremists' use of social media, and secondly, the social media companies response to the extremist presence on their platforms.

Each of these two topics are essential in developing a fundamental understanding, necessary in acquiring the tools to critically consider the topics that can be found in the remainder of this sub-chapter and the dissertation as a whole. The aim of the 'extremists' use of social media' section is to build an understanding of the factors surrounding extremists' use of the internet and more specifically social media. Therefore, the following questions will be covered: Why do extremists use these platforms, to what extent are they utilised, how are they used, how the usage has changed over time, and what is the understanding around the scale of the threat? The findings drawn from the literature in this section will leapfrog to the topics discussed in the

subsequent social media extremist content removal section. The latter section aims to briefly identify the methodologies adopted by large social media companies to counteract the hijacking of their sites' intended function and the repercussions associated with them. Following this, the section identifies the apparent differences between several social media companies' approaches to extremist content removal. The likes of which are surfacing as a result of the differences in political stances i.e., upholding the principles of freedom of speech and Privacy. In conjunction, beyond looking at these responses by each company separately, the united front developed by social media companies under the Global Internet Forum to Counter Terrorism (GIFCT) collective will also be discussed and analysed. Finally, this section identifies social media hybrid human-automated extremist content removal. In doing so the two-pronged approach this method adopts will be specified and and identified within the context of Extremism, prior to its analysis in Chapter 5. As a result, this section will identify the issues faced by social media companies and additionally, the specified approach adopted in response to the said challenges.

In covering the contents found in both of these sections, this chapter will take the position that extremists only seek to benefit from the adoption of the internet and more specifically social media. To date social media companies are consistently having to lead from the front in mitigating the benefits experienced by extremists, namely through account suspension and content removal. Although these methods have yet to have the desired effect of eradicating extremist content entirely, it is arguably a step in the right direction in what will almost certainly be a fight that continues for as long as social media itself continues to thrive. In the meantime, social media companies have to account for the Censorship issues associated with Grey-zone Content which walks the line between political speech and extremist narratives [52]. Censorship issues in this area are then exacerbated by the currently undocumented levels of false positives or false negatives that are currently unavoidable in the content removal process. In many ways modern day online extremists that use social media are largely embodied by the concept of the many-headed hydra. To cut off the head will not resolve the problem just as banning accounts and blocking posts will achieve the same. However, it is the game that must be played in order to uphold human rights and part take in ethical processes necessary to maintaining a civil society.

2.2 Extremists' Use of Social Media

2.2.1 Introduction

Prior to analysing the numerous factors surrounding extremist's use of the internet, it is first worth identifying the expectations versus the reality regarding such conduct. The threat of extremists' use of the internet when first considered by countless academics, experts, and politicians alike, to an extent, created a Moral Panic [43]. The focus of this attention was mainly directed toward the internet and social media being used to plan or carry out an attack. However, this attention has more recently been directed towards these online spaces as grounds to engage in Radicalisation and inspire vulnerable individuals to be indoctrinated into extremist groups [19, 20]. This is where fears have been raised around an evolved form of extremism, the likes of which can ultimately lead to a cyber-terrorist attack. Such an attack would entail an act within the remit of gaining access and damaging a governmental system, which would result in a devastating effect on the economy an essential industry (i.e. health care, military, or financial) [114]. These fears stem from a number of concepts, the most notable being the understanding that with offline counter-terrorism practices becoming more robust this could make online attacks both the path of least resistance as well as being the most devastating [95]. The potential threat posed by the concept of cyber-terrorist attacks and online extremism has provoked considerable alarm. Numerous security experts, politicians, and others have publicised the potential danger caused by such an event. However, despite these predictions, there has yet to be a single case of cyber-terrorism. Regardless, there are numerous experts predicting imminent extremist cyber-attack catastrophe [71, 75, 115, 116]. However, there have been several events and attacks that have been conducted by hackers which were at one point mistaken for Terrorism. Despite this moral panic and extreme expectation of extremism online, with the current reality of extremist activity online being more along the lines of propaganda radicalisation, this for the most part takes a far less immediate and aggressive form than cyber-terrorism. However this notion is certainly no less harmful or dangerous.

2.2.2 Context

The endless stream of the negative media portrayal of extremists has painted a picture of the individuals that fall within the parameters of this label as being something outside of the ordinary. Although their ideologies and dogmas provide a stark contrast to the ideologies found in those who fall outside of the extremist label, there are many examples where the line becomes

blurry and more similarities may be shared with the wider population than one may be led to believe. A prime example of this is the use of the internet and the adoption of social media. Since the establishment of Web 2.0 in 2004, the number of internet and social media users have continually increased every year and continues to do so. For example, between 2005-2015 65% of adults are using social media, which is ten times that seen in the previous decade [88]. Extremists, akin to the rest of society, have become ever-dependable on the internet and the many resources and opportunities it offers. In the context of the adoption of the internet and social media, extremists are neither behind nor ahead of the curve set by the wider society. As evidenced in a study conducted by VoxPol in 2016 [49] which comprised 272 convicted UK extremists, terrorists, and attack plotters, primarily male (96%) primarily jihadists (89%), and the remainder was constituted of right-wing extremists (11%). One of the findings taken from this study evidenced how 54% of this sample utilised the internet's resources to learn something which contributed towards the intended action -planning an attack- before 2012. From 2012 onward, this number dramatically increased by 22% up to a total increase of 76%. From these statistics, two things can be determined. Firstly, just like the broader population, extremists have increasingly adopted the internet since Web 2.0. And secondly, in line with the finding on the expectations versus the realities of online extremism, the use of the internet by extremists on the surface level is far less sophisticated than what was immediately feared and hypothesised. As can be found in the previous example, whereby, internet usage used for planning an attack could constitute something as simple as typing locations into a search engine or identifying public transport mechanisms. In conjunction, extremists are using the internet for research as opposed to coordinated devastating cyber-attacks, as was and still is feared and hypothesised by several reputable individuals and organisations [6, 62, 77]. Which as a result, leaves members of the public in an unnecessary and disproportionate state of fear regarding extremists' use of the internet.

2.2.3 The Left-Wing

With the fear of a crippling cyber-attack for the most part not constituting extremists' use of the internet, this begs the question of what exactly is constituted. This is a challenging topic to cover, as there are both countless extremist groups internationally and countless differences between them. In regard to extremists' use of the internet and more specifically social media platforms, their broad motivations range across spreading hateful narratives, distributing propaganda, enhancing their financing and fundraising, recruitment, radicalisation and sharing

operational information [50]. In fact, it can be understood that there is such diversity between extremist groups that the most consistent and unifying factor between each of these groups is that they are are all labelled as extremists. For the most part, the academic literature on this matter partitions into two categories, one addresses the left-wing i.e., Al Qaeda, ISIS, Boko Haram, etc. and the other entails the right-wing i.e., Generation Identity, Britain First, Reclaim Australia, etc. The former in many ways pioneered the intense utilisation of social media. With ISIS establishing and prioritising an aggressive social media strategy, namely on larger social media sites such as Twitter and Facebook. Where their social media status and impact began to grow in 2006 and then reached its peak following the declaration of the Caliphate in June 2014 [39]. It was estimated that ISIS supporters on Twitter accumulated between 46,000 and 90,000 accounts between September and December 2014 [8]. Their strategy encompassed the delivery of their high quality, professionally developed propaganda images, videos, and text. In the wake of this propaganda strategy, ISIS used their status to begin messaging out to vulnerable individuals susceptible to radicalisation [8]. The radicalisation process may have been simplified due to the scale of the operation and the support they accumulated, which in large part, meant they could dictate a significant amount of their media coverage through the use of alarming and graphic visual content.

However, new disruption measures were set out by several mainstream social media companies, which meant that ISIS supporters on these platforms began being disincentivize. Twitter for example began in mid-2014 with low level and sporadic ISIS content removal and account interception and by 2016 said measures were significantly more effective and robust. As evidenced by a quoted 15 to 18 thousand IS-supporting accounts being suspended every month between mid-2015 and the beginning of 2016. And an average of 40 thousand suspensions took place each month from mid-February to mid-July 2016 [109, 110]. As a result, by 2017 ISIS supporters began to question whether the risk associated with supporting the group online was worth going through a process of being censored, and ultimately whether losing their social media accounts was worth the reward of social status. Thus, since ISIS's temporarily successful social media campaign was -for the most part- brought to an end through a larger social media platform trend of developing effective and robust content removal and account banning measures. As a result, in the context of Twitter, the company has since resolved their ISIS problem [28]. Consequently, attempts at radicalisation, recruitment, and planning attacks on Twitter have likely decreased as a result of the newly adopted and refined measures.

2.2.4 The Right-Wing

Remaining on the topic of Twitter, despite this relative success in mitigating ISIS's presence on the platform, the company has more recently fallen under fire as a result of the failure to have the same success with right-wing extremists on their platform [29]. Although it is certainly not the only major social media platform facing such issues and falling under criticism, it has, however, been one of the more readily criticised social media platforms despite other platforms such as Facebook having a larger user base [99]. By extension, with Facebook having a larger database, it may lead one to presume the scale of the problem would also be larger compared to Twitter. However, despite this presumption, Twitter has been the more widely criticised platform. The primary difficulty faced by Twitter and other social media companies is that unlike extreme left-wing content, extreme right-wing is notoriously difficult to identify. And central to Twitters core beliefs is the protection privacy and freedom of speech which can clash against the process of content removal when considering grey-zone content [111]. As identified by the definition of Right-Wing Extremism, the ideologies that fall within its broad scope are almost synonymous with politics. The recent history of the political landscape of the UK has become increasingly affiliated with right-wing politics. For example, the political group the British National Party has been the only Radical Right political party in the UK to date [118], as well as the emergence of more fringe and aspiring political parties such as the English Defense League and Britain First.

Despite the current trend of right-wing extremists adopting and hijacking social media, their prolific use of online platforms has long been established. On sites and forums such as Stormfront, online platforms have offered right-wing extremists a safe space to form and grow a sense of community. Despite Stormfront having originally been established in 1996 and having been through several take-downs, it is still used by right-wing extremists at present [17, 70, 81]. The scale amassed by the platform is equally deserving of note due to the accumulation of over 13 million total posts as of 2017 [81,90]. The user base grown by the site largely stems from the growing community which in turn provides the means for a shared ideology to continually exist and flourish [17]. Similarly to the methods used by ISIS's online strategy, the extreme right-wing use these online platforms such as Reddit, 4chan and 8chan among others, as a catalyst to engage in the process of radicalising the individuals engaging in their content [42, 51]. Only this time, there was already a thoroughly developed online community just waiting to jump platforms and adopt the new benefits offered by social media.

As put across by former right-wing extremist Brad Galloway, online spaces (Stormfront being the most common at the time) have a crucial role pivotal for the extreme right movement. Although Galloway emphasises the role of Stormfront and its chat groups, it is the interpretation of this paper that social media is the contemporary equivalent. In outlining the strength of the platform, Galloway identifies its extremely inclusive nature, which utilises discussion groups which would in turn reinforce sub-communities within what is already a niche community in itself [42]. Thus, reinforcing community strength and extreme right-wing ideology. What is understood by the former right-wing extremist on the defining principles of community-centric platforms (Stormfront during his time as an extremist and social media currently) is that they offer insight into the users, which is then used for recruitment. With the mid-term aim of conducting an offline and in-person meeting. The byproduct of which is where the greatest effect may be felt as expressed by Galloway. Through viewing the frequency of their online engagements, how outspoken and risk-averse the member is, each factor contributed to the vetting process the extremist recruiters use on these sites. The key differentiation being made in this process is whether this user meets the required standards or whether they fall under keyboard warrior status [42, 82]. What this identifies is that while content removal struggles to keep up with online extremists, social media platforms offer a wealth of information to those who look to exploit its services.

Having identified the adoption of social media by extremists, one of the common terms coined in the literature is the process of radicalisation. Research shows that social media is just one of the platforms used, with left-wing extremists moving to platforms such as Telegram [14] and right-wing extremists using platforms such as Gab, Reddit, 4chan, and 8chan [42, 85]. Each of these platforms is broadly used for the same principle: finding a community with like-minded individuals. In such a controlled setting, the radicalisation process is allowed to manifest through what is commonly referred to as an echo chamber [96]. An echo chamber -in this context- refers to space where extremist thoughts and ideas are disseminated, reinforced, met without resistance and as a result, ideologies become gradually more extreme [3]; the effects of which are only emphasised by the presence of a virtual "imagined community" [96]. The significance of this may be seen to a greater effect on message boards, dedicated sights, and forums as opposed to social media. As social media generally offers a larger user base and so a larger opportunity for contrasting opinions. However, the effect remains the same if private messaging and private community pages are utilised and adopted. In the context of the most contemporary threat posed by right-wing extremists. By having an already established

community that has migrated to social media platforms, counter-narratives may be less likely to have any effect if not force a more extreme narrative and the publicity only works in the favour of the extremist recruitment strategy.

2.2.5 Conclusion

Despite inherent differences in extremist groups, one of the unifying factors that can be seen between them is the profound adoption of the internet and more specifically social media. A cross-group motivation for this widespread adoption varies between groups; looking at the left-wing extremist group ISIS, their methodologies provide a stark contrast in some ways to the right-wing. With the left-wing using fear and the right-wing using memes [26], both have had temporary success despite vastly different methodologies. And although the desired outcome of their both group's intentions is malicious, extremists have adopted uniquely humane motivations for adopting the internet and social media. This motivation in most cases is not to hack the government and conduct devastating cyberattacks, but to seek community, develop a sense of belonging, and to meet like-minded individuals. The consequences of extremists finding like-minded individuals are, however, a cause for concern. And so social media companies have found themselves in a position where the onus has been left on them. And with extremism and terrorism being a longstanding trend throughout history, the task they have been left with and the hand they have been dealt, to put it lightly is dire. In recent years this has increasingly proven to be the case with the rise of right-wing discourse on social media and the subsequent grey-zone content makes the removal of extremist content not as clear-cut as one would imagie.

2.3 Extremist Content Removal

2.3.1 Introduction

As insinuated previously in this chapter, social media companies were left on the back-foot when ISIS quickly stormed their platforms with a well put together publicity strategy. The success and effectiveness of this platform hijacking sent waves through both the political, industrial, and academics communities. Among many other questions, the regulatory responsibility that come as a result of extremists' use of social media-generated pressing debate and concern. Currently, the regulatory responsibility social media companies face is primarily self-imposed. However, this may not be the case in the future, as countries such as the United Kingdom and

Germany have begun drafting new regulatory frameworks in the form of a white paper, which will be enforced by national law [101, 120]. However until this is the case on an international basis, the primary mode of checks and balance for social media regulatory responsibility manifests through national legal systems. Where countries are limited to conduct legal cases against online company's or the users themselves in being too slow to act or sharing defamatory/extremism content [35]. Both the significance of this and the applicability are unprecedented. This topic is applicable to not only civil society but to government, and the private sector equally in addition to the social media governance itself [53]. And when conducted should be balanced and should take into careful consideration whether infringements of human rights in regard to the freedom of expression are in proportion to the act. In doing so, academics have proposed a multi-pronged approach including content regulation and moderating in addition to developing effective counter-narratives [43]. Thereby, in order to conceptualise the issues surrounding this topic, this subchapter will consider two points. The first identifies the actions taken by governments and the second entails social media governance and its response to the extremist content removal problem.

In June 2017, ISIS' presence and their accompanying propaganda machine had, for the most part, been eradicated from social media. Right-wing extremists had seemingly taken their place and social media companies had since developed their account suspending and content removing methodology. However, in the wake of the London Bridge attack in the UK, the nation's then Prime Minister Theresa May took a stand. In doing so she claimed that social media platforms are extremist "safe spaces" and that there was a need to reverse their tolerance for extremism and instead be transformed into hostile environments for said extremists [87]. However, before exploring the onus that has been placed on social media companies to remove extremist content, it is first worth looking at the other side of the coin. Which pertains to the responsibility by a state's or governments alike to protect its citizens and regulate extremist content.

2.3.2 Governmental Social Media Content Removal

A 2020 report produced by VoxPol examined legal responses to online extremism in six countries: France, Germany, Israel, Spain, the UK, the US in addition to the UN and the EU [92]. The six countries were selected by meeting the criteria that they had recent dealings with terrorism and had recently developed legislation which in some way regarded online extremism. For the countries that had already established counter-terrorism legal tools, the study found that

minor adjustments needed to be made to account for the factors that come with contemporary technology. This was deemed the case as previous conflicts brought these countries to having stricter limitations on rights when it comes to freedom of speech. Examples of this include the UK with its historic conflicts with Ireland in the 1920's, whereby as a result its citizens had limitations attached to freedom of expression; which today is defined as a qualified right, loosely meaning that it is a right that is subject to several limitations. A similar case study can be seen in Germany after 1945 as a result of the impacts of the Nazi Germany propaganda machine. In regards to the latter, the German legal system incorporates several limitations on freedom of speech, for example, if said speech engages in the support or glorification of terrorism, hate speech or unconstitutional propaganda. As can be found in the parliamentary approved act 'Network Enforcement Act' or NetzDG ratified in 2017 [101]. This provides a stark contrast to the US, whereby as a result of freedom from the severe restrictions imposed by the British Empire, this meant that in the wake of the nation's liberation there were very little limitations on human rights including freedom of speech. As set out by the US Constitution in 1787, freedom of speech is an extremely protected right which exists with almost no direct legislation that imposes any limitations on it. As is evidenced by the opening remarks of the first amendment of the US constitution [25]:

> "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech..."

Although the governments in this report have sufficient means to punish terrorists, they - along with most other nations- are less adept at dealing with online extremism. With this being the case, the countries that have attempted to mitigate online extremism have done so through preventive measures. Although these measures are numerous and varied between countries they can be categorised into the following four extremist content limiting techniques: the blocking and removal of online content; the surveillance of online activity; the criminalising of certain online public expressions; and the use of online content as a justification for applying restrictive administrative measures [92]. Each of these measures will have varying impacts depending on the platform. As previously noted, right-wing extremists are using platforms such as Gab, Reddit, 4chan, and 8chan. Both Gab and Reddit require an email, a unique username, and a password, very little in the way of a unique identifier. Both 4chan and 8chan, however, simply require a name that does not have to be unique, which is nothing in the way of an effective identifier. This means a number of things, tracing the owner of the online account

is difficult, and therefore to criminalise any online public expressions may have a limited effect as users can quickly and effortlessly create a new account. In addition, applying restrictive administrative measures and surveying online activity would heed little progress as someone can use a different username or create another email address and username for another account. Thereby, the most relevant of these methods, and thus the one that will be considered below is the blocking and removal of online content. However, this understanding does not apply to the dark web for the obvious reason that social media sites are not applicable to this area of the internet. Although, it has been made exceptionally clear that terrorists have been known to utilise the encrypted features that the dark web offers for functions such as communication and the exchanging of funds [79, 117]. Despite not specifically pertaining to social media it is worth identifying that as content removal on social media becomes increasingly adopted and refined the effects of Displacement are likely to take place. Through processes such as content removal which consider the disruption of propaganda distribution [11] social media companies can begin to push extremists to other platforms as can be seen with ISIS in 2016 following the Twitter crackdown [94]. In that, it pushes the extremist users to more niche sites are areas of the internet where the effects of an echo chamber are likely to be more severe. Thus, conducting effective content removal whilst paying due regard to the right to privacy is essential. In doing so this outlook would uphold elements of responsible innovation through limiting harm; not only to the stakeholders of social media companies but to the wider population who may suffer the consequences of pushing extremists to the fringes of the online environment.

