International Journal of Law in Context (2019), **15**, 183–197
doi:10.1017/S1744552319000119

§

ARTICLE

Regulating terrorist content on social media: automation and the rule of law

Stuart Macdonald*, Sara Giro Correia and Amy-Louise Watkin

College of Law and Criminology, Swansea University
*Corresponding author. E-mail: s.macdonald@swansea.ac.uk

Abstract

Social-media companies make extensive use of artificial intelligence in their efforts to remove and block terrorist content from their platforms. This paper begins by arguing that, since such efforts amount to an attempt to channel human conduct, they should be regarded as a form of regulation that is subject to rule-of-law principles. The paper then discusses three sets of rule-of-law issues. The first set concerns enforceability. Here, the paper highlights the displacement effects that have resulted from the automated removal and blocking of terrorist content and argues that regard must be had to the whole social-media ecology, as well as to jihadist groups other than the so-called Islamic State and other forms of violent extremism. Since rule *by* law is only a necessary, and not a sufficient, condition for compliance with rule-*of*-law values, the paper then goes on to examine two further sets of issues: the clarity with which social-media companies define terrorist content and the adequacy of the processes by which a user may appeal against an account suspension or the blocking or removal of content. The paper concludes by identifying a range of research questions that emerge from the discussion and that together form a promising and timely research agenda to which legal scholarship has much to contribute.

Keywords: terrorism; social media; regulation; propaganda; artificial intelligence

1 Introduction

In January 2018, in a speech to the World Economic Forum, Prime Minister Theresa May declared that technology companies need 'to move further and faster in reducing the time it takes to remove terrorist content online', adding that ultimately such content should be 'removed automatically' (May, 2018). The following month, the UK Home Office announced that it had developed a new tool that – by analysing the audio and images of a video file during the uploading process – could detect and reject 94 percent of propaganda of the so-called Islamic State (IS) with a 99.99 percent success rate. The tool would be made available to all Internet platforms (Greenfield, 2018).

The terms of service of the leading social-media companies stipulate that terrorist content is forbidden. For example, Facebook's *Community Standards* state that organisations that are engaged in terrorist activity are not allowed a presence on the platform. The use of Facebook to facilitate or organise criminal activity that causes physical harm or financial damage to people or businesses is also prohibited, as is threatening physical harm to individuals and celebrating a crime that one has committed. The *Twitter Rules* forbid the use of the service for any unlawful purposes or in furtherance of illegal activities. This includes threatening or promoting terrorism, as well as affiliation with organisations that use or promote violence against civilians to further their causes. And YouTube's *Community Guidelines* state that terrorist organisations may not use the platform for any purpose, including recruitment. Content related to terrorism, such as content that promotes terrorist acts, incites violence or celebrates terrorist attacks, is also prohibited.

Given both the sheer volume of online content (every minute, 350,000 tweets are posted,[1] 300 hours of video are uploaded to YouTube[2] and, on Facebook, 510,000 comments are posted, 293,000 statuses are updated and 136,000 photos are uploaded[3]) and the variety of conduct that may be removed (from sexual exploitation and hate speech to spam and bullying), it is unsurprising that each of these platforms already utilises artificial intelligence (AI) to assist in the enforcement of their rules and standards (in addition to referrals from users, law enforcement and governments, which are responsible for only a small minority of suspensions and take-downs (Bickert and Fishman, 2018; Twitter, 2017; Wojcicki, 2017)). Facebook's use of AI includes image matching (so that, if someone tries to upload a photo or video that matches a photo or video that has previously been identified as terrorist, they are prevented from doing so), language understanding (analysing text that has been removed for praising or supporting terrorist organisations in order to develop text-based signals that can go into machine-learning algorithms to detect similar future posts), removing terrorist clusters (using algorithms to work outwards from pages, groups, posts or profiles that have been identified as supporting terrorism, employing signals such as whether an account is friends with a high number of accounts that have been disabled for terrorism) and tackling recidivism (detecting new accounts created by repeat offenders) (Bickert and Fishman, 2017). Twitter also uses algorithms to identify accounts that may be violating its rules (Twitter, 2016) such that, from 1 August 2015 to 31 December 2017, a total of 1,210,357 accounts were suspended for the promotion of terrorism (Twitter, 2018a) – a stark contrast to 2013–14, when Twitter's account suggestion algorithm[4] was having the effect of connecting individuals at risk of radicalisation with extremist propagandists (Berger, 2013). Meanwhile, from June to December 2017, YouTube removed over 150,000 videos for violent extremism, with 98 percent of these flagged by machine-learning algorithms. Nearly 70 percent were taken down within eight hours of upload (and nearly half within two hours) (Wojcicki, 2017). In November 2017, YouTube also drastically reduced its video archive of the jihadist preacher Anwar al-Awlaki. Although some of this content did not violate YouTube's *Community Guidelines*, the decision was made that al-Awlaki should be treated similarly to a terrorist organisation with an outright ban (Shane, 2017). Prior to the ban, a search for 'Anwar al-Awlaki' returned over 70,000 videos. Now it returns fewer than 20,000 and these are overwhelmingly news reports about his life and death, debates over the legality of his killing and refutations of his work by scholars. Video-fingerprinting technology is being used by YouTube to flag videos of al-Awlaki automatically before anyone sees them (Shane, 2017).

Each of the measures identified in the previous paragraph were taken by one social-media company acting independently. This meant that terrorists could 'jump from platform to platform' (Bickert and Fishman, 2017), reposting content that had been removed from one platform on a different one or posting on multiple platforms in the first place. In an effort to address this, in December 2016, Facebook, Twitter, YouTube and Microsoft announced the creation of a shared industry database of hashes (i.e. unique digital fingerprints) for content produced by or in support of terrorist organisations (Facebook, 2016). These hashes allow member companies to identify and remove matching content that violates their policies – and sometimes block such content before it is even posted. As of June 2018, the database contained more than 88,000 hashes and nine further companies had joined the consortium (including Ask.fm, Cloudinary, Instagram, Justpaste.it, LinkedIn, Oath and Snap) (GIFCT, 2018).

The various methods that have been outlined share a common goal: to remove terrorist content as swiftly as possible or, ideally, to prevent it from appearing in the first place.[5] They are thus examples of

[1]Twitter usage statistics. Available at http://www.internetlivestats.com/twitter-statistics/ (accessed 10 April 2018).

[2]37 mind blowing YouTube facts, figures and statistics – 2018. Available at https://fortunelords.com/youtube-statistics/ (accessed 10 April 2018).

[3]The top twenty valuable Facebook statistics – updated April 2018. Available at https://zephoria.com/top-15-valuable-facebook-statistics/ (accessed 10 April 2018).

[4]Detailed at https://help.twitter.com/en/using-twitter/account-suggestions (accessed 10 April 2018).

