Lex Informatica and freedom of expression: reflections on the regulation of Internet trolling behaviors

David McMenemy [0000-0002-3203-9001]

Information Studies, University of Glasgow, Glasgow, Scotland
david.mcmenemy@glasgow.ac.uk

Abstract. This paper revisits the concepts of *lex informatica* (Reidenberg) and *code is law* (Lessig), both early theories related to how regulation of behaviors on the internet might be managed, By focusing on the context of internet trolling in the United Kingdom, we consider the nature of trolling, how it has manifested in the UK in terms of actual notable incidences, and reflect on whether the law of the state within the regulatory regimes envisioned by Reidenberg and Lessig are in fact the best fit for managing these behaviors. Abusive behavior online can have profound impacts on people, causing significant fear and leading to an impact on private and family life. In balancing the qualified nature of freedom of expression rights versus other human rights we consider whether the laws of peoples, the regulatory capabilities of social media companies, and the wider culture of digital society are best placed collectively to create respectful norms on the Internet.

Keywords: trolling, freedom of expression, harassment, law, social media,

1 Introduction

This paper explores a subset of Internet regulation by investigating practice with regards to the extent of trolling/offensive behaviors on Twitter, with emphasis on the United Kingdom.

As early as 1999, Lawrence Lessig claimed that the regulation of the Internet could be relatively straightforward, and although involving differences to regulation in the real world, could in essence be even *more* effective and potentially regressive than regulation in the real world. Lessig argued that what would emerge would be, "an architecture that will perfect control and make highly efficient regulation possible" spurred on by both governments and corporations [1]. Lessig's work was inspired by the earlier work of Reidenberg, who stated in a 1997 piece that, "the set of rules for information flows imposed by technology and communication networks form a "Lex Informatica" that policymakers must understand, consciously recognize, and encourage" [2].

What we have then is a digital domain where the laws of society have their place, but ultimately that is also controlled by the systems and architectures, rules, and norms of behaviors in the digital spaces we occupy. For Lessig these constraints on the digital world could be classified under four modalities of control: laws, norms, markets, and

code (or architecture) [1]. We can see the viability of Lessig's four modalities clearly as still being of significant relevance. On the one hand social media companies demand all users sign up to a set of terms and conditions (T and Cs) before they are given an account, and these invariably include elements of regulation of behavior related to trolling and harassment. This is the element of regulation that is applied by the *market* itself, in the hope of encouraging acceptable *norms*. In addition, the system *architecture* itself can be utilized to control trolling behaviors by analyzing words and tone of posts and blocking content and/or restrict access to users who breach norms. On the other hand, the *law* also takes an interest in trolling activity, with the Crown Prosecution Service (CPS) in England and Wales producing clear guidelines and a typography for prosecutors on how to deal with alleged trolling behavior from a legal standpoint [3].

2 Social media trolling: an ever-growing problem?

Social media trolling and bullying affects one in four young people, with targets disproportionately coming from disabled groups and ethnic minorities [4]. In recent years leading Members of Parliament (MPs) in the UK such as Jess Philips [5] and Yvette Cooper have also been victims of the activity, with Cooper calling for action to specifically prevent the abuse of women on social media [6]. In terms of creating a safe and welcoming space, "trolling has been framed as a major, if not the major, impediment to online community formation" [7]. In terms of Twitter trolling activity can quickly become a major news story, notwithstanding any legal implications which may follow. There have increasingly been incidents whereby public figures have been exposed to significant amounts of abuse online. MP Stella Creasy reportedly installed a panic button in her home after death threats on Twitter, and Caroline Criado-Perez reported impacts on her wellbeing and state of mind as a result of messages received. At the height of the abuse, Criado-Perez was receiving up to 50 offensive tweets per hour [8].

In 2017 a campaign utilizing the hashtag #ReclaimTheInternet was backed by MPs Yvette Cooper and Maria Miller. The campaign aimed to counter sexist trolling of women on social media and was launched by both MPs to significant publicity. Cooper raised the issue that much activity identified as trolling is not necessarily illegal:

> ..in most cases, online abuse isn't a crime. So, the question is what responsibility all of us have - as individuals, through campaigns, through schools, workplaces, unions, and through publishing platforms and social media - to challenge it and change attitudes [9].

In the words from Yvette Cooper, we can see reflected Lessig's modalities, with a call for better norms through individual behaviors, but also the responsibilities of the markets themselves to build better platforms. There is a crucial need, then, when discussing the regulation of trolling to consider both the potentially illegal elements of the behaviors, alongside the kinds of behaviors that, though not illegal, may be able to be dealt with by social media companies T and Cs.