All of the six countries excluding the US have developed a legal basis to demand platforms such as social media companies to take down or block extremist content. Many of the law enforcement agencies within these countries have developed Internet Referal Units (IRU's) to moderate online content against their national legislation. However, their role is to flag this content and justify its removal to social media companies as opposed to removing it themselves. In doing so, leaving the final decision to the social media companies who will consider the removal against their own terms of service, the impact of these units is hindered considerably [63, 113]. However, the total sum of annual requests to online platforms by these countries is in the tens of thousands, which at a glance may seem significant. However, given the scale of the issue, this is equivalent to a drop in a bathtub. To quantify the scale of extremist activity on social media is difficult as a result of 'extremism' not being directly identified in all social media companies' terms of service and thereby not always addressed in annual reports. Although, in the context of Facebook, extremist activity may be represented between two cat-

egories, the organised hate category, and the terrorism category. In the first quarter of 2020 Facebook removed 4.7 million pieces of organised hate content and in the fourth quarter of 2019, they took down 7.5 million pieces of terrorism content [36]. Despite this disparity in efforts made between national agencies and social media companies, in September 2018 EU member states proposed new legislation that places harsher penalties for not removing extremist and terrorist content from their platforms within an hour of a notice being issued [15, 33]. If the terms are not met then member governments are eligible to issue a fine to the company of up to 4% of their global annual revenue [89]. From this it can be derived that governments have been fast to attribute responsibility on social media companies, they have statistically provided little in the way of support, and have demanded results with little in regard paid to the negative repercussions. Many of which have plagued these platforms since more aggressive extremist content removal has been in place, such as wrongful removals, an inadequate appeals process, impacting political narratives in a biased way, and hindering academic research [92]. On the contrary, another factor to be understood is that although this may be viewed as Governments sifting the blame, it may also be seen as the most effective route. In that, given the current standings, nobody is better equipped at dealing with the threat than the social media companies themselves. Thus, on the back of harsher penalties for social media companies regarding content removal, in certain circumstances, it can also be seen that governments praise social media companies. Governments such as UK governments are praising efforts by social media companies for providing evidence in cases, publishing 'clear community guidelines', and building new technologies to make their online platform space safe from extremism [67]. This being a seemingly contradictory relationship between governments and social media companies. The result of which places a harsh and unforgiving burden on the platforms which are already taking accountability and responding proportionately. The increasing demand with higher stakes -as will become evidenced- may lead social media companies to cutting corners in order to meet these demands.

From this discussion, it is apparent that on an international scale governments are quick to put the blame on social media companies and even quicker to push for harsh standards and penalties for not meeting their standards. Although when it comes to offering support, what is provided by Governments through their IRU's is lackluster and not scalable to meet the demand that the governments themselves have set. However, what has been provided is a legal gateway to allow social media sites a more conflict-free approach to content removal. Through the establishment of 'remove and block' legislation, the legal precedent acts as a catalyst for

social media companies who can best utilise their terms of service and the affiliated content removal terms to enforce them.

2.3.3 Social Media Content Removal

Governments have almost unanimously placed the responsibility on social media companies to effectively moderate their platforms and the extremist narratives that are present on them. Said platforms have seemingly taken these demands on board and although they can be slow to act, there has been little in the way of resisting them. Therefore, having established where the primary responsibility lies it is worth considering how this responsibility may be interpreted and manifested. Akin to governments, there is a significant disparity in how social media companies have approached limiting freedom of speech and expression through their values and terms of service [28]. Give for example Twitter, as previously referred to, one of the company's core values is to defend and respect their user's voice [111]; however, when looking for Facebook's core values to date nothing to this effect can be found. What can be determined from this is that a company's terms of service (TOS) is likely to fall in line with their core values. And their TOS and policy guidelines are essentially what dictates the parameters of content removal on a given platform. Consequently, if a company like Twitter's core values include defending a user's voice, content moderation is less likely to be so scrupulous and intrusive. This may well explain the point made previously in this chapter which stated that Twitter is one of the more heavily criticised social media platforms for being slow to regulate right-wing extremist content.

However, when it comes to policy and legislation, words like extremism and terrorism have a way of building a united front against a common enemy. In the case of social media content removal, this united front has flowered under the umbrella of the GIFCT [45]. Akin to any source of response to extremism, whether it be community, governmental or non-governmental organisation, there will be a different approach with varying parameters, and the same can be said for social media companies. What GIFCT provides is an environment of common ground and harmony. Where social media companies can determine not what the definitions are and not the boundaries may be, but rather what they know is unacceptable and intolerable. GIFCT was established with a top-down approach where bigger companies including Facebook, Microsoft, Twitter, and YouTube provide a means to support smaller platforms become hostile environments for extremists. Although the details on this are not readily available, it is understood that this includes sharing technologies, practices, databases and partnerships [27,45].

What is understood is the broad five-part structure that GIFCT is comprised of, as illustrated in Figure 1 below. The first component of this structure is the 'Independent Advisory Committee' which guides the subsequent 'Operating Board' through producing annual reports and conducting performance reviews. Said Operating Board is tasked with selecting the 'Executive Director', setting the operational budget, and maintaining GIFCT's alignment with its mission. Thirdly, the Executive Director is tasked with providing leadership and coordination when it comes to the forum's operations for example program implementation and fund-raising. Then there are the strategic pillars, "Prevent" "Respond" and "Learn" which categorises the forum's aim and provides avenues of work programs for maximal transparency. And within each of the three strategic pillars are working groups that conduct pillar-specific projects and advise. Finally, there is the 'Multi-stakeholder Forum' which includes social media companies, civil society members, and governments committed to upholding and respecting human rights and preventing terrorists from exploiting digital platforms [48]. One of the most notable achievements by this forum was announced in July 2017, where a shared 'hash' database was announced. This database can be understood as a collection of unique identifiers of extremist propaganda, including videos and images shared between all of the forum's members [47]. Said database has amassed over 200,000 hashes since July 2019 [46]. Unquestionably this is a step in the right direction, although, how this is implemented is another area of significant contention.

Taking into consideration both the scale of content being posted in addition to the international governmental pressure on social media platforms. Social media companies are becoming increasingly dependent on automated technologies such as AI to remove extremist content to match the scale of the problem. The GIFCT hash system addressed above is a prime example of how AI and other automated systems are being applied to this area. When a user of any of the platform members or stakeholders of the forum for example uploads an image, said image is run through the database. If said image matched any of the proscribed images in the shared database the post will not upload. This prevents the content from ever reaching the platforms. When such a database is applied to the likes of ISIS propaganda, it can be deemed extremely effective. For example, several months following the establishment of GIFCT Facebook announced in a blog post that through the use of AI the company was removing 99% of ISIS and Al-Qaeda-related terror content before it was flagged [12]. These kinds of numbers are achieved by repeat uploads. Whereby, once an image is uploaded it is then replicated by other users and fake accounts to then be re-uploaded for thousands of iterations. Thus, once a hash

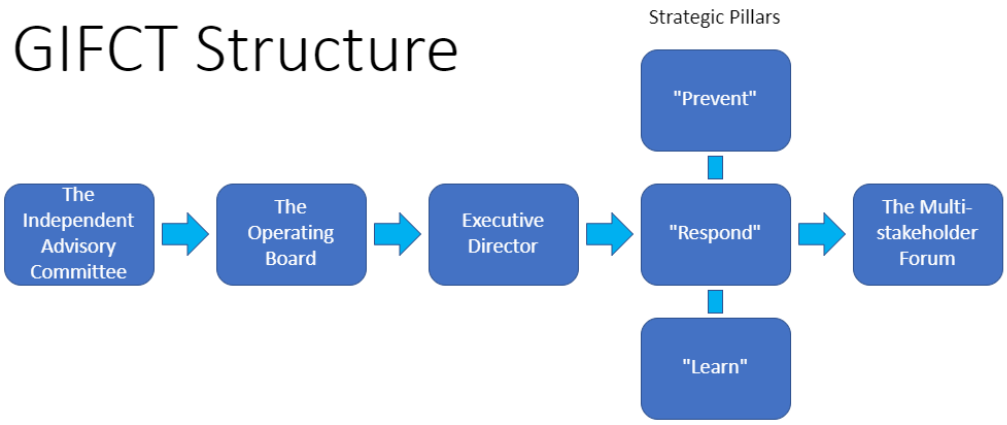

Figure 2.1: GIFCT Structure

is created for the original image the automated AI system can detect and block the re-upload. However, this is solely limited to ISIS and Al-Qaeda who are more easily identified through their branding and labeling of content. The likely reason that the same success is not shared when it comes to right-wing content removal is that it is more closely intertwined with politics and is, therefore, more difficult to clearly justify it as extremist content. In addition if a small change is made to this image i.e., applying a filter, converting it to a gif or video, the original hash's effectiveness is hindered. When factoring this difficulty in with the inevitable presence of data bias, false positives and negatives, operation in a pre-criminal space, and difficult appeals processes, the use of AI and automated decision-making in this context is not without significant limitation [78].

Having identified the successes and failures of implementing AI and automated decision-making in this context, it is worth briefly considering the second most prominent approach. This approach considers human moderation in regard to content removal. When dealing with the most prominent contemporary threat in online extremism being right-wing extremists it can be argued that the human element becomes integral to effective content removal. When regarding models used in AI systems, this embodies dealing with big data and coding, not complex arguments with meaning [55]. As previously stated, this is sufficient in regard to ISIS content but not for right-wing extremism due to the latter's close-knit political nature

of far-right ideologies [28]. Thereby, to unveil these complexities social media companies employ content moderators who deal with flagged content that is yet to be comprehended by AI. Thus, the solution to extremist content removal may lie in the form of a developed hybrid human-automated decision-making [113]. However, online extremist content removal extends beyond effectiveness. As throughout this literature review due regard has been paid to the legal standing and platform terms of service. An inherent gap in this discussion is the absence of a discussion of the various ethical issues associated with technological management [22]. Determining which of these three approaches are the most effective is explored in more detail in chapter five of this dissertation.

2.3.4 Conclusion

This sub-chapter has sought to develop a critical understanding of social media extremist content removal. In doing so, it became apparent this process is complex, open to perspective, and relies on a lot of parties doing their job well. But when compiled together a linear process begins to shine through. The process begins with a terrorist or extremist event, this instigates a political response which in turn leads to more human rights infringing legislation, which as a result provides a legal basis for social media companies to have more aggressive terms of service in which both humans and AI are left to delegate what content is removed. The AI and human elements of the hybrid human-automated decision-making both have limitations. However, due to the AI dealing with significantly more cases than the human content moderators, this is where the pressing cause for concern exists. And this cause concern is the absence of robust and specific ethical AI principles in which social media companies abide by.

2.4 Chapter Summary

This section has covered the topics that fall within social media hybrid human-automated extremist content removal. In doing so it became clear that for the most part, online platforms such as social media act as a catalyst for radicalisation, recruitment, and the spread of ideologies and propaganda as opposed to creating an evolved form of terrorism. This is not to say, however, that the former is not potentially dangerous and cannot contribute to significant harm. On this matter, the goalposts never stop shifting; the tactics continually evolve and so do the groups with the most popularity. In this sense, social media provides opportunities for all, from the spread of extremist content to sharing your favorite holiday pictures.

The response to the threat of online extremism is, however, more convoluted. As has historically been the case in creating change, it takes a significant event for things to move in the political sphere. And since the beginning of the millennium things have shifted considerably. From the findings in this research, it becomes increasingly clear that social media companies are the ones being held increasingly accountable. Governments are essentially forcing the big social media companies to solve a problem that no other government has ever solved: to eradicate extremism and fast. Under growing pressure, social media companies have been fast at developing new tactics and technologies and as a result, GIFCT has grown as a united front. In its wake hybrid human-automated decision-making is the tip of the spear in countering the threat. However, when producing results in such a short time frame corners are cut, and stones are left unturned. And in this case, it is a big stone, this refers to the implementation of ethical methods. Whereby, in implementing these big data automated systems, taking into consideration AI ethical principles compared to legal regulations is an angle that has yet to be explored in this context. And when dealing with freedom of speech and expression on such a scale the consequences and repercussions are significant. The likes of which can inform which content removal methodologies are most appropriate. Furthermore, the inherent importance of the users voice is is yet to be identified and examined. And thus, in the chapters following the methodology, these notions issue be explored and analysed in order to critique extremist content moderation on social media.

Chapter 3

Methodology

3.1 Introduction

The methodological approach adopted by this dissertation considers the amalgamation of several technical techniques. The overarching theme between all of these approaches is referred to as a mixed-methods research methodology. This approach is defined as "studies that are products of the pragmatist paradigm and that combine the qualitative and quantitative approaches within different phases of the research process" [100]. In line with this definition, the research employs two qualitative methods (a literature survey and a critical appraisal) and two quantitative research methodologies (an empirical study which includes a thematic analysis). Due to the complex and multifaceted nature of the topic of this paper, using different kinds of methods are essential in exploring and understanding the complexities that cannot be uncovered using just one methodological approach. The qualitative methods will are more commonly seen in this research area, whereas the quantitative methods are rarely applied to this context. Furthermore, the strengths of each of these approaches both complement one another and to an extent mitigate their disadvantages and limitations. In order to evidence this understanding, the methodological approaches section of this section will explain and justify the use of each of the methodologies in the content of this research. In doing so, each approach will evidence its necessity to relay an applicable and well justified methodological approach to the research. The combination of these approaches when applied to this research topic offers a unique and robust research project template which is not yet seen in the academic literature.

3.2 Methodological Approaches

As previously identified there are four methodological approaches that together constitute the mixed methods approach deployed for this research paper. These four approaches can be partitioned into two categories, quantitative and qualitative, thus, this format will be used to group the methods in the following pages. Each of these methods will be explained, justified, and applied to this paper in order to rationalise each of the methodology's inclusions.

3.2.1 Qualitative Methods

3.2.1.1 Literature Survey

The function of a literature survey is to develop and display an understanding of the relevant literature in relation to the research topic [21]. When referring to the surveying of literature sources, the following inclusions of written work were applied: books, academic papers, articles, and journal submissions; in addition to various governmental sources, news sources, and non-governmental/ charity sources. This approach's role and purpose are inherent to this research paper with regard to the multi-disciplinary nature of this research which combines social and computer science. As a result, surveying these varying forms of literature is necessary to best communicate the unique knowledge and information each of them offers. This mainly refers to the literature review along with Chapters 4 and 5 which required in-depth knowledge from a multitude of fields and sources.

In order to conduct this methodology effectively, several literature search techniques were employed. The first of which was to know where to find reliable and relevant academic literature. Thus, particular search engines were utilised including Google Scholar, JSTOR, IEEE, and VoxPol among others akin to these examples as they are reputable tools used by experts in the field. On said search engines, several specific search terms were used to find specific literature [80]. Selecting specific search terms in the context of this paper include 'AI Ethics' and 'Ethical social media content removal' were used and strengthened further by the adoption of Boolean operators and 'wildcard symbols' [91]. Boolean operators can be understood as terms used between keywords used to narrow the scope of results, for example using terms such as 'and', 'or', 'not' [23]. For example searching for 'AI ethics' and 'Online extremism' to search for both topics simultaneously. Wildcard symbols refer to applying variations of terms such as varying tenses or spelling variations [5]. This included using americanised teminology to broaden the scope of the search such as replacing a 's' with a 'z' in radicalise. Thus concluding

the techniques and methods used throughout the literature survey element of the research.

3.2.1.2 Technical Appraisal

The technical appraisal element of this research applies to the 'Analysing Effective Extremist Content Removal' chapter. This methodological approach refers to conducting a technical review of a project which in this case is the process of content removal conducted by social media companies. The process of this reviewal technique draws light to the numerous and various parameters that contribute to the extremist content removal process used by social media companies to some extent. In the case of social media content removal, this pertains to the technology being used, the storage capacity of the systems performing this task, financial funding of a social media company to conduct this process, and the scale of the workforce of human content moderators among other factors in order to ensure technical feasibility. In doing so, this approach offers an effective mode of analysing literature in a more specific and singular purpose when compared to the aforementioned literature survey. As it offers a specialised template to provide more of a critical analysis of a given process or model. Which to reiterate, in the content of this paper refers to the process of social media content removal practices.

3.2.2 Quantitative Methods

3.2.2.1 Empirical Study

To complement the qualitative element of this research, the use of empirical research was adopted. Empirical studies can be understood as "a type of research methodology that makes use of verifiable evidence in order to arrive at research outcomes" [41]. Thereby, this methodology solely pertains to evidence-based research, however, it is not strictly qualitative or quantitative by nature. The survey used in this research is a frequently used method of data gathering when conducting empirical research. As a result, its presence in this paper is as a result of its robust nature, as is reinforced by the wealth of its used in other academic literature. Surveys are a commonly used technique employed for data gathering, commonly seen in the form of a set of closed and open-ended questions regarding the given subject area. However, in the context of this research paper, the same cannot be said. Extremist research very rarely works with data, and there is yet to be an example in the research that considers social media users and stakeholders perception of extremist content removal. Under normal circumstances, surveys can be

presented either on or offline. However, due to the COVID-19 pandemic during the time of data collection, the safest and most regulation-compliant way of conducting this research was to use online methods. Thus, the survey was created using the online platforms 'Google Forms'. To share the survey and receive responses, it was posted on a number of social networking sites such as Facebook, WhatsApp, and email. In turn, this also aided in developing that anonymous element of the survey, which would have been hindered by handing out physical copies.

3.2.2.2 Thematic Analysis

Presenting the data collected as a result of the aforementioned empirical study was organised into two sections, closed-ended and open-ended questions. The process of identifying themes in the closed-ended data spoke for itself through the use of simple visualisations. However, analysing the open-ended question required a specific empirical approach to label and code the broad spectrum of written responses. The technique utilised in this research is referred to as a thematic analysis. The work of Braun and Clarke was used to conduct said thematic analysis. Whereby, their six-step process for evaluating open-ended questions was simulated [18], these steps are as follows:

1. Familiarising yourself with your data,
2. Generating initial codes,
3. Searching for themes,
4. Reviewing themes,
5. Defining and naming themes,
6. Producing the report.

This process was selected over the methods posed by other variations of the technique due to its strength in rigorously processing qualitative data as apposed to quantitative data. However, this process in not void of limitations, as by using this method can result in a game of bias mitigation. Thus, in making the survey anonymous, ensuring the open-ended questions were not leading participants or forcing any narrative. This ensured that participants' responses conveyed their own feelings and not informing the survey of what they believe it wants to hear. By developing the survey with no expectations, this also aided in the processing of data with limited bias when it came to coding the written responses.

3.3 Ethical Consideration

To conduct this research, a light-touch ethical approval was submitted for survey component of this project. In addition the very nature of the topic of discussion for this dissertation i.e., extremism may be deemed sensitive by nature [73]. Thus, in combining both of these factors, undergoing an ethical approval process to some extent was deemed both appropriate and necessary for the completion of this research project. However, due to the lack of risk posed to the researcher and research participants, full ethical approval from the College of Science at Swansea University was deemed unnecessary. This was also reinforced by the limited scale and scope of the survey. In turn, ethical approval for this study was granted by the Swansea University College of Science Ethics Committee (STU_CSCI_143462_130820153253_2). Thereby, evidencing the ethical consideration made and acted upon within the context of this research

3.4 Conclusion

To conclude, this methodology section has analysed the adoption and employment of a mixed methods methodology. In doing so, the four elements that comprise this methodology have been explained, justified, and applied within the context of this research paper. These four individual techniques work to mitigate the limitations of one another, and in doing so provide a cohesive and effective methodology that combines quantitative and qualitative methods. One that has yet too be applied to the new and unfolding nature of the topic of this research. The necessity and the integral role this methodology plays for integrating both secondary and primary data is necessary to the process of produce rigorous research. In addition, the ethical component of this research was identified and met in order to account for the collection of primary data and research being conducted on a sensitive topic through achieving ethical approval. In combing all of these processes, this evidences the sound methodological considerations identified, adopted, and applied throughout this research paper which sets it apart from the rest.

Chapter 4

Analysing the Legal and Ethical Regulations of AI

4.1 Introduction

When it comes to extremist content removal through automated systems, the regulatory responsibility that social media companies face are for the most part self-imposed, as previously identified. However, it was noted that governments could contribute to an extent if there were any infringement on national or international legislation, the likes of which is commonly encompassed within the remits of domestic hate speech, extremism, or terrorism legislation. However, this is to look at extremist content removal on a surface level. If the scope of this topic were to be broadened and looked at on a macro scale there are other means of regulating such content. A domestic example of this in the UK is the Data Protection Act which was, ratified, implemented, and most recently revised and brought into effect in 2018 [65]. Both this legislative framework and its predecessor (ratified in 1998) were amended and updated in response to directives and laws set out by the European Union. Despite the Data Protection Act originally coming into force in 1988 [57] solely from a national incentive and not as a result of EU legislation. Regardless, the Data Protection Act 1998 was a domestic manifestation of the EU's Data Protection Directive, set out in 1995 [103]. And the Data Protection Act 2018 was implemented in line with the EU's General Data Protection Regulation (GDPR) which was also put into effect in 2018 [104]. Each of these legislative frameworks pertain to the same subject matter, the processing of personal data in addition to the free movement of such data. Both of these legislative frameworks ensure that the legal protections of privacy are not

a secondary objective, but rather protections of 'privacy by design' and 'privacy by default' as clearly stated in the GDPR [104]. Thus, the Data Protection Act 2018 can be understood as the UK's implementation of the GDPR. This ultimately translates to organisations including social media companies having to plan how a user's personal data will be processed for it to be able to pass through the platform both safely and securely. The importance and relevance of data protection legislation is paramount due to its relevance to the deletion of personal data used in AI extremist content removal. Although legislation has significant remit over Data Privacy and its close link to data protection, there are also ethical considerations in play. AI ethical principles fill a middle ground between a companies policies and a particular government's legislative framework. And act as a voluntary gesture of goodwill; in this context, such an ethical consideration could take the form of a collection of ethical principles that state measurable ways and factors that contribute to the ethical process of removing extremist content using AI. Although each of these approaches have been considered individually, comparing the two and applying them to an extremist content removal context is an angle yet to be explored in the academic literature,

Thereby, this chapter will consider the protections of privacy from the aforementioned legislative frameworks followed by the legal protections freedom of expression which is synonymous with data privacy when considering content removal of any variety. This will be followed by an analysis of the UK's national ethical AI principles being applied to the processing of personal data and protection of privacy. In doing so, identifying both the legal and the ethical regulations of processing personal data which consider the implementation of an AI framework within the context of protection of privacy and freedom of expression. As a result, this chapter will argue that although legal regulation has a substantial effect in regulating social media companies when it comes to AI content removal; as previously stated in this paper, there is perhaps nobody better suited than the social media companies themselves. Consequently, adopting and/or developing robust AI ethics principles in conjunction with meeting legal requirements may be the optimal route to conducting extremist content removal. Thereby, adopting and promoting transparent ethical considerations when conducting extremist content removal through AI systems may be deemed essential in the protection of privacy and freedom of speech.