[5]AI has also been used to divert individuals searching for IS propaganda online away from such content. The 'Redirect Method', created by Google's think-tank Jigsaw, uses the targeting tools of Google's advertising service Adwords to redirect

what Hildebrandt (2008) labels 'constitutive', as opposed to 'regulative', interventions (p. 174). Facebook's *Community Standards*, the *Twitter Rules* and YouTube's *Community Guidelines* each seek to engage users' 'moral register' (Brownsword, 2011, p. 1328). The former two emphasise the importance of users feeling safe when using the platforms, whilst the latter emphasises that respect for the guidelines ensures that the platform is 'fun and enjoyable for everyone'. This moral reasoning is supplemented with warnings that users may report – or flag – content that it is felt to violate the relevant policies. These warnings appeal to a user's 'prudential register' (Brownsword, 2011, p. 1329): they seek to deter the posting of inappropriate content by appealing to the user's own interests. What moral and prudential signals have in common is that they are normative: they say that X ought, or ought not, to be done. By contrast, the uses of AI by social-media companies outlined above, and encouraged by the UK government, are non-normative. By seeking to prevent terrorist content from ever reaching their platforms, the signal to users is 'you cannot do this; you *really* cannot do this' (Brownsword, 2015, p. 6, emphasis in original).

This paper examines social-media companies' use of AI to block terrorist content, against the backdrop of the wider societal shift towards constitutive, non-normative regulation. Building on recent scholarship – in particular the work of Roger Brownsword (2011; 2015; 2016), which argues that forms of non-normative regulation should be regarded as both within the province of legal scholarship and subject to rule-of-law principles – the paper focuses on three sets of issues: the challenges presented by the displacement effect of blocking terrorist content on the biggest social-media platforms; the clarity with which social-media companies define terrorist content; and the adequacy of the processes by which a user may appeal against an account suspension or the blocking or removal of content. Before turning to these issues, the paper first explains why the use of non-normative regulatory instruments by private technology companies should be regarded as being subject to the rule of law.

Two caveats should be highlighted at the outset. First, for reasons of space, the paper focuses on three platforms: Facebook, Twitter and YouTube. As will become clear, this should not be taken as indicating that other platforms are not also important. Second, the paper's focus on three sets of issues related to the rule of law should not be taken as indicating that there are not also other important issues that merit further examination. In particular, the accuracy and false positive rates of the AI that identifies and blocks terrorist content warrant closer scrutiny.

2 Automation, terrorist content and the rule of law

At first, the deployment of AI to block terrorist content from social-media platforms seems some way removed from discussions of the rule of law, in two respects: the mode of regulation and the responsible actors. Starting with the latter, the rule of law is traditionally associated with public institutions, not private technology companies. Yet, in the current context, a steadfast public–private distinction is difficult to maintain. The importance of public–private partnership in responding to terrorists' use of the Internet is widely recognised (Kavanagh *et al*., 2017). In addition to take-down requests issued by government agencies, there are a number of multi-stakeholder initiatives at both the national level (including the technology developed by the UK Home Office mentioned earlier) and the international level (including the EU Internet Forum, which brings together governments, Europol and technology companies and Tech Against Terrorism, a project that supports technology companies (particularly start-ups and small companies) that is mandated by the UN Counter Terrorism Committee Executive Directorate (UN CTED) and implemented by the ICT4Peace Foundation). More fundamentally, states have a responsibility to protect national security. As Lord Hope stated in *A v. Secretary of State for the Home Department*: 'It is the first responsibility of government in a democratic society to protect and safeguard the lives of its citizens.'[6] At the same time, Lord Hope continued, the rights and

such individuals to curated YouTube videos that aim to discredit and debunk the content that the individual searched for (Jigsaw, 2017).

[6][2004] UKHL 56, at [99].

liberties of individuals must be respected. Indeed, as the UK's CONTEST strategy emphasises, respect for human rights and the rule of law is in itself an important component of an effective counter-terrorism strategy (HM Government, 2018). It is important, therefore, that the policies and practices that technology companies adopt to counter terrorist activities also reflect this commitment to human rights and the rule of law.

In terms of the mode of regulation, the rule of law would traditionally be associated with governance by rules, such as those prescribed in statute and expounded in case-law. However, 'From the fact that law is conceived of as a normative phenomenon, it does not follow that the ambit of jurisprudence should be limited to legal norms' (Brownsword, 2015, p. 10). As Brownsword has argued, to limit the ambit of jurisprudence to legal norms would be to exclude other kinds of normative order that are found in society (such as religious and moral codes) and which have an important influence on 'what makes things tick in the social world [and] how legal norms fit into the full array of normative signals' (Brownsword, 2015, p. 12). Equally, it would be mistaken to limit the domain of jurisprudence to the exclusively normative:

> 'To give ourselves the chance of understanding and assessing a radical transformation in the way that the State channels human conduct, we need to work with a notion of the regulatory environment that includes both normative and non-normative instruments …. For jurists to turn away from the use of technological instruments for regulatory purposes is to diminish the significance of their inquiries and to ignore important questions about the way that power is exercised and social order maintained.' (Brownsword, 2015, pp. 13–14)

A 'regulatory environment' should thus be understood 'as an action-guiding environment in which regulators direct the conduct of regulatees with a view to achieving a particular regulatory objective' (Brownsword, 2011, p. 1327).

What makes the use of AI to prevent terrorist content from reaching social-media platforms appealing is its greater effectiveness in comparison to the moral and prudential signals found in these companies' terms of service. At the same time, however, it is important that this 'brute instrumentalism … is conditioned by principles that give it legitimacy' (Brownsword, 2016, p. 138). So, regardless of the different regulatory form, rule-of-law principles are of continuing relevance:

> 'On the one side, it remains important that governance – now in the form of power exercised through technological management – is properly authorised and limited; and, on the other, although citizens might have less opportunity for "non-compliance", it is important that the constraints imposed by technological management are respected.' (Brownsword, 2016, pp. 106–107)

In short, 'it is the ideal of legality together with the Rule of Law that stands between us and a disempowering techno-managed future' (Brownsword, 2016, p. 138).

3 Imperfect impossibility and tactical displacement

The study of the population of IS supporters on Twitter conducted by Berger and Morgan (2015) found that, during October and November 2014, there were no fewer than 46,000 overt IS supporter accounts on Twitter – and possibly as many as 90,000. The average number of followers of these accounts was 1,004 and each account posted an average of 7.3 tweets per day over its lifetime. Since this 'Golden Age' (Conway *et al.*, 2017, p. 28), IS's presence on Twitter has been reduced significantly. Towards the end of 2014, Twitter began an aggressive campaign of suspensions. Berger and Morgan (2015) found that, by February 2015, IS supporters on Twitter were having to devote far more time to rebuilding their networks (as opposed to disseminating propaganda, recruiting and other activities). A follow-up study conducted by Berger and Perez (2016) also found that suspension activity had a significant disruptive effect. Individual users who repeatedly created new accounts after being suspended 'suffered devastating reductions in their follower counts' and declines in

networks persisted even when suspension pressure eased, 'suggesting that suspensions diminish activity in ways that extend beyond the simple removal of accounts' (p. 4). Those IS supporters who persisted in their use of the platform resorted to the use of countermeasures, such as locking their accounts so that they were no longer publicly accessible, using an innocuous image or the default egg as the avatar image and selecting a random combination of letters and numbers as the user handle or screen name. But, since 'A conscious, supportive and influential virtual community is almost impossible to maintain in the face of the loss of access to such group or ideological symbols and the resultant breakdown in commitment', today IS's Twitter activity 'has largely been reduced to tactical use of throwaway accounts for distributing links to pro-IS content on other platforms, rather than as a space for public IS support and influencing activity' (Conway *et al.*, 2017, p. 30) – and these throwaway accounts are also suspended before they gain much of a following (Grinnell *et al.*, 2017; 2018).