2.1 Free speech and trolling: the liberal dilemma

A key aspect of the freedom desired by many on the Internet is that related to free speech, or freedom of expression. From an American perspective the cherished First Amendment to the Constitution has meant that free speech rights are guarded, and the emphasis falls most often on the right of the speaker unless they are defaming a person, breaching their target's intellectual property rights, or otherwise injuring them via some form of *fighting words* (words designed to provoke or incite violence).

The United Kingdom has no written constitution to protect free speech, however a rich tradition has evolved related to the utility of free speech for society, and the introduction of the European Convention on Human Rights via the 1998 *Human Rights Act* enshrined a *qualified* legal right to freedom of expression, balanced against other rights. The balancing of freedom of expression versus other human rights remains the crucial qualification. As Nussbaum argues, "none of the major philosophical theories gives us reason to think that repeated slurs, or cyber bullying, are high-value speech" [10]. The qualification stated in Article 10 is, "since [freedom of expression] carries with it duties and responsibilities, [it] may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society" [11].

2.1.1. The nature of communication on the internet

Spinello has observed, "democratization of speech and information may well be the greatest legacy of the … Internet era" [12]. However, "unencumbered by generally accepted social norms [people] are more prone to say and do things that they perhaps would not under their real identities" [13]. McGoldrick suggests that "the empowerment provided by the internet has proved intoxicating and led individuals to issue communications as though they were within a 'Wild West' type, law-free, zone in Cyberspace" [14].

2.2 Arguments for and against free speech in the online space

Barlow's "Internet Manifesto" declaration laid down a bold vision for free speech on the Internet, one that was grounded in a libertarian adherence to First Amendment principles. Barlow's cyberspace was "a world where anyone, anywhere may express his or her beliefs, no matter how singular, without fear of being coerced into silence or conformity" [15]. As Danielle Citron has observed, however: "The Internet is a double-edged sword. While it can facilitate the empowerment of people who often face discrimination, it can also be exploited to disenfranchise those very same individuals" [16]. It is an obvious assertion, but the Internet as it was envisioned by Barlow and the early pioneers did not foresee dilemmas such as online trolls, threats of physical violence, and death threats, and the types of discriminatory interactions as alluded to by Citron. With the distance of time passed since Barlow, there is an argument that could be posited that the free speech values of 1996 cyberspace are not readily applicable to 2022, assuming they ever were.

The 1997 Supreme Court decision in *Reno v ACLU* offered that the Internet enabled anyone to become "a pamphleteer…a town crier with a voice that resonates farther than

it could from any soapbox" [17]. Important in the decision of the court was that levels of protection appropriate for children should not be the norm for adults using the Internet, and that while protections must rightly extend in the areas of obscenity or child pornography, free speech was too important a right to be toyed with. As was acknowledged, however, there are forms of expression that pose challenges to the notion of free speech on the Internet, and this poses "a contentious moral problem" [12].

Reno v ACLU encapsulated much of what is at stake in the free speech debate with regards to the Internet. A contemporary reflection on the case by Lipschultz offered that the decision of the court represented an adherence to the marketplace of ideas approach to free speech [18]. Within this philosophical approach to free speech the notion is that all ideas should be expressible, and the market (or the community) will decide which have efficacy or not, allowing all ideas to be tested. Critics of the marketplace of ideas approach to free speech would argue that some people have more access to distribute their ideas than others. As Barendt suggests, "the marketplace is not in practice open to everyone who wants to communicate his ideas……Differences in the availability of ideas have little to do with their truth" [19]. Others have posited that often what is expressed when hate speech is the topic are not ideas but merely passions that should not be supported in a civilized society [20].

The marketplace of ideas does not take into account that the expression of some types of view actually may work in such a way that those ideas drive others away through fear. Again, in the context of online trolling, Citron argues that:

> When individuals go offline or assume pseudonyms to avoid bigoted cyber-attacks, they miss innumerable economic and social opportunities. They suffer feelings of shame and isolation. Cyber mobs effectively deny people the right to participate in online life as equals [16].

