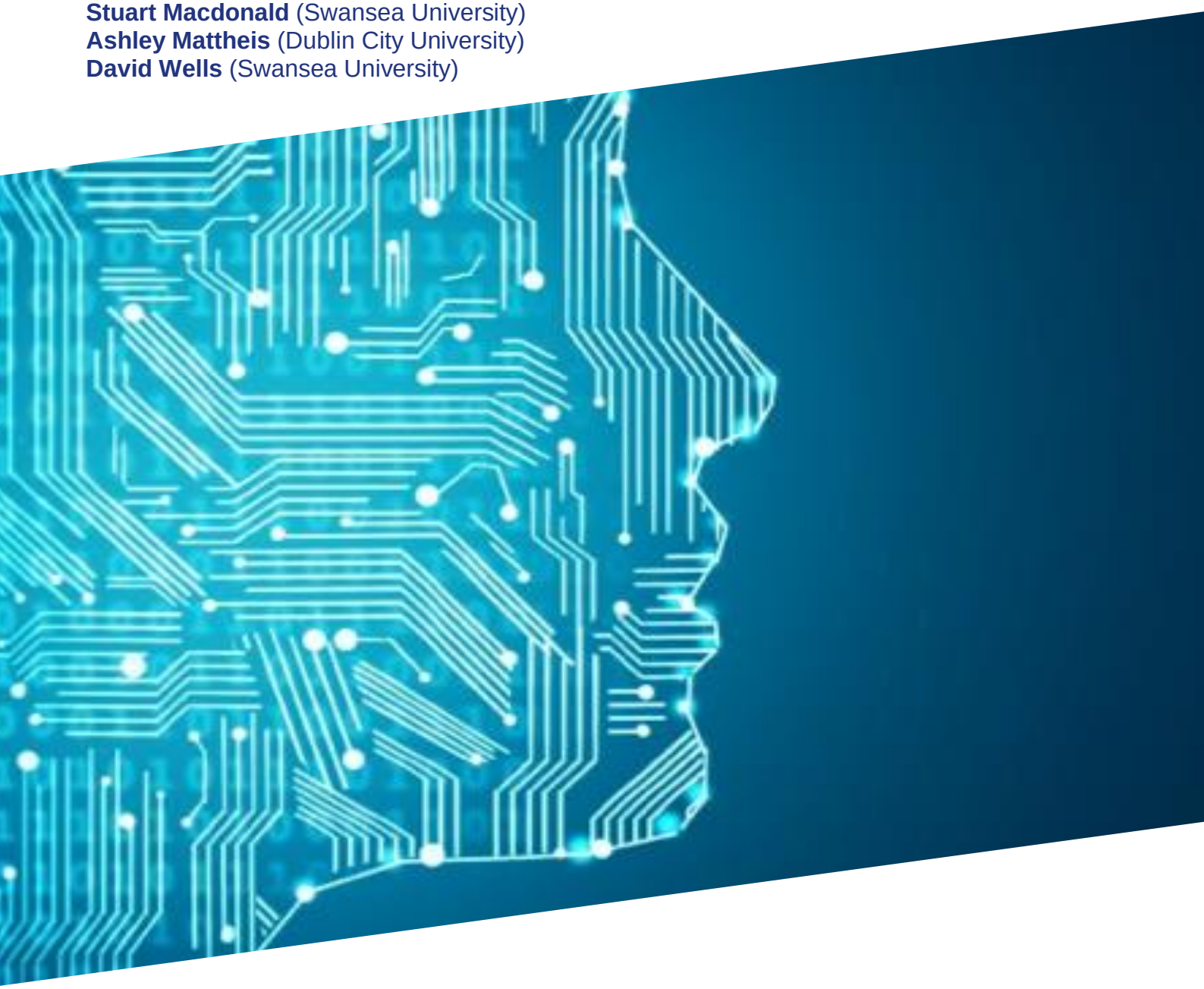
tech against terrorism europe

Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online

Stuart Macdonald (Swansea University)
Ashley Mattheis (Dublin City University)
David Wells (Swansea University)



DCU

Swansea University
Prifysgol Abertawe

Funded by the European Union

Stuart Macdonald is Professor of Law at Swansea University, UK. Stuart is Co-Director of the University's Cyber Threats Research Centre (CYTREC), the lead organiser of the biennial #TASMConf (Terrorism and Social Media Conference) and the Co-ordinator of the VOX-Pol Network. Stuart's research interests lie in criminal law and counterterrorism, particularly cyberterrorism and terrorists' use of the internet. His most recent work has examined violent jihadist narratives, their dissemination via online platforms, and legal and policy responses. Stuart has held visiting scholarships in the US, Australia and France and in 2016/17 was the holder of a Fulbright Cyber Security Award

Ashley A. Mattheis, PhD, is a Postdoctoral Researcher in the School of Law and Government at Dublin City University, Ireland. Her work brings together cultural studies, media studies, and visual rhetorical criticism, through the lens of feminist theory to explore cultural production and consumption online with a focus on extremist digital cultures including the Manosphere, the Far/Alt-Right, #Trad, and QAnon. Her publications have explored the rhetorical and persuasive force of extremist propaganda materials, gendered communicative approaches, and media circulation practices; the material effects of socio-technical systems; as well as digital research ethics, methodology, and methods.

David Wells is a global security consultant focused on emerging terrorism and counter-terrorism challenges. He is also an Honorary Research Associate at Swansea University's Cyber Threats Research Centre, and a non-Resident Scholar at the Middle East Institute. Between 2017 and 2022, he was Head of Research and Analysis at the UN Counter-Terrorism Directorate, managing the team responsible for monitoring terrorism trends for the Counter-Terrorism Committee of the Security Council. David began his career coordinating investigations into international terrorist networks for UK intelligence agency GCHQ, worked for multiple agencies in the Australian intelligence community, and has collaborated with the leading research and academic institutions focused on counter-terrorism and CVE.

Table of contents

Executive summary

The enormous quantity of content that is posted to online platforms every minute must be assessed for compliance with a diverse range of prohibitions, from the promotion of terrorism and violent extremism to spam and violation of intellectual property rights. Since it is not possible to manually inspect every individual item, there has been a concerted effort to develop automated tools for the identification and removal of violative content. The focus of this report is the use of automated content-based tools – in particular those that use artificial intelligence (AI) and machine learning – to detect terrorist content.

In broad terms, automated content-based tools for identifying terrorist content online follow one of two approaches. The first approach is **matching-based** (see section 3.1). This involves comparing new images or videos to an existing database of images and videos that have previously been identified as terrorist in nature. Matching-based approaches rely on a technique known as hashing, in which items of content are converted into a string of data intended to uniquely identify that specific item. The most secure form of hashing is cryptographic hashing. Cryptographic hashes appear to be random, so that they reveal nothing about the content from which they are derived. While this is beneficial in terms of privacy, it is vulnerable to attempts to evade detection since even tiny alterations to the content will generate a completely different hash value. For this reason, tech companies have tended to rely on perceptual hashing. This focuses on patterns in salient features of the hashed content and disregards changes that would go unnoticed by human viewers. While this is more resilient to attempts to circumvent content moderation, the values of perceptual hashes reveal something about the underlying input and are vulnerable to both reversal attacks (where the hash is used to generate the original image) and poisoning attacks (where an image is generated that has the same hash value as, for example, a corporate logo, and the generated image is added to the hash database, preventing the logo from being uploaded to the platform).

The second approach is **classification-based** (see section 3.2). This approach often employs machine learning and other forms of AI, such as Natural Language Processing (NLP). Classification-based approaches typically involve using a large corpus of texts, which have been manually annotated by human reviewers, to train the AI to recognise the features of specified categories (such as terrorist content). The AI is then able to predict whether a new item of content belongs to one of these categories.

Classification-based approaches raise three sets of issues:

- First, attempts to compile a dataset to train the AI face a number of challenges, including collecting a dataset that is both sufficiently large and also representative of the data on which the algorithms will be deployed, as well as cleaning and labelling the data – which is a time-consuming and resource-intensive task.