Whilst considerable progress has been made in disrupting the use of Twitter by IS supporters, Conway *et al.* (2017) found that supporters of other jihadist groups – such as Hay'at Tahrir al-Sham (HTS), Ahrar al-Sham, the Taliban and al-Shabaab – experienced significantly less disruption. Whilst more than 30 percent of pro-IS accounts were suspended within 48 hours of their creation, the equivalent figure for these other pro-jihadist accounts was less than 1 percent.[7] The latter were also able to post six times as many tweets, follow four times as many accounts and gain thirteen times as many followers as the pro-IS accounts (Conway *et al.*, 2017).[8]

The criminological literature on situational crime prevention highlights the fact that no pre-emptive scheme is perfect (Rosenthal, 2011). As Rich (2013) observes: 'Though the goal of an impossibility structure is to make the commission of a crime practically impossible, the reality is that the ingenuity of potential criminals is inexhaustible' (p. 823). One technique that has been used by IS supporters is out-linking, namely the use of Twitter as a 'gateway' (O'Callaghan *et al.*, 2013, p. 276), to other websites, social-media platforms and content-hosting sites. In Conway *et al.*'s (2017) study, 7,216 (12.5 percent) of the 57,574 tweets analysed contained out-links. The platform with the most out-links from Twitter was YouTube. Interestingly, however, Facebook did not appear in the top ten, whilst the less-known justpaste.it, sendvid.com and archive.org all featured in the top six.[9] The use of these smaller platforms appears to be an attempt to 'exploit an overlapping ecosystem of services', taking advantage of the fact that smaller companies 'don't have the scale or resources to handle the challenge on their own' (Tech Against Terrorism, 2017). Justpaste.it, for example, is a free content-sharing service that allows content to be posted within seconds with no registration required. Owned by Mariusz Zurawek, who runs the site out of his home in Poland, the content posted on Justpaste.it began to include IS propaganda in early 2014. By March 2015, Zurawek estimated that he had removed up to 2,000 posts at the request of London Metropolitan Police (Stalinsky and Sosnow, 2016). Since then, he has received a large volume of take-down requests from all over the world. This poses challenges in terms of identifying what content is legal and responding to take-down requests in other languages, as well as capacity and resources (Tech Against Terrorism, 2017).

Impossibility structures may also lead to 'tactical displacement' (Rich, 2013, p. 824). Besides the use of throwaway Twitter accounts out-linking to propaganda hosted on other platforms, IS supporters have largely moved their community-building activities to other platforms, in particular Telegram. Telegram is a cloud-based instant-messaging service, providing optional end-to-end encrypted messaging. Features include: a self-destruct timer that permanently deletes messages and media after they are viewed; group chats, which users can only join when invited to do so; and channels that are public and so can be used to broadcast messages to large audiences. As IS's presence on Twitter diminished, a

[7]Similarly, 85 percent of pro-IS accounts were suspended within the first sixty days of their life; the equivalent figure for the other pro-jihadist accounts was 40 percent.

[8]A further challenge is the removal of historical content. Facebook has reported that, in the first quarter of 2018, its 'historically focused technology found content that had been on Facebook for a median time of 970 days' (Bickert and Fishman, 2018).

[9]The other two platforms in the top six were Google Drive and Google Photos.

migration to Telegram occurred. By early 2016, Telegram was being used to share content produced by official IS channels and IS members and supporters were pushing out more than 30,000 Telegram messages each week (Prucha, 2016). Telegram chat-rooms and channels also possess psychologically addictive qualities, which aids efforts to build virtual communities as well as the promotion of so-called self-starter terrorism (Bloom *et al*., 2017). These uses of Telegram form part of a wider movement towards the use of more covert methods (UN Security Council, 2017). As well as Telegram, other encrypted messaging services, including WhatsApp, have been used by jihadists for communication and attack-planning (Malik, 2018). Websites have also been relocated to the Darknet: the day after the 2015 Paris attacks, for example, IS's media division Al-Hayat Media Center launched a new website on the Darknet to host its propaganda, including a video celebrating the attacks (Ragan, 2015). Darknet platforms are, by definition, more difficult to police than the surface and deep webs, meaning they have the potential to function as a jihadist 'virtual safe-haven' (Malik, 2018, p. iv).

An important aspect of the rule of law is enforcement: rule *by* law is a necessary condition for the rule *of* law (Spigelman, 2003). What the examples discussed in this section illustrate is that, if the aim of blocking terrorist content online is to be realised, regard must be had to the whole social-media 'ecology' (Conway *et al*., 2017, p. 32). This involves broadening the focus beyond just the biggest social-media companies. It has been reported that, in 2016, the UK's Counter Terrorism Internet Referral Unit reported content across 300 different services (Tech Against Terrorism, 2017). The platforms offered by smaller companies are also susceptible to exploitation, and the challenges and barriers to these companies self-regulating may be quite different to the social-media giants. Meanwhile, calls such as those from the UK's Home Secretary for 'backdoors' to be built into encrypted messaging services also raise complex issues around privacy and security (in the sense of both protection against terrorism and protection against the backdoor being exploited) (Haynes, 2017). Similarly, it is necessary to broaden the focus beyond IS, not just to other jihadist groups, but also to other forms of violent extremism. Extreme right-wing groups, for example, also have a significant online presence (O'Callaghan *et al*., 2013). And, whilst steps have been taken to disrupt their presence on the surface and deep webs (e.g. Facebook's decision to ban Britain First from its platform (Nouri, 2018)), these groups also appear to be migrating to the Darknet (Deep Dot Web, 2017).

4 Non-normativity and definitional clarity

For scholars of counter-terrorism legislation and policy, the potential for tension to exist between rule *by* law and rule *of* law will be familiar. There is a significant body of literature that examines the analogous notion that security and liberty should be balanced and, in particular, the suggestion that, if the (perceived) threat of terrorism increases, some measure of liberty should be sacrificed for the sake of enhanced security (Macdonald, 2009a; 2009b; Posner, 2006; Posner and Vermeule, 2007; Waldron, 2003; Zedner, 2005). Importantly, it has been pointed out that security and liberty are interdependent: whilst security may in some sense be regarded as a prerequisite for the enjoyment of liberties, in another sense, the protection of liberties is necessary to ensure security (against the power of the state) (Barak, 2002). Moreover, as noted above, it is widely acknowledged that respect for human rights is an important component of an effective counter-terrorism strategy. So, whilst regulators must have regard to the whole social-media ecology in order to realise the objective of blocking terrorist content online, ensuring rule *by* law is only a necessary, and not a sufficient, condition for compliance with rule-*of*-law values. In this section and the one that follows, we accordingly examine two issues – the clarity with which social-media companies define terrorist content and the adequacy of the processes by which a user may appeal against an account suspension or the blocking or removal of content – that engage these wider rule-of-law values.