Abah reiterates this view, stating: "the marketplace stops functioning as a marketplace of different ideas and becomes a monopoly if certain people are driven off from the center to the periphery, or chased completely off by intimidation and threats, not because their ideas and contributions are bad, but because their ideas contradict and challenge the status quo or are contrary to the presumed mainstream ideas" [21]. Free speech proponents might counter that the people who feel driven away by trolls have a right of reply, but this does not take into account where the victims may experience fear or harassment to such an extent that they feel a need to withdraw themselves or restrict their online interactions [22]. Such withdrawal has consequences for victims over and above any fear or harassment felt in terms of missed opportunities.

2.3 The concept of trolling

The Dictionary of Contemporary Slang defines troll as: "a malicious, anonymous online presence" [23]. As the Guardian newspaper commented in September 2011, '…technically speaking, a troll isn't someone who is merely offensive…They're people who purposefully drag an online conversation off-topic – often by being offensive, but sometimes just by being needlessly pedantic or bizarre' [24]. At the heart of the traditional meaning of troll, as it is applied to the Internet, is related to a sub-culture who sought to gain social capital among members by disrupting message board threads.

This was seen as a challenge, and the more disruptive a troll could be, the more they earned respect in the eyes of the sub-culture. In this early phase of the Internet trolls were just that, a sub-culture, that might have a nuisance value to users of forums they had disrupted, but they were not in any way a societal problem. As Lovink highlights, "the problem of trolling can easily be isolated to individual cases. Trolls are figures of exception" [25].

Increasingly in the modern era, the term troll has been used extensively for a range of behaviors on the Internet, including significant incidences of harassment. In her research on the trolling phenomena, Whitney Phillips found significant frustration amongst those who identify as trolls in the "traditional" sense with the "increasingly fuzzy popular definition of trolling, which in mainstream media circles has been attributed to such a wide variety of behaviors that it has been rendered almost meaningless" [7, p.158].

2.4 The culture of trolling in the modern era

There is little doubt that the idea of the Internet troll has transcended its sub-culture boundaries and become a staple of popular culture. In many ways this provides a cachet the perpetrator does not deserve. Another emerging element of trolling and online harassment in recent years has been its identification with the political movement known as the Alt-Right. This rise has been mapped by Angela Nagle in a recent book, and she suggests the link between trolling of vulnerable groups and the alt-right community is a significant one:

> One of the things that linked the often nihilistic and ironic chan culture to a wider culture of the alt-right orbit was their opposition to political correctness, feminism, multiculturalism, etc., and its encroachment into their freewheeling world of anonymity and tech [26].

While on one hand the troll subculture might argue their activities do not constitute genuine threats, the merging of elements of that subculture with the alt-right, and the white supremacist elements within it, suggest that trolling and politics have become more integrated. It therefore asks a lot for a victim of trolling to be able to make the distinction between the high jinks of "legitimate" trolls versus the received message that may constitute a significant threat. As Diaz has offered:

> Detecting when one is being "trolled" on the Internet is often an impossible task considering the anonymity of the speaker, and the ambiguity of text.' A jest may appear as a threat, sarcasm as defamation, or criticism as bullying [27].

It is unreasonable to ask digital citizens who are not part of a sub-culture to respect abuse and/or harassment as a joke when it appears to be identical to the real thing. The reality of the troll on social media is that they are often merely bullies, often targeting vulnerable groups and causing them distress at worst, and nuisance value at best:

> Anonymous mobs employ collaborative technologies to terrorize and silence women, people of color, and other minorities. The harassment typically includes threats of sexual violence, postings of individuals' home addresses alongside the suggestion that they should be raped, and technological attacks that shut down

> blogs and websites. Cyber mobs brand targeted individuals as inferior beings and as sexual objects [16, Kindle Locations 379-381].

The anonymity element of cyberbullying/trolling cannot be underestimated. In research undertaken in 2015 by the Pew Research Center, the statistics on this were stark, highlighting the extent of anonymous harassment received:

> [O]f those who have experienced online harassment 38%, said a stranger was responsible for their most recent incident and another 26% said they didn't know the real identity of the person or people involved. Taken together, this means half of those who have experienced online harassment did not know the person involved in their most recent incident [28].