- The next set of issues concern the temporal, contextual and cultural limitations of machine learning algorithms. The algorithms reflect the time period during which the training data were collected, which is problematic in a dynamic and evolving field like online extremism. They also have difficulty understanding context – including such things as subtlety, irony, and sarcasm – and in discerning the intention of humans, which is at odds with the centrality of intention to definitions of 'terrorist content'. And they lack cultural sensitivity, including variations in dialect and language use across different groups.
- Together, the two previous sets of issues result in a danger that automated content moderation tools will produce incorrect and inconsistent outcomes. There have been reported instances of failures to remove hate speech, while at the same time there has been overenforcement that has curbed users' freedom of expression, including of activists and journalists. This generates resentment and mistrust and has led to claims that content moderation enforces Western values upon users from the Global South.

So, while automated tools and techniques are essential given the volume of online content, **human input remains necessary** (see section 4.1). Matching-based approaches require an ongoing manual effort to identify items of terrorist content and add these to the hash database. Classification-based approaches require human input to prepare a large dataset and train the machine learning algorithms. Moreover, most classification-based tools are only used to flag items for human review, so that human moderators are left to make contextual judgements, assess nuance and intention, and consider social, cultural, historical, and political factors. It is important that companies employing human moderators not only ensure that the moderators have the necessary expertise, but also that adequate provision is made for their health and wellbeing. ***A set of minimum standards, including examples of best practice and provision for moderators' wellbeing, should be developed for those employing content moderators (recommendation 1(a))***.

There is also the potential to **develop AI to address some of these challenges** (see section 5.3). Generative AI approaches can be used to create new items of content, in order to correct biases in training datasets and ensure that they are more representative. AI can also be used to reduce the harmful effects on human moderators, such as by identifying and blurring out areas of images and using visual question answering to enable moderators to reach a decision by asking the system questions about the content without viewing it directly. ***Further development of AI tools for safeguarding the wellbeing of content moderators should be promoted (recommendation 1(b)).***

It is crucial to ensure that **appropriate oversight mechanisms** are in place (see section 4.2). As well as providing a way to correct errors, this also helps to ensure accountability, improve the quality of decision making, and build trust and legitimacy. One form of oversight mechanism is an appeals process. This must adhere to standards of due process and ensure that the user's opportunity to appeal is effective. Other forms of oversight include transparency mechanisms (including the publication of relevant policies and transparency reports), algorithmic auditing and access to data. Many of these requirements have been formalised by the EU's Terrorist Content Online Regulation and Digital Services Act.

Developing automated tools, recruiting human moderators with the necessary expertise, providing them with wellbeing support, and ensuring appropriate oversight mechanisms are in place all requires a **significant investment of resource**. One way in which some companies address this is to purchase content moderation capabilities from a third-party provider (see section 5.1). This has a number of potential benefits, though there are some important issues to consider. These include the quality and relevance of the data on which the off-the-shelf product has been trained and the risk of biased or discriminatory decisions against the user base. In addition, oversight of third-party providers, including in terms of human rights compliance, is more limited than for tech platforms. There are also collaborative initiatives that offer capacity-building and knowledge-sharing services (see section 5.2). These include the capacity-building programme offered by Tech Against Terrorism Europe, the Knowledge-Sharing Platform and Terrorist Content Analytics Platform provided by Tech Against Terrorism and the hash-sharing database maintained by the Global Internet Forum to Counter Terrorism (GIFCT).

Small platforms should (1) assess any off-the-shelf content moderation solution carefully, (2) explore the opportunity to make use of Tech Against Terrorism Europe's capacity-building programme, the knowledge-sharing platform and Terrorist Content Analytics Platform offered by Tech Against Terrorism, and the hash-sharing database maintained by GIFCT, and (3) where GIFCT membership is not possible, seek other potential forms of collaboration to bolster their content moderation resources (recommendation 2). Alongside this, international organisations and governments should support the development of openly available automated content moderation tools by NGOs. In addition, the largest tech platforms should develop and make openly available automated content moderation tools, accompanied by a good practice guide that explains how the tool works, its limitations, and how it can be integrated into a platform. Large platforms should also consider developing multiple models for collaboration, taking into account both their need to vet partners and protect IP whilst also enabling increased access to tools and collaboration for small to medium platforms (recommendation 3).

1. Introduction

There have been many developments in artificial intelligence (AI) and machine learning in recent years, across a number of different fields. One area in which there has been a concerted effort to realise the benefits of these advances is content moderation. This is unsurprising, given the sheer volume of content that is posted to online platforms. On average, every minute Facebook users share 694,000 stories, X (formerly Twitter) users post 360,000 posts, Snapchat users send 2.7 million snaps and YouTube users upload over 500 hours of new content.[1] And these large platforms form just part of a much wider ecosystem, with Wikipedia for example listing a total of 260 active social networking services. The volume of data generated is growing exponentially and is currently estimated at 120 zettabytes every day.[2] This must be assessed for compliance with a diverse range of prohibitions that includes the extremist, violent, sexually explicit and fraudulent, as well as that which constitutes sexual exploitation, the promotion of self-harm, spam, a violation of intellectual property rights and the trade of restricted goods and services. Self-evidently, such a vast amount of content cannot all be manually inspected to check adherence to these standards.

Although bold claims have been made about the ability of automated content moderation tools, including by the leading social media companies,[3] this is a complex field. For a start, each type of prohibited content raises a distinct set of issues and challenges, such that each content type requires its own set of automated tools with their own distinct architecture.[4] In this report, the focus is the use of automated tools to identify terrorist content online, in particular ones that use AI and machine learning.

Online terrorist propaganda has been an important policy concern for at least the past decade.[5] Against a backdrop of calls for tech companies to do more to ensure the resilience of their platforms against exploitation by terrorists, regulatory regimes have been implemented at both the national level (e.g., Germany's Network Enforcement Act and the UK's Online Safety Act) and the transnational level. An example of the latter is the EU's Terrorist Content Online Regulation,[6] the provisions of which include the power for member state competent authorities to issue hosting service providers with an order requiring them to remove or disable access to identified items of

1 Jimit Bagadiya, '500+ Social Media Statistics and Facts of 2023' (SocialPilot, 2 October 2023) <https://www.socialpilot.co/blog/social-media-statistics> accessed 27 October 2023.

2 Petroc Taylor, 'Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025' (Statista, 16 November 2023) <https://www.statista.com/statistics/871513/worldwide-data-created/> accessed 20 November 2023.

3 Tarleton Gillespie, 'Content moderation, AI, and the question of scale' (2020) 7 Big Data & Society https://doi.org/10.1177/2053951720943234.

4 Cambridge Consultants, Use of AI in Online Content Moderation (OFCOM, 2019).

5 Anne Aly, Stuart Macdonald, Lee Jarvis and Thomas M Chen, 'Introduction to the Special Issue: Terrorist Online Propaganda and Radicalization' (2017) 40 Studies in Conflict & Terrorism 1.

6 Regulation 2021/784 on addressing the dissemination of terrorist content online (29 April 2021).

terrorist content 'as soon as possible and in any event within one hour of receipt'.[7] Alongside this, the EU Commission launched a call for proposals for projects aimed at supporting small companies in implementing the Regulation. Three projects were funded under this call. This report forms part of one of these projects, which is entitled Tech Against Terrorism Europe.[8]

It is important to note at the outset that the focus of this report is the use of AI and machine learning to identify terrorist content online using content-based approaches. Accordingly, the following are outside the scope of the report:

- The moderation of so-called borderline content, i.e., content that does not violate a platform's Terms of Service but which is nevertheless regarded as potentially harmful.[9]
- The identification of individuals on a radicalisation trajectory, which is a different – and even more difficult – task;[10] and,
- The use of behaviour-based cues, such as abnormal posting volume and use of unrelated, trending hashtags, to identify accounts that are sharing terrorist content.[11] This includes approaches based on recidivism.[12]

The report begins, in section 2, by explaining the terms AI, machine learning and terrorist content online. Readers that are already familiar with these concepts may wish to move straight to section 3, which discusses the two main content-based approaches to the automated identification of terrorist content online: matching-based approaches and classification-based ones. Having explained the limitations of each approach, section 4 details two ways in which it is necessary to supplement automated tools. Section 5 then addresses issues of resource, before the report concludes with three recommendations.

7 ibid, Article 3(1).

8 See <https://tate.techagainstterrorism.org> accessed 20 November 2023.

9 See further Stuart Macdonald and Katy Vaughan, 'Moderating Borderline Content while Respecting Fundamental Values' (2023) Policy & Internet https://doi.org/10.1002/poi3.376 and 'Sanitising Extremism: "Borderline Content" and Antisemitism Online' (Tech Against Terrorism Podcast, 27 April 2023) <https://podcast.techagainstterrorism.org/1684819/12711788-sanitising-extremism-borderline-content-and-antisemitism-online> accessed 20 November 2023.