The challenges involved in defining 'terrorism', and the lack of an internationally agreed-upon definition, are well-documented (see e.g. Hardy and Williams, 2011; Saul, 2006). Summed up by the slogan 'One person's terrorist is another's freedom fighter', perhaps the most controversial definitional issue is that of just cause. In his report on the UK's statutory definition of terrorism, the

then-Independent Reviewer of Terrorism Legislation Lord Carlile recognised the initial attractiveness of a just-cause exception but concluded that it would create 'real difficulties' (Carlile, 2007, p. 44). Not only is there 'far from total international or domestic agreement as to which regimes are/are not in breach of international humanitarian law' (Carlile, 2007, p. 44), but 'Those who opt for terror always believe their cause is just' (Fletcher, 2006, p. 906). According to Hodgson and Tadros, this results in a trilemma:

> 'Horn 1: Define terrorism narrowly to exclude from the definition all attacks on the state and its officials. In doing so terrorism law will not be suitable for the purposes that we have for it. Horn 2: Define terrorism broadly to include all attacks on the state and its officials. In that case terrorism law will in principle, and probably in practice, apply to legitimate freedom fighters. Horn 3: Define terrorism in a way that discriminates between legitimate and illegitimate attacks on the state and its officials. This involves a range of legal actors making political judgments that they have inadequate expertise to make.' (Hodgson and Tadros, 2013, p. 496)

This issue arose for consideration in *R. v. F.*[10] In this case, the Court of Appeal upheld the conviction for terrorism offences of a Libyan man who was found in possession of information on explosives and notes on how to overthrow Colonel Gaddafi. He had fled to the UK in 2002 and been granted asylum in 2003, after members of his family and friends had allegedly been killed by Gaddafi's regime. Delivering the judgment of the Court, Sir Igor Judge P. stated: 'the legislation does not exempt, nor make an exception, nor create a defence for, nor exculpate what some would describe as terrorism in a just cause … Terrorism is terrorism, whatever the motives of the perpetrators' (para. [27]). Commenting on the judgment, Hodgson and Tadros argue that it amounts to a recognition that 'the law must apply to all such cases because the court cannot adequately discriminate between legitimate and illegitimate uses of violence, acknowledging that there may be cases where this leads to serious injustice' (2013, p. 517).

The issue was revisited in *R. v. Gul*,[11] in which the Supreme Court was asked to determine whether the UK's definition of terrorism contains an exception for military attacks by a non-state group against the armed forces of a state in the course of a non-international armed conflict. Whilst the definitions of some other countries do contain such an exemption, the Supreme Court held that the 'natural, very wide' wording of the UK's definition could not be read down so as to contain one (para. [38]). UK law is thus 'impaled' on the second horn of Hodgson and Tadros's (2013, p. 517) trilemma.

Even without a just-cause exception, a pro-democracy activist may be undeterred: the paradigmatic example is Nelson Mandela. Yet one of the concerns that commentators have expressed about technological management is that it 'might compromise the possibility of engaging in responsible moral citizenship' (Brownsword, 2015, p. 36; see also Rosenthal, 2011). Morozov explains:

> 'Laws that are enforced by appealing to our moral or prudential registers leave just enough space for friction; friction breeds tension, tension creates conflict, and conflict produces change. In contrast, when laws are enforced through the technological register, there's little space for friction and tension – and quite likely for change.' (Morozov, 2013, p. 36)

The movement towards non-normative regulation could thus have significant repercussions for activists living in oppressive regimes. Gerbaudo's (2012) analysis of the role played by Internet communication and social media in contemporary activism argues that, against the backdrop of a crisis of public space, 'social media have become emotional conduits for reconstructing a sense of togetherness among a spatially dispersed constituency, so as to facilitate its physical coming together in public space' (p. 159). An example is the Facebook page 'We Are All Khaled Said' (KKS), which promoted

[10][2007] EWCA Crim 243.
[11][2013] UKSC 64.

the protests that led to the overthrow of the Mubarak regime in the Egyptian revolution of 2011 (Alaimo, 2015). Gerbaudo (2012) explains that the KKS page was used to build a community of activists in a 'Choreography of Assembly' (p. 12). Choreographing consists of 'the mediated "scene-setting" and "scripting" of people's physical assembling in public space' (p. 40) and, in this particular instance, included the use of 'feeder marches' (p. 64) advertised on the KKS Facebook page and the advance publication of an online booklet explaining to people how they should behave during the protests.

However, the growing use of AI and increasing government pressure to block terrorist content means that today social media companies' 'policies and the architecture of their products will increasingly complicate collective action efforts' (Youmans and York, 2012, pp. 325–326). A stark example is the removal from YouTube of thousands of videos documenting atrocities in Syria, following the introduction of new technology to automatically remove content that potentially breached YouTube's *Community Guidelines* (Browne, 2017). These videos provided important evidence of human rights violations and some existed only on YouTube, since not all Syrian activists and media can afford an offline archive. The channel of the monitoring group Violation Documentation Center (VDC) – which housed over 32,000 videos of the conflict – was also taken down, although VDC was later able to regain access to its channel and videos (Kayyali and Althaibani, 2017).

The Syria example demonstrates the importance of human review. Facebook, Twitter and YouTube each has processes by which users may appeal decisions to remove content. They also recognise that AI cannot always deduce 'what supports terrorism and what does not … and algorithms are not yet as good as people when it comes to understanding this kind of content' (Bickert and Fishman, 2017). But, whilst a discretionary 'safety valve' (Roth, 2016, p. 1287) may be both unavoidable and necessary, it is also important to be mindful of rule-of-law values. This is evident in discussions of the UK's statutory definition of terrorism. On the one hand, concerns have been expressed about the definition's (over-) breadth (Anderson, 2014); on the other hand, the UK government has highlighted the 'complexity and fluidity of changing political events in the terrorism context and its ability to evolve and diversify at great speed' – which, it is argued, means that a 'flexible statutory framework' is required in order to 'ensure that our law enforcement and intelligence agencies can continue to disrupt and prosecute those who pose a threat to the public' (HM Government, 2015, p. 8). Commenting on this, the Supreme Court in *R. v. Gul* stated:

> 'The Crown's reliance on prosecutorial discretion is intrinsically unattractive, as it amounts to saying that the legislature, whose primary duty is to make the law, and to do so in public, has in effect delegated to an appointee of the executive, albeit a respected and independent lawyer, the decision whether an activity should be treated as criminal for the purposes of prosecution …. Further, such a device leaves citizens unclear as to whether or not their actions or projected actions are liable to be treated by the prosecution authorities as effectively innocent or criminal – in this case seriously criminal.' (para. [36])

Whilst there may be obvious differences between drafting legislation that criminalises particular conduct and terms of service that specify when content may be removed from social-media platforms, the concerns identified by the Supreme Court remain relevant. Publicly defining what will be deemed to constitute terrorist content restricts and guides the discretion of individual reviewers, reducing the risk of inconsistent – or even inappropriate – decision-making. And, even if there are already internal procedures or policies[12] – or other, extra-legal factors (such as institutional culture or practice) – that

[12]The distinction between decision rules (addressed to officials) and conduct rules (addressed to the public) was discussed by Dan-Cohen (1984), who argued that the selective transmission of the former can in some circumstances be compatible with the rule of law. Even if selective transmission (in the sense used by Dan-Cohen) could be justified in the current context – and we are sceptical as to whether it could be – there remain important rule-of-law concerns about the conduct rules (the terms of service), as we explain above.

already limit how reviewers exercise this discretion in practice,[13] defining terrorism in the terms of service serves an important communicative function. It provides users with information needed to understand their rights and responsibilities when using the platform (Citron, 2018). Here, the 'principle of maximum certainty' (Horder, 2016, p. 85) is important; a definition that is unduly vague or ambiguous will not enable users to make informed decisions about their use of the platform, and also leaves open the possibility of 'censorship creep' (Citron, 2018, p. 1050).