The impact of trolling or cyberbullying can be significant on a victim. As Rosewarne has argued, "in the context of cyberbullying, a single electronic attack can in fact have recurrent and repetitious effects" [29]. The ubiquity of words on the Internet means that the victim may feel there is no escape from the repeatedly experiencing the bullying, since anyone with a computer can see it, and unless deleted or the perpetrator blocked, the victim may always see it in their timeline or on their account. The act of retweeting, a technique which forwards an initial tweet, means that the followers of the person retweeting will also then be aware of the initial act. Compounding this is the belief that perhaps the offending tweet will always be there, with the use of search engines to archive twitter data it may well always be on the Internet, following the victim and forever reminding them of the interaction. As Carrabis and Haimovitch argue, in the "new online world, victims have no safe haven to retreat from these public malicious acts of cyberbullying" [30]. As Delgado and Stefancic offer:

> [M]uch material posted on the Internet will remain there indefinitely, becoming "a permanent or semi-permanent part of the visible environment in which our lives, and the lives of vulnerable minorities, have to be lived." If the hate message "goes viral," it may attract millions of viewers and remain in cyberspace, perhaps forever [31].

Notwithstanding the ability of the victim to remove themselves from the perpetrator via blocking mechanisms and the like, the impact of trolling or cyberbullying can have profound effects on the person on the receiving end.

We will explore this further when we discuss the criminalization of trolling, but examples that challenge the victim/perpetrator narrative from the point of view of desert can also be regularly found and are of vital importance in considering the complexity of Internet culture. In Ronson's *So you've been publicly shamed*, the author explores the mass shaming on the Internet of people who have been believed to have transcended a norm. Such shaming exercises follow remarkably similar patterns as other cyberbullying and trolling examples. X says something that Y or Group Y on the Internet believes to be wrong, and the responses in opposition then are sent. Interestingly in Ronson's exploration, the group shaming was seen to be virtuous by those taking part in it. He highlights the case of Justine Sacco who tweeted a tasteless joke before boarding a flight to South Africa: 'Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!' [32, p.64]. Sacco arrived at the other side completely oblivious to the reality that her tweet had gone viral. Most of the tweets were genuine shock and outrage at something they deemed to be racist, however many of the tweets did seem to be akin

to the type one would associate with trolling/cyberbullying. There is a sense in some of the public shaming that if people are so sure about their world view, they do not tend to see that what they are doing might constitute trolling or bullying. The notion that the victim *deserves* the treatment is one that all bullies may allow themselves to be justified by. Commenting on Ronson's work on public shaming, Peter Bradshaw of *The Guardian* suggested: "Twitter-shaming allows people who complacently think of themselves as basically nice to indulge in the dark thrill of bullying – in a righteous cause. Perhaps Ronson's article will cause a questioning of Twitter's instant-Salem culture of shame" [33].

A final point related to online harassment/trolling is the gendered elements that relate to it. While the Pew study found that men receive more harassment related to public embarrassment and the calling of offensive names, it also highlighted that females, and young females 18-24 especially, receive the most severe forms of online harassment, such as stalking, physical threats, and sustained harassment [28]. Clearly these more severe forms of harassment move beyond a simple freedom of expression justification into areas of wrong doing that have genuine implications for womens' safety online. In this context it is important to highlight that while harassment online is meted out to all, concern must be expressed at the types aimed primarily at women. Vitak et al undertook a study of undergraduate women across US universities related to their experiences of online harassment, and a key finding suggested that women were becoming almost resigned to abuse being part of the online experience if they want to remain a part of it: "Even when women do not retreat from online spaces, a disheartening trend exposed in both anecdotal work and this study is the general sense that women are tolerant of these behaviors because they have become part and parcel of interacting online" [34].

3 Criminality of trolling

Social media is at its root an example of electronic communication, and the law is quite clear that electronic communications technology "can be used to incite, encourage or assist another in the commission of an offence, or to form a conspiracy" [35]. Jones organizes his treatment of the topic under four categories of offence, which closely relate to the CPS guidelines on prosecuting social media offences, as we will see further below:

- Threats of violence or damage to property.
- Harassment of an individual.
- Breaches of court orders.
- Communications that are grossly offensive, indecent or obscene

Jones points out that a threat to kill someone is unlawful under s.16 of the *Offences against the Person Act 1861* regardless of the medium from which that threat is received, and anyone doing so shall be guilty of an offence and liable on conviction on indictment to imprisonment for a term not exceeding ten years." Importantly, Jones points out that the *mens rea* (guilty mind) must be clear and that a threat delivered as a

joke would not succeed: the intent must be to make the victim fear that the threat would be carried out. In an online scenario where over 50% of trolling behavior comes from an unknown person, how is one to tell a joke from the real thing?