10 Miriam Fernandez and Harith Alani, 'Artificial Intelligence and Online Extremism: Challenges and Opportunities' in John McDaniel and Ken Pease (eds), Predictive Policing and Artificial Intelligence (Routledge, 2021).

11 Isabelle van der Vegt, Paul Gill, Stuart Macdonald and Bennett Kleinberg, Shedding Light on Terrorist and Extremist Content Removal (GRNTT, 2019).

12 On which, see further TG Thorley and E Saltman, 'GIFCT Tech Trials: Combining Behavioural Signals to Surface Terrorist and Violent Extremist Content Online' (2023) Studies in Conflict & Terrorism https://doi.org/10.1080/1057610X.2023.2222901.

2. Defining key terms

2.1 Artificial Intelligence and machine learning

There are various definitions of AI. Some focus on thinking like a human (such as '[t]he automation of activities that we associate with human thinking, activities such as decision-making, problem solving, learning'[13]) or acting like a human (such as '[t]he art of creating machines that perform functions that require intelligence when performed by people'[14]). Other definitions focus on thinking rationally (such as '[t]he study of the computations that make it possible to perceive, reason, and act'[15]) or acting rationally (such as '[a] field of study that seeks to explain and emulate intelligent behaviour in terms of computational processes'[16]). From a practical and strategic perspective, the focus tends to be on the behavioural performance of AI systems at an equal or better than human level of accuracy, speed or decision quality.[17] Indeed, AI is commonly divided into three tiers:

- artificial narrow intelligence (ANI) - machine intelligence that equals or exceeds human intelligence for specific tasks;
- artificial general intelligence (AGI); machine performance meeting the full range of human performance across any task;
- artificial superintelligence (ASI) - machine intelligence that exceeds human intelligence across any task.[18]

Machine learning lives at the intersection of computer science, statistics, and data science. It can be thought of narrow AI, in the sense that machine learning systems can learn to carry out specific functions intelligently.[19] While traditional approaches to programming rely on hardcoded rules which set out how to solve a problem step-by-step, machine learning allows computers to detect patterns from examples, data and experience in order to learn how best to achieve the desired output. Today, people interact with machine learning-driven systems on a daily basis – from spam filtering and product recommendations to facial recognition and predictive text – and there is significant future potential in fields such as healthcare, transport and education.

13 Richard Bellman, An Introduction to Artificial Intelligence: Can Computers Think? (Boyd & Fraser Publishing Company, 1978).

14 Ray Kurzweil, The Age of Intelligent Machines (MIT Press, 1990).

15 Patrick Henry Winston, Artificial Intelligence (3rd edn, Addison-Wesley Publishing Company, 1992).

16 Robert J Schalkoff, Artificial Intelligence: An Engineering Approach (McGraw-Hill Education, 1990).

17 Stephan De Spiegeleire, Matthijs Maas and Tim Sweijs, Artificial Intelligence and the Future of Defense: Strategic Implications for Small- and Medium-Sized Force Providers (Hague Centre for Strategic Studies, 2017).

18 ibid.

19 Machine learning: the power and promise of computers that learn by example (Royal Society, 2017).

Figure 1 provides an overview of the machine learning process.[20] The first stage of the process is to define what the machine learning algorithms will predict or estimate, and how this should be measured. Here, a distinction can be drawn between supervised and unsupervised machine learning.[21] In supervised machine learning, data points are categorised into groups and labelled (e.g., mouse, cat, dog). These labelled data are then used to train the system, so that it can predict the categories of other data points. This may be contrasted with unsupervised machine learning, in which the data are not labelled. Instead, the system creates clusters of data that share similar characteristics. This clustering can be an end in itself, or it can constitute the first step towards a supervised approach.

The next stages are data collection, data cleaning, and a summary statistics review. To realise the benefits of machine learning, a sufficiently large dataset must be compiled. The dataset must also be representative of the real-world data on which it will later be deployed.
Having collected the dataset, any missing or incorrect values should be identified and addressed. The dataset should also be reviewed for any outliers that may cause concern about generalisability.

At this point the dataset is split into a training dataset and a test dataset. The training dataset is used for the machine learning algorithm to learn the optimal predictive rules, while the test dataset is used to assess its accuracy and performance. An important consideration at this stage is how much of the data should be allocated to the training dataset, and how much to the test dataset. The larger the training dataset, the greater the chance that the algorithm will learn predictively useful rules. But a smaller test dataset will mean more uncertainty about how well the algorithm's performance might generalise to data other than those on which it was trained.

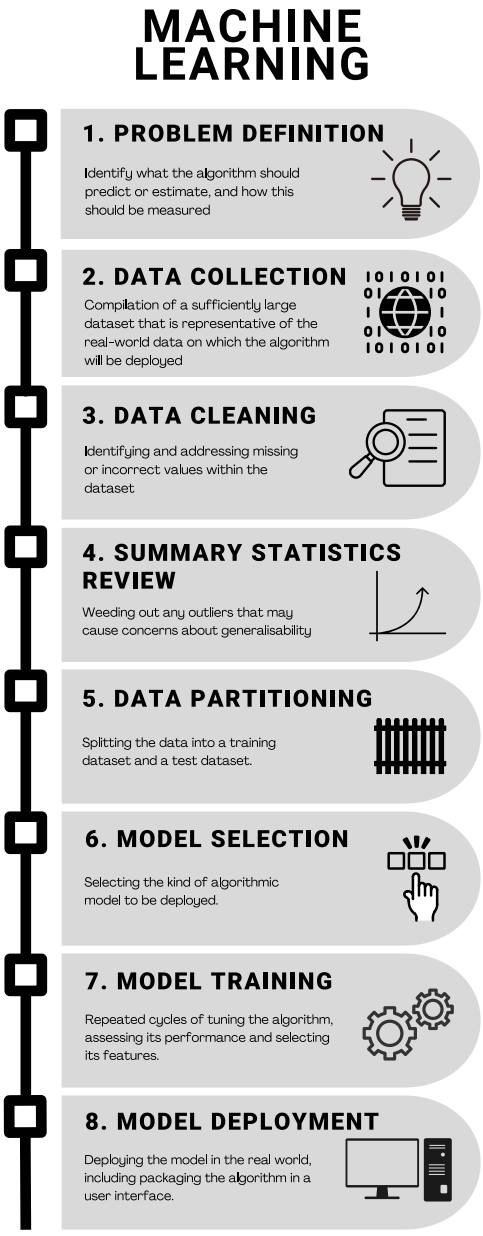


Figure 1: The stages of machine learning

20 Figure 1 is based on the description contained in David Lehr and Paul Ohm 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning' (2017) 51 University of California, Davis Law Review 653.

21 Machine learning: the power and promise of computers that learn by example (n 19). There is also a third branch of machine learning, which lies somewhere between supervised and unsupervised learning: reinforcement learning. Here, the system interacts with its environment, making sequential decisions so as to maximise future rewards (e.g., learning which moves were important in winning a game).

The next stages are model selection and training. A number of factors will influence the choice of the kind of algorithmic model. These include the type of outcome (e.g., a binary classification, placement on a continuous scale), the importance of explainable decisions, resource limitations and asymmetric cost ratio (importing how different types of error are viewed normatively, so that some types of error are made more often than others). The learning part of the process – model training – can then begin. Training consists of tuning, assessment and feature selection. Tuning the algorithm may involve determining how to assess its accuracy, giving effect to an asymmetric cost ratio and striking a balance between bias and variance,[22] while feature selection involves trimming down the algorithm's set of input variables. Repeated cycles of tuning the algorithm, assessing its performance and selecting its features will often be required.

The final stage is deployment of the model. This will often require packaging the algorithm into some kind of user interface. Machine learning algorithms running at scale may also be regularly and automatically retrained upon the collection of new data. This allows the performance of the system to continue to improve in real time in response to real-world data. However, it also means that there is no opportunity for human checking of the consequences of updates to the model before users are exposed to them.[23]

Although there have been significant advances in machine learning in recent years, there remain some important limitations.[24] Some of these are particularly relevant to efforts to use machine learning to identify terrorist content online. Some machine learning algorithms require large datasets. Not only can accessing such datasets be difficult, but the data also need to be labelled – which can be resource-intensive and time-consuming. It is also difficult to develop machine learning systems that have an understanding of context and that are able to understand the intention of humans. The significance of these challenges for automated efforts to identify terrorist content online will be discussed further in section 3.