4.2 The Legal Protection of Privacy

4.2.1 Introduction

The legal protection of a users privacy on a digital platform can be presumed to be a contemporary issue. However, it has been 36 years since the UK's Data Protection Act 1984 [57] was ratified which did just that, before swiftly being updated and superseded by the Data Protection Act in 1988 [57]. As previously noted this act has since been revised on several occasions to deal with the complexities beyond just introducing rules of registration and rights of access to data that regard said individual. The most recent of which is largely a national manifestation and supplementation of the GDPR. Thus, the following section will identify the more notable features in the European Union's 2018 GDPR, and where appropriate and relevant reference the UK's variation of the regulation.

4.2.2 General Data Protection Regulation

The GDPR came into effect in 2018 where it implemented the toughest privacy and security law in the world [105]. Despite its coming into effect in the EU, its remit stretches throughout the rest of the world on the basis that a person or organisation targets or collects data related to someone within the EU. It is worthy of note that this regulation is not to be taken lightly, as failure to abide by its legally binding regulations as a fine can be levied against an organisation of up to 4% of its annual turnover. Which when applied to a large company such as Facebook or Twitter translated to tens of millions of pounds. As a result, social media companies are documenting their efforts to show their compliance with the GDPR as data controllers [38,112].

Throughout the GDPR several key terms are referred to. In order to understand the regulation and the following application of it to automated extremist content removal, it is necessary to identify these. Thus, the following list found below comprises several of these key terms necessary to this paper that can be found in Article 4 of the regulation:

- Controller: the company/person who decides how and why personal data is processed (in the context of this dissertation it would be the social media companies, i.e., Facebook and Twitter)

- Data subject: an identifiable living individual (in this content the data subjects would the users of the social media platforms)

- Personal data: any information relating to a data subject (this could consist of image, or text post on a given platform)
- Processing: doing anything in relation to personal data (for example, passing an image through an AI system before it is allowed to be uploaded to the platform)
- Processor: the company/person who processes personal data on behalf of the controller (this would also be the social media companies, i.e., Facebook and Twitter)
- Recipient: a company/person to which personal data is disclosed (this could include other social media companies if data is shared or through a unified body such as the aforementioned GIFCT)

To rephrase where this regulation applies using this GDPR terminology, the remit of the GDPR includes anyone who is processing and/or holds the personal data of a data subject within the European Union as stated in Article 2. And as AI and automated systems pose several data privacy challenges, it falls under the remit of the GDPR. Through using the previously identified hash system GIFCT uses to identify extremist images as an example. The most notable privacy challenge for this example is that it has to collect additional personal data to the database which is then used in the AI technology if it is deemed to be extremist. However, the GDPR states through Articles 13 and 14 that data subjects have the right to be informed. These articles state that when using and collecting personal data, users must be informed about what the data is being used for and to not use it for an alternative purpose. Therefore, using AI systems to remove extremist content suggests an issue of scope. This may be deemed the case as an effort must always be taken to minimise the personal data held by a social media or in this case GIFCT, at every step of a given process.

In addition to limiting data collection and storage, the GDPR also sets out the requirement that said data cannot be held indefinitely. Data storage must be put in place with limitations on the duration of its storage, as set out in Article 5 (e). Furthermore, following data collection a data subject can request information regarding what personal data of theirs is being held, and what it is being used for. It is the responsibility of the data controller to provide said information or even to remove said data upon request. This is accounted for in the scope of Article 17 which entails the right to be forgotten and erasure without delay. However, in the case of the Data Protection Act 2018 the definition of deletion cannot be found. The importance of deletion is essential in complying with the right to erasure in addition to ensuring personal data is not

used for any additional purpose beyond what was originally stated. However, deleting digital data can be difficult with information being stored in several places, unlike traditional paper stores where the original document can be located and simply incinerated. Using the GIFCT database as an example, GIFCT may have their own database, however each of the companies that use this system may have that same database in their own servers. Thus, deleting data on one does not necessarily guarantee its erasure on another. Thus, the ICO (who regulate data protection in the UK) will allow data controllers -in the case that the deletion of data may not be possible- to put said data 'put beyond use', as long as it can be accurately identified [69]. However, it is unclear if a user should be made aware of this caviate or if these is a loophole in storing an individuals data without having to inform them upon request. Thus, it may be the case that the GDPR places undue stress on companies trying to develop effective content removal systems, as it puts into place almost unattainable and undeniably difficult measures for efficient content removal methodologies such as automation and AI. To expand, if an image used to train a model needs to be erased, every time this takes place the AI may have to be retrained. Which as a result is both demanding of resources and can contribute to hampering the effectiveness of a content removal model.

As a result, data controllers must be able to identify and access this data for a single individual even if a system is operating on a global level such as Facebook, who as of March 2020 have accumulated an average of 1.73 billion daily active users [37]. In other words, Facebook must find a needle in a haystack with 100% accuracy on demand. Furthermore, through this inherent limitation on privacy and by extension data, the quality of security achieved by automated extremist content removal is certainly hindered to an extent. This may be deemed the case for several reasons. Firstly, AI and automated systems seek to benefit from accumulating as large and quantity and as high a quality of a training data set as possible. As the quality of a systems training data quality reflects how much it can learn thus, the more accurate and effective the results that it can return. The more dimensions said data set acquires, the stronger the model, thus, the longer the data is held the more any historic correlations can be explored. The likes of which is an important and telling dimension and data quality to consider in the context of extremism [93]. Therefore, by extension of the defense of privacy achieved by the GDPR, its impact might stretch to resulting in the simplification of the automated methods such as AI by social media companies. Which by extension is likely to result in non-optimal extremist content removal. Thus, this clearly lays out the dichotomy between privacy (achieve through privacy protection legislation) and security (achieve by automated content removal methods).

With the intended nature of the GDPR being to protect privacy, acquiring transparency from data controllers is necessary by extension, as set out by Article 5(1) of the GDPR. This places undue pressure on those looking to implement AI and automated systems [30]. To increase AI transparency with the GDPR may mean to hinder the quality of AI. To continue using the GIFCT hash database as an example the quality of the data as previously established is hindered by how long the data can be held, how it is acquired, and how accessible it is. However, AI in this case is likely to be explainable as images are removed if they match the banned images list found in the database. However, if this AI were to be expanded and developed to remove content based on learning from the images in the database using a Black Box Model model for example, the GDPR would limit its effect. As the decision to remove said content is required to be explainable, which in the case of black box models is far from an attainable requirement (as insinuated by the nondescript nature of the catch-all term). This limitation placed on those looking to implement AI and other automated system is also extended to the features they can extract. This limitation is put in place by Article 9 of the GDPR which prohibits discrimination through the use of data for the following:

> "... the processing of personal data revealing characteristics such as racial or ethnic origin, political opinions, religious or philosophical beliefs..."

All of which may be significant factors in a large variety of research areas. Not only to the area of identifying online extremist content that this research focuses on but also research in the medical sector and other avenues of the security sector alike. Thus, continuing the theme that the GDPR in principle makes the implementation of automated content removal methods to a broad spectrum of fields unquestionably difficult.

However, the effect the GDPR has on the implementation of automated methods for extremist content removal is currently a matter of how this regulation is interpreted [30]. With an inherent lack of case law to reference, the interpretations are left unclear and with room for interpretation comes room for misinterpretations. One could argue that the legal protection of privacy, in this case, proposes a dilemma. A clear trade-off between privacy (provided by the GDPR) and security (provided by effective automated extremist content removal). This is not to say, however, that the GDPR prohibits the use of AI. There are exemptions to several of these requirements, including where providing the information to the individual would render impossible or seriously impair the achievement of the objectives of the processing, where it would be impossible or provide disproportionate effort. However, akin to the regulations themselves

the exemptions are subject to interpretation. Therefore, these exemptions are equally limited in scope as the regulations themselves until case law begins to establish a broader understanding. To develop on from a theme uncovered in the previous chapter; to meet the requirements of the exemptions set by the GDPR will depend on who the technology is applied to. For example, these exemptions may be more easily met when it comes to the a group such as Al-Qaeda and ISIS, but this may be less the case when it comes to right-wing extremists such as Britain First and Generation Identity, where guaranteed outcomes are not so easily determined, the proportion of effort is difficult to determine and risk of harm is variable. However, more clarity on this matter will begin to present itself in time and when the potency of these regulations will become clear as the courts produce their verdicts and the matter is then clarified. In the meantime, where content is being removed freedom of expression (as it is understood in the UK) may be under infringement. Thereby, not only is extremist content removal subject to Privacy legislation, it also finds itself at loggerheads with freedom of expression legislation.

4.2.3 Protecting Freedom of Speech and Expression

Beyond the protection of privacy, extremist content removal also has to tread lightly when it comes to infringing on freedom of expression. social media companies essentially walk the fine line between free speech and hate speech. The difficulty in walking this line is that one protects the safety of other and the other limits and infringes on a person's human rights. In the UK article 10 of the Human Rights Act 1998 states that freedom of expression is a limited or qualified human right [58]. This right can be interfered with if it meets a legitimate aim, such as the protection of other people's rights, for national security, public safety purposes or it prevents crime. Thereby, there is a clear legal basis to limit an individual's freedom of speech in the UK that provides a safety net and legal basis for extremists content removal systems. However, in using this legal basis, social media companies and governments can be seen to be taking liberties in moderating content without transparency, notice, or due process [106].

In addition to these exclusions and exemptions, freedom of expression can also be hampered if it is to be categorised as 'hate speech' in the UK. Hate speech, however, is a catch-all term in this content. The is no hate speech act that defines its parameters, rather it can be understood as the collection of limitation on freedom of expression for committing a crime. The most recent account of hate speech legislation can be found in The Terrorism Act 2006 [60] which criminalises "encouragement of terrorism". Additionally, Section 127 of the Communications Act 2003 [59] makes sending a message through a public electronic communications

network that is considered grossly offensive, or of an indecent, obscene or menacing character a criminal offense. And finally, section 4 of the Public Order Act 1986 [56] makes "threatening, abusive or insulting words or behaviour that causes, or is likely to cause, another person harassment, alarm or distress" a prosecutable offense. All of these examples individually hold significant implications for freedom of expression that is not to be taken likely given the scale of this context [106].

What each of these examples provides are ways in which social media companies can govern their own content moderation on their respective platforms without the essential element of transparency that is necessary to a achieving fair and effective content removal process. This is not to say that social media platforms should not moderate the content on their sites. But platforms can all too easily censor valuable freedom of speech and expression without due process. Combining this hate speech legislation with the limitation on freedom of expression in addition or aggressive or non-specific content removal terms of service, social media companies are free to limit freedom of expression without persecution. This becomes even more apparent when operating in the pre-criminal space. While the UK government has set out these limitation, balance is achieved by those charged with infringing this civil liberty with a fair trial that offers due process. Social media companies on the other hand get to remove content and thus limit freedom of expression before it has been uploaded, without delay and on a global scale. The majority of society wants a platform free of extremist content, however nobody is a winner when social companies censor online speech without transparency, notice, or due process [44].

4.2.4 Conclusion

The legal protection of privacy quite clearly plays a significant role in regulating the implementation of AI and other automated systems. The GDPR and by extension the UK's Data Protection Act puts in place strict measures with even stricter punishments/incentives depending on your outlook. It can be understood without question that the GDPR is effective and concise about how it maintains users privacy and lives up to its 'privacy by design' priority. However, it has become clear that the strict regulations it puts in place reinforce an already present narrative that suggests you cannot have maximal privacy and maximal security simultaneously. In other words, to have more of one means less of the other. And in the case of these regulations and the resulting legislation's, when it comes to the legal aspect of automated content removal, privacy is quite clearly placed before accuracy, safety, and security. This is not to say however that all forms of regulation of AI and privacy may be one in the same, ethical

regulation of AI may have a different story to tell. Conversely, the strict protection of privacy juxtaposes the protection of freedom of expression, which is far less thoroughly safeguarded. Which may be to the dismay of the right of the user, although the benefit of effective content removal to some may out-weight this cost.

4.3 The Ethical Protection of Privacy

4.3.1 Introduction

It is without question that ethical considerations should always take place when considering anything with an impact on humans. And the inherent trade-off between privacy and security found in the legal protections of privacy makes ethical considerations all the more necessary. Thus, this avenue will be explored specifically in reference to AI ethics principles which now populate discussions around the various implementations of AI.

4.3.2 AI Ethics Guidelines and Policies

Following on from the legal protection of privacy and regulations of AI it is necessary to also consider the ethical protections and regulations. Looking at computation through an ethical lens is not a new angle in academic literature, however, the recent surge in AI ethics principles over the last five years is a modern manifestation of these long-established considerations. Naturally, this recent trend of AI ethics documents being produced by companies, governments, and unions alike has attracted attention in the academic and research communities. As has been drawn from the previous section, when it comes to the regulation of AI through legal processes, a lot is left to be desired in terms of clarity of how such regulation is to be interpreted and what approaches governments and the companies themselves expect. Thus, supplementary AI ethical guidelines may prove the necessary vessel to develop clarity and uncover the themes that are present in legal regulations.

The purpose of AI ethics is both human-centered, multifaceted, and in many ways diverse. Depending on which document is being referenced it is likely to have a different intended audience to the next with different motivations and end-users in mind. Prior to uncovering the broad narrative shared by these documents, it is first worth understanding what is meant by AI ethics more broadly; the said definition can be found below which effectively defines the notion [2]:

> "AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies."

In regard to AI ethics principles more specifically, these regulations can more commonly be seen and intended as a guide for the responsible innovation, design, and implementation of AI systems. Whether it is reflecting on their own ethical processes or recommending them to others, these guides often feature values, principles, and guidelines to develop transparency, aid in privacy, and assist the deployment of ethical AI systems both safely, and responsibly [2]. Such documents acknowledge the potential successes and harms that come with AI. Therefore, in the search for the optimum route regarding public benefit when it comes to both safety and privacy, AI ethics principles may be the vessel of choice in meeting this aim.

One of the key and most apparent differences between the legal and the ethical responses to regulating AI is who can produce such regulations. Legal frameworks have long been established on an international scale, however, there is no national ethics framework that produces all of the ethical guidelines. Equally, there is no vetting process for the release of said guidelines, thereby, when initially considering ethical regulations the authenticity and robustness are already called into question. By extension, this equally calls into question what the intention is for organisations of all varieties in developing and sharing such documents. Especially when considering the broad array of publishers of such AI ethics principles e.g., Microsoft, Google, the Alan Turing Institute, the UK House of Lords Select Committee, the European Commission's High Level Expert Group on AI, IBM, and Springer. The recent boom in the publishing of the AI ethics principles is undeniable. In a study on AI ethics principles documents in 2019 that identified 84 documents matching the description [72]. One of the more notable trends was the boom in the number of publications being released with 88% being released after 2016. Over 50% were made up of the top three categories which were private companies (22.6%), governmental agencies (21.4%), and academic and research institutions (10.7%) almost entirely from the western regions of the globe. This surge clearly evidences the embryonic nature of such regulations, which provides a stark contrast to the development of legislation. Although the reputation of those producing such documents may outweigh the new, untested and likely fluid development of such policies and regulations.

Despite the development of such documents being a clear step in the right direction, the integrity of such documents needs to be called into question. As identified in the AI ethics

definition above, there is an innate presumption that the policies in these documents have the correct qualities that would translate into practical implementation, however, this is often not the case [10]. Several academics have noted this nature of impracticality in such documents and have claimed they are a method of keeping stakeholders content whilst simultaneously working to delay companies from adopting practical AI regulations [83]. In addition, a 2019 paper that explored the ethical nature of AI ethics principle documents, one of the several findings in the paper determined that regardless of the organisation which produces a said document, there are almost unanimously always vague and superficial principles present [54]. As previously identified these documents have different stakeholders and end-users in mind and despite this, these principles that are designed to inform the boundaries of ethical guidelines and frameworks are commonly found without implementable or practical measures, checks, or balances for those developing, creating or regulating creating ethical AI systems. It can, therefore, be deduced that there is anywhere upward of 84 AI ethics documents that do little for the protection of privacy and the development of ethical AI, and instead provide nuanced terms and topics that keep governments and stakeholders content. Ultimately this brings into question both the purpose and benefit of such documents and to what extent are their contents produce abstract problems with the absence of technical solutions. Despite the negative perception that some academics have taken to these documents it certainly does not make them redundant. The medical sector for example provides an adequate case of where ethics principles are implemented into practical principles that are used to inform and regulate such processes [16]. What this suggests is that the embryonic nature of these principles in this field is apparent. Thus, developing these documents to make them both more measurable and implementable is an essential next step. Carrying out measures such as these to strengthen the documents may leads to bringing the gap between the harsh demands set out by legislation and the realities associated with effective AI implementation.

It may be determined that there is a slight shift in focus compared to the legal and ethical approaches to regulating AI. Where the GDPR engages in privacy by design, AI ethics principles seemingly engage more in humans (or political motives) first and privacy by extension. Consequently, in an effort to uncover what AI ethical principles are being put in place the following section will follow the same theme as that found in the previous section. Whereby, both the UK and the EU's AI ethics principles will be analysed in their regulation of AI in addition to the implications of the protection of privacy. In doing so, each of the respective organisation's principles will be considered and compared in addition to uncovering the measures put

in place to monitor the effect of each principle as and where said these sections are present. As a result, this will determine the effectiveness and impact of these regulations on AI extremist content removal.

4.3.3 AI Ethics Principles

Having conceptualised the recently emerged field of AI ethical principles, both strengths and limitations have been identified. Thus, in line with the legal protection of users' privacy using the UK and EU as examples, the same sources will be considered for the ethical protections of privacy. Baring these in mind, the following three chapters will consider each of the two different AI ethics guidelines. As a result, the ethical principles will be identified, compared, and analysed in terms of what they may represent in terms of the protection of privacy. This will be structured in three key sections, the stated purpose and audience of the principles, the principles themselves, and the application and measurement of said principles.

4.3.3.1 Purpose and Audience

In the document produced by the UK Government [64], AI ethics are referred to as the ethical building blocks needed for the responsible delivery of an AI project. The foundations of these principles are built on the understanding that AI ethics emerged from the need to address the harms AI systems can cause namely, applications that invade users' privacy [66]. The principles are aimed at anyone involved in the design, production, and deployment of an AI project including but not limited to data scientists, data engineers, domain experts, delivery managers, and departmental leads. The document sets out four principles that were developed in correspondence with the Alan Turing Institute's public policy program [1] who themselves have their own AI ethic principles [2]. In partnering the principles with other documents as referenced above in addition to the UK's AI ethical guidance document [64], the government implies the difficulties in feasibly protecting privacy through AI principles.

The document produced by the EU is a far more diverse and multifaceted approach to ethical AI than what is commonly seen in such documents [34]. To achieve trustworthy AI there are three components, which are to be met throughout the entirety of the system's entire life cycle: which in broad terms are to be lawful, ethical, and robust. Thus, the ethical principles found in this document are supplementary as opposed to being the fundamental quality as they make up a third of the total requirements to achieve trustworthy AI. The principles are primarily aimed at the developers, deployers, and end-users of the system, in addition to the wider society

which in other terms means almost everyone. However, each category of stakeholder has a specific role, with developers the expectation as that they implement and apply these principles, deployers are required to make sure that the systems uphold principles and end-users should be informed about the principles. Of which there are four, although this could be expanded by a further seven ethical requirements, however, this is not in the scope of this section.