General threats that might cause distress to a victim are an offence in England and Wales under s.1 of the *Malicious Communications Act* 1988 and s127 of the *Communications Act 2003*. s1 of the *Malicious Communications Act* states that a person may be guilty if they send "a letter or other article" which conveys a threat, and he is guilty of an offence if "distress or anxiety to the recipient or to any other person to whom he intends that it or its contents or nature should be communicated." Importantly there is a defense available which states that a person is not guilty if the threat was used to reinforce a demand, or he believed the threat was a legitimate means of doing so.

In *DPP v Collins* the network utilized for transmitting offensive content was the telephone network, with the defendant leaving a series of offensive telephone messages on the answering service of his MP [36]. The messages used several extremely strong racial epithets which were said to have extremely offended the MP's staff, none of whom were from the ethnic minorities targeted by the insults. The defendant was initially acquitted on the basis that his messages were offensive, but not grossly so. This was held on appeal initially; however, he was finally convicted on subsequent appeal when the Law Lords decided that the offensive messages would grossly offend a reasonable person in a multicultural society, and he should have been convicted under s127 of the *Communications Act 2003*. The *actus reus* (guilty act), then, was deemed to be sufficient to warrant conviction under s127, and the defendant should have been aware that his messages may offend, whether he intended them to or not.

Perhaps the most influential case related to social media behavior in UK law date has been *Chambers v DPP* [36] in which the key legal question related to the intention of the tweeter, and therefore the *mens rea*. In this case, which the defendant won on appeal, the defendant sent out a now infamous tweet to his followers about Robin Hood Airport in England, where he was due to leave for a trip with a love interest, and the fact that it was closed for snow. The tweet said:

> *Crap! Robin Hood Airport is closed! You've got a week and a bit to get your shit together otherwise I am blowing the airport sky high!!*

On viewing the tweet later, an airport employee informed the police and Chambers was arrested and charged under s127(1)(a) of the *Communications Act 2003* with sending a message of "a menacing character" on a public electronic communications network. Chambers was convicted at the original trial, despite him claiming that the tweet was an obvious joke and that no *mens rea* existed. Nevertheless, the court took the stance that s127(1)(a) was a strict liability offence, and that sending the tweet was sufficient to convict [36].

An initial appeal was lost by the defendant, but a later Divisional Court found in his favor, with the court taking the stance that the legislation was not designed to chill freedom of expression. The judgement summarized that: "We would merely emphasise that even expressed in these terms, the mental element of the offence is directed exclusively to the state of the mind of the offender, and that if he may have intended the message as a joke, even if a poor joke in bad taste, it is unlikely that the mens rea

required before conviction for the offence of sending a message of a menacing character will be established" [36].

Chambers v DPP has provided a high threshold for prosecutions under s127, erring on the side of freedom of expression over offense. Convictions are still clearly possible and probable under the statute, but the dangers of a chilling wind of jokes being perceived as threats receded somewhat with the decision. Murray suggests that the case provided a rap on the knuckles for lower courts in England and Wales who failed to consider both the *actus reus* and *mens rea* of the offence [37]. The case arguably highlights a double-edged issue, in that technology is feared as a mechanism for delivering offensive content, but also that content related to terrorism, even bad jokes, is treated with little patience by some courts. Judge Bennett in the original appeal classified the tweet as being "being of a menacing nature in the context of the times in which we live" which suggests there was as much about fear of terrorism in the original decision as there was in a reasonable approach to s127 of the *Communications Act* [38]. The Chambers case led to the development of guidelines by the CPS for prosecuting social media cases, which we cited earlier.

R v Stacey was another case where the *Public Order Act 1988* was used to prosecute a Twitter troll [39]. In this case the defendant posted an offensive tweet with regards to a footballer, Fabrice Muamba, who had collapsed on the pitch gravely ill during a match. When taken to task by other Twitter users on the offending post, Stacey replied with further tweets that were racially offensive. Stacey was charged and convicted under elements of the *Public Order Act* related to racially offensive speech, rather than any of the legislation related to the online elements. This case reinforced that while the medium of the message may be at the heart of the ability to commit the crime, crimes that have real-world equivalences can be charged under existing legislation where it is deemed to be fit to do so.

A final important case to discuss is *R v Nimmo and Sorley* [40]. In this case Caroline Criado Perez and Stella Creasy MP were on the receiving end of harassing tweets from both defendants related to a campaign they had been involved in to have Jane Austen appear on a bank note. Tweets traced to an account operated by Sorley were summarised in the sentencing report as:

> "F*** off and die...you should have jumped in front of horses, go die; I will find you and you don't want to know what I will do when I do... kill yourself before I do; r*** is the last of your worries; I've just got out of prison and would happily do more time to see you berried; seriously go kill yourself! I will get less time for that; r***?! I'd do a lot worse things than r*** you" [40].