22 Bias is the distance between a model's predictions and the true values. Bias errors result from oversimplification and not learning the patterns. In contrast, a high rate of variance means that the model pays so much attention to the training data that it does not generalise from it sufficiently. There is thus a bias-variance trade-off. This trade-off is relevant to the complexity and speed of an algorithm's learning. More complex, slower learning tends to yield less bias but more variance.

23 Machine learning: the power and promise of computers that learn by example (n 19).

24 ibid.

2.2 Terrorist content

Different legal instruments offer different definitions of terrorist content. The focus here is on the definition offered by the EU's Regulation on addressing the dissemination of terrorist content online (the 'TCO Regulation').

According to the TCO Regulation, the term 'content' covers a range of formats, including text, images, sound recordings and videos, as well as live transmissions of terrorist attacks.[25] Article 2(7) explains that content is to be regarded as 'terrorist' in nature if it falls within one of five categories.

The first of these categories, found in Article 2(7)(c), focuses on participation in the activities of a terrorist group. A terrorist group is a group of more than two people that has been established for a period of time and which acts in concert to commit terrorist offences. It does not need to have continuity of membership, formally defined roles for its members, or a developed structure, but must be more than a collection of individuals that is randomly formed for the immediate commission of an offence.[26] Any content that solicits a person or a group of persons to participate in the activities of such a group is terrorist content.[27] Participation is defined as including the provision of funding and the supply of information or material resources in the knowledge that this will contribute to the group's criminal activities.[28]

The other four categories of terrorist content focus on inciting, soliciting, threatening, and providing instruction for the commission of a terrorist offence. These categories consist of two requirements, which will be considered in turn: first, that the content incites, solicits, threatens or provides instruction for the commission of an act; second, that the act in question would constitute a terrorist offence.

First, the content must incite, solicit, threaten or provide instruction for the commission of an act. Each of these four categories is explained further in Article 2(7) of the TCO Regulation. The categories do overlap – in particular, incitement and solicitation – but there are also some significant differences between them.

- The content **incites** the commission of a terrorist offence (Article 2(7)(a)). For content to amount to incitement, it must advocate the commission of the act. This may be either direct or indirect. An example of indirect incitement is the glorification of acts of terrorism. In addition, by advocating the commission of the act the content must cause a danger that the act may be committed.
- The content **solicits** the commission of a terrorist offence (Article 2(7)(b)). While there are similarities between the concepts of incitement and solicitation, this paragraph has three

25 Regulation 2021/784 (n 6), para 11.

26 Directive (EU) 2017/541, Article 2(3).

27 Regulation 2021/784 (n 6), Article 2(7)(c).

28 Directive (EU) 2017/541, Article 4(b).

differences to the previous one. First, it states that the content must have solicited 'a person or a group of persons.' This suggests that the paragraph applies only where the content addresses a specific person or group of people. Second, it is enough that the person(s) was solicited to *contribute* to the commission of the act. It is not necessary that they were solicited to commit the act themselves. Third, this paragraph does not require that the solicitation cause a danger that the act will be committed.

- The content **provides instruction** for the commission of a terrorist offence (Article 2(7)(d)). The paragraph encompasses a wide range of instruction. As well as the 'making or use of explosives, firearms or other weapons or noxious or hazardous substances', other 'specific methods or techniques' are also included. The key restriction is that the instruction must have been provided for the purpose of committing a terrorist offence, or for the purpose of contributing to the commission of a terrorist offence. So, for example, training materials on communication security could fall within this paragraph, but only if the material was provided for the purpose of committing, or contributing to the commission, of a terrorist offence.
- The content constitutes a **threat** to commit a terrorist offence (Article 2(7)(e)).

The act that was incited/solicited/threatened, or for which instruction was provided, must constitute a terrorist offence. Here, Article 3 of Directive (EU) 2017/541 must be applied.[29] This states that the following five conditions must be satisfied for an act to qualify as a terrorist offence.

- The act must involve one of the actions listed in paragraphs (a) – (i) of Article 3(1) of Directive (EU) 2017/541. The actions specified in these nine paragraphs extend beyond killing, and include: attacking the physical integrity of a person; kidnapping and hostage-taking; causing extensive destruction to a transport system or government, public or infrastructure facility that is likely to endanger life or result in major economic loss; seizure of aircraft, ships or other public or goods transport; manufacturing, possessing, transporting, supplying or using explosives or weapons, including chemical, biological, radiological or nuclear weapons; endangering human life by releasing dangerous substances, causing fires, floods or explosions, or interfering with or disrupting the water or power supply; illegal interference with an information system or computer data.
- The act must be intentional.
- The act must constitute an offence under the relevant national law.
- Given its nature or context, the act may seriously damage a country or an international organisation.
- The act must be performed with one of the aims listed in Article 3(2) of Directive (EU) 2017/541. These are: (1) to seriously intimidate a population; (2) to unduly compel a government or an international organisation to perform or abstain from performing any act; or, (3) to seriously destabilise or destroy the fundamental political, constitutional, economic or social structures or a country or an international organisation.

29 Regulation 2021/784 (n 6), Article 2(6).

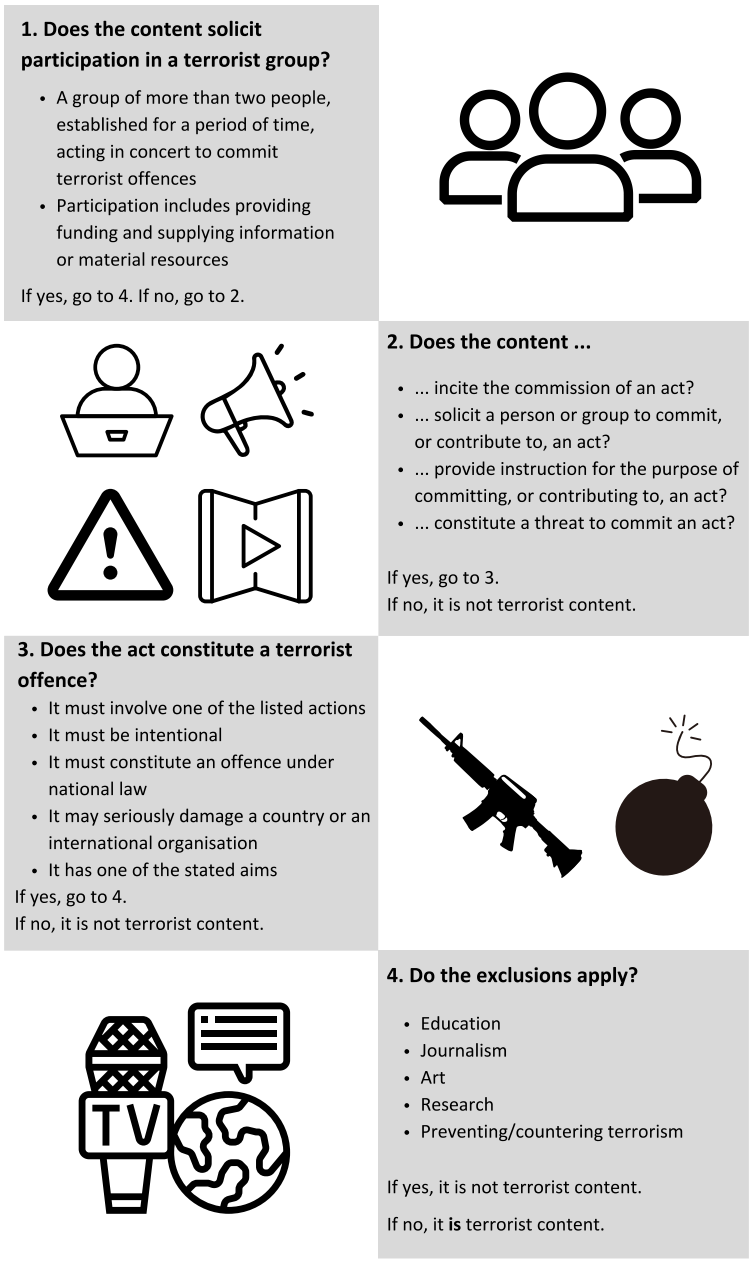


Figure 2: The TCO Regulation definition of terrorist content

If an item of content falls within one of the five categories contained in Article 2(7), the final step is to consider whether the Article 1(3) exclusions apply. According to Article 1(3), content should not be regarded as terrorist in nature if it is disseminated to the public for one of the following five purposes: education; journalism; art; research; or, to prevent or counter terrorism. This includes the dissemination of material 'which represents an expression of polemic or controversial views in the course of public debate'. In making this assessment, the right to freedom of expression and information – including the freedom and pluralism of the media and the freedom of the arts and sciences – should be taken into account.[30]

Figure 2 summarises the preceding paragraphs and provides an overview of the process for identifying terrorist content under the TCO Regulation. When assessing an item of content, competent authorities and hosting service providers should not only take into account the nature and wording of the material, but also the context in which the material was posted and its potential to lead to people's security and safety being harmed.[31] The TCO Regulation also identifies one further 'important factor', namely, whether the material was produced by, is attributable to, or is disseminated on behalf of a person, group or entity on the EU's designation list.[32]

30 ibid, para 12.