4.3.3.2 Principles

The four principles previously referred to in the UK AI ethics principles are as follows: fairness, accountability, sustainability, and transparency. The first principle of fairness is made up of four subcategories including data, design, outcome, and implementation fairness in order to meet a minimum level of discriminatory non-harm. The second principle of accountability requires that AI systems are completely answerable and auditable. This is to be achieved through an ongoing chain of responsibility with accompanying oversight. The third principle recognises sustainability, this principle can be met through achieving safety, accuracy, reliability, security, and robustness in order to address the real-world impact of the AI system. The last principle addresses transparency; firstly through producing explainable model performance and then to justify its performance based on ethics, trustworthiness, and nondiscrimination/harm. These points are then expanded for further reference in the Alan Turing Institute's guidance on AI ethics and safety [1].

The four principles previously referred to in the EU's AI ethics principles are as follows: respect for human autonomy, prevention of harm, fairness, and explicability. The first principle of respect for human autonomy addresses the need for AI systems to augment, complement and empower human cognitive, social, and cultural skills. And not subordinate, coerce, deceive, manipulate, condition, or herd humans. The next principle regarding the prevention of harm entails ensuring dignity as well as mental and physical integrity are always maintained. And that AI systems are robust enough to ensure that they cannot perform their function for malicious purposes, however, there is no mention of privacy on this matter. The third principle of fairness largely entails developing systems free of bias, discrimination, and stigmatisation and to promote societal fairness through equal opportunity without deception or impairment. However, the recommendation that practitioners should respect the principle of proportionality between means and ends is notable as it leaves an open door when it comes to the battle between privacy and security. Finally, there is explainability, which requires that AI systems should be as transparent as possible in order to be understood both by those directly and indi-

rectly affected by its function.

4.3.3.3 Application and Measurement

When analysing the application of the four principles set out by the UK government, it is first worth reinstating their purpose. The guidelines were not developed as a standalone publication of processes or as a vessel for transparency but to provide others the tools and building blocks to implement their own ethical AI. The principles very much achieve this through their signposting to other topics in a way that makes a starting point but by design not an ending point. The recommended template provides a step-by-step for each governance action having a designated member of staff, targeted considerations, time-frames, clear and well-defined protocols. However, the only principle that may not negatively impact extremist content removal through AI is the sustainability principle. Whereas the fairness principle rules out the inclusion of discrimination which may ultimately lead to increased efficiency in this content. Give for example the correlation between white males associated with far right groups and Arabic males associated with groups such as ISIS and Al-Qaeda. Furthermore, both the accountability and sustainability principles require systems to be answerable and transparent, both of which is likely to the detriment of an AI system if it uses a black box model. In addition, little is mentioned in regard to privacy or to measuring the impact of these policies. Thereby, the purpose of the four principles is clearly broad with questionable benefit, and where these principles are more specific it may be argued that it is to the detriment of the removal of extremist content through AI.

In regard to the principles set out by the EU, the developers, deployers, and end-users of an AI system are the intended recipients. Having addressed each of the four ethical principles, how this aim is to achieve is raised into question. It becomes unclear how certain categories of the intended audience are to digest and act on the information. Thus, either the scope could have been limited with more specificity or several separate/more descriptive resources could have been developed for each intended recipient. In each of these principles, it becomes clear that privacy is a secondary concern as it can be found in the ethical requirements and not in the principles. In line with the UK's ethical principle, the EU principles may also find themselves between privacy and security. The third principle of fairness and the fourth principle of transparency are potentially counterintuitive to removing extremist content. As removing such content may favour a models accuracy over 'fairness' and discrimination. And revealing the decision making processed by these tools may give online extremists the upper hand. And in

regard to measurability, there is nothing in the way of implementation beyond hiring staff.

4.3.4 Sub-Conclusion

Based on the assessment of AI ethical principle documents, the ethical protection of privacy plays a nuanced and secondary or arguably tertiary to a non-existant role in the development of AI. In exploring the ethical impact on privacy, a harsh truth regarding the authenticity and utility of AI ethics principles has been called into question. Simply to recommend prioritising user privacy in such documents would be almost meaningless. With the black cloud that follows the legitimacy and motive of these documents in conjunction with the inherent lack of measurable and practical implementations. A page can be taken out of the book of the medical sector in having robust principles that hold their own unique obligation. It is the opinion of this paper that these documents have potential utility and benefit, however, they are in a significant need of reinventing with a more practical outlook that pushes privacy towards the top of the heap. As a result, their application to the implementation of AI and automated systems would be virtually obsolete; leaving the majority of the weight on self regulation and legal regulation.

4.4 Conclusion

To conclude this chapter on the analysing the legal and ethical implementation of AI and other automated content removal methods to the developmental stage of this area is necessary to highlight. Both pairs of legal and ethical regulations of AI have been released in the past five years. Thus, they are subject to the imperfections that come with regulations in a new field. When considering the legal regulations, privacy is clearly a priority, however, as a result, the quality of AI is likely to be hampered in order to meet these requirements. Convesly, freedom of expression may be infringed within legal remit which may tip the scale in the favour of security but away from the preservation of human rights. From this, it becomes apparent that supplementary AI ethical guidelines would in theory prove necessary in developing clarity and expanding on the notions the themes that are present in legal regulations. Which to develop on from, lacks the necessary case law to be fully understood in regard to the differences between what the law says and how it is applied and enforced. Essentially in this context, the law shows its age and leaves a lot of clarity to be desired in regards to users' protections of privacy and freedom of expression. The hypothesis was that ethical guidelines would fill this gap. However, in practice, these documents leave those who engage with them with more questions

than when they started. And do little beyond paying homage to the protection of privacy. The adoption of nuances, nonspecific, non-implementable, non-measurable principles does little in the way of developing clarity. Although, this is not to say that a more thorough example of these principles could not achieve this aim. If the producers of these principles would hear such criticisms and adopt a more practical approach to ethical principles, the likes of which can be seen in the medical sector. Then the pairing between legal and ethical regulations of AI and the protection of privacy still has the potential of being a harmonious amalgamation of those fields looking to achieve the same goal. However, the current state of affairs between these two fields is unclear, exists with questionable intentions and reinforces an age-old problem which is the trade-off between privacy and security. Thus, as a result of these conflicting regulations on extremist content removal, the following chapter will consider how this process can most effectively be conducted in regard to said regulations.

Chapter 5

Analysing Effective Extremist Content Removal

5.1 Introduction

Having analysed where the large portion of the responsibility lies in addition to both the legal and ethical constraints of social media extremist content removal; it is the purpose of this chapter to analyse how this content removal can most effectively be conducted in light of these constraints. As previously identified, the three key categories of content removal consist of automated, human, and the combination of the two hybrid-automated extremist content removal. Each of these three categories will be explored individuals, identifying, and developing on the practical advantages and disadvantages of each approach. These factors will include the required resources to perform, any legal or ethical implications, and requirements in addition to recommendations based on these factors. In doing so this chapter will evidence that identifying the most effective approach to removing extremist content from social media is not so clear cut. Rather, it relies on larger companies sharing resources with smaller companies, the justification of potential harms and the unique ethical implications associated with each approach. In short the correct method of removing online extremist content removal is a complex balancing act, where there is yet to be a perfect technique.

5.2 Automated Extremist Content Removal

The adoption of automated extremist content detection systems appeals to both ends of the social media spectrum. Small companies cannot afford to employ a large human workforce to moderate their content and large companies require scalable content moderation to deal with the hundrends of millions and billions of users that frequent their platforms. Automated systems can perform a number of functions, for example, the aforementioned Facebook systems that utilises AI to remove 99% of ISIS and Al-Qaeda content before it is even uploaded, in addition, such a content detection system can flag suspect messages, posts, and accounts. Not only are automated systems more feasible for smaller companies to adopt, but such systems and their corresponding databases can also be shared to strengthen them. A prime example of this can be seen in the previously identified hash system developed by GIFCT, which shares a database developed by the larger social media companies with smaller social media companies that join the forum. What this essentially looks to mitigate are the disadvantages that smaller platforms face which is the time it takes them to develop their own automated systems that come as a result of their limited financial resources and manpower. And with a topic that ebbs and flows like extremism, time has significant value, as a system that takes a relatively long time to develop may no longer be as effective upon the time of its implementation. However, pairing these technologies with automated systems that analyse behavioural cues may provide an effective way of combining automated systems to combat extremist content [113]. Behavioural cues can consists of factors such as an abnormal posting volume and hashtag hijackings. In this context, hijacking a hashtag refers to a process where a post uses trending hashtags to increase its range and viewability beyond the sources immediate audience. One of the stronger aspects of using automated systems is that the quoted accuracy ratings that support a given system as referred to above. Naturally, companies are chasing down the potentially unattainable 100% accuracy, however, how this statistic is calculated has not yet reached consensus [113].

The two aforementioned examples of the GIFCT and Facebook automated systems are representative examples of both real-time and retrospective automated filtering techniques. The GIFCT system is a clear-cut example of a real-time filtering system; whereby, in the time between a social media post is being uploaded and before it is officially posted the content is passed through the GIFCT extremist filter. The benefits of this approach is that it prevents known extremist content from ever being posted, however, when the platform is flooded with content, platform users will notice an increase in delay when posting content. Conversely,

retrospective filtering comes in to play after content has been posted, where the filter passes through all social media content in order to identify any extremist content base on a more up to date understanding. This approach is strengthened as it offers the most up to date filter to existing content, however, for this to work it means that for a period of time extremist content will exists on a given platform. There are clear and significant benefits to both of these approaches, however, when combined, the strengths of each of these methodologies is significant. However, the development of a system which mitigates the limitations may be deemed necessary. Certainly there is scope for future work on this topic which combines retrospective and real-time filtering to create hybrid filtering which adjusts depending on set characteristics. For exampl, if the filtering system were to be applied uniquely to each user based on a set of characteristics such as a calculated trust value. Where low trust value users are more frequently subject to retrospective filtering. Or, in reference to the previously identified behavioural cues, where a users posts an abnormally high load of content real time filtering could be applied and when low to normal levels of content is posted retrospective filtering can be applied later to deal with latency and capacity limitation. However these concepts are just that... concepts.

When companies quote 99% effectiveness on automated extremist content removal systems, this leads one to presume that 1 piece of extremist content would pass through the moderation net for every other 99 pieces that would be caught. However, this fails to factor in content that is falsely marked as being extremist (false positives) and content that is falsely marked as not being extremist (false negatives). Thus, 99% effectiveness fails to represent the content that the system is producing false postives and negatives constantly for the billions of data entries that run through it. In a study that anonymously interviewed GIFCT partner companies and law enforcement based internet referral units, two methods of measuring accuracy were proposed. The first considered examining the appeals rate, whereby the number of successful appeals would represent the average false positives, thus, representing the quality of the system. However, this fails to address the issue of calculating how many people appeal their content being removed and the quality of the appeal process itself. The alternative considers an internal reviewing process that uses a range of human moderators to re-label a random sample of the removed content to calculate the accuracy [113]. However, this is not applicable to platforms solely using an automated system. Thereby, even in cases where systems perform at remarkable accuracy rates, companies have not yet avoided primarily removing innocent content. When this is combined with the absence of a consensus method of measuring accuracy

and the different understandings of what can be identified as a true and false negatives, such statistics are to be taken lightly.

In the frequent case where false positives occur, the GDPR sets out the right to lodge a complaint with a supervisory authority in article 77 [105], in other words the right to appeal a decision made by a social media company. The difficulty in this setting is that without human moderation how such appeals can be dealt with is a relative unknown. This becomes emphasised when considering the scale of content seen by larger social media companies. And if said system was a black box, then explaining the decision making process may not prove possible. Considering this through an ethical lens may also prove troublesome due to certain systems such as the GIFCT shared hash database which acts in the pre-criminal space. Thus, when using solely automated content removal techniques, transparency may be the most crucial element of conducting the practise effectively, ethically and legally.

5.3 Human Extremist Content Removal

Many social media companies are hesitant to disclose any specifics on how they moderate the content on their platforms. However, in a series of blogs referred to as 'Hard Questions', Facebook outlined a broad narrative of how they are striving to counter terrorism [13]. Said blog series, briefly touches on the strengths and the limitations of automated non-human content removal (as discussed above) and thus, justifies the necessity for a human element to extremist content removal. As of March 2020, the most recent estimation is that in the U.S. Facebook has accumulated 15,000 content moderators, the likes of which are commonly employed via several third-party contracting companies [107]. Of this number, the most recently quoted statistic by Facebook states that just 150 of these are dedicated to terrorist content [13]. Which, if this is still an accurate figure would clearly evidence a significant disparity, that is disproportionate to the threat posed by online extremism and terrorism in proportion to the number of human content moderators. However, despite this disparity, Facebook has set the highest bar on this front, with companies such as Twitter failing to uniquely identify extremist or terrorist content as a separate matter in the content moderation efforts. In addition, 10,000 people jointly moderate YouTube and Google products; and Twitter have employed around 1,500 moderators. This may shine a more positive light on efforts made by Facebook, however, given the daily volume of contents being posted on these platforms, these numbers are simply beyond inadequate [4].

Facebook identifies the role of its content moderators as a necessary supplement to their

larger AI strategy. The broad job description of the 'content moderator' largely consists of reviewing the millions of posts which can take the form of written posts, images, and videos that have been flagged by users of the platform. The three possible outcomes of this reviewal process are to decide whether to ignore, escalate (forward to management), or delete these posts depending on whether or not they violate Facebook's terms of service [13]. The decision making in this process is informed and supplemented by two key components. Firstly, content moderators must attend a two-week training course and secondly, moderators are provided with manuals with the companies policies on all matters and topics that are likely to be encountered [68]. What this effectively provides is a complex decision-making tool provided by humans, to what is an equally complex topic which is seen as essential to the task of removing extremist content [27, 68].

However, despite the strengths of tasking humans with understanding the complexities associated with removing extremist content, the realities of this role are far less reassuring. One of the more glaring concerns with human content moderation is its scalability. On social media platforms such as Facebook, Twitter, and YouTube in a single day billions of posts are accumulated. Through the user's flagging system and AI screening systems, Facebook gathers over three million accounts of reported content [4, 74]. An interview with a Facebook content moderator revealed that moderators will likely review posts up to 400 times per day [84]. If every content moderator were to meet this target every shift, then Facebook would require 7500 content moderators to work every single day. However, this statistic factors AI screening to flag posts. If this were not factored in and humans had to review an average of the 350 million images along being published every single day according to Omnicore [86]. Then this would require 875,000 content moderators to review 400 pieces of content every single day. Not only is this not financially feasible, but it is also not the silver bullet when it comes to accuracy. Current estimations predict that for every 10 posts reviewed, one is acted on incorrectly [74]. Thus, if this workforce were to be gathered, it would mean that every day, from the 350 million posts 35 million would not be responded to accurately. This excludes stories, posts, videos, comments, and the various other forms of uploads that the site uses. In the case of inaccurate results and an appeal by a user, this presence of a human conducting this process is necessary for deciding where the line is to be drawn in the grey areas posed by a companies terms of service. However, it may, therefore, be fair to say that this is not an effective method of regulating extremist content for large platforms. However, even if this method were to be adopted by a smaller social media platform, it is unlikely that said platform

would have the finances to sustain such an operation.

When comparing the role of human moderators to automated content removal one of the main advantages is the complexities that human decision making can comprehend to a standard to yet to be achieved by machines systems. However, when companies are claiming 99% accuracy in removing Al-Qaeda content, this provides a stark comparison to the 90% accuracy achieved by human moderators. Although calculating this accuracy figure is subject to scrutiny and debate, there are several human-centered factors at play on this matter to explain/justify this scrutiny. Firstly, the vast majority of human content moderators do not focus on one topic, rather they are exposed to all forms of extreme content from child pornography, animal abuse to terrorism [24]. Thus, there is a broad array of topics that these individuals have to cover make achieving higher accuracy levels more challenging. In addition, with experts struggling to determine what extremist content is, an employee with a manual and a two week training program is not likely to stand a chance. Secondly, where a social media platform user may come across extreme content on an infrequent basis due to content moderation, the content moderators themselves are continually exposed to such content for long arduous hours on a weekly basis. It is without question and beyond a reasonable doubt that this may be deemed unethical in certain lights as it can and does cause actual harm to those that perform this role. Not only will this reduce accuracy but it is also to the detriment of the health to those individuals who conduct this role [24, 68, 84]. The results of the long-term psychological strains associated with this role leaves an individual in some cases with PTSD and various other mental illnesses, but mid to long-term exposure to extreme content could hypothetically change how the threshold of what content moderators deem as extreme enough to be acted upon [84]. Workers have claimed to notice a change in their sense of humor as content that is not normal becomes normalised, and so their views become increasingly unpopular and extreme which as a result interferes with their decision making [31]. Despite the harms associated with this difficult role, the workers at Facebook are looked at as a lesser tier of workers by their peers and are also poorly financially compensated [84]. This low level of appreciation may also take an effect on their mental health and as a result on their accuracy levels. And to work effectively amongst all of these detrimental factors, content moderators at Facebook are provided with a short two-week training program [68]. This calls into question the ethical nature of social media companies that use human content moderators. The ethical concern in this context is referred to as John Mill's principle of harm [108], which raises the question of whether the inherent and unquestionable harm done to content moderators out-weight the prevention

of harm on far larger platform user-base. In addition the removal of harm may fall into the category of being a 'beneficence' (i.e., doing good for others) which may be separately applied. What this seeks to outline is that even with humans who have a far larger grasp of extremist content removal, the process of content removal is a ethical balancing act.

5.4 Hybrid-Automated Extremist Content Removal

Having analysed both automated and human content moderation individually, hybrid-automated content removal essentially considers the amalgamation of both of these methods, pairing human content moderation with automated systems. On the surface what this does is it combines two imperfect methods which as a result effectively doubles the number of methodological limitations. In addition this approach effectively two methods that compliment one another work to limit the disadvantages of the respective approach. For example, as previously stated automated systems can detect suspect messages, posts, and accounts. Instead of using such systems to work independently to remove or suspend such content, it is instead flagged to human moderator to undergo a reviewing process. Thus providing integrated checks and balances through a process conducted by humans, if a decision is made to remove the content and this decision is appealed by a user, this process can also more appropriately be conducted by humans. Who are for more adept at reasoning with grey-zone content, that by its nature verges on violating user guidelines. Furthermore, with automated systems providing the first line of defense, this ultimately limits the harms that human content moderators are subjected to. Due to these attributes the scaling of this form of content removal can be both up and down-scaled to the needs of a given social media platform.

However, akin to the previous methodologies this approach too is no silver bullet. As previously highlighted, content moderation is a resource-taxing requirement which the majority of smaller platforms struggle with. Most smaller social media platforms are not using any automated systems for content moderation and struggling to facilitate the large scale employment of human moderators [113]. The concept of combining the pair of these approaches puts twice the amount of stress on the smaller companies. Whom without support struggle to effectively adopt even one of these methodologies. This burden on smaller platforms is then partially offloaded on to the larger platforms who by sharing their resources as a collective can come to the aid of the smaller scale companies. By sharing automated systems and databases to develop automated content removal, this is yet another example of the benefit of the GIFCT hash sys-

tem. The benefit of sharing databases developed by larger companies with smaller companies can be applied to the 'Justpaste.it' case. In this case, the small information-sharing website ran by a single polish student was hijacked by ISIS supporters [40]. As a result of the crackdown on extremist content removal by Twitter a process of dispersion took place, whereby, groups like ISIS were forced off of Twitter and on to smaller platforms such as Justpaste.it and Telegram where content removal is either not prioritised or far less developed. Thus allowing ISIS supporters to upload images of executions, beheadings and massacres as a essential part of the group's social media operation with limited resistance in two quick clicks. Consequently, making the sharing of resources a growing need in the fight against online extremism.

Hybrid-automated extremist content removal unlike solely human or automated approach can be applied in varying proportions. Either prioirtising automated content removal, human moderator or equally applying both. Prioritising AI with human supervision may be deemed the most effective approach in order to play to the strengths of the scaling of both of these methods. However there is a gap in the literature concerning how content that has been flagged by automated systems is prioritised. For example whether a human content moderator receives content that is ordered in a chronological order, if certain teams are to be allocated certain topics, if categories like extremism and child pornography are to be prioiritised or if content is entered into a pool and randomised to varying extents; or finally prioritised by a metric of extremism. These are factors that are yet to be explored in the academic research, which by extension leaves room for future research on this matter. Having spent this chapter critiquing the broader advantages and disadvantages of these three approaches, more answers may be found in a micro level analysis.