Nimmo's tweets were in a similar vein, threatening, and coming from several accounts all linked to the defendant. The sentencing comments feature a victim report, and it makes for stark reading. Miss Criado-Perez stated that the tweets received had been life-changing in terms of putting her in fear. The report goes on to state:

> She feared the abusers would find her and carry out the threats. She felt hunted. She remembers feeling terror every time the doorbell rang. She has had to spend substantial time and money ensuring she is as untrackable as possible [40].

Creasy informed the court of the impact on her life, including installing a panic button at her home, as well as the fear it instilled in both her family and her staff.

R v Nimmo and Sorley brought the topic of Twitter abuse into the public consciousness even more strongly, especially around the gendered elements of the abuse. Despite one of the defendants in the case being female, the anonymity offered by Twitter accounts meant that the victims had no idea who made the threats, or where they were. This uncertainty and trepidation were clear in the comments made in the sentencing report, and mirrors the research discussed above by Citron. The fear instilled by harassment on social media is real, and notwithstanding times when the purpose is to trick or joke, the impact on victims is clearly stark in many circumstances. In *R v Nimmo and Sorley* an element of the case highlighted that both defendants were to some extent social misfits, and the technology allowed them to vent in what they perceived to be anonymity and without repercussions. This belief that such behavior on social media is unlikely to be traced back might explain some of the worst of the incidences, however as discussed earlier, the troll subculture may also mean that for some, trolling brings some kind of social benefit or kudos from within specific sub-communities that makes it worth the risk.

Harassment of an individual: Under s.2 and/or s.2A *Protection from Harassment Act 1997* two or more messages sent to a victim can constitute an offence. Citing *Majrowski v Guy's and St Thomas's NHS Trust* we are reminded by the courts that day to day life involves a range of situations where we will come across annoyances:

> Courts are well able to recognise the boundary between conduct which is unattractive, even unreasonable, and conduct which is oppressive and unacceptable. To cross the boundary from the regrettable to the unacceptable the gravity of the misconduct must be of an order which would sustain criminal liability [41].

In terms of social media, then, the messages received by the victim would need to meet this threshold to be liable to prosecution under the *Protection from Harassment Act*. In all of the potential prosecutions of social media interactions, the courts are reminded that the need to balance free speech with the rights of victims is paramount. Courts are especially reminded of the chilling effect on free speech of criminalizing behaviors that may well be unsavory but may well have to be permitted to occur in a free society. However, in the cases cited it is difficult to see how the laws passed by the state should not take a role in trying to punish such behaviors.

Can this be left simply to the *market*? It must be noted that the difficulty of regulating social media from the point of view of trolling and freedom of expression is a challenging task, given freedom of expression is such a culturally located value. Social media companies based in the USA and built on First Amendment values are essentially offering to the world a service that transcends the rights and values of even some liberal democracies in Europe. *Law* within a jurisdiction can attempt to address this, but the *code or architecture* of the social media companies may well be best placed to do so.

4 Regulation by social media

Outwith the differing legal systems in which companies such as Twitter and Facebook operate, there is also increasing pressure for the companies to be effective regulators of

their own services, with a great deal of this work undertaken by code applied to the content setting standards, or norms, of behavior expected.

The relationship between the social media platform and the user forms a contract between them: the user agrees to adhere to behavioral norms, the social media company punishes anyone who deviates from these norms. Both sides living up to their side of the bargain ultimately would make regulation straightforward. Nevertheless, even if social media companies were successful in regulating most offensive content, some people would likely still be on the receiving end of content that could be deemed illegal.

In February/March 2017 Twitter introduced new initiatives to attempt to reduce the incidences of harassment some members were receiving. Included was the ability to ignore unverified accounts, set up safer searching, and provide more control over who can contact you (i.e. limiting access to you from people you don't follow, or new accounts) [42]. More recently, additions have included the ability to set who can respond to a tweet you send, as well as the ability to create a curated set of followers within your overall number. In addition, functions like blocking users who you may not wish to interact with, and reporting activity you feel breaks Twitter rules are open to anyone receiving (or seeing) abusive content.

Reidenberg highlighted that the regulation of content is a basic dilemma of policy, and that it "poses intricate philosophical, practical, and political complications" [2]. As social media companies operate in a global marketplace, the pressure on them to censor within specific jurisdictions and not others has become a pressing one. As Reidenberg summarizes, "network service providers may opt for the overly cautious route of self-censorship and opt policies of 'when in doubt, take it out'" [2].