31 ibid, para 11.

32 ibid. The list is available at: https://www.consilium.europa.eu/en/policies/fight-against-terrorism/terrorist-list/.

3. Automated identification of terrorist content online

A vast amount of terrorist content is posted across the online ecosystem. From November 2020 to January 2023, Tech Against Terrorism identified terrorist content on a total of 187 different online platforms.[33] Of these, 78 were small or micro platforms.[34] On the biggest platforms, meanwhile, in 2022 alone Facebook removed more than 56 million items of terrorist propaganda,[35] and YouTube removed 275,261 videos for the promotion of violence and violent extremism,[36] while in 2021 X (formerly Twitter) suspended 78,6687 accounts for the promotion of terrorism.[37] Most of this content was detected proactively using automated tools. On Facebook, the proportion of terrorism-promoting content that is detected proactively, before being reported by users, is roughly 98%.[38] The proactive detection rate on YouTube and X is also above 90%.[39]

Several types of content-based moderation processes and tools have been used for the identification of terrorist content online. Of these, a distinction may be drawn between systems that aim to *match* content and systems that aim to *classify* content. While the distinction between approaches based on matching and those based on classification may sometimes be more a matter of degree than of kind, there are important differences between the two.[40] This section provides an overview of each approach.

3.1 Matching

A matching-based approach to detecting terrorist content online works by comparing new images or videos to an existing database of images and videos that have previously been identified as terrorist content. Such systems rely on what is commonly referred to as hashing: a process in which items of content are converted into a string of data intended to uniquely identify that specific item.[41] One of the benefits of this approach is that it enables the sharing of hash values between

33 Patterns of Online Terrorist Exploitation (Tech Against Terrorism, 2023) <https://26492205.fs1.hubspotusercontent-eu1.net/hubfs/26492205/260423%20TCAP%20INSIGHTS%20-%20FINAL.pdf> accessed 20 November 2023.

34 This is, in part, a product of the TCAP collection methodology, which focuses on smaller platforms.

35 'Community Standards Enforcement Report – Dangerous Organizations: Terrorism and Organized Hate' (Meta Transparency Center) <https://transparency.fb.com/data/community-standards-enforcement/dangerous-organizations/facebook/#content-actioned> accessed 30 October 2023.

36 'YouTube Community Guidelines Enforcement' (Google Transparency Report) <https://transparencyreport.google.com/youtube-policy/removals> accessed 30 October 2023.

37 'Rules Enforcement' (X Transparency) <https://transparency.twitter.com/en/reports/rules-enforcement.html> accessed 30 October 2023.

38 'Community Standards Enforcement Report – Dangerous Organizations: Terrorism and Organized Hate' (n 35).

39 'YouTube Community Guidelines Enforcement' (n 36); 'Rules Enforcement' (n 37).

40 Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7 Big Data & Society https://doi.org/10.1177/2053951719897945.

41 ibid.

tech platforms without any transfer of users' personally identifiable information (PII).[42] Hashes are also easy to compute and often smaller in size than the original item of content, making it easier to check whether a hash matches any other in a large database of existing hashes.[43]

Cryptographic hashes are the most secure form of hashing. They aim to create hashes that appear to be random, meaning that they reveal nothing about the content from which they are derived.[44] The drawback, in terms of detecting terrorist content online, is that they require an exact match: any changes – even just changing the colour of one pixel in an image – will result in a completely different hash value. This renders approaches based on cryptographic hashes vulnerable to attempts to evade or circumvent content moderation by making minor modifications to the image or video. That this is an important policy consideration is illustrated powerfully by the 2019 Christchurch attacks. Facebook has reported that the video was viewed fewer than 200 times during the live broadcast.[45] The first user report on the original video arrived 29 minutes after it started, and 12 minutes after the live broadcast ended, by which time a user on 8chan had already posted a link to a copy of the video on a file-sharing site.[46] The video was subsequently shared on YouTube, as well as a number of smaller platforms.[47] In the 24 hours after the attack, Facebook blocked more than 1.2 million videos of the attack at upload. A further 300,000 copies were removed after they were posted.[48] One of the reasons why these additional copies were not detected by Facebook's image and video matching technology was the proliferation of different variants of the video: more than 800 'visually-distinct variants' were in circulation.[49] Some of these were the product of 'a core community of bad actors working together to continually re-upload edited versions of this video in ways designed to defeat our detection'.[50]

For this reason, other forms of hashing are generally used: in particular, perceptual hashing. For example, Facebook's content moderation systems use the photo-matching algorithm PDQ and the video-matching technology TMK+PDQF.[51] Perceptual hashing focuses on patterns in salient features of the hashed content and disregards changes that would go unnoticed by human

42 Thorley and Saltman (n 12).

43 Gorwa, Binns and Katzenbach (n 40).

44 ibid.

45 Guy Rosen, 'A Further Update on New Zealand Terrorist Attack' (Meta Newsroom, 20 March 2019) <https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/> accessed 14 October 2023.

46 ibid.

47 Tech Against Terrorism, 'Analysis: New Zealand attack and the terrorist use of the internet' (26 March 2019) <https://techagainstterrorism.org/news/2019/03/26/analysis-new-zealand-attack-and-the-terrorist-use-of-the-internet> accessed 14 October 2023.

48 Rosen (n 45).

49 ibid.

50 ibid.

51 Antigone Davis and Guy Rosen, 'Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer' (Meta Newsroom, 1 August 2019) <https://about.fb.com/news/2019/08/open-source-photo-video-matching/> accessed 15 October 2023.

viewers.[52] On this approach, the hash value reveals something about the underlying input. The benefit of this, in terms of content moderation, is that it facilitates the detection of visually similar images with mathematically proximate hash values.[53] The downside is that perceptual hashing is vulnerable to both reversal attacks (where the hash is used to generate the original image) and poisoning attacks (where an image is generated that has the same hash value as, e.g., a corporate logo, and the generated image is added to the hash database, preventing the logo from being uploaded to the platform).[54] By contrast, hash collisions (where two different items of content share the same hash value) are very unlikely in cryptographic hashing.[55]

Two further observations should be noted. First, while hashing involves the use of technological tools, human input is also required. For hash matching to be effective, an ongoing effort is required to identify prohibited items of content and add these to the hash database. As Gillespie notes, 'at least for now, the overwhelming majority of what is being automatically identified are copies of content that have already been reviewed by a human moderator'.[56] In addition, a matching-based approach like hashing may be insensitive to the use of the same item of content in a different context – such as journalism or academic research[57] – and so an appeals process involving human review is also necessary. Second, while hashing involves the use of automation, 'it is hardly AI, except under the broadest possible definition'.[58] This may be contrasted with classification-based approaches, to which we now turn.

3.2 Classification

The tools used for classification-based content moderation often employ machine learning, as well as other forms of AI such as Natural Language Processing (NLP).[59] Classification typically involves using a large corpus of texts, which have been manually annotated by human reviewers, to train the AI to recognise the features of specified categories (such as terrorist content, hate speech, etc).[60] The AI is then able to predict whether a new item of content belongs to one of these categories. For example, Facebook uses machine learning to assess posts that may signal support for IS or al-Qaeda. The tool produces a score that indicates the likelihood that the post violates Facebook's counterterrorism policies. Posts with a very high confidence score are

52 Gorwa, Binns and Katzenbach (n 40).

53 Thorley and Saltman (n 12).

54 Nick Locascio, 'Black-Box Attacks on Perceptual Image Hashes with GANs' (Towards Data Science, 3 March 2018) <https://towardsdatascience.com/black-box-attacks-on-perceptual-image-hashes-with-gans-cc1be11f277> accessed 15 October 2023.