5.5 Conclusion

It becomes immediately clear that following the critical analysis of each of these three technical methods of extremist content removal there is no silver bullet. In its place is a delicate balancing act of maximising the strengths of these approach's in order to limit each of their individual limitations. In order to do this effectively, as a a result of this critical analysis this paper finds hybrid-automated extremist content removal to be deemed the lesser evil if conducted effectively. To utilise one approach without the other is ultimately to the detriment of the platforms user base. Which by extension raises several ethical considerations due to it causing actual harm to people. Although it is unclear what exactly is necessary to improve

this process -which to be clear is in need of improving- remains uncertain. However, when regarding a topic which exists to facilitate human interactions a critical component has yet to be identified. The solution to this problem may be found by asking the right question to the end-user of the technology. Beyond the legal, ethical and accuracy-centred considerations, perhaps the largest stone has been the one left unturned, the human-centred stone. Thereby, by asking social media users on their opinion of content removal, their understandings of extremist content and experiences with it may inform how hybrid-automated extremist content removal is conducted and which elements are prioritised.

Chapter 6

Social Perceptions of Extremist Content Removal on Social Media

6.1 Introduction

Having analysed the legal, ethical, and practical implications of extremist content removal, this chapter seeks to explore the inherent human element that has yet to be accounted for in this context. Following the conclusion that hybrid-automated extremist content removal may be deemed the most optimal route for social media content moderation, identifying which method is prioritised and what public expectations and perceptions are essential in determining this. Thus, this section considers an online anonymous user study consisting of a combination of 14 optional open and closed-ended questions, completed by 90 individuals. The structure of which can be located in Appendix B at the end of this document. Unless stated otherwise all of the participants would have answered the question being analysed. The purpose of this is not to identify correlations between views and opinions with age groups and ethnicity. Rather it is to gauge a broad scope of views and opinions regarding the matters identified in this dissertation, including privacy and security, defining extremism and content removal accuracy ratings, etc. This is done to make informed decisions about societally accepted technical solutions that take a human-centred approach to be explored and possibly implemented in future. Thus this section will consider the results of each of these questions individually followed by a discussion sub-chapter to consider these findings within a proportionate scope based on the scale of this study.

6.2 Survey Analysis

Question 1: Out of the following four definitions of extremism, which if any do you think is the most accurate?

This question provided five potential answers, four definitions of extremism, and then a 'none of the above' option. Out of the 90 responses, only 1 selected the 'none of the above' option (1%). This reinforces that the four definitions were adequate in meeting people's understanding of the term. These definitions are as follows:

- Definition 1: Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order [97].

- Definition 2: Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms [32].

- Definition 3: Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist [61].

- Definition 4: People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behavior is called extremism [119].

Despite or as a result of the inherent differences between each of these four definitions, there is a relatively even split of votes between each of them. In order of frequency, definition one was the most common with 26 votes (29%), definition two closely followed with 2 votes (28%), definition three was the third most common with 17 votes (23%) and definition four was the least common with 17 votes (19%). Therefore, from the beginning of the survey, it

can be stated that this sample offers a broad range of view, understandings, and perspectives on extremism.

Question 2: Do you use social media at all (e.g., Facebook, Twitter, etc.)?

Out of all 90 participants, 78 of them stated that they themselves use social media in some facet (87%). Whereas 12 stated that they do not, which constitutes 13% of the sample. This could be interpreted as 13% of the sample being detrimental to the results, however, due to the scope and relevance of social media during this time the insights these people have to offer with a perspective from outside of the social media box is just as necessary as the majority who dwell within said box. Furthermore, this question does not ask whether a participant has ever used social media, thus, the 13% who do not use social media may have used it in the past and/or have observed the use of social media through their environments and various social groups. Finally, this also confirms that the participant's sample is not disproportionate to the wider society, the majority of which uses social media to some extent as is effectively represented in Figure 6.1 located below.

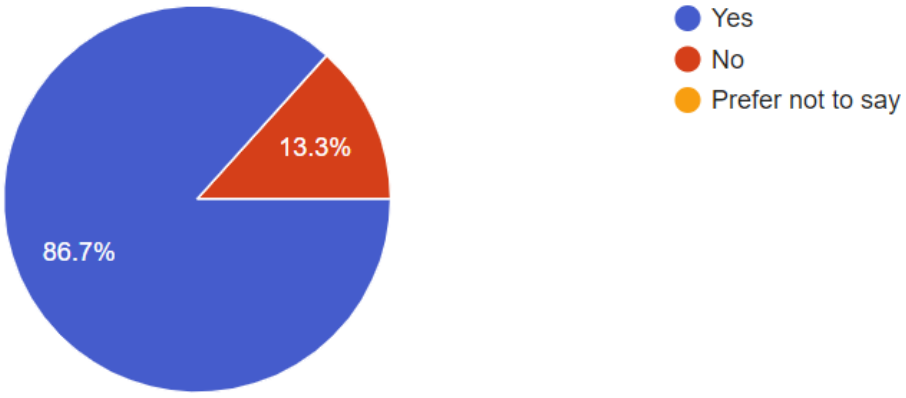

Figure 6.1: Question 2

Question 3: How often do you use social media?

All 90 participants responded to this question, however, 1 answer had to be discarded as the answer could not be categorised, leaving this question with a total of 89 responses. The resounding majority of these fell into the 'Daily category' which totalled 59 out of the 89 respondents (66%). In order of frequency, this is followed by 'Hourly' with 15 respondents

(17%), 'Never' with 12 respondents (13%, the same number of people which stated they do not use social media in the previous question). The remaining 4 votes are made up of 2 who elected 'Weekly' (2%), 1 who chose 'Every other day' (1%), and one who picked 'Monthly' (1%). Akin to the previous questions, these results are in line with the understanding that those who use social media do so frequently, perhaps more than they are aware of. Which both validates the opinion of the survey participants and shows the representative scope of participants in this survey.

Question 4: In the last year, have you seen extremist content based on your understanding of the term?

Naturally with a concept as difficult to understand and define as extremism this question is subject to people's memories and understandings. Despite these factors, out of the 90 respondents, 54 stated that they had (60%). The remaining 40% is made of the 19 (21%) respondents who had not, the 12 respondents that could not remember (13%), the 4 respondents who selected 'I don't know' (4%), and the 1 participant who elected not to say (1%). As is clearly displayed in Figure 5.2 found below. The purpose of this question was not designed to uncover the true percentage of people who encounter extremist content. Rather it is to identify the social perception of the problem social media companies face. It is very possible that the true statistic could be in the top or bottom 10%. However, for as long as social media users perceive that there is extremist content on these platforms, the more pressure and responsibility is placed on the companies to convince the users and stakeholders that something is being done about it.

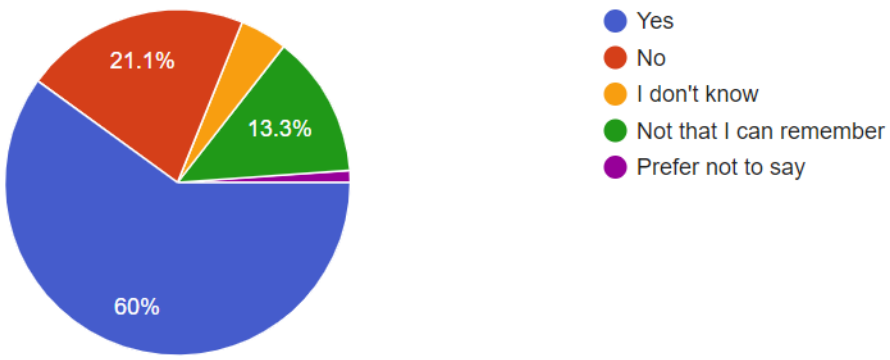

Figure 6.2: Question 4

Question 5: How would you feel if your social media content was removed after being flagged for containing extremist content?

Out of the 90 participants that took part in this study, 80 of them answered this open-ended question to varying lengths and detail. From these answers, four key themes/categories surfaced: understanding, emotional, anti-censorship / freedom of expression, and appeal/justification.

With the 'understanding' category, users' responses communicated that as long as it was justified then they would accept this outcome, or if their post was removed as a false positive, they would tolerate this to an extent. For example, the following response combines both of these attributes "If it were a one-off I would be understanding". A juxtaposition to this stance is the 'emotional' category, which uses personal language to communicate that they are offended about their content being rightly or wrongly categorized as extremists. An example of this would be responses like the following "Offended that my personal beliefs were not accepted by social norms". The 'anti-censorship / freedom of expression' category refers to the responses that suggest that censorship of any form should not be in place or that such actions infringe their freedom of expression. Give for example the following two statements, "I do not agree with censoring any content including my own" and "Disappointed. I'm a firm believer in free speech which essentially permits all view, extreme or not". An interesting example in this category is offered by a non-social media user who states "I do not use social media but firmly believe that (legal) extremist content has a place in the public eye". Finally, there is the 'appeal/justification' category. This grouping of responses refers to the need for justifications of such actions. For example "If it was correct I'd be okay with it, otherwise I'd appeal".

What immediately comes across in analysing this qualitative data is the varying viewpoints offered by the participants. In many ways, this reflects the reality social media companies face, and the difficulties in providing a one size fits all to extremist content removal. But what can be taken from this are the collective needs conveyed by this sample. Thus, it can be determined that having your content removed can be an emotional experience. The possible follow up to this is an appeal. Therefore, following this action with an effective and functional appeals process may be deemed essential for people to feel that they're rights are not at stake.

Question 6: Do you ever report social media content?

Out of the 88 participants that responded to this question 50 selected 'No' (57%) and 38

selected 'Yes' (43%). What these statistics suggest is that there is a relatively even representation of individuals who do and do not report social media content that leans slightly towards not removing content. This essentially communicates that the sample is split in opinions and prioritisation in responding to extremist content which is a concept that will be explored in the following survey discussion section. As is effectively represented in Figure 5.3 below.

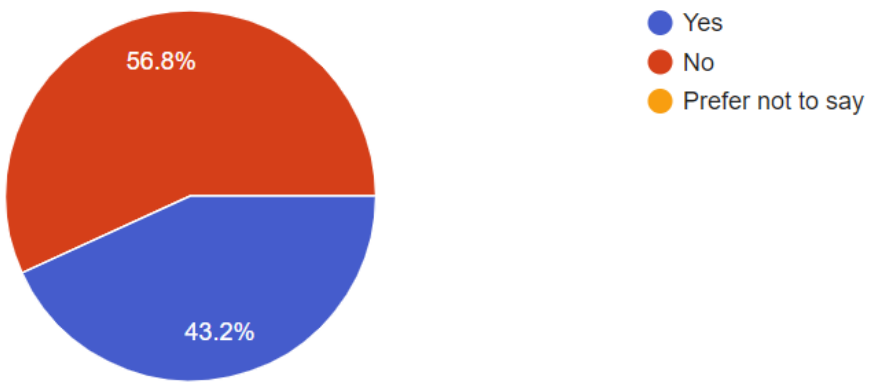

Figure 6.3: Question 6

Question 7: What reasons would make you report content?

From the 81 participants that answered this question, 72 answers were considered in the analysis of the answers. As 9 of the 81 did not provide answers which could be conceptualised or categorised. What this question proposed to the survey participant is in other words, what would make you interfere with the content available in the public domain, what would make you put your content moderating hat on. For the most part, the answers to this question are binary, there are those who would not report content regardless of the context and there are the others who would report content if it caused harm of some form.

In the first category, participants directly communicate that they simply do not report content. Regardless of their views on content removal, this is communicated as a principle and a conscious choice not to assume the role of a voluntary content moderator. A key example of this is the following "I generally avoid reporting content in order to not infringe on people's right to freedom of speech. I would potentially report footage of graphic violence" or as another participant more simply put "I wouldn't". Whether it is out of principle or as a result of a personal belief, there are those that use social media that have no intention of moderating

content. However, this provides a stark contrast to the more common theme of those who report content that they deem to be harmful in some way. In regard to those who report content if it caused harm of some form, this category can be partitioned into two sub-categories, those who report content as they please and those who see it as a last choice. In regard to the former, the following two quotes provide a prime example "Anything that I genuinely do not believe should be put into an essentially public domain or I believe would cause harm" and "If I felt it was lies to try to invoke certain feelings in people or if it was nasty and disgusting content". These both communicate an extremely low threshold to removing content. Whereas the following example and other responses like it communicate a more strict approach "I normally report content if I get something very violent on my feed which is normally some kind of suggested follow or something" or "I generally avoid reporting content in order to not infringe on people's right to freedom of speech. I would potentially report footage of graphic violence".

What these perspectives offer is on the surface a binary response to justifications of content removal, but when looked at more deeply the varying spectrum of views on limitations of freedom of expression beginning to flower. Even in a study with certain bias in participants, human perspective still shows significant variability when addressing the topics around content removal such as data privacy, censorship, and freedom of expression.

Question 8: Would you find it justifiable if a certain percentage of social media content would be removed if this meant that all extremist content was identified and removed?

All 90 participants answered this question, and from this sample 58 selected 'Yes' (64%), 18 chose 'I don't know' (20%) and 14 selected 'No' (16%). The broad indication from this question is that social media users are largely receptive and understanding to the limitations of current content removal methods. In addition, with the second-largest choice being the 'I don't know' category, this indicates that there is little understanding of current content removal methods and practices. Furthermore to the 16% that selected no, this reinforces the said notion, as this choice implies that they are unaware that the proposed scenario in this question is the current state of affairs. This is more appropriately visually represented in Figure 5.4 on the page below.

Question 9: In light of the greater good, what is the lowest level of accuracy you would tolerate for an AI system that removes extremist content?

The top three answers for this question made up the majority of the 90 responses. These in order of the most frequency to the least these categories consisted of '90%', '80%', and

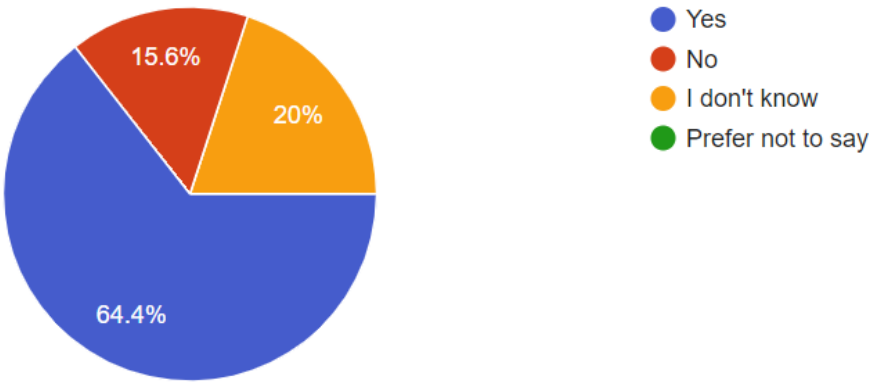

Figure 6.4: Question 8

'100%', the three highest percentage categories are given in the survey as is seen in Figure 5.5. The '90%' category accumulated 31 responses (34%), 25 chose the '80%' category (28 %) and 19 selected '100%' (21%). Out of the remaining 15 responses 8 chose '50% or lower' (9%), 6 selected '60%' (7%) and 1 picked '60%' (1%). What this communicates is that social media users for the most part have high expectations and standards for social media content removal. A large population of which hold these standards to a currently unattainable standard of 100% content removal accuracy. This reinforces the limited understanding that social media users have of content removal and the immense pressure placed on the companies as a result.

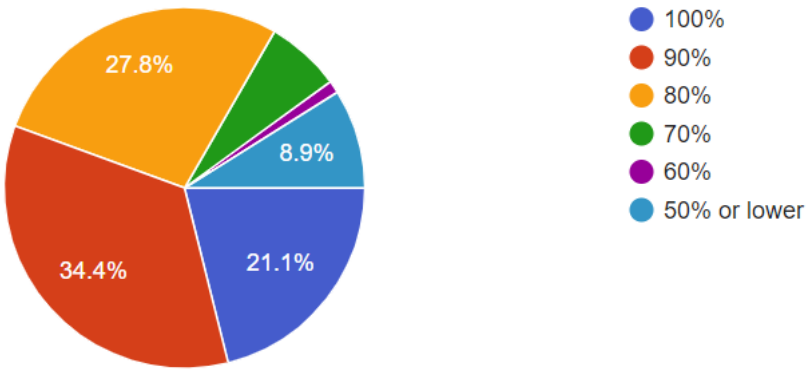

Figure 6.5: Question 9

Question 10: How comfortable are you with privacy levels being reduced for security levels to be increased?

On a number line of whole numbers from 1 through to 10, 1 being very uncomfortable and 10 being very comfortable, the most common answer by a considerable margin voted by 20 was the most neutral answer of '5' (22%). This is then followed by the 12 participants (13%) that voted '1' and the following three categories that had 9 votes (10%) each '10', '6', and '2'. The Numbers '8' and '7' both received 8 votes (9%), number '3' received 7 (8%), '4' received 5 (6%) and '9' received 3 votes (3%) as is shown in the number line in Figure 5.6. The simple fact that in order of frequency, numbers 5, 1, and 10 were in the top three most chosen categories speaks volumes. With the first most prevalent figure being 5, this suggests that people have no preference and/or viewpoint on this matter. With the following two most prevalent numbering being 1 and 10, this evidences that people have contrasting opinions on this matter and there is no clear favourite. The preference shown in this data suggests that social media users desire a balance between data privacy and data security.

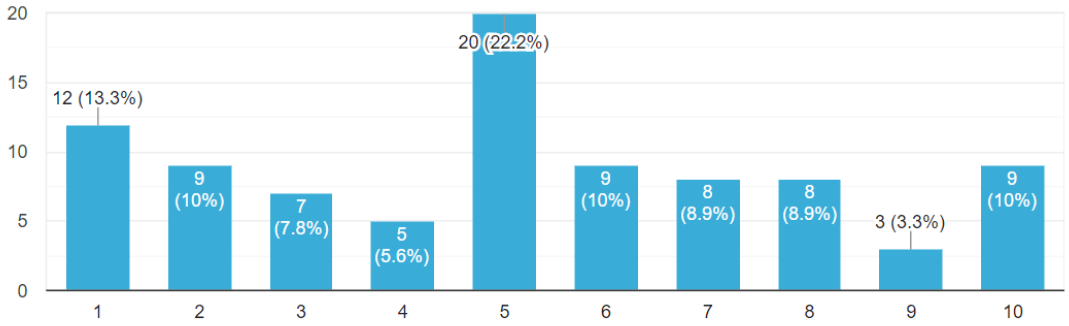

Figure 6.6: Question 10

Question 11: What are your thoughts on social media companies mostly self-regulating the content on their platforms?

Out of the 85 participants that answered this question, 80 of the responses were considered in the analysis of the answers. As 5 of the 81 did not provide answers which had insufficient responses which could not be conceptualised. The broad theme for this question is scepticism and opposition, there are very few examples of participants for self-regulation of social media content by social media companies. Thus the categories include scepticism, opposition, in dependant body/third party, and no alternative.

With the overwhelming reception of this question being in some form negative, the natural largest category was ‘scepticism’. In this category, participants question the integrity of social media companies and how they moderate the content not ethically but rather for their own gain. Give for example the following quotes "I wouldn’t trust that social media platforms would regulate content outside of their own aims and benefit. I feel they should be externally regulated" and, "I think that most companies will have their own agenda for what they want people to see and many allow people to pay their way to show what they want to so I wouldn’t trust them to regulate themselves". In this sense, it can be interpreted that people would be more susceptible to agreeing to social media self-regulation if it were far more transparent. In addition, this theme has an element of cross-pollination with the ‘independent body/third party’ category. Whereby, people strictly state that something along the lines of "Would rather an independent body regulate". There are also those who are more severely opposed to the concept, one participant stated "Self-regulation never works. The turkeys will not vote for Christmas. They either pander to a demographic like Twitter or abdicate all responsibility like Facebook". Finally, there is the ‘no alternative’ category, whereby, people oppose self-regulation but are also less motivated to give this responsibility to governments. The following two quotes encapsulate this sentiment "Social media companies are very poor at moderating their own content although I am loathed to allow the government further power to regulate free expression on online forums" and "It would be more reassuring if they demonstrated habitual care over their own behavior. But the alternative of having state regulation is probably no more reassuring".