The *Twitter Rules* document contains the regulatory information with regards to behaviors on the platform. At the preamble to the document, it is stated clearly that Twitter stresses the important of the user experience and the safety of its users. It requires all members to adhere to the rules, which include provisions related to abuse of copyright as well as specific behaviors related to harassment. The main elements that relate to the regulation of trolling behaviors (there are nine Twitter Rules in total) are:

- Violence
- Abuse/Harassment
- Hateful Conduct [43]

All the functionality in the world will not stop trolls from harassing victims when they simply need to create new accounts to perpetuate the attacks (such as in *R v Nimmo and Sorley*) if they are blocked and/or banned by the victim. Indeed, the increasing volumes of abuse received because of the use of multiple accounts is likely to make the victim feel even more vulnerable, since it is not clear if the harassment is coming from the same original troller, or if multiple others are joining the attacks. This enhanced fear is not something that is immediately within the power of the social media companies to prevent, unless the victim removes themselves from social media, and thus the obvious remedy does seem to necessitate recourse to the law.

5 Conclusions

This paper has sought to explore some legal and regulatory issues of Internet trolling and offensive behavior from a UK perspective. While there are calls on social media companies to do more to prevent abuse and harassment online, ultimately the UK state is of a mind to step in to regulate trolling via both new laws and existing laws, all the while being mindful of the clear potential for impacting on freedom of expression rights. At the time of writing in the UK, the government has introduced the *Online Safety Bill* which seems to place more onus on social media companies to regulate the content that is hosted by them [41]. The key emphasis in the bill is on companies to improve their architecture and procedures to limit exposure to harmful content, empowering users to be able to place more and easier limitations on what they are exposed to.

In closing, Lessig's four modalities remain a key paradigm of how trolling and abusive behaviors can be managed: the Internet can be regulated via the traditional justice system (law), which can be supported via the algorithms that govern usage of the services provided (code) leading to the third modality (markets) setting the parameters for the fourth (norms). Lovink suggests, the extent of how the modalities can deal with the problem is not straightforward and will say a lot about the society we live in: "Editors, programmers and, eventually, the Law will deal with the unstoppable deviant Other…. The way society deals with those who cross invisible lines tells us a lot about the limits of the rhetoric of tolerance, openness, and freedom" [25, p.163].

References

1. Lessig L, Code: and other laws of cyberspace (Basic Books 1999) p.4.
2. Reidenberg, J.R. Lex Informatica: The Formulation of Information Policy Rules through Technology , 76 Texas Law Review.. 553 (1997-1998) Available at: https://ir.lawnet.fordham.edu/faculty_scholarship/42
3. Crown Prosecution Service. *Guidelines on prosecuting cases involving communications sent via social media*. 2018. Available from: https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media
4. Gani, Asha, "Internet trolling: quarter of teenagers suffered online abuse last year" *The Guardian*. 9th February 2016. https://www.theguardian.com/uk-news/2016/feb/09/internet-trolling-teenagers-online-abuse-hate-cyberbullying
5. Allegretti, Aubrey. Jess Phillips Responds To Trolls Who Sent Her Rape Threats On Twitter. *The Huffington Post*. Avaible from: http://www.huffingtonpost.co.uk/entry/jess-phillips-rape-threats-twitter_uk_574d95c6e4b03e9b9ed6262c
6. Labour's Yvette Cooper explores trolling aimed at women. *BBC News*. 17th December 2015. http://www.bbc.co.uk/news/av/uk-politics-35120551/labour-s-yvette-cooper-explores-trolling-aimed-at-women
7. Phillips, W. This is why we can't have nice things: mapping the relationship between online trolling and mainstream culture. Cambridge, MIT Press. 2015. p.16.
8. Carter, C. 'Twitter troll jailed for 'campaign of hatred' against Stella Creasy.' The Daily Telegraph. 24th September 2014. Available from: http://www.telegraph.co.uk/news/uknews/crime/11127808/Twitter-troll-jailed-for-campaign-of-hatred-against-Stella-Creasy.html