At present, YouTube's *Community Guidelines* state that the platform does not permit 'terrorist organisations to use YouTube for any purpose' and 'strictly prohibits content related to terrorism', but without defining the terms 'terrorist organisation' and 'terrorism'. In terms of Hodgson and Tadros's trilemma, this leaves it uncertain which of the three approaches the company takes and – on the assumption that it does not seek to block content related to a just cause – leaves it unclear how it adjudicates which causes are just. By contrast, Facebook's *Community Standards* define a terrorist act as 'a premeditated act of violence against persons or property carried out by a non-government actor to intimidate a civilian population, government or international organisation in order to achieve a political, religious or ideological aim', whilst the *Twitter Rules* offer the following definition (of violent extremism):

> 'We consider violent extremist groups to be those which meet all of the below criteria:
>
> - identify through their stated purpose, publications, or actions, as an extremist group
> - have engaged in, or currently engage in, violence (and/or the promotion of violence) as a means to further their cause
> - target civilians in their acts (and/or promotion) of violence
>
> Exceptions will be considered for groups that have reformed or are currently engaging in a peaceful resolution process, as well as groups with representatives elected to public office through democratic elections. This policy does not apply to military or government entities.'

These two definitions share some similarities and, as a result, raise some common issues, such as whether the reference to 'violence' is both too broad (it is not limited to serious violence) and too narrow (it would apparently not encompass a cyber terrorist attack that caused serious economic or environmental harm).[14] There are also several dissimilarities. Only Facebook's definition stipulates that the violence must have been premeditated, whilst only Twitter's definition raises the possibility of exceptions for groups that have reformed, are engaged in a peace process or have democratically elected representatives (though the discretion it retains in this regard is left undefined). Like the definitions of many national legislatures (Hardy and Williams, 2011), Facebook's definition contains an intention requirement ('to intimidate a civilian population, government or international organisation') and a motive requirement ('political, religious or ideological aim'), whilst Twitter's definition instead asks whether the group identifies 'as an extremist group'. The latter seems inadequate to meet the demands of the principle of maximum certainty, given the possibility of articulating more precise intention and motive requirements. Lastly, Twitter's definition requires that the violence that was employed or promoted targeted civilians, whereas Facebook's definition contains no such limitation.[15] The *Twitter Rules* thus take the first of Hodgson and Tadros's three approaches – excluding from its definition

[13]It has been reported that some companies use international, regional or national sanctions lists to guide their decision-making on this issue (ICT4Peace and Counter-Terrorism Committee Executive Directorate, 2016).

[14]These issues were considered by the UK government during the consultation process that resulted in the existing definition in s. 1 of the Terrorism Act 2000 (Home Office, 1998).

[15]An intention to intimidate a civilian population will suffice for the intention requirement in Facebook's definition, but the conduct requirement simply refers to a premeditated act of violence against 'persons or property'.

attacks on the state and its officials, with the concomitant risk of under-inclusivity – whereas, on the face of it, Facebook, like the UK government, takes the second approach, with no express exception made for just causes – leading, therefore, to either over-inclusivity and unfairness or the exercise of an unarticulated discretion.

It is also noteworthy that, for the purposes of the shared industry database of hashes, the companies have stated that the database will only contain hashes 'of the most extreme and egregious terrorist images and videos' (Facebook, 2016). This not only leaves terrorism undefined, which is significant given the companies' differing approaches outlined above; it also adds another layer of ambiguity, since 'what constitutes extreme and egregious terrorist content is unclear' (Citron, 2018, p. 1053).

Concocting a satisfactory definition of terrorism has been likened to the quest for the Holy Grail (Levitt, 1986). Whilst social-media companies cannot be expected to achieve the impossible, the rule of law requires that they explain to their users what content will be deemed terrorist and blocked as clearly as they are able. Looking to the existing definitions of national legislatures may provide some assistance, although many of these definitions have also been criticised on rule-of-law grounds (Hardy and Williams, 2011). Citron (2018) accordingly argues that, in the first instance, a multi-stakeholder group that includes both academics and human rights groups should be established to help the companies articulate more clearly what constitutes terrorist material, whilst, in an open letter to Facebook Chief Executive Officer Mark Zuckerberg, the current UN Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism has recommended that Facebook adopt the model definition of terrorism devised by her predecessor (Ní Aoláin, 2018).

5 Prior restraint and technological due process

Machines work with data and code; they do not attribute meaning (Hildebrandt, 2018). Inevitably, therefore, the algorithms that identify and block terrorist content cannot be expected to operate with 100 percent accuracy. Accordingly, Facebook, Twitter and YouTube each offers an appeals process. Facebook's appeal process has three steps: first, the user will receive a notification and be given the option to request additional review; a human reviewer will then conduct the review, hopefully within twenty-four hours; finally, if a mistake has been made, the user will be notified and the content restored (Bickert, 2018). In April 2018, Facebook announced that, for the first time, it was giving users the opportunity to appeal decisions to remove individual posts, photos or videos that had been deemed to violate its *Community Standards*; prior to this, users had only been able to appeal against the removal of a profile or page (Bickert, 2018). Twitter's most severe enforcement action is permanent account suspension. In this situation, the user may appeal through the platform interface or by filing a report, outlining the reasons why they believe they did not violate the *Twitter Rules*. If the suspension is deemed to be valid, the user will receive information on the policy that the account violated (Twitter, 2018b). Whilst YouTube reserves the right to terminate an account for a single case of severe abuse (such as predatory behaviour), for most violations of its *Community Guidelines*, it operates a three-strike policy. A user may appeal against both an account termination and the issuance of a strike. When a strike is issued, the user receives an e-mail and an alert in their channel settings. A strike remains in place for a period of three months. A strike may only be appealed once and, if the appeal is unsuccessful, the user will not be able to appeal a different strike for sixty days. If the appeal is successful, the strike will be removed and the video will normally be restored.[16] Three strikes within a three-month period result in account termination (YouTube, 2018).

The increasing resort to automated decision-making has led to concerns being expressed about 'technological due process' (Citron, 2008). One specific issue in the current context is the impartiality of the human reviewers considering appeals against account suspension or the blocking/removal of content who may, for example, be inclined to defer to their company's language-understanding algorithms. An additional concern is whether sufficient information is available to users who wish to lodge

[16]YouTube reserves the right to keep the video down or to reinstate it behind an age restriction.

an appeal. This is all the more important given that the blocking of content is a form of prior restraint and, as such, demands especial scrutiny.[17]

The first type of information regards the AI that detects and blocks violating content. This resonates with wider calls for 'algorithmic transparency' (Chandler, 2017, p. 1024). Pasquale (2015) uses the analogy of a black box – that is, 'a system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other' (p. 3) – and suggests that agencies 'ought to be able to "look under the hood" of highly advanced technologies like … algorithms' (p. 165) as a way to police such behaviour. This is all the more important since '"black box" processes, because of their mechanical appearance and apparently simple output, have a veneer of objectivity and certainty' (Roth, 2016, pp. 1269–1270). However, these calls for transparency face a number of difficulties. There may be commercial and proprietary objections to disclosure. There may also be concern that the information will be exploited: 'The process for deciding which tax returns to audit, or whom to pull aside for secondary security screening at the airport, may need to be partly opaque to prevent tax cheats or terrorists from gaming the system' (Kroll *et al.*, 2017, p. 639). And, perhaps most fundamentally, 'machine learning tends to create models that are so complex that … even the original programmers of the algorithm have little idea exactly how or why the generated model creates accurate predictions' (Rich, 2016, p. 886; see also Desai and Kroll, 2017).[18] Kroll *et al.* explain:

> '[I]nspecting source code is a very limited way of predicting how a computer program will behave. Machine learning, one increasingly popular approach to automated decision making, is particularly ill-suited to source code analysis because it involves situations where the decisional rule itself emerges automatically from the specific data under analysis, sometimes in ways that no human can explain. In this case, source code alone teaches a reviewer very little, since the code only exposes the machine learning method used and not the data-driven decision rule.' (Kroll *et al.*, 2017, p. 638)

Moreover, whilst techniques do exist that offer some form of reason-giving, these do not provide reasons for individualised predictions, but instead operate on an algorithm-wide scale, describing how important an input variable was to the algorithm's accuracy during training across many individuals (Lehr and Ohm, 2017).