55 Gorwa, Binns and Katzenbach (n 40).

56 Gillespie (n 3), 3.

57 Emma Llanso, 'Platforms Want Centralized Censorship. That Should Scare You' (Wired, 18 April 2019) <https://www.wired.com/story/platforms-centralized-censorship/> accessed 15 October 2023.

58 Gillespie (n 3), 3.

59 United Nations Office of Counter-Terrorism, Countering Terrorism Online With Artificial Intelligence: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia (UNOCT, 2021).

60 Gorwa, Binns and Katzenbach (n 40).

automatically removed; other posts with high scores are prioritised for review by human moderators.[61] A feedback loop is employed, with the machine learning algorithms analysing text that has previously been removed for promoting terrorism in order to improve their future performance.[62]

The use of classification-based tools to identify terrorist content raises three sets of issues. The first of these concerns the challenges involved in compiling a training dataset. At the data collection stage, there are no existing datasets that are publicly available and there are obstacles to the sharing of data, including privacy legislation such as the EU's General Data Protection Regulation (GDPR). While it might be possible for platforms to collate posts that they remove for violating terrorism-related prohibitions, this may not generate a sufficiently large dataset. Even if it does yield a large dataset, as in the case of the biggest social media platforms, it is unlikely that this will be fully representative of the data on which the algorithms will be deployed – for the reasons detailed below. Researchers working in this field have employed a variety of tactics in an effort to produce large datasets. One approach has been to search for the use of particular terms (e.g., 'dabiq' or 'rumiyah', the titles of Islamic State online magazines). The difficulty with this is that the data may exclude discussion of other relevant topics or entities. It may also omit algospeak, i.e., terms that have been developed to avoid content moderation. Another approach has been to collect posts from accounts that expressly support terrorist organisations (such as those which are explicitly pro-IS or white supremacist), or to collect posts from channels that expressly support such organisations. However, this approach requires that all posts that are not terrorist content are removed from the dataset and, even then, there remains the question of the extent to which any findings can be generalised more widely. The data may only encompass the terminology of a particular (sub)group and/or a particular language.[63]

There are also challenges at the next stage of the process: cleaning and labelling the data. This is a time-consuming and resource-intensive task.[64] As a consequence, data that were collected using the methods described in the previous paragraph are in many instances either not verified at all, or only partially verified.[65] Moreover, to label the data accurately may require subject-matter expertise, cultural understanding and/or proficiency in other languages. These attributes may be lacking, such as where data annotation is crowdsourced,[66] and the labelling may reflect the demographics and cultural biases of the labellers.[67]

---

61 Monika Bickert and Brian Fishman, 'Hard Questions: What Are We Doing to Stay Ahead of Terrorists?' (Meta Newsroom, 8 November 2018) <https://about.fb.com/news/2018/11/staying-ahead-of-terrorists/> accessed 16 October 2023.

62 Thorley and Saltman (n 12).

63 Fernandez and Alani (n 10).

64 Machine learning: the power and promise of computers that learn by example (n 19).

65 Fernandez and Alani (n 10).

66 Ibid.

67 Spandana Singh, Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content (New America, 22 July 2019) <https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/> accessed 17 October 2023.

As a result of the difficulties in collecting, cleaning, and labelling the data, the dataset may contain false positives (items that have been included but should not have been) and exclude false negatives (items that have not been included but should have been). Not only might this impact the summary statistics review, but, if the algorithms have been trained on erroneous data, they will not in fact be performing at the reported levels of accuracy. This problem is exacerbated by the fact that the training datasets are not made publicly available, meaning that it is not possible for others to verify the data that they contain.[68]

The second set of issues concerns the inherent limitations of machine learning algorithms. First, there are temporal limitations. Datasets are normally collected during a particular time period. Fernandez and Alani explain:

> Data is therefore biased to the world events happening during those particular months (i.e., particular terror attacks, regions of the world, political and religious figures, etc.). Classifiers may therefore learn that naming certain political or religious figures, or locations, are reliable indicators to determine whether a piece of content, or a user account, is radical. However, as time evolves, those locations, those popular figures, those events, may not be relevant or even discussed any longer. In certain cases, they may even become discriminative of the opposite class (e.g., locations under control by a radical group that become liberated).[69]

This problem is particularly acute in the context of terrorist content online, given its dynamic and evolving nature. Extremist groups and movements engage with the latest events and popular culture to try and engage users.[70] They invent new terms and expressions, the meanings of which are sometimes known only to members of the community that created them.[71] They may also deliberately adjust the terminology they employ in order to try and circumvent content moderation systems.[72] As Barnes observes, '[i]n general, AI is built on what already exists, but innovations in wording, phrasing, targets, tactics, and much else are always on the horizon, limiting its capabilities'.[73] To try and address this, the machine learning algorithms may be retrained after deployment using the real-world data that have been collected. This may improve the performance of the system, but it allows no opportunity for human checking of the consequences of the updates to the model before users are exposed to these.[74]

---

68 Fernandez and Alani (n 10).

69 ibid, 147.

70 Stuart Macdonald, Kamil Yilmaz, Chamin Herath, JM Berger, Suraj Lakhani, Lella Nouri and Maura Conway, The European Far-Right Online: An Exploratory Twitter Outlink Analysis of German and French Far-Right Online Ecosystems (Resolve Network, 2022).

71 Fernandez and Alani (n 10).

72 Cambridge Consultants (n 4).

73 Michael Randall Barnes, 'Online Extremism, AI, and (Human) Content Moderation' (2022) 8 Feminist Philosophy Quarterly Article 6, 14.

74 Machine learning: the power and promise of computers that learn by example (n 19).

The second limitation of machine learning algorithms is their difficulty understanding context and nuance. An example is Google's AI AlphaGo. Although the system is able to identify the best possible moves in the board game Go – and defeat the world's best players[75] – it is nonetheless 'unable to explain the context for such moves, or that it is in fact playing a game, or why one would even want to play one'.[76] This limitation means that machine learning algorithms have difficulty accounting for such things as subtlety, irony and sarcasm.[77] This is especially important for certain types of content, such as memes.[78] And, as section 2.2 above showed, an understanding of context is required in order to apply the TCO Regulation's definition of terrorist content. It is necessary to determine the *intention* with which content was posted (was it to *incite* the commission of an act, to *solicit* a person or group to commit an act, etc?). Where content solicited participation in a group, the *purpose* of that group must be assessed. Where instruction was provided for the commission of an act – or where an act was incited, solicited, or threatened – it is necessary to determine the underlying *objective* of that act. And to decide whether any of the exclusions apply, it must be asked whether the content was disseminated for the *purpose* of education, journalism, art, research or to prevent or counter terrorism. Making these contextual judgements and inferences of intention is a complex task which is more suited to human assessment than algorithmic determination.[79]

As well as context, there are also linguistic and cultural limitations. Many NLP tools are only effective for English language text.[80] According to the Facebook papers that were leaked in 2021, 77% of Arabic-language content that had been removed for promoting terrorism had been removed incorrectly.[81] Company insiders were also aware of inadequate coverage of local languages in many countries, such as Myanmar, Afghanistan, India, Ethiopia and much of the Middle East.[82] Few platforms will have the resources to employ specialist teams with such a diverse mix of global dialects.[83] While it may be possible to use AI to help address this difficulty by providing translations,[84] this is also not without its limitations. Schroeter asks, 'Some languages have no grammatical explicit future tense, so what would a threat, which implies the future, even

75 D Silver, A Huang, C Maddison et al, 'Mastering the game of Go with deep neural networks and tree search' (2016) 529 Nature 484.

76 Marie Schroeter, Artificial Intelligence and Countering Violent Extremism: A Primer (GNET, 2020), 8.

77 Gorwa, Binns and Katzenbach (n 40); Gillespie (n 3).

78 Cambridge Consultants (n 4).

79 van der Vegt, Gill, Macdonald and Kleinberg (n 11).