9. Cooper, Y. 'Why I'm campaigning to reclaim the internet from sexist trolls. The Telegraph. 26th May 2017. Available from: http://www.telegraph.co.uk/women/politics/why-im-campaigning-to-reclaim-the-internet-from-sexist-trolls/
10. Levmore, S. X. & Nussbaum, M. C. The offensive Internet: speech, privacy, and reputation (Harvard University Press, 2010).
11. Council of Europe. European Convention on Human Rights. (1950). Available from: https://www.echr.coe.int/documents/convention_eng.pdf
12. Spinello RA, Regulating Cyberspace: The Policies and Technologies of Control. p.109.
13. Gaus, A. Trolling attacks and the need for new approaches to privacy torts. USFL Rev. 47 (2012): 353.
14. McGoldrick, D. The Limits of Freedom of Expression on Facebook and Social Networking Sites: A UK Perspective. 13 Hum. Rts. L. Rev. 125, (2013) p.130.
15. Barlow, J.P. A Declaration of the Independence of Cyberspace. (1996).
16. Citron. In Levmore SX and Nussbaum MC, The offensive Internet: speech, privacy, and reputation. 2010. Kindle Locations 378-379
17. Reno Vs ACLU 521 U.S. 844 (1997)
18. Lipschultz JH, Free expression in the age of the Internet: social and legal boundaries. p. 127.
19. Barendt, E.M. Freedom of Speech. Oxford. (1985). p.12
20. Group Vilification Reconsidered, 89 Yale L.J. 308, 332 (1979)
21. Abah,, Legal Regulation of CSR: The Case of Social Media and Gender-Based Harassment 5 U. Balt. J. Media L. & Ethics 38, 55 (2016)
22. Marshak, E. Online Harassment: A Legislative Solution, 54 Harv. J. Legis.. 501, 515– 17 (2017)
23. troll2. In T. Thorne, *Dictionary of contemporary slang* (4th ed.). London, UK: Bloomsbury. (2014) Retrieved from http://search.credoreference.com/content/entry/acbslang/troll2/0?institutionId=396
24. Trolls – pass notes No 3,268. *The Guardian*. 22nd October 2012. https://www.theguardian.com/technology/shortcuts/2012/oct/22/pass-notes-trolls
25. Lovink, G. Social media abyss: critical internet cultures and the force of negation. Wiley. (2016). p.163.
26. Nagle, A. Kill All Normies: Online Culture Wars From 4Chan And Tumblr To Trump And The Alt-Right Zero Books. (2017) (p. 16).
27. Diaz, F.L. Trolling and the First Amendment: Protecting Internet Speech in the Era of Cyber Bullies and Internet Defamation. U. Ill. J.L. Tech. & Pol'y 135, 160 (2016) p.137.
28. Pew Research Center, October 2014, "Online Harassment" Available at: http://www.pewinternet.org/2014/10/22/online-harassment/
29. Rosewarne, L. *Cyberbullies, Cyberactivists, Cyberpredators : Film, TV, and Internet Stereotypes.* p.83
30. Carrabis, A.B. and Haimovitch, S.D., Cyberbullying: Adaptation from the Old School Sandlot to the 21st Century World Wide Web—The Court System and Technology Law's Race To Keep Pace 16 J. Tech. L. & Pol'y 143. p.173
31. Delgado, R. and Stefancic, J. Four Observations About Hate Speech. 49 Wake Forest L. Rev. 319, 344 (2014)
32. Ronson, J. So You've Been Publicly Shamed. Pan Macmillan. (2015).
33. Bradshaw, P. 'Here come the Oscars: still a cruel joke in a cruel town' *The Guardian*. 18th February 2015. https://www.theguardian.com/commentisfree/2015/feb/18/oscars-awards-night-winner-tv

34. Vitak J and others, ‘Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment’ (Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. p.1239.
35. DPP v Collins [2006] UKHL 40
36. Chambers v DPP [2012] EWHC 2157
37. Murray, Ian. *Information technology law: the law and society*. p.151.
38. D. Allen Green, ‘The High Court Is Unable to Agree on Twitter Joke Trial Appeal’, New Statesman, 28 May 2012.
39. R v Stacey Appeal No: A20120033
40. https://www.judiciary.gov.uk/judgments/r-v-nimmo-and-sorley-judgment/
41. [2006] UKHL 34; [2007] 1 A.C. 224
42. Constine, J. Twitter lets you avoid trolls by muting new users and strangers. *Tech Crunch*. 10th July 2017. Available from: https://techcrunch.com/2017/07/10/twitter-mute/
43. Twitter Rules. https://help.twitter.com/en/rules-and-policies/twitter-rules
44. Department for Culture, Media, and Sport. Online Safety Bill: Factsheet. 2022. https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/online-safety-bill-factsheet