Much of the wider literature about algorithmic transparency refers to the use of AI to generate future predictions, such as a person's creditworthiness (Citron and Pasquale, 2014) or the risk of criminality (Joh, 2016). In the current context, by contrast, AI is being used to detect violations of the platforms' terms of service. Here, understanding *how* the alleged violation was detected is less important to the user wishing to lodge an appeal than understanding *what* violation is alleged to have occurred. This takes us back to the importance of the principle of maximum certainty; definitional clarity not only provides users with fair warning before they post content – it also provides them with the information needed to have an effective opportunity to challenge alleged violations. For example, a pro-democracy activist who wishes to challenge the removal of content may be hampered if they have no indication whether, and how, the platform distinguishes between just and unjust causes. Similarly, the *Twitter Rules* forbid 'promoting terrorism'. This leaves it unclear whether the fact that content would be understood by others as promoting terrorism is sufficient to justify removing it or whether the promotion of terrorism must also have been the user's purpose (Macdonald and Lorenzo-Dus, forthcoming).[19] There is also a further danger here: in rule-of-law terms, congruence. Examinations of terrorism-related precursor offences have shown that there is a discrepancy within these offences between the offence definition and the wrong that is being targeted (Edwards, 2010; Tadros, 2007). As a result of these offences' overly broad wording, which encompasses conduct that is not being targeted for prosecution, only some

[17] *Yildirim v. Turkey* (Application no. 3111/10), 18 December 2012; *Near v. Minnesota* 283 US 697.

[18] It has been reported, for example, that Google engineers do not fully understand RankBrain (Schwartz, 2016).

[19] Empirical research suggests that Twitter only suspends accounts where it was the user's purpose to promote terrorism (Grinnell *et al.*, 2017).

of those who fall within the offence definition will be prosecuted. Yet, at trial, the factors that led to the defendant being selected for prosecution will not be in issue; the question will instead be whether the requirements set out in the offence definition are satisfied. The effect is to deprive the trial court of the opportunity to adjudicate on the actions that the offence is targeting (Macdonald, 2014). By analogy, if the wording of social-media platforms' terms of service encompasses significantly more content than is in fact selected for removal, and as a result does not reflect the factors that cause the AI to block/remove content, then the effect will be to deprive human reviewers – whose task is to apply the terms of service – of the opportunity to adjudicate on the factors underlying the automated decision. Some form of reason-giving, of the type mentioned at the end of the previous paragraph, would go some way towards addressing this problem.

As well as the provision of sufficient information, an additional concern is the possibility of bias. Whilst describing an algorithm as biased might seem to perpetuate the 'homunculus fallacy' – 'the belief that there is a little person inside the program who is making it work – who has good intentions or bad intentions, and who makes the program do good or bad things' (Balkin, 2017, p. 1223) – it is important to recognise that 'automated systems are not free of bias simply because they are executed by logical machines' (Chandler, 2017, p. 1045). There are a number of reasons why an algorithm might impact different groups of individuals disparately, such as disadvantageous definition of the outcome variables and non-representative data collection (Lehr and Ohm, 2017). For this reason, it is important to examine the actual outcomes of algorithmic decision-making (Chandler, 2017; Kim, 2017). This is particularly important in the current context, given that claims of anti-Muslim prejudice are utilised by jihadist radicalisers who deploy an us vs. them discourse to Other the West (Lorenzo-Dus and Macdonald, 2018). Indeed, a study of IS Twitter activity found that suspension played an important role in community-building, with the majority of the accounts studied referring to Twitter's use of suspension as a specific tool to persecute Muslims (Pearson, 2017). Examples included the tweets 'There has been a major Twitter purge targeting Muslims' and 'I got #Radicalised when me and my people started to get suspended on every social media platform'. This is not to suggest that the suspension of these accounts was not in accordance with the *Twitter Rules*; rather, it is to underline how important it is that the AI is equally effective at blocking and removing violating content relating to all forms of violent extremism.

6 Conclusion

This paper began by arguing that social-media companies' efforts to use AI to block terrorist content should be subject to the rule of law. The ensuing discussion highlighted the relevance of a number of rule-of-law principles, including maximum certainty, congruence, non-discrimination and enforceability, and in the process raised a variety of questions that together form a promising and timely research agenda. These questions include: How should smaller technology companies be encouraged to protect their platforms against exploitation and supported in this task? How can the challenges presented by encrypted messaging services be addressed, whilst respecting the privacy of individual users? When defining terrorism, which of the definitional approaches should social-media companies take? Should they create an express exemption for those with a just cause? If so, how should they seek to differentiate between just and unjust causes? Could – and should – more information be provided to users as to the factors that cause the AI to block content? Could the wording of terms of service more closely reflect these factors? And is the AI equally effective in identifying content from different forms of violent extremism and, if not, how can the perception of discrimination be mitigated?

This paper has taken a first step in developing answers to some of these questions, but much further work remains to be done. Our overarching objective has been not only to show the societal importance of these questions as part of national and international efforts to counter contemporary terrorism, but also to demonstrate that these questions fall squarely within the expertise of legal scholars. As such, legal scholarship has much to contribute to both these discussions and those of the trend towards non-normative forms of regulation more generally.

References

Alaimo K (2015) How the Facebook Arabic page 'We are all Khaled Said' helped promote the Egyptian revolution. *Social Media + Society* **1**, 1–10.

Anderson D (2014) *The Terrorism Acts in 2013: Report of the Independent Reviewer on the Operation of the Terrorism Act 2000 and Part 1 of the Terrorism Act 2006*. London: The Stationery Office.

Balkin JM (2017) The three laws of robotics in the age of big data. *Ohio State Law Journal* **78**, 1217–1241.

Barak A (2002) Foreword: A judge on judging: the role of the Supreme Court in a democracy. *Harvard Law Review* **116**, 19–162.

Berger JM (2013) Zero degrees of al Qaeda: how Twitter is supercharging jihadist recruitment, *Foreign Policy*, 14 August.

Berger JM and Morgan J (2015) *The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter*. Washington, DC: The Brookings Institution.

Berger JM and Perez H (2016) *The Islamic State's Diminishing Returns on Twitter: How Suspensions Are Limiting the Social Networks of English-speaking ISIS Supporters*. Washington, DC: George Washington University Program on Extremism.