Overwhelmingly there is resistance to the notion of self-regulation content by social media companies on their platforms. Whether people question the intention of social media or outright oppose it, the difficulties lie with whom is best suited to reliably, effectively, transparently, and without bias conduct such a task, which unimaginable consequences. Many participants are quick to scrutinise social media, however, several point out that governments conducting such a task is no viable alternative. What is offered in these findings is that whoever is conducting this role, the participants of this study highlight the essential requirement of transparency by design.

Question 12, 13 and 14: What is your age range, gender, and ethnicity?

As identified in the introduction, the purpose of this study was not to explore trends between different factors due to the smaller scale. However, optionally requesting the age range, gender, and ethnicity of the survey participants is necessary to identify key areas of bias in the

data.

The age range of this user study is primarily made up of the '25-34' and the '18-24' age ranges, the former constituted of 35 responses (38%) and the latter was made of up 31 responses (34%). Out of the remaining responses 14 elected the '45-54' range (34%), 4 elected the '55-64' range (34%), 3 chose '>64' (34%), 2 chose '<18' (34%) and 1 elected the 'Prefer not to say' option (34%).

The gender split of this study is 54% female (totalling 49 study participants), 42% male (totalling 38 study participants) and the remaining 4% constitutes the three study participants who elected not to select the 'Prefer not to say' option.

The ethical category significantly weighted towards the 'English / Welsh / Scottish / Northern Irish / British' which totaled 77 out of the 90 responses (86%). Out of the remaining 13 responses 7 identified as 'Irish' (8%), 3 selected 'Prefer not to say' (3%), and of the remaining 3 constituted one each of the following three ethnicity's, 'Polish' (1%), 'Asian / Asian British' (1%) and lastly 'Ashkenazi Jewish' (1%).

In summary, what these three questions offer is the fact that there is a relatively even split in gender, and a predominantly younger demographic in the 25-34 and 18-24 age ranges and an overwhelming bias towards the 'English / Welsh / Scottish / Northern Irish / British' in this data. To reiterate the sentiment shared in this chapter, these questions were implemented to identify bias as opposed to embarking on the impossible task of eradicating it altogether

6.3 Survey Discussion

The results of this survey appropriately communicate the complexities that shroud effective extremist content removal by introducing this human element to the equation. Firstly through questions 12-14, the key sources of bias are addressed in this survey. The survey is mostly made up of predominantly younger demographic in the 25-34 and 18-24 age, slightly female-dominated, and an undeniable bias towards the 'English / Welsh / Scottish / Northern Irish / British' ethnicity grouping. There are, however, other areas in which bias can form, however steps have been made to mitigate this. For example, participants are likely to give the answers they believe the survey creator wants to hear, as a result, an anonymous online survey was chosen over in-person interviews to get as close to unbiased opinions as possible. In doing so, bias has both been mitigated and where it clearly exists, identified.

From this survey it has become unquestionably clear that there are tensions between the users and thereby the stakeholders with social media companies; for the survey participants and social media, users have not had their voice heard. Interestingly when selecting the four definitions used in question one, they were each categorised as a type of definition. Definition 1 was deemed neutral (a common strain and variation of extremist definitions) definition 2 was categorised as a liberal interpretation of the term, definition 3 was interpreted as a nationalist definition referring to concepts such as 'our fundamental values' and lastly, definition 4 was found to be broad and nonspecific. These definitions were chosen due to there innate differences in order to represent the different sets of opinions and beliefs that the survey participants held. The relatively even split between the number of people who elected these definitions helps conceptualise the scope of personal interpretation that is bound to this term. This also helps to conceptualise what the results of Question 4 convey. Furthermore, from the sentiments conveyed by survey participants, it is not that most people have seen extremist content, but that most people perceive that they have had exposure to extremist content. This alone reinforces the need not only for greater extremist content removal practices but for greater clarity and transparency on what is meant by the term.

In light of the understanding that most survey participants believe they have seen extremist content, question six evidences that despite exposure to harmful content most people do not report social media content. Thus, for the content that passes through the automated content removal net, a significant amount of this extremist content is likely to go unreported by most social media users. And with approaches to content removal partially relying on the flagging of content by users, this approach can only be as effective as the community that engages in it. However, this is not because social media users in this sample are lazy or feel no responsibility. In some cases, it is a matter of belief and principle that they themselves feel that they should not have a role in censorship by interfering with another's freedom of expression, regardless of context. And most participants would only remove 'harmful' content, and as previously identified in question 1, people have different notions of extremism. Thereby, it is not always the case that an individual may find extremist content to be harmful enough to warrant its removal. For the most part, this survey identified that participants perceive that extremist content removal on social media is in large part the responsibility of the social media company/regulating body and not the service user.

The hesitation in removing extremist content from social media is continued in question 8. Where 35.6% of participants voted that they would not find it justifiable if a certain percentage

of social media content would be removed if this meant that all extremist content was identified and removed or said that they were unsure. In conjunction with question 1, the results indicate that not all survey participants believe that extremist content should be removed, especially when at the cost infringing on non-extremist user's freedom of expression. This may, however, be in part a reflection of a limited understanding of automated content removal. Whereby 19 of the 90 participants selected the '100%' as the lowest level of accuracy you would tolerate for an AI system that removes extremist content. This is yet to be an attainable figure, especially when considering the broad array of topics that fall under the extremist banner. An additional contributing factor to the survey participants' hesitation in extremist content removal may be a reflection of who is conducting said process. With the majority of responses to question 11 conveying a distrust in social media companies and governments to dictate what is acceptable and what is not, and to do so with human safeguarding, and not their own interests in mind. From these findings, it becomes clear why there might limited literature that considers the human component of extremist content removal because the responses call for significant alteration, reform, and transparency.

In large part, what this conveys is a spectrum of beliefs that to varying extents value privacy and freedom of expression over or under effective extremist content removal. There are the intolerant who are completely against censorship/content removal, there are the tolerant who are for censorship/content removal, and there are those who fit in between. In response to the tolerant, there is a subcategory of those who were concerned/unhappy about unjustified removal of their own/the posts of others. With this category constituting the more popular sentiment it is worth briefly discussion what can be done in this situation in providing scope for future research. Naturally this first action could be to report when a users content is incorrectly classified as extremist. However, an escalation system could be developed to assign trust value/rating to each user depending on how much of their content is violating a company's TOS regarding extremism. However, depends on accounts that are repeatedly used. In addition this also leaves room for the difficulties in dealing with illegal trade of accounts which are sold with a positive trust value. The likes of which is commonly seen with gaming platforms such as Steam. Although this is not a perfect solution, this concept certainly leaves scope for future research on this topic matter.

6.4 Conclusion

Having critically analysed the results of the survey discussed in this section several conclusions and findings have been drawn. Firstly, the survey itself -despite scope for a larger variation to be conducted in the future- actively fills a gap in the literature. There is an inherent lack of exploration into the human user-base which is ultimately at the mercy of the systems and process being analysed. Bearing the scope of this study in mind it is fair to comment that the participants of this study are generally critical of extremist content removal on social media. There is an underlying theme of the protection of provacy and defending the freedom of expression from censorship that is communicated throughout. Pairing this with a limited understanding of the methodologies involved, the entities that conduct said processes, there is a call of a shift in the protection of human rights and civil liberties in addition to the call for these processes to be conducted transparently. An unbiased third party to conduct content removal more broadly may be an avenue of exploration for future research as well as developing a larger user study to gauge a larger sample for more conclusive findings.

Chapter 7

Conclusions and Future Work

7.1 Review of Dissertation

The aim of this section is to summarise the findings of this dissertation and issue the concluding remarks. In doing so, this section will reflect on the dissertation project's title 'Extremist Content Removal on Social Media: A Process of Cutting Corners' and reiterate the understandings drawn from this topic. The likes of which seek to appropriately meet the research aim set out at the beginning of this research. This was to identify the factors which contribute to how extremist content removal is conducted and gauge how users understand and acknowledge this process. With the follow up aim being to achieve a reliable, balanced, and impartial critical analysis of the said process. From said analysis, the increasingly apparent themes of trust and transparency shined throughout. Whereby, with the absence of transparent disclosure of how extremist content removal paired with a self-interest oriented reputation, users are hesitant to trust these processes in the hands of social media companies.

As set out by the literature review when regarding extremist content removal it is first necessary to understand extremism. Although this may seem simple at a glance, the absence of a consensus definition of the term creates a level of uncertainty and skepticism around anything that builds on it. However, this is not to say that it invalidates the topic altogether. But rather to be vigilant when considering the scope. The presence of online extremism has become a normality in recent years, however, it has and continues to change form as different groups with different motives flourish in these online spaces. As is always the case with fighting crime, everyone is always one step behind the criminal. However, this more commonly refers to law enforcement who is bound to offer due process. Whereas when privately-owned

companies moderate their own content the same cannot be said. With varying degrees of flexibility and room for interpretation in data privacy and freedom of expression legislation, social media companies are to an extent left to their own devices. However, with recent legislation being passed in Europe, all that seems to matter is how quickly and how accurately content is removed. Thus, in recent years content removal has become a numbers game and a race to achieve high accuracy content removal scores. This dissertation has evidenced how this quickly becomes a dangerous game of cutting corners. Where social media companies are subject to laws which regardless of the limitations are expected to remove content increasingly quickly, seemingly regardless of the consequences.

As a result, little heed is being paid to the ethical nature of this conduct and the even less attention that is paid to the users and stakeholders' views and opinions. Although it was not in the scope of this dissertation to produce compare international legislation or ethical principles on a larger basis, key themes were extracted. In addition, the conclusions drawn from the user study does not claim to be representative of a global collective. However, the understandings drawn from these resources create a clear-cut narrative. Which expresses that there is an inherent imbalance where the only winner is the extremists. Whereby, as a result of governmental and users/stakeholder pressure increasing in severity about removing extremist content with increasing accuracy in a shrinking time frame. Social media companies are seemingly pushing users' rights and ethical practices to the side. How avoidable this is, is a relative unknown and excellent concept worth exploring. However, if stopping people from seeing extremist content means removing people's rights and ethical practice then who is the winner? Because despite all of these sacrifices extremism still takes place online and there is little to suggest that it will end any time soon. Bearing this in mind, social media companies have a responsibility to uphold ethical values and practice and facilitate human rights regardless of any extremist narrative. For if online extremists continue to exist and human rights and ethical practices cease to upheld then what are human rights other than the optional choice given to social media companies that have an increasing role in shaping narratives found in modern society.

7.2 Contributions

To reiterate, the primary contributions of this research stated in the introduction and achieved in this document can be summarised in the following points:

- **Addressing AI regulation based on ethical principles**

This element of the dissertation addressed the new and booming surge of AI ethics principles and applied them to extremist content removal. Reviewing such documents is an angle that has seen growth in the academic literature, however, applying it to this context has yet to be covered. In doing so the integrity of such documents was brought into question. This contribution has left scope for future research on this topic matter.

- **Analysing regulatory impacts on the capabilities of extremist content removal**

  Reviewing social media content removal techniques is not a new angle in the literature. However, exploring how legal and ethical regulations benefit, detriment, and create uncertainty over social media extremist content removal is conducted is an unexplored narrative. How each of these factors shape and mold content removal creates questions regarding who is to be held accountable for the current practices which are subject to more than its fair share of criticisms.

- **Exploring human perspective on extremist content removal**

  This component of the dissertation shines a light on a significant angle that is yet to be explored in academic research. In that, the opinions and views of social media users and stakeholders are not necessarily reflected in the ways extremist and content removal, in general, is conducted. In addition, this leaves a significant scope for a wealth of future research on this subject matter.

7.3 Future Work

Throughout the course of this paper, there have been references to scope for future research. This final section will reference each of the three primary areas where future research is necessary along with examples of how these could and need to be conducted.

The first and to the viewpoint of this study, the most necessary call for future research regards AI ethics principles. Currently, the utility of these documents is questionable in terms of its lack of specific implementable and measurable principles. Future work could look at the medical sector as a point of reference to make principles more of a specific resource. The angle that this research should consider is how to make these documents more specific in order to draw away from nuances notions such as 'fairness'. The breadth of documents referred to would also need to be increased in order to gain a greater understanding of themes and trends

found in these documents, in order to know what ethical principles are in need of becoming more feasibly implementable.

A separate example is the proposed idea of assigning users an individual trust value to inform extremist content removal methods and processes. The focus of a study on this topic could include modifying existing and creating a new trust model that adjusts a user's rating based on the frequency of a user's content being flagged or removed to inform their trustworthiness. Furthermore, this study could consider adjusting retroactive and real-time content filtering based on said user ratings. For example, if a user has a poor rating real-time filter would be applied, whereas is an account has a high trust value then retroactive filtering would be applied to deal with capacity and latency issues. A system such as this is yet to exist in the academic literature, and so it could mitigate the blanket concerns with limitations on freedom of expression by rewarding those who do not violate a platform's terms of service and by more thoroughly filtering those who post potentially extremist content.

The third and final call for future research based on this study refers to the public perception of who is best suited to conduct content removal. With sentiments from the user study found in this research expression largely encompassing distrust towards social media companies. A more specific user study with a larger participant number should look at where people would choose to place their trust regarding extremist content removal. With an unbiased third party to conduct content removal hypothetically being the obvious choice, future work could explore the realities of such a concept.

Bibliography

[1] Alan Turing Institute. 2019a. Public Policy. Website. (2019). Retrieved July 30, 2020 from https://www.turing.ac.uk/research/research-programmes/public-policy.

[2] Alan Turing Institute. 2019b. Understanding artificial intelligence ethics and safety. Website. (2019). Retrieved July 30, 2020 from https://doi.org/10.5281/zenodo.3240529.

[3] I. Awan. 2017. Cyber-Extremism: Isis and the Power of Social Media. *Society* (03 2017), 1–12. DOI:http://dx.doi.org/10.1007/s12115-017-0114-0

[4] P. Barrett. 2020. Who Moderates the Social Media Giants? Website. (2020). Retrieved August 20, 2020 from https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf.

[5] Nina Baur and Stefanie Ernst. 2011. Towards a Process-Oriented Methodology: Modern Social Science Research Methods and Norbert Elias's Figurational Sociology. *The Sociological Review* 59, 1_suppl (2011), 117–139. DOI:http://dx.doi.org/10.1111/j.1467-954X.2011.01981.x

[6] BBC. 2017. Cyber-security threat to UK 'as serious as terrorism' - GCHQ. Website. (2017). Retrieved July 20, 2020 from https://www.bbc.co.uk/news/uk-41547478.

[7] BBC. 2020. Kim Kardashian West joins Facebook and Instagram boycott. Website. (2020). Retrieved July 20, 2020 from https://www.bbc.co.uk/news/entertainment-arts-54171526.

[8] J.M. Berger and J. Morgan. 2015. The ISIS Twitter Census Defining and describing the population of ISIS supporters on Twitter. Website. (2015). Retrieved July 10, 2020 from https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf.

[9] John M Berger. 2018. *Extremism*. MIT Press.

[10] J Bergh and D. Deschoolmeester. 2010. Ethical Decision Making in ICT: Discussing the Impact of an Ethical Code of Conduct. *Communications of the IBIMA* (03 2010). DOI: http://dx.doi.org/10.5171/2010.127497

[11] A. Beutel, S.M. Weine, A. Saeed, A.S. Mihajlovic, and A. Stone. 2016. Guiding Principles for Countering and Displacing Extremist Narratives. 7, 3 (2016), 35–49. DOI: http://dx.doi.org/10.15664/jtr.1220

[12] M. Bickert and B Fishman. 2017a. Hard Questions: Are We Winning the War On Terrorism Online? Website. (2017). Retrieved July 13, 2020 from https://about.fb.com/news/2017/11/hard-questions-are-we-winning-the-war-on-terrorism-online/.

[13] M. Bickert and B. Fishman. 2017b. Hard Questions: How We Are Countering Terrorism. Website. (2017). Retrieved August 20, 2020 from https://about.fb.com/news/2017/06/how-we-counter-terrorism/.

[14] M. Bloom, T. Tiflati, and H. Horgan. 2019. Navigating ISIS's Preferred Platform: Telegram1. *Terrorism and Political Violence* 31, 6 (2019), 1242–1254. DOI:http://dx.doi.org/10.1080/09546553.2017.1339695

[15] D. Boffey. 2020. Remove terror content or be fined millions, EU tells social media firms. Website. (2020). Retrieved July 10, 2018 from https://www.theguardian.com/media/2018/sep/13/social-media-firms-could-face-huge-fines-over-terrorist-content.

[16] C. Bosk. 2010. Bioethics, Raw and Cooked: Extraordinary Conflict and Everyday Practice. *Journal of Health and Social Behavior* 51, 1_suppl (2010), S133–S146. DOI: http://dx.doi.org/10.1177/0022146510383839 PMID: 20943578.

[17] L. Bowman-Grieve. 2009. Exploring "Stormfront": A Virtual Community of the Radical Right. *Studies in Conflict & Terrorism* 32, 11 (2009), 989–1007. DOI:http://dx.doi.org/10.1080/10576100903259951

[18] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[19] R. Briggs. 2011. Radicalisation, the Role of the Internet. Website. (2011). Retrieved July 13, 2020 from https://www.isdglobal.org/wp-content/uploads/2016/07/StockholmPPN2011_BackgroundPaper_FOR20WEBSITE.pdf, annote = Website URL,.

[20] R. Briggs and S. Feve. 2014. POLICY BRIEFING : COUNTERING THE APPEAL OF EXTREMISM ONLINE. Website. (2014). Retrieved July 14, 2020 from https://www.dhs.gov/sites/default/files/publications/Countering%20the%20Appeal%20of%20Extremism%20Online_1.pdf, annote = Website URL,.

[21] British Council. 2020. Literature Surveys: Structure 1. Website. (2020). Retrieved September 13, 2020 from https://learnenglish.britishcouncil.org/skills/writing/writing-purpose/literature-surveys-structure-1.

[22] R Brownsword. 2016. Technological management and the Rule of Law. *Law, Innovation and Technology* 8, 1 (2016), 100–140. DOI:http://dx.doi.org/10.1080/17579961.2016.1161891

[23] David Byrne and Charles C Ragin. 2009. Case-based methods. *SAGE Publications India Pvt Ltd* (2009).

[24] A. Chen. 2014. The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. Website. (2014). Retrieved August 20, 2020 from https://www.wired.com/2014/10/content-moderation/.

[25] Constitution Annotated. 2020. Constitution of the United States: First Amendment. Website. (2020). Retrieved July 21, 2020 from https://constitution.congress.gov/constitution/amendment-1/#:~:text=Congress%20shall%20make%20no%20law,for%20a%20redress%20of%20grievances.

[26] M. Conway. 2020. Routing the Extreme Right. *The RUSI Journal* 165, 1 (2020), 108–113. https://doi.org/10.1080/03071847.2020.1727157

[27] M. Conway and M. Courtney. 2017. VIOLENT EXTREMISM AND TERRORISM ONLINE IN 2017:THE YEAR IN REVIEW. Website. (2017). Retrieved July 13, 2020 from https://www.voxpol.eu/download/vox-pol_publication/YiR-2017_Web-Version.pdf.

[28] M. Conway, M. Khawaja, S. Lakhani, J. Reffin, A. Robertson, and D. Weir. 2019a. Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts. *Studies in Conflict & Terrorism* 42, 1-2 (2019), 141–160. DOI:http://dx.doi.org/10.1080/1057610X.2018.1513984

[29] M. Conway, R. Scrivens, and L. Macnair. 2019b. Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends. (11 2019), he International Centre for Counter–Terrorism – The Hague. DOI:http://dx.doi.org/10.19165/2019.3.12

[30] E. Day. 2020. AI and GDPR - what do you need to know. Website. (2020). Retrieved July 30, 2020 from https://www.innovation-academy.co.uk/resource/ai-and-gdpr-what-do-you-need-to-know/.

[31] E. Dwoskin, J. Whalen, and R. Cabato. 2019. Content moderators at YouTube, Facebook and Twitter see the worst of the web — and suffer silently. Website. (2019). Retrieved August 20, 2020 from https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-\philippines-generation-workers-is-paying-price/.

[32] European Commission. 2015. STRIVE FOR DEVELOPMENT: Strengthening Resilience to Violence and Extremism. Website. (2015). Retrieved July 30, 2020 from https://rusi.org/sites/default/files/mn0115566enn.pdf.

[33] European Commission. 2018. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Website. (2018). Retrieved July 30, 2020 from https://ec.europa.eu/commission/sites/beta-political/files/soteu2018-preventing-terrorist-content-online-regulation-640_en.pdf.