80 Natasha Duarte, Emma Llanso and Anna Loup, Mixed Messages? The Limits of Automated Social Media Content Analysis (Centre for Democracy & Technology, 2017).

81 Mark Scott, 'Facebook did little to moderate posts in the world's most violent countries' Politico (Arlington, VA, 25 October 2021) https://www.politico.com/news/2021/10/25/facebook-moderate-posts-violent-countries-517050 accessed 19 October 2023.

82 Isabel Debre and Fares Akram, 'Facebook's language gaps let through hate-filled posts while blocking inoffensive content' Los Angeles Times (Los Angeles, 25 October 2021) <https://www.latimes.com/world-nation/story/2021-10-25/facebook-language-gap-poor-screening-content> accessed 17 October 2023.

83 Thorley and Saltman (n 12).

84 Cambridge Consultants (n 4).

look like? If automated filtering is to be scaled up, it has to face these design failures.'[85] These linguistic challenges are exacerbated by AI's lack of cultural sensitivity. NLP tools have trouble with variations in dialect and language use across different groups of English speakers.[86] One study, for example, highlighted the difficulties that machine learning content moderation algorithms had in 'assessing culturally shaped English usage in countries in the global South (here, India and Kenya) in extreme speech contexts'.[87] Others have suggested that some systems display racial dialect bias.[88] More generally, there are regional variations both in terms of what national laws prohibit and what is regarded as socially acceptable. Sometimes, words that are used within a community take on a different meaning when they are targeted at members of that community.[89] Interpreting content requires an understanding of societal, cultural, historical, and political factors – which is challenging even for human, as well as automated, moderators.[90]

The third set of issues is a product of the previous two: errors in the training dataset and the limitations of machine learning algorithms mean that there is a danger that content moderation policies will be applied incorrectly and inconsistently. Importantly, as the previous paragraph indicated, this risk is not distributed evenly.[91] In particular, there is a disproportionate impact on users in the Global South.[92] This has a number of important effects. On the one hand, failures to remove hate speech in countries such as Ethiopia[93] and Romania[94] have contributed to real-world violence. On the other hand, overenforcement has curbed important forms of expression.[95] For example, Egyptian opposition activists have reported having their Facebook pages repeatedly banned and their livestreams shut down.[96] Syrian activists have campaigned against the takedown of anti-Assad Facebook accounts and pages that since 2011 have documented war

85 Schroeter (n 76), 13.

86 Duarte, Llanso and Loup (n 80).

87 Sahana Udupa, Antonis Maronikolakis and Axel Wisiorek, 'Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence' (2023) 10 Big Data & Society https://doi.org/10.1177/20539517231172424, 7.

88 Barnes (n 73).

89 ibid.

90 Cambridge Consultants (n 4).

91 Gillespie (n 3).

92 Farhana Shahid and Aditya Vashistha, 'Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony?' in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (ACM, 2023) https://doi.org/10.1145/3544548.3581538.

93 Zecharias Zelalem and Peter Guest, 'Why Facebook Keeps Failing in Ethiopia' (Rest of World, 13 November 2021) <https://restofworld.org/2021/why-facebook-keeps-failing-in-ethiopia> accessed 19 October 2023.

94 Delia Marinescu, 'Facebook's Content Moderation Language Barrier' (The Thread, 8 September 2021) <https://www.newamerica.org/the-thread/facebooks-content-moderation-language-barrier> accessed 19 October 2023.

95 Duarte, Llanso and Loup (n 80).

96 Dania Akkad, 'Revealed: Seven years later, how Facebook shuts down free speech in Egypt' (Middle East Eye, 30 January 2018) <https://www.middleeasteye.net/news/revealed-seven-years-later-how-facebook-shuts-down-free-speech-egypt> accessed 19 October 2023.

crimes.[97] In 2020, the Facebook accounts of more than 60 Tunisian activists, journalists and musicians were disabled.[98] And, during the May 2021 crisis in Israel and Palestine, Facebook deleted content reporting on Israeli forces storming the Al-Aqsa Mosque because #AlAqsa had been added to a hashtag block list.[99] This error was later attributed to two possible factors: Arabic classifiers may have higher error rates for Palestinian Arabic; and, potentially violating Arabic content may not have been routed to content reviewers who spoke or understood the specific dialect of the content.[100] All of this fosters resentment and mistrust, as illustrated by a study of Bangladeshi social media users that had received restrictions for violating Facebook's community standards. In their eyes, Facebook's content moderation was inconsistent, biased and enforced Western values upon users from the Global South, thus operating as a form of 'digital colonialism'.[101]

4. Supplementing automated tools

Having evaluated matching-based and classification-based approaches to content moderation, section 4 of this report considers two ways in which the use of automated tools should be supplemented: integrating human input and ensuring appropriate oversight mechanisms.

4.1 'Human-in-the-loop' processes

As this report has shown, even where automated tools are used human input is still required. For matching-based approaches, an ongoing effort is required to identify items of terrorist content and add these to the hash database. For classification-based approaches, it is necessary to collect, clean and label a large dataset, and to train the machine learning algorithms. Most classification-based tools are also only used to flag items for human review. Human moderators are more able to make contextual judgements, assess nuance and intention, and consider social, cultural, historical and political factors. Appeals processes are also required – such as for when matching-based approaches do not recognise differences in context or when classification-based approaches identify false positives – and human moderators are needed to consider these appeals. In short, it is not possible to fully automate effective content moderation.[102] 'Human-in-

97 'Facebook Deletes Accounts of Assad Opponents' (The Syrian Observer, 8 June 2020) <https://syrianobserver.com/news/58430/facebook-deletes-accounts-of-assad-opponents.html> accessed 19 October 2023.

98 Marwa Fatafta and Rima Sghaier, 'Transparency required: is Facebook's effort to clean up "Operation Carthage" damaging free expression in Tunisia?' (Access Now, 12 June 2020) <https://www.accessnow.org/transparency-required-is-facebooks-effort-to-clean-up-operation-carthage-damaging-free-expression-in-tunisia> accessed 19 October 2023.

99 Marwa Fatafta, 'Facebook is bad at moderating in English. In Arabic, it's a disaster' (Rest of World, 18 November 2021) < https://restofworld.org/2021/facebook-is-bad-at-moderating-in-english-in-arabic-its-a-disaster> accessed 19 October 2023. The following year the Facebook page of a Palestinian news agency was temporarily suspended due to its coverage of Israeli forces again raiding the Al-Aqsa Mosque: 'Facebook briefly suspends Palestinian news page following al-Aqsa raid coverage' (Middle East Eye, 16 April 2022) <https://www.middleeasteye.net/news/israel-palestine-facebook-suspends-news-page-aqsa-raid-coverage> accessed 19 October 2023.

100 BSR, Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021 (BSR, 2022).

101 Shahid and Vashistha (n 92), 11.

102 Cambridge Consultants (n 4).

the-loop' processes are necessary,[103] with humans and automated tools performing different, complementary roles.[104]

Companies employing human moderators should be mindful of two important considerations. The first is capacity, both in terms of the volume of content and the necessary expertise. This encompasses subject matter expertise (e.g., an understanding of the latest terms and expressions being used by particular extremist groups and movements) and linguistic and cultural understanding (reflecting the location of the user base). As one influential report concluded, '[r]esponsible global companies have people on the ground where they do business. A social media platform should be no different. Facebook, YouTube, and Twitter should have offices in every country where users can access their sites'.[105]

The second consideration is the harmful effect of viewing extremist and terrorist content on the health and wellbeing of moderators. Many tech companies, including the largest platforms, have outsourced a significant proportion of the human elements of their content moderation systems to third-party vendors.[106] In some instances, outsourcing has been used even when the moderators work on the same site as the companies' other employees – who have access to perks and benefits that the moderators do not – leading some to suggest that outsourcing is being used as a shield against potential liability issues.[107] Individuals working for these vendors have consistently reported suffering from significant mental health issues, with an absence of meaningful programs to help address the consequences of regularly viewing large volumes of the most graphic and harmful content.[108] Recent reporting suggests that generative AI companies (who are offering their services in the content moderation field) may be similarly outsourcing content moderation and training to low-paid and inadequately resourced vendors,[109] resulting in similar mental health challenges for the moderators. Any tech platform considering outsourcing their content moderation to a third-party provider must consider the ethical and moral impact of this decision.