Bickert M (2018) Publishing our internal enforcement guidelines and expanding our appeals process, *Facebook News*, 24 April. Available at https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/ (accessed 22 February 2019).

Bickert M and Fishman B (2017) Hard questions: how we counter terrorism, *Facebook News*, 15 June. Available at https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/ (accessed 22 February 2019).

Bickert M and Fishman B (2018) Hard questions: how effective is technology in keeping terrorists off Facebook?, *Facebook News*, 23 April. Available at https://newsroom.fb.com/news/2018/04/keeping-terrorists-off-facebook/ (accessed 22 February 2019).

Bloom M, Tiflati H and Horgan J (2017) Navigating ISIS's preferred platform: Telegram, *Terrorism and Political Violence*. Available at https://doi.org/10.1080/09546553.2017.1339695 (accessed 22 February 2019).

Browne M (2017) YouTube removes videos showing atrocities in Syria, *New York Times*, 22 August.

Brownsword R (2011) Lost in translation: legality, regulatory margins, and technological management. *Berkeley Technology Law Journal* **26**, 1321–1365.

Brownsword R (2015) In the year 2061: from law to technological management. *Law, Innovation and Technology* **7**, 1–51.

Brownsword R (2016) Technological management and the rule of law. *Law, Innovation and Technology* **8**, 100–140.

Carlile A (2007) *The Definition of Terrorism*, Cm 7052. London: The Stationery Office.

Chandler A (2017) The racist algorithm. *Michigan Law Review* **115**, 1023–1045.

Citron DK (2008) Technological due process. *Washington University Law Review* **85**, 1249–1314.

Citron DK (2018) Extremist speech, compelled conformity, and censorship creep. *Notre Dame Law Review* **93**, 1035–1071.

Citron DK and Pasquale F (2014) The scored society: due process for automated decisions. *Washington Law Review* **89**, 1–33.

Conway M *et al.* (2017) *Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts*. Dublin: VOX-Pol Network of Excellence.

Dan-Cohen M (1984) Decision rules and conduct rules: on acoustic separation in criminal law. *Harvard Law Review* **97**, 625–677.

Deep Dot Web (2017) After being shut down by Google and GoDaddy, major neo-Nazi site moves to Darknet, *Deep Dot Web*, 31 August. Available at https://www.deepdotweb.com/2017/08/31/shut-google-godaddy-major-neo-nazi-site-moves-darknet/ (accessed 22 February 2019).

Desai DR and Kroll JA (2017) Trust but verify: a guide to algorithms and the law. *Harvard Journal of Law & Technology* **31**, 1–64.

Edwards J (2010) Justice denied: the criminal law and the ouster of the courts. *Oxford Journal of Legal Studies* **30**, 725–748.

Facebook (2016) Partnering to help curb spread of online terrorist content, *Facebook News*, 5 December. Available at https://newsroom.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content/ (accessed 22 February 2019).

Fletcher G (2006) The indefinable concept of terrorism. *Journal of International Criminal Justice* **4**, 894–911.

Gerbaudo P (2012) *Tweets and the Streets: Social Media and Contemporary Activism*. London: Pluto Press.

GIFCT (2018) Global Internet Forum to Counter Terrorism: an update on our efforts to use technology, support smaller companies and fund research to fight terrorism online, *Global Internet Forum to Counter Terrorism Press*, 8 June. Available at https://gifct.org/press (accessed 22 February 2019).

Greenfield P (2018) Home Office unveils AI program to tackle ISIS online propaganda, *The Guardian*, 13 February.

Grinnell D *et al.* (2018) *Who Disseminates* Rumiyah*? Examining the Relative Influence of Sympathiser and Non-sympathiser Twitter Users*, paper presented at the 2nd European Counter Terrorism Centre (ECTC) conference on online terrorist propaganda, 17–18 April 2018, at Europol Headquarters, The Hague. Available at https://www.europol.europa.eu/publications-documents/who-disseminates-rumiyah-examining-relative-influence-of-sympathiser-and-non-sympathiser-twitter-users (accessed 22 February 2019).

Grinnell D, Macdonald S and Mair D (2017) *The Response of, and on, Twitter to the Release of* Dabiq *Issue* 15, paper presented at the 1st European Counter Terrorism Centre (ECTC) conference on online terrorist propaganda, 10–11 April 2017, at Europol Headquarters, The Hague. Available at https://www.europol.europa.eu/publications-documents/response-of-and-twitter-to-release-of-dabiq-issue-15 (accessed 22 February 2019).

Hardy K and Williams G (2011) What is 'terrorism'? Assessing domestic legal definitions. *UCLA Journal of International Law and Foreign Affairs* **16**, 77–162.

Haynes J (2017) Backdoor access to WhatsApp? Rudd's call suggests a hazy grasp of encryption, *The Guardian*, 27 March.

Hildebrandt M (2008) Legal and technological normativity: more (and less) than twin sisters. *Techne* **12**, 169–183.

Hildebrandt M (2018) Law *as* computation in the era of artificial legal intelligence: speaking law to the power of statistics. *University of Toronto Law Journal* **68**(Suppl 1), 12–35.

HM Government (2015) *The Government Response to the Annual Report on the Operation of the Terrorism Acts in 2013 by the Independent Reviewer of Terrorism Legislation*, Cm 9032. London: The Stationery Office.

HM Government (2018) *CONTEST: The United Kingdom's Strategy for Countering Terrorism*, Cm 9608. London: The Stationery Office.

Hodgson JS and Tadros V (2013) The impossibility of defining terrorism. *New Criminal Law Review* **16**, 494–526.

Home Office (1998) *Legislation Against Terrorism: A Consultation Paper*, Cm 4178. London: The Stationery Office.

Horder J (2016) *Ashworth's Principles of Criminal Law*, 8th edn. Oxford: Oxford University Press.

ICT4Peace and Counter-Terrorism Committee Executive Directorate (2016) *Private Sector Engagement in Responding to the Use of the Internet and ICT for Terrorist Purposes: Strengthening Dialogue and Building Trust*. Geneva: ICT4Peace.

Jigsaw (2017) *The Redirect Method: A Blueprint for Bypassing Extremism*. Available at www.redirectmethod.org (accessed 22 February 2019).

Joh E (2016) The new surveillance discretion: automated suspicion, big data, and policing. *Harvard Law & Policy Review* **10**, 15–42.

Kavanagh C *et al.* (2017) Terrorist use of the Internet and cyberspace: issues and responses. In Conway M *et al.* (eds), *Terrorists' Use of the Internet: Assessment and Response*. Amsterdam: IOS Press.

Kayyali D and Althaibani R (2017) Vital human rights evidence in Syria is disappearing from YouTube, *VOX-Pol Blog*, 22 November. Available at http://www.voxpol.eu/vital-human-rights-evidence-syria-disappearing-youtube/ (accessed 22 February 2019).

Kim PT (2017) Auditing algorithms for discrimination. *University of Pennsylvania Law Review Online* **166**, 189–203.

Kroll JA *et al.* (2017) Accountable algorithms. *University of Pennsylvania Law Review* **165**, 633–705.

Lehr D and Ohm P (2017) Playing with the data: what legal scholars should learn about machine learning. *University of California, Davis, Law Review* **51**, 653–717.

Levitt G (1986) Is 'terrorism' worth defining? *Ohio Northern University Law Review* **13**, 97–116.