[34] European Commission. 2019. ETHICS GUIDELINES FOR TRUSTWORTHY AI. Website. (2019). Retrieved July 30, 2020 from https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf.

[35] European Court of Human Rights. 2013. Delfi AS v. Estonia ECtHR 64669/09. Website. (2013). Retrieved July 21, 2020 from https://hudoc.echr.coe.int/fre#{%22itemid%22:[%22001-126635%22]}.

[36] Facebook. 2020a. Community Standards Enforcement Report. Website. (2020). Retrieved July 10, 2020 from https://transparency.facebook.com/community-standards-enforcement#dangerous-organizations.

[37] Facebook. 2020b. Facebook Reports First Quarter 2020 Results. Website. (2020). Retrieved July 30, 2020 from https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-First-Quarter-2020-Results/default.aspx.

[38] Facebook. 2020c. What is the General Data Protection Regulation (GDPR)? Website. (2020). Retrieved July 30, 2020 from https://www.facebook.com/business/gdpr.

[39] A. Fernandez. 2015. Here to stay and growing: Combating ISIS propaganda networks. Website. (2015). Retrieved July 10, 2020 from https://www.brookings.edu/wp-content/uploads/2016/07/IS-Propaganda_Web_English_v2.pdf.

[40] C. Fishmwick. 2014. How a Polish student's website became an Isis propaganda tool. Website. (2014). Retrieved August 20, 2020 from https://www.theguardian.com/world/2014/aug/15/-sp-polish-man-website-isis-propaganda-tool.

[41] Formplus. 2020. What is Empirical Research Study? [Examples Method]. Website. (2020). Retrieved September 13, 2020 from https://www.formpl.us/blog/empirical-research.

[42] B. Galloway and R. Scrivens. 2018. THE HIDDEN FACE OF HATE GROUPS ONLINE: A FORMER'S PERSPECTIVE. Website. (2018). Retrieved July 10, 2020 from https://www.voxpol.eu/hidden-face-hate-groups-online-formers-perspective/.

[43] B. Ganesh and J. Bright. 2020. Countering Extremists on Social Media: Challenges for Strategic Communication and Content Moderation. *Policy Internet* 12, 1 (2020), 6–19.

[44] G. Gebhart. 2019. Who Has Your Back? Censorship Edition 2019. Website. (2019). Retrieved August 20, 2020 from https://www.eff.org/wp/who-has-your-back-2019#executive-summary.

[45] GIFCT. 2017. Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism. Website. (2017). Retrieved July 13, 2020 from https://gifct.org/press/facebook-microsoft-twitter-and-youtube-announce-formation\-global-internet-forum-counter-terrorism/.

[46] GIFCT. 2018a. GIFCT. Website. (2018). Retrieved July 13, 2020 from https://gifct.org/press/global-internet-forum-counter-terrorism-update-our-progress\-two-years/.

[47] GIFCT. 2018b. Global Internet Forum to Counter Terrorism: an update on our efforts to use technology, support smaller companies and fund research to fight terrorism online. Website. (2018). Retrieved July 13, 2020 from https://gifct.org/press/global-internet-forum-counter-terrorism-update-our-efforts\-use-technology-support-smaller-companies-and-fund-research\-fight-terrorism-online/.

[48] GIFCT. 2020. Global Internet Forum to Counter Terrorism: Evolving an Institution. Website. (2020). Retrieved July 24, 2020 from https://www.gifct.org/about/#:~:text=Core%20Structure,be%20drawn%20from%20industry%20contributions.

[49] P. Gill. 2016. Online Behaviours of Convicted Terrorists. Website. (2016). Retrieved July 10, 2020 from https://www.voxpol.eu/download/vox-pol_publication/Online-Behaviours_FINAL.pdf.

[50] P. Gill, E. Corner, M. Conway, A. Thornton, M. Bloom, and J. Horgan. 2017. Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns, and Processes. *Criminology Public Policy* 16 (02 2017). DOI:http://dx.doi.org/10.1111/1745-9133.12249

[51] P. Gill, E. Corner, and A. Thornton. 2015. WHAT ARE THE ROLES OF THE INTERNET IN TERRORISM?: MeasuringOnline Behaviours of Convicted Terrorists. Website. (2015). Retrieved July 10, 2020 from https://www.voxpol.eu/download/vox-pol_publication/What-are-the-Roles-of-the-Internet-in-Terrorism.pdf.

[52] R. Golemanova. 2019. How To Handle Content Moderation With The Human Factor In Mind. Website. (2019). Retrieved August 20, 2020 from https://imagga.com/blog/how-to-handle-content-moderation-with-the-human-factor-in-mind/.

[53] R. Gorwa. 2019. The platform governance triangle: conceptualising the informal regulation of online content. *Internet Policy Review: Journal on Internet Regulation* 8, 2 (2019), 1–22. DOI:http://dx.doi.org/10.14763/2019.2.1407

[54] T. Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* (02 2020). DOI:http://dx.doi.org/10.1007/s11023-020-09517-8

[55] M Hildebrandt. 2018. Primitives of Legal Protection in the Era of Data-Driven Platforms. *SSRN Electronic Journal* (01 2018). DOI:http://dx.doi.org/10.2139/ssrn.3140594

[56] H.M.Government. 1986. The Public Order Act. Legislation. (1986). Retrieved July 26, 2020 from https://www.legislation.gov.uk/ukpga/1986/64/contents.

[57] H.M.Government. 1988. The Data Protection Act. Legislation. (1988). Retrieved July 26, 2020 from https://www.legislation.gov.uk/ukpga/1998/29/contents.

[58] H.M.Government. 1998. The Human Rights Act 1998. Legislation. (1998). Retrieved August 03, 2020 from https://www.legislation.gov.uk/ukpga/1998/42/section/12.

[59] H.M.Government. 2003. The Communications Act. Legislation. (2003). Retrieved July 26, 2020 from https://www.legislation.gov.uk/ukpga/2003/21/contents.

[60] H.M.Government. 2006. The Terrorism Act. Legislation. (2006). Retrieved July 26, 2020 from https://www.legislation.gov.uk/ukpga/2006/11/contents.

[61] H.M.Government. 2015. Counter-Extremism Strategy. Website. (2015). Retrieved July 12, 2020 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/470088/51859_Cm9148_Accessible.pdf.

[62] H.M.Government. 2016. NATIONAL CYBER SECURITY STRATEGY 2016-2021. Website. (2016). Retrieved July 20, 2020 from http://data.parliament.uk/DepositedPapers/Files/DEP2016-0790/National_Cyber_Security_Strategy_v20.pdf.

[63] H.M.Government. 2017. Hate crime: abuse, hate and extremism online. Website. (2017). Retrieved July 12, 2020 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/668676/CCS207_CCS1217596236-1_Cm_9556_COVER_AND_TEXT_-_BASE_-_WEB.pdf.

[64] H.M.Government. 2018a. Data Ethics Framewrok. Website. (2018). Retrieved July 30, 2020 from https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework.

[65] H.M.Government. 2018b. The Data Protection Act. Legislation. (2018). Retrieved July 26, 2020 from https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted.

[66] H.M.Government. 2019. Understanding artificial intelligence ethics and safety. Website. (2019). Retrieved July 30, 2020 from https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety#who-this-guidance-is-for.

[67] H.M.Government. 2020. Hate and abuse on social media. Website. (2020). Retrieved July 14, 2020 from https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/60904.html.

[68] N. Hopkins. 2017. Facebook moderators: a quick guide to their job and its challenges. Website. (2017). Retrieved August 20, 2020 from https://www.theguardian.com/news/2017/may/21/facebook-moderators-quick-guide-job-challenges.

[69] ICO. 2020. Deleting Personal Data. Website. (2020). Retrieved July 30, 2020 from https://ico.org.uk/media/for-organisations/documents/1475/deleting_personal_data.pdf.

[70] ISD. 2018. 'Alternative' Social Media. Website. (2018). Retrieved July 10, 2020 from http://www.isdglobal.org/wp-content/uploads/2018/07/UK-Insight-Report_Volume-4_FINAL.pdf.

[71] A. L. James. 2002. *Assessing the risks of cyber terrorism, cyber war and other cyber threats*. Center for Strategic & International Studies Washington, DC.

[72] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sep 2019), 389–399. DOI:http://dx.doi.org/10.1038/s42256-019-0088-2

[73] Urfan Khaliq. 2006. Islamic State Practices, International Law and the Threat from Terrorism: A Critique of the 'Clash Of Civilisations' in the New World Order by Javaid Rehman. *Journal of Law and Society* 33 (05 2006), 324 – 330. DOI:http://dx.doi.org/10.1111/j.1467-6478.2006.00360.x

[74] J. Koetsier. 2020. Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day. Website. (2020). Retrieved August 20, 2020 from https://www.forbes.com/sites/johnkoetsier/2020/06/09/

300000-facebook-content-moderation-mistakes-daily-report-says/#1dd9d35654d0.

[75] I Lachow. 2009. Cyber terrorism: Menace or myth. *Cyberpower and national security* (2009), 434–467.

[76] D Lowe. 2019. Christchurch Terrorist Attack, The Far-Right and Social Media: What can we learn? *The New Jurist* (2019).

[77] E. MacAskill. 2010. Countries are risking cyber terrorism, security expert tells first world summit. Website. (2010). Retrieved July 20, 2020 from https://www.theguardian.com/technology/2010/may/05/terrorism-uksecurity.

[78] S. Macdonald, S. Correia, and A. Watkin. 2019. Regulating terrorist content on social media: automation and the rule of law. *International Journal of Law in Context* 15, 2 (2019), 183–197. DOI:http://dx.doi.org/10.1017/S1744552319000119

[79] N. Malik. 2018. TERROR IN THE DARK HOW TERRORISTS USE ENCRYPTION, THE DARKNET, AND CRYPTOCURRENCIES. Website. (2018). Retrieved July 25, 2020 from https://henryjacksonsociety.org/wp-content/uploads/2018/04/Terror-in-the-Dark.pdf.

[80] Noortje Marres and Carolin Gerlitz. 2016. Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology. *The Sociological Review* 64, 1 (2016), 21–46. DOI:http://dx.doi.org/10.1111/1467-954X.12314

[81] C. Maura. 2018. VIOLENT EXTREMISM AND TERRORISM ONLINE IN 2018:THE YEAR IN REVIEW. Website. (2018). Retrieved July 10, 2020 from https://www.voxpol.eu/download/vox-pol_publication/Year-in-Review-2018.pdf.

[82] J. Mullhall. 2019. Modernising and Mainstreaming: The Contemporary British Far Right. Website. (2019). Retrieved July 10, 2020 from https://www.hopenothate.org.uk/wp-content/uploads/2019/07/HnH-Briefing_Contemporary-British-Far-Right_2019-07-v1.pdf.

[83] P Nemitz. 2018. Constitutional Democracy and Technology in the age of Artificial Intelligence. *Royal Society Philosophical Transactions A* (02 2018). DOI:http://dx.doi.org/10.1098/rsta.2018.0089

[84] C. Newton. 2019. THE TRAUMA FLOOR. Website. (2019). Retrieved August 20, 2020 from https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-\working-conditions-arizona.

[85] L. Nouri, N. Lorenzo-Dus, and A. Watkin. 2019. Following the Whack-a-Mole: Britain First's Visual Strategy from Facebook to Gab. Website. (2019). Retrieved July 24, 2020 from https://rusi.org/sites/default/files/20190704_grntt_paper_4.pdf.

[86] Omnicore. 2019. Facebook by the Numbers. Website. (2019). Retrieved August 20, 2020 from https://www.omnicoreagency.com/facebook-statistics/.

[87] G. Parker. 2017. Theresa May warns tech companies: 'no safe space' for extremists. Website. (2017). Retrieved July 10, 2020 from https://www.ft.com/content/0ae646c6-4911-11e7-a3f4-c742b9791d43.

[88] A. Perrin. 2015. Social Media Usage: 2005-2015. Website. (2015). Retrieved July 20, 2020 from https://www.pewresearch.org/internet/2015/10/08/social-networking-usage-2005-2015/.

[89] J. Porter. 2019. UPLOAD FILTERS AND ONE-HOUR TAKE-DOWNS: THE EU'S LATEST FIGHT AGAINST TERRORISM ONLINE, EXPLAINED. Website. (2019). Retrieved July 10, 2018 from https://www.theverge.com/2019/3/21/18274201/european-terrorist-content-regulation-extremist-terreg-\upload-filter-one-hour-takedown-eu.

[90] D. Prisk. 2017. The Hyperreality of the Alt Right: How Meme Magic Works to Create a Space for Far Right Politics. (Nov 2017). DOI:http://dx.doi.org/10.31235/osf.io/by96x

[91] Keith F Punch. 2013. *Introduction to social research: Quantitative and qualitative approaches*. sage.

[92] N. Ramati. 2020. THE LEGAL RESPONSE OF WESTERN DEMOCRACIES TO ONLINE TERRORISM AND EXTREMISM. Website. (2020). Retrieved July 10, 2020 from https://www.statewatch.org/media/documents/news/2020/apr/vox-pol-legal-response-western-democracies-online\-terrorism-extremism-4-20.pdf.

[93] A. Reed and J. Dowling. 2018. The Role of Historical Narratives in Extremist Propaganda. *Defence Strategic Communications* 4 (2018), 79–104.

[94] A. Reed, Haroro J.I., and J. Whittaker. 2017. Countering Terrorist Narratives. Website. (2017). Retrieved July 26, 2020 from https://icct.nl/publication/countering-terrorist-narratives/.

[95] J. Rollins and C. Wilson. 2005. *Terrorist Capabilities for Cyberattack: Overview and Policy Issues*. Technical Report. http://www.dtic.mil/docs/citations/ADA444928

[96] L. Schlegel. 2019. CHAMBERS OF SECRETS? COGNITIVE ECHO CHAMBERS AND THE ROLE OF SOCIAL MEDIA IN FACILITATING THEM. Website. (2019). Retrieved July 10, 2020 from https://www.voxpol.eu/chambers-of-secrets-cognitive-echo-chambers-and-the-role-of\-social-media-in-facilitating-them/.

[97] A. Soltar. 2004. Some Problems with a Definition and Perception of Extremism within a Society. Website. (2004). Retrieved August 20, 2020 from https://www.ncjrs.gov/pdffiles1/nij/Mesko/208033.pdf.

[98] C Sonderby. 2019. Update on New Zealand. Website. (2019). Retrieved September 20, 2020 from https://about.fb.com/news/2019/03/update-on-new-zealand/.

[99] Statista. 2020. Most popular social networks worldwide as of April 2020, ranked by number of active users. Website. (2020). Retrieved July 10, 2020 from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[100] Abbas Tashakkori and Charles Teddlie. 2008. Introduction to mixed method and mixed model studies in the social and behavioral sciences. *The mixed methods reader* (2008), 7–26.

[101] The Bundestag. 2017. Network Enforcement Act. Website. (2017). Retrieved July 24, 2020 from https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2.

[102] The Bussola Institute. 2020. Extremism in the time of COVID-19. Website. (2020). Retrieved September 20, 2020 from https://www.bussolainstitute.org/wp-content/uploads/2020/07/Extremism-in-the-time-of-COVID-19-V2.pdf.

[103] The European Union. 1995. The Data Protection Directive. Website. (1995). Retrieved July 26, 2020 from https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML.

[104] The European Union. 2018. The General Data Protection Regulation. Website. (2018). Retrieved July 26, 2020 from https://gdpr-info.eu/.

[105] The European Union. 2020. What is GDPR, the EU's new data protection law. Website. (2020). Retrieved July 30, 2020 from https://gdpr.eu/what-is-gdpr/.

[106] The Week. 2020. Hate Speech vs. Free Speech: UK Laws. Website. (2020). Retrieved August 03, 2020 from https://www.theweek.co.uk/97552/hate-speech-vs-free-speech-the-uk-laws.

[107] Z. Thomas. 2020. Facebook content moderators paid to work from home. Website. (2020). Retrieved August 20, 2020 from https://www.bbc.co.uk/news/technology-51954968#:~:text=Facebook%20has%20approximately%2015%2C000%20content,content%20is%20inappropriate%20or%20harmful.

[108] P. Turner. 2014. "Harm" and Mill's Harm Principle. *Ethics* 124, 2 (2014), 299–326. http://www.jstor.org/stable/10.1086/673436

[109] Twitter. 2016. Combating Violent Extremism. Website. (2016). Retrieved July 10, 2020 from https://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html.

[110] Twitter. 2016. An Update on our Efforts to Combat Violent Extremism. Website. (2016). Retrieved July 10, 2020 from https://blog.twitter.com/2016/an-update-on-our-efforts-to-combat-violent-extremism.

[111] Twitter. 2020a. Defending and respecting the rights of people using our service. Website. (2020). Retrieved July 13, 2020 from https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice.

[112] Twitter. 2020b. Welcome to Twitter's GDPR Hub. Website. (2020). Retrieved July 30, 2020 from https://gdpr.twitter.com/.

[113] I. Van der Vegt, P. Gill, S. Macdonald, and B. Kleinberg. 2019. Shedding Light on Terrorist and Extremist Content Removal. Website. (2019). Retrieved July 12, 2020 from https://rusi.org/sites/default/files/20190703_grntt_paper_3.pdf, annote = Website URL,.

[114] G. Weimann. 2004. CYBERTERRORISM: HOW REAL IS THE THREAT ? (Article). *U.S. INSTITUTE OF PEACE (USIP)* (2004).

[115] G. Weimann. 2005. Cyberterrorism: The Sum of All Fears? *Studies in Conflict & Terrorism* 28, 2 (2005), 129–149. DOI:http://dx.doi.org/10.1080/10576100590905110

[116] Weimann, G. 2004. *Cyberterrorism: How real is the threat?* Vol. 31. United States Institute of Peace.

[117] Weimann, G. 2016. Going Dark: Terrorism on the Dark Web. *Studies in Conflict & Terrorism* 39, 3 (2016), 195–206. DOI:http://dx.doi.org/10.1080/1057610X.2015.1119546

[118] A. Widfeldt and H. Brandenburg. 2018. What Kind of Party Is the UK Independence Party? The Future of the Extreme Right in Britain or Just Another Tory Party? *Political Studies* 66, 3 (2018), 577–600. DOI:http://dx.doi.org/10.1177/0032321717723509

[119] Wirral Safeguarding Children Partnership. 2020. Extremism and Radicalisation. Website. (2020). Retrieved July 12, 2020 from https://www.wirralsafeguarding.co.uk/extremism-and-radicalisation/.

[120] H. Woodhouse. 2020. Social Media Regulation. Website. (2020). Retrieved July 21, 2020 from https://commonslibrary.parliament.uk/research-briefings/cbp-8743/.