Lorenzo-Dus N and Macdonald S (2018) Othering the West in the online jihadist propaganda magazines. *Inspire* and *Dabiq: Journal of Language, Aggression and Conflict* **6**, 79–106.

Macdonald S (2009a) The unbalanced imagery of anti-terrorism policy. *Cornell Journal of Law and Public Policy* **18**, 519–540.

Macdonald S (2009b) Why we should abandon the balance metaphor: a new approach to counterterrorism policy. *ILSA Journal of International and Comparative Law* **15**, 95–146

Macdonald S (2014) Prosecuting suspected terrorists: precursor crimes, intercept evidence and the priority of security. In Jarvis L and Lister M (eds), *Critical Perspectives on Counter-terrorism*. Abingdon: Routledge, pp. 130–149.

Macdonald S and Lorenzo-Dus N (forthcoming) Purposive and performative persuasion: the linguistic basis for criminalising the (direct and indirect) encouragement of terrorism. In Fuller C and Finkelstein C (eds), *Using Law to Fight Terror: Legal Approaches to Combating Violent Non-state and State-sponsored Actors*. Oxford: Oxford University Press.

Malik N (2018) *Terror in the Dark: How Terrorists Use Encryption, the Darknet, and Cryptocurrencies*. London: The Henry Jackson Society.

May T (2018) Theresa May's Davos address in full, *World Economic Forum*, 25 January. Available at https://www.weforum.org/agenda/2018/01/theresa-may-davos-address/ (accessed 22 February 2019).

Morozov E (2013) *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems that Don't Exist*. London: Allen Lane.

Ní Aoláin F (2018) Mandate of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, OL OTH 46/2018, 24 July. Available at https://www.ohchr.org/Documents/Issues/Terrorism/OL_OTH_46_2018.pdf (accessed 22 February 2019).

Nouri L (2018) Britain first and Facebook: banned but not solved, *Europe Now*, 2 October. Available at https://www.europe-nowjournal.org/2018/10/01/britain-first-banned-on-facebook-but-not-solved/ (accessed 20 March 2019).

O'Callaghan D *et al.* (2013) Uncovering the wider structure of extreme right communities spanning popular online networks, *WebSci* '13, Proceedings of the 5th Annual ACM Web Science Conference, New York, 276–285.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

Pearson E (2017) Online as the new frontline: affect, gender, and ISIS-take-down on social media. *Studies in Conflict & Terrorism*. Available at https://doi.org/10.1080/1057610X.2017.1352280 (accessed 22 February 2019).

Posner EA and Vermeule A (2007) *Terror in the Balance: Security, Liberty, and the Courts*. Oxford: Oxford University Press.

Posner RA (2006) *Not a Suicide Pact: The Constitution in a Time of National Emergency*. Oxford: Oxford University Press.

Prucha N (2016) IS and the jihadist information highway: projecting influence and religious identity via telegram. *Perspectives on Terrorism* **10**, 48–58.

Ragan S (2015) After Paris, ISIS moves propaganda machine to Darknet, *CSO News*, 15 November. Available at https://www.csoonline.com/article/3004648/security-awareness/after-paris-isis-moves-propaganda-machine-to-darknet.html (accessed 22 February 2019).

Rich M (2013) Should we make crime impossible? *Harvard Journal of Law and Public Policy* **36**, 795–848.

Rich M (2016) Machine learning, automated suspicion algorithms, and the Fourth Amendment. *University of Pennsylvania Law Review* **164**, 871–929.

Rosenthal D (2011) Assessing digital preemption (and the future of law enforcement). *New Criminal Law Review* **14**, 576–610.

Roth A (2016) Trial by machine. *Georgetown Law Journal* **104**, 1245–1305.

Saul B (2006) *Defining Terrorism in International Law*. Oxford: Oxford University Press.

Schwartz B (2016) Google's Paul Haahr: we don't fully understand RankBrain, *Search Engine Roundtable*, 8 March. Available at https://www.seroundtable.com/google-dont-understand-rankbrain-21744.html (accessed 22 February 2019).

Shane S (2017) In 'watershed moment,' YouTube blocks extremist cleric's message, *The New York Times*, 12 November.

Spigelman JJ (2003) The rule of law and enforcement. *University of New South Wales Law Journal* **26**, 200–209.

Stalinsky S and Sosnow R (2016) The jihadi cycle on content-sharing web services 2009–2016 and the case of Justpaste.it: favored by ISIS, Al-Qaeda, and other jihadis for posting content and sharing it on Twitter – jihadis move to their own platforms (Manbar, Nashir, Alors.Ninja) but then return to Justpaste.it, *MEMRI Inquiry & Analysis Series No* 1255, 6 June. Available at https://www.memri.org/reports/jihadi-cycle-content-sharing-web-services-2009-2016-and-case-justpaste-it-favored-isis-al (accessed 20 March 2019).

Tadros V (2007) Justice and terrorism. *New Criminal Law Review* **10**, 658–689.

Tech Against Terrorism (2017) UK launch of tech against terrorism at Chatham House, *Tech Against Terrorism*, 12 July. Available at https://www.techagainstterrorism.org/2017/07/12/tat-at-chatham-house/ (accessed 22 February 2019).

Twitter (2016) Combating violent extremism, *Twitter blog*, 5 February. Available at https://blog.twitter.com/official/en_us/a/2016/combating-violent-extremism.html (accessed 22 February 2019).

Twitter (2017) New data, new insights: Twitter's latest #Transparency report, *Twitter Public Policy*, 19 September. Available at https://blog.twitter.com/official/en_us/topics/company/2017/New-Data-Insights-Twitters-Latest-Transparency-Report.html (accessed 22 February 2019).

Twitter (2018a) Expanding and building #TwitterTransparency, *Twitter Public Policy*, 5 April. Available at https://blog.twitter.com/official/en_us/topics/company/2018/twitter-transparency-report-12.html (accessed 22 February 2019).

Twitter (2018b) Our range of enforcement options, *Twitter Help Center*. Available at https://help.twitter.com/en/rules-and-policies/enforcement-options (accessed 22 February 2019).

United Nations Security Council (2017) *Fourth Report of the Secretary-General on the Threat Posed by ISIL (Da'esh) to International Peace and Security and the Range of United Nations Efforts in Support of Member States in Countering the Threat*, S/2017/97. New York, NY: United Nations.

Waldron J (2003) Security and liberty: the image of balance. *Journal of Political Philosophy* **11**, 191–210.

Wojcicki S (2017) Expanding our work against abuse of our platform, *YouTube Official Blog*, 4 December. Available at https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html (accessed 22 February 2019).

Youmans WL and York JC (2012) Social media and the activist toolkit: user agreements, corporate interests, and the information infrastructure of modern social movements. *Journal of Communication* **62**, 315–329.

YouTube (2018) Community Guidelines strike basics, *YouTube Help*. Available at https://support.google.com/youtube/answer/2802032?hl=en-GB (accessed 22 February 2019).

Zedner L (2005) Securing liberty in the face of terror: reflections from criminal justice. *Journal of Law and Society* **32**, 507–533.

Cite this article: Macdonald S, Correia SG, Watkin A-L (2019). Regulating terrorist content on social media: automation and the rule of law. *International Journal of Law in Context* **15**, 183–197. https://doi.org/10.1017/S1744552319000119