Appendix A

Supplementary Survey Data

Out of the following four definitions of extremism, which if any do you think is the most accurate?	Do you use social media at all (e.g., Facebook, Twitter, etc.)?	How often do you use social media?	In the last year, have you seen extremist content based on your understanding of the term?	How would you feel if your social media content was removed after being flagged for containing extremist content?	Do you ever report social media content?	What reasons would make you report content?	Would you find it justifiable if a certain percentage of social media content would be removed if this meant that all extremist content was identified and removed?	In light of the greater good, what is the lowest level of accuracy you would tolerate for an AI system that removes extremist content?	How comfortable are you with privacy levels being reduced for security levels to be increased?	What are your thoughts on social media companies mostly self-regulating the content on their platforms?	What is your age range?	What is your gender?	What is your ethnicity?
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	Understandable if it was extreme	Yes	Racial, sexual or anything to do with children	Yes	50% or lower	10	Would rather an independent body regulate	45-54	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	No	I do not use it	Not that I can remember	If it were a one-off I would be understanding	No	It would have to identify me or my family in a negative way	Yes	0.9	4	I was unaware of this, I think it is unacceptable and needs reform.	18-24	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	Yes	Not sure	Yes	Racism, child exploitation, child pornography, violence	Yes	1	1	There needs to be greater regulating outside of self regulation	45-54	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	No	if The content is not allowed then it's right that it should be removed.	Yes	Hateful Speach or inappropriate images.	I don't know	1	10	They should be regulated by independent persons.	45-54	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Hourly	Not that I can remember	I would feel like my views (that I have the right to show people) are being diminished and ignored as if it doesn't matter.	Yes	If i found it offencive to me or others	I don't know	0.7	6	It may not be regulated very efficiently	<18	Male	English / Welsh / Scottish / Northern Irish / British

People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Hourly	I don't know	Shocked and upset with myself	No	If I seen something disturbing, or bullying	Yes	0.8	5	Its worrying as each individual social media company will have different levels of tolerance, different beliefs etc. that will effect how they react to content. What one company might tolerate another many not.	25-34	Female	Irish
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	Offended that my personal beliefs were not accepted by social norms.	No	Actively dangerous content, physically or emotionally.	Yes	0.6	8	Potentially dangerous allowing corporate and individual bias to be taken as facts	18-24	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	No	Shocked as I wouldn't purposefully do anything like that	No	If its inappropriate	I don't know	0.9	5	If they can get it right	<18	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	No	I would be quite confused as I wouldn't believe that my content would fall in that section but if it did offend someone in that way then I would be happy for it to be removed.	Yes	I normally report content if I get something very violent on my feed which is normally some kind of suggested follow or something.	Yes	0.7	7	I think that most companies will have their own agenda for what they want people to see and many allow people to pay their way to show what they want to so I wouldn't trust them to regulate themselves.	18-24	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	No	Not at all.	I don't know	Offended	No	Don't know.	Yes	1	10		>64	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	No	Surprised	No	None	Yes	0.8	5	Tighter control needed	25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	No	0	Yes	As long as justification is given, world safety is more important than personal opinion .	No	Offencive content that condones violence towards any individuals.	Yes	0.8	6	They should have no place in it. The process should be regulated externally.	45-54	Prefer not to say	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	No	Never	Prefer not to say	Not applicable	No		Yes	0.8	5	They do not seem to do a good job	>64	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	Offended that my personal beliefs are perceived as less important than social norm.	No	Nothing.	No	1	2	They own the system. They can control the content. Other systems fiction differently and can easily be found.	25-34	Male	Asian / Asian British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	Yes	I don’t have any reason to write extremist content.	No	Not sure I would.	I don't know	1	8	It’s not consistent	45-54	Female	English / Welsh / Scottish / Northern Irish / British

People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	No		No	Racist remarks	Yes	0.8	10		18-24	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	No	I have only started using WhatsApp since the lockdown, checking messages as they come through.	Not that I can remember	Embarrassed if anything I had written/forwarded was deemed to contain extremist under tones.	No	Something hateful or unlawful, or unacceptable or inappropriate content that was regularly sent.	Yes	0.9	8	It was acceptable in the early days of social media. Now that social media companies are an integral part of every aspect of the modern globalized world we live in they must be regulated by governments ,EU,in etc, as other organisations /businesses are.	45-54	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	I would be frustrated and upset because I believe in free speech.	Yes	If the content is explicit I report it. I wouldn't want a child to accidently see it.	I don't know	0.9	4	When there is a lot of control from one body I think that can be potentially dangerous. It would be good to have an independent body that could regulate.	18-24	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	No	Never	Yes	I do not use social media but firmly believe that (legal) extremist content has a place in the public eye.	No	Explicit and illegal content posted to inappropriate platforms.	No	0.9	10	No organisation should be self-regulating, even if you built the platform yourself.	18-24	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	No	Depends on the content. If I had distributed something hateful denying a group their rights, something violent or pornagraphic I would understand. If I had just expressed a controversial opinion I would be aggrieved.	No	Homophobia, inciting violence, pornagraphic , violence.	No	0.9	2	Self-regulation never works. The turkeys will not vote for Christmas. They either pander to a demographic like Twitter or abdicate all responsibility like Facebook.	25-34	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Weekly	Yes	I would be annoyed and frustrated	Yes	Indecent content/fake accounts	I don't know	0.8	6	In theory it sounds great, but in practice it doesn't seem very effective. It needs to be improved as it misses some extremist content and removes some content that isn't extremist.	18-24	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Yes	As long as it was justified based on the guidelines set out I would understand, however could see how flyby doing this people could feel like views they had were being branded as extremist.	Yes	Anything that I genuinely do not believe should be put into an essentially public domain or I believe would cause harm	Yes	0.9	2	I think it is a dangerous precedent that they set in which they control millions of people's personal information and provides large companies a good platform to push their beliefs on their users. Cherry picking content has always been a part of social media	25-34	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	I would be confused as I do not post political or overtly offensive content and generally avoid discussing politics online.	No	I generally avoid reporting content in order to not infringe on people's right to freedom of speech. I would potentially report footage of graphic violence.	I don't know	1	3	It is important to acknowledge that many of these social media companies may use the guise of regulating extremist content to remove content that is not extremist in nature but clashes with their agenda or the popular political narrative such as expression of conservative views and opinions	18-24	Male	English / Welsh / Scottish / Northern Irish / British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	Yes	Upset	Yes	Illegal content or discriminatory based on people's protected characteristics	No	0.9	7	It doesn't work well	45-54	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Yes	I would have to consider the subject matter of the content and revisit it to understand if it was extreme or not. Based on the sort of content I usually share on SM, I would be rather surprised to discover something had been taken down for extremism.	No	If I were to find something illegal or perverse then I would probably report it.	I don't know	0.9	7	There should be an independent body to regulate the content in order to insure integrity and objectivity.	25-34	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	Disappointed. I'm a firm believer in free speech which essentially permits all view, extreme or not.	No		No	1	3	They should either be treated as publishers, and therefore editors of their content; or not regulate it themselves at all. There is too much fear that extreme content will harm the population. Most of us will see it for what it is. Censorship tends to remove beneficial content as well as the intended target. The cure for extreme content is reasoned argument.	Prefer not to say	Prefer not to say	Prefer not to say

People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	Yes	I do not agree with censoring any content including my own	No	I wouldn't	No	50% or lower	1	Abuse of power and should be illegal and prosecuted	18-24	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	Confused. I do not tend to post content with strong personal views.	No	If I deemed it extremist etc.	Yes	0.8	6	Hard to tell how seriously they would take it compared to an external company.	18-24	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Yes		Yes	Outwardly promoting something harmful.	I don't know	0.9	5		25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	Understandable if it was extreme	Yes	Racial, sexual or anything to do with children	Yes	50% or lower	10	Would rather an independent body regulate	45-54	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	No	Terrible	No	If I believed the content to be bullying, abusive, harmful or extreme.	Yes	1	9	I don't believe they do a good enough job.	18-24	Female	English / Welsh / Scottish / Northern Irish / British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	No	Fine, it should be removed	No	Harmful	Yes	0.9	5	Not sure	18-24	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Hourly	Not that I can remember	Not bothered	Yes	Abuse	Yes	1	5	Neither here or there	25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	No	Shocked	No	Threatening behaviour	Yes	0.9	2	They have a right as long as they meet the appropriate laws of that country and show how they are self regulating their platforms	18-24	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	No		No		Yes	0.9	3		25-34	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	No	Shocked	No	Threatening content	I don't know	0.8	1	As long as a balance is provided in keeping with laws of country. Self regulation should be transparent.	45-54	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Not that I can remember	I would be upset and want a thorough explanation.	Yes	Content that is offensive and hateful.	I don't know	0.8	4	I would be happy with self regulation if this meant actually taking action, but I feel this is largely not the case, especially due to the transient nature of social media, so would support more government intervention.	25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Yes	Frustrated.	Yes	If it was inciting hate or violence, or further marginalising people/ communities.	Yes	0.8	7	They’re terrible at it.	18-24	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Hourly	Yes	I do not have extreme views, so not relevant	No		Yes	0.8	1	They do not do a very good job of it, one persons ideal of extremity is different from someone else's point of view	45-54	Male	English / Welsh / Scottish / Northern Irish / British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Hourly	Yes	Surprised but I would want to find out why and educate myself	Yes	Antisemitism	Yes	0.9	10	Ok as long as it is thorough	18-24	Female	Ashkenazi jewish
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	Yes	Terrible	Yes	Normally gruesome images but sometimes it has been over hateful things	Yes	0.8	7	I dont think they are good enough or care enough	45-54	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	Yes	Confused as it's unlikely that I'd post anything that I would regard as extremist	Yes	Highly inappropriate or illegal content	Yes	0.8	7	I wouldn't trust that social media platforms would regulate content outside of their own aims and benefit. I feel they should be externally regulated.	25-34	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Not that I can remember	I wouldnt like it as my content isnt considered as extremist content.	No	Seeing racist, harmful content or anything showing any types of abuse	Yes	0.8	8	I feel that they will be influenced by things that will help their own company or influence users in a bad way.	25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	No	Never	Yes	Confused as I do not hold extremist views, values, beliefs or thoughts	No	Racism, extremism, sexism, general harmful content	Yes	0.9	1	Should be regulated by politically, socially, unbiased body. As per a set of rules set not to promote or discourage political views. As per a set of rules controlling conversation and discussions. Allowing them to take place but stopping short of becoming dangerous, racist, sexist	25-34	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Yes	I don't believe I would ever post content that would be taking this way but if it hurt or caused upset to others I would understand.	Yes	Usually for child or animal abuse.	Yes	0.8	5	If the companies were vetting the individuals who were monitoring that content and ensure that they are being checked out psychologically on a regular basis I would be okay with that.	25-34	Female	Irish
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	Yes	I do not agree with censoring any content including my own	No	I wouldn't	No	50% or lower	1	Abuse of power and should be illegal and prosecuted	18-24	Male	English / Welsh / Scottish / Northern Irish / British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	Yes	Understanding	Yes	If I feel the post incites hatred, includes fake news, or discriminates	Yes	0.9	10	Social media should be more regulated - if platforms are self-regulating they will naturally have bias and intentions of increasing engagement.	25-34	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	Yes		Yes	Abusive posts	I don't know	0.8	6	Regulations on social media are required for the safety of its users	18-24	Female	English / Welsh / Scottish / Northern Irish / British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	No	I would ask for the reason why then accept that it's against the content rules etc of the social media platform	No	hate crime, bullying, unnecessary content, violent acts towards men or women,	Yes	0.8	5	They themselves should be more aware and keep an eye on the content that is being put through social media, also there should be age limits to access, whereas you have to at least show a form or ID to be able to access them.	25-34	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	If it was a post which was voicing an opinion and opening a topic for proper discussion, I would not be happy. If it was a post which was directed at individual / groups and intentionally causing harm or upset, I would accept that another person did not see this as being suitable for social media.	Yes	Threats of violence, obvious attacks to slander a person / group which can not be justified, sexualising children.	Yes	0.8	7	Whilst they do attempt to, they are not able to regulate all users and often content can go months before any action is taken. Many wait for it to be reported before taking any action over content not being suitable.	25-34	Female	English / Welsh / Scottish / Northern Irish / British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Weekly	Not that I can remember	Confused.	No	Something offensive or videos of abusive.	Yes	0.9	1	I think any regulation is good but it may be better if an external non biased body.	18-24	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	No	Never	Yes	N/A	No	Violent content	Yes	0.8	10	Needs to be externally regulated	25-34	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	Yes	Shocked	Yes	Something that hugely differs from what my moral compass says is acceptable. Violence, racism, sexism, homophobia, completely factually incorrect information designed to divide.	Yes	0.9	3	They won't - My personal opinion is that money and power govern what the general public see on all social media platforms.	25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Yes		Yes	Outwardly promoting something harmful.	I don't know	0.9	5		25-34	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Yes	As long as it was justified based on the guidelines set out I would understand, however could see how flyby doing this people could feel like views they had were being branded as extremist.	Yes	Anything that I genuinely do not believe should be put into an essentially public domain or I believe would cause harm	Yes	0.9	2	I think it is a dangerous precedent that they set in which they control millions of people's personal information and provides large companies a good platform to push their beliefs on their users. Cherry picking content has always been a part of social media	25-34	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes	If it was a post which was voicing an opinion and opening a topic for proper discussion, I would not be happy. If it was a post which was directed at individual / groups and intentionally causing harm or upset, I would accept that another person did not see this as being suitable for social media.	Yes	Threats of violence, obvious attacks to slander a person / group which can not be justified, sexualising children.	Yes	0.8	7	Whilst they do attempt to, they are not able to regulate all users and often content can go months before any action is taken. Many wait for it to be reported before taking any action over content not being suitable.	25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Hourly	Yes	Devastated.	No	If I felt no one else had reported it.	No	0.7	2	I don't think they are capable of doing it on the current format. People should have to provide more verification to use social media.	25-34	Male	English / Welsh / Scottish / Northern Irish / British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	No	Fine, it should be removed	No	Harmful	Yes	0.9	5	Not sure	18-24	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Hourly	Yes	Depends on the context, however, given that my views on most topics are relatively mainstream I would take issue with my posts being removed as I would perceive it as a breach of my freedom of expression.	Yes	Insulting language directed at an individual or group. Slanderous or otherwise inflammatory comments that breach the platform's terms of service.	No	0.9	2	Social media companies are very poor at moderating their own content although I am loathed to allow the government further power to regulate free expression on online forums.	18-24	Male	English / Welsh / Scottish / Northern Irish / British
People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	Yes	If it was correct I'd be okay with it, otherwise I'd appeal.	No	I don't recall a time I've reported something but it would have to be offensive or bullying/ racism	Yes	50% or lower	5	All designed to promote addiction to the digital world and ludicrous one can self govern what you see and how/who you interact with.	25-34	Male	English / Welsh / Scottish / Northern Irish / British

People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	No	I would ask for the reason why then accept that it's against the content rules etc of the social media platform	No	hate crime, bullying, unnecessary content, violent acts towards men or women,	Yes	0.8	5	They themselves should be more aware and keep an eye on the content that is being put through social media, also there should be age limits to access, whereas you have to at least show a form or ID to be able to access them.	25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Not that I can remember	Annoyed, I am not an extremist and I do have extremist views.	Yes	If the content is harmful.	No	1	1	They're not doing a good enough job.	25-34	Male	English / Welsh / Scottish / Northern Irish / British

People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	Yes	If it was correct I'd be okay with it, otherwise I'd appeal.	No	I don't recall a time I've reported something but it would have to be offensive or bullying/ racism	Yes	50% or lower	5	All designed to promote addiction to the digital world and ludicrous one can self govern what you see and how/who you interact with.	25-34	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Hourly	Yes	Devastated.	No	If I felt no one else had reported it.	No	0.7	2	I don't think they are capable of doing it on the current format. People should have to provide more verification to use social media.	25-34	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	Yes		Yes	Abusive posts	I don't know	0.8	6	Regulations on social media are required for the safety of its users	18-24	Female	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Hourly	Not that I can remember	Not bothered	Yes	Abuse	Yes	1	5	Neither here or there	25-34	Female	English / Welsh / Scottish / Northern Irish / British
Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Daily	Yes		Yes	If the content was inappropriate, violent or racist	I don't know	0.9	5	I feel that this is okay in most cases but does allow social media companies to push or hide information depending on their views and interests.	18-24	Male	English / Welsh / Scottish / Northern Irish / British

People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Daily	Yes	Concerned! And would like to know why / what was breached	Yes	Inappropriate, extreme, worthy of warning for other users	Yes	0.7	5	I feel that social media companies do not do a good job of self regulation and it is more than likely the users themselves who report. They have such a large financial base and do not do enough.	25-34	Male	English / Welsh / Scottish / Northern Irish / British
Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.	Yes	Daily	Yes	Shocked	Yes	Something that hugely differs from what my moral compass says is acceptable. Violence, racism, sexism, homophobia, completely factually incorrect information designed to divide.	Yes	0.9	3	They won't - My personal opinion is that money and power govern what the general public see on all social media platforms.	25-34	Female	English / Welsh / Scottish / Northern Irish / British

People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.	Yes	Hourly	Yes	If it was an extremist view I'd be happy it's been removed	Yes	Anything that targets a group of people based in things they cant change, or mob mentality and bullying	Yes	0.9	6	Nonsense. Should be a universal measure. Certain companies provide a platform for extremism allowing poorly constructed opinions based on false information to be aired, which leads to widespread misinformation and hate. Even twitter banned Katie Hopkins.	18-24	Male	Prefer not to say
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Hourly	Yes	Depends on the content and context	No	If it is excessively criminal	I don't know	1	3	Its not consistent	18-24	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Not that I can remember	Would not happen as I do not upload any extremist content	No	If I felt it was lies to try to invoke certain feelings in people or if it was nasty and disgusting content	Yes	0.8	1	Don’t agree with that.There needs to be a higher lawful power dealing with social media companies.	25-34	Female	Irish
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	Yes	It's whether it contains genuine, harmful content or not. Many things are flagged as dangerous when they aren't, while other genuinely dangerous and harmful content is not removed. Social media users have a wide range of views on topics such as politics, religion and more. A view considered ethically and morally right by one group may be considered extreme and inappropriate by another group. To answer the question: no I probably wouldn't like social media to dictate to me what is extreme and what isn't. If there is something that I don't approve of or I find harmful, I will remove it myself.	Yes	Blatantly prejudiced content, sexually exploitative content especially of minors, threatening messages directed toward individuals or groups, doxxing.	I don't know	1	3	Unlike news papers, social media sites like Facebook are not publishers, they are public forums. They have users numbered in the billions across every continent, making them an integral part of the lives of a significant percent of the planet. With the influence they have, they have an immense amount of responsibility, and that responsibility should be exercised very cautiously when it comes to	18-24	Male	English / Welsh / Scottish / Northern Irish / British

Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.	Yes	Hourly	Not that I can remember	Quite annoyed, I would pursue further action to determine exactly what aspect of the content was deemed extremist.	No	Any calls of violence, or deliberate doxing.	No	0.9	9	They are private companies and should have the right to, however, some of the guidelines can be quite ambigous which can result in banning of content based along political or ideological reasons.	18-24	Male	Polish
Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.	Yes	Daily	No	I am not an extremist so if content was removed for containing extremist content I would be extremely confused! However I wholeheartedly agree that all extremist content should be removed from social media.	No	If it was hateful, racist, bullying, illegal, concerning	Yes	0.8	4	Concerning. Most companies have external regulatory bodies and to have such a prolific platform there really must be non biased, independent regulator monitoring content.	45-54	Female	English / Welsh / Scottish / Northern Irish / British

Appendix B

Survey Template

1. Out of the following four definitions of extremism, which if any do you think is the most accurate?

   O Extremism is the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist.

   O Extremism is generally understood as constituting views that are far from those of the majority of the population. Extremist views are not necessarily illegal and do not automatically lead to violence or harm; indeed those who chose to observe extreme practices with no impact on the civil liberties of fellow citizens are rightly protected under fundamental freedoms and human rights norms.

   O People who have certain beliefs about politics or religions which are hateful, dangerous or against the law are often known as extremists. This harmful behaviour is called extremism.

   O Extremism is essentially a political term which determines those activities that are not morally, ideologically or politically in accordance with written (legal and constitutional) and non-written norms of the state; that are fully intolerant toward others and reject

democracy as a means of governance and the way of solving problems; and finally, that reject the existing social order.

O None of the above.

2. Do you use social media at all (e.g., Facebook, Twitter, etc.)?
   O Yes
   O No
   O Prefer not to say

3. How often do you use social media?
   O Hourly
   O Daily
   O Every other day
   O Weekly
   O Monthly
   O Other
   If 'Other' please state.....

4. In the last year, have you seen extremist content based on your understanding of the term?
   O Yes
   O No
   O I don't know
   O Not that I can remember
   O Other
   If 'Other' please state.....

5. How would you feel if your social media content was removed after being flagged for containing extremist content?
   .....
   .....
   .....

.....

6. Do you ever report social media content?
   O Yes
   O No
   O Prefer not to say

7. What reasons would make you report content?
   .....
   .....
   .....
   .....

8. Would you find it justifiable if a certain percentage of social media content would be removed if this meant that all extremist content was identified and removed?
   O Yes
   O No
   O I don't know
   O Prefer not to say

9. In light of the greater good, what is the lowest level of accuracy you would tolerate for an AI system that removes extremist content?
   O 100%
   O 90%
   O 80%
   O 70%
   O 60%
   O 50% or lower

10. How comfortable are you with privacy levels being reduced for security levels to be increased?

Very uncomfortable 1 2 3 4 5 6 7 8 9 10 Very uncomfortable

11. What are your thoughts on social media companies mostly self-regulating the content on their platforms?

    .....

    .....

    .....

    .....

12. What is your age range?

    O <18

    O 18-24

    O 25-34

    O 35-44

    O 45-54

    O 55-64

    O >64

13. What is your gender?

    O Male

    O Female

    O Prefer not to say

    O Other

    If 'Other' you can optionally state.....

14. What is your ethnicity?

    O English / Welsh / Scottish / Northern Irish / British

    O Irish

    O Gypsy or Irish Traveller

    O White and Black Caribbean

    O White and Black African

    O White and Asian

    O Asian / Asian British

O Indian

O Pakistani

O Bangladeshi

O Chinese

O African

O Caribbean

O Arab

O Prefer not to say

O Other

If 'Other' please state.....