4.2 Oversight mechanisms

Oversight mechanisms are necessary for the correction of errors, as this report has shown. They are also essential for a variety of other reasons. These include compliance with Article 10 of the TCO Regulation, as well as ensuring accountability, improving the quality of platforms' decision

103 Gorwa, Binns and Katzenbach (n 40), 12.

104 Thorley and Saltman (n 12).

105 Paul M. Barrett, Who Moderates the Social Media Giants? A Call to End Outsourcing (NYU Stern Center for Business and Human Rights, 2020), 25.

106 ibid.

107 Barnes (n 73).

108 Natasha Bernal, 'Facebook's content moderators are fighting back' (Wired, 11 June 2021) <https://www.wired.co.uk/article/facebook-content-moderators-ireland> accessed 22 October 2023.

109 Niamh Rowe, '"It's destroyed me completely": Kenyan moderators decry toll of training of AI models' (The Guardian, 2 August 2023) <https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai> accessed 22 October 2023.

making, spreading expertise across the sector, improving policymaking, building trust and legitimacy, increasing public understanding, and promoting multistakeholder collaboration.[110] Three specific mechanisms should be highlighted.

4.2.1 Appeal processes

The inevitability of erroneous content moderation decisions means that processes should be put in place for users to be able to appeal against the removal of their content. These processes should adhere to standards of due process, out of respect for the autonomy of users and to ensure that the opportunity to appeal is effective.[111] According to the Santa Clara Principles,[112] users should be notified when their content is removed, informed of the reason for the removal, and given details of the appeal process. The appeal process should be clear and easily accessible, include a review by a person or panel of people who were not involved in the original decision and who possess the necessary linguistic and cultural understanding, and provide the user with an opportunity to present additional information in support of their appeal.

4.2.2 Transparency mechanisms

In response to sustained calls for greater transparency around (automated) content moderation,[113] tech companies have developed two forms of transparency mechanism in particular. The first is the publication of relevant policies that describe the types of content that are not permitted and define key terms (such as terrorism). This is now mandated by the EU's TCO Regulation and Digital Services Act (DSA).[114] The second is the publication of transparency reports containing statistical data and breakdowns. The TCO Regulation and DSA formalise transparency reporting obligations for all platforms, requiring annual, publicly available, easily comprehensible reports containing information on content moderation policies and practices; any use made of automated means for the purpose of content moderation; and data on, among other things, orders received from authorities in Member States, referral notices and complaints received and responses to these.[115]

4.2.3 Auditing and access to data

Recent legislative initiatives aimed at ensuring online safety have imposed algorithmic auditing requirements. For example, the DSA obliges 'very large' online platforms and search engines to

110 Macdonald and Vaughan (n 9).

111 Stuart Macdonald, Sara Giro Correia and Amy-Louise Watkin, 'Regulating terrorist content on social media: automation and the rule of law' (2019) 15 International Journal of Law in Context 183.

112 A set of principles that were developed by a coalition of civil society organisations and academic experts and endorsed by 12 major companies, including Apple, Meta, Google, Reddit, Twitter and Github. See further 'The Santa Clara Principles on transparency and accountability in Content Moderation' <https://santaclaraprinciples.org> accessed 22 October 2023.

113 See, e.g., Macdonald, Correia and Watkin (n 111).

114 Regulation 2021/784 (n 6), Article 7; Regulation 2022/2065 on a Single Market For Digital Services and amending Directive 2000/31/EC (19 October 2022), Article 14.

115 Regulation 2021/784 (n 6), Article 7; Regulation 2022/2065 (n 114), Article 15.

conduct annual, independent audits, including access to all relevant data and premises.[116] The DSA will also require very large platforms and search engines to provide access to vetted researchers at academic institutions for research that 'contributes to the detection, identification and understanding of systemic risks in the Union' and assesses the 'adequacy, efficiency and impacts' of the companies' risk mitigation measures.[117]

5. Resources

Content moderation requires a significant investment of resource. As developers with the skillset needed to create and implement automated tools are much sought after, they are difficult and expensive to recruit.[118] Acquiring a suitable training dataset for the development of machine learning tools takes time and effort. Human moderation needs to be properly resourced, including provision for reviewers' wellbeing and mental health. Relevant legal frameworks also need to be navigated. All of these demands bring with them an opportunity cost. In a fast-moving and competitive technological landscape, there will be pressure to prioritise product development. Accordingly, this section looks at three ways in which these resources challenges might be addressed.

5.1 Off-the-shelf products

Given the broad range of challenges associated with developing in-house content moderation capabilities, many tech platforms instead purchase this capability from a third-party provider. Outsourcing the responsibility for content moderation has some significant potential benefits, particularly given the cost and resourcing required to provide an equivalently scaled effort in-house, and the complexities and human impact of administering such a service. As a result, off-the-shelf AI and machine learning driven content moderation solutions are growing in popularity and availability, with estimates suggesting that the market could be worth $32 billion by 2031.[119]

However attractive these services may at first appear, there are some important issues that a tech platform should consider before outsourcing its content moderation activities. As explained above, the effectiveness of any system that uses machine learning algorithms is dependent on the quality and relevance of the data on which it has been trained. In the context of a third-party solution, small tech platforms are unlikely to have transparency regarding the datasets that have been used to train the machine learning algorithms, and how in turn these might be reflected in decisions which are biased or discriminatory against their user base.

116 Regulation 2022/2065 (n 114), Article 37. Article 33(1) defines 'very large' as more than 45 million average monthly users of the service in the EU.

117 ibid, Article 40(4).

118 Cambridge Consultants (n 4).

119 'Content Moderation Solutions Market to Cross US$32 Bn by 2031, TMR Report' (Transparency Market Research, 30 March 2022) <https://www.prnewswire.com/news-releases/content-moderation-solutions-market-to-cross-us-32-bn-by-2031-tmr-report-301514155.html> accessed 22 October 2023.

Facebook employs 15,000 content moderators,[120] yet this has not been sufficient to ensure the necessary levels of linguistic and cultural understanding, particularly in the Global South. It is unlikely that most third-party providers would have equivalent human resources to dedicate to content moderation, thus increasing their reliance on automated and machine learning driven decision making, with all the risks associated with this approach. As explained above, this risk disproportionately impacts some (often already marginalised) groups, including human rights activists, journalists and civil society groups who may see their content removed incorrectly, which jeopardises their right to freedom of expression. Yet third-party providers are not currently subject to the same level of oversight as tech platforms themselves, including in respect of human rights compliance.

A further complicating factor for small tech platforms with users in multiple jurisdictions are the different legal frameworks relating to illegal content they might be operating under. National governments and the European Union have levied specific requirements on tech platforms relating to content that is illegal or terrorist in nature (including in relation to the time that the content is permitted to remain online before being removed). There will also be instances where national authorities request that content be removed, or where liaison with relevant agencies may be required to resolve sensitivities regarding online speech by or about a particular group. The practicalities of using a third-party provider to navigate this complex series of potentially conflicting requirements are challenging, particularly when incorrect decisions have legal implications, such as the platform's content moderation being seen to have broken the law or private lawsuits brought by survivors of the families of terrorist attack victims.

5.2 Collaborative initiatives

There are existing collaborative initiatives driven by organisations working within the online counterterrorism space that offer capacity-building and knowledge-sharing services. Two key examples are Tech Against Terrorism and the Global Internet Forum to Counter Terrorism (GIFCT).

5.2.1 Tech Against Terrorism

Tech Against Terrorism is an initiative launched and supported by the United Nations Counter Terrorism Executive Directorate, working with the global tech industry to tackle terrorist use of the internet whilst respecting human rights. The interdisciplinary team consists of counter-terrorism experts and developers, who offer tech companies practical and operational support to help implement effective mechanisms to respond to terrorist use of the internet.

Tech Against Terrorism offers its Knowledge-Sharing Platform (a collection of resources that small platforms can use to improve their content moderation tools) and a capacity-building

120 Barrett (n 105).

programme, via the EU-funded project Tech Against Terrorism Europe.[121] It also maintains the Terrorist Content Analytics Platform (TCAP). The TCAP is a secure and transparent online tool which helps platforms to detect and action verified content produced by designated terrorist entities.[122] It provides real-time, targeted alerting with the context necessary to action removal of terrorist content online. This provides platforms with rapid notice of the presence of terrorist content and enables moderators to take immediate action, reducing the spread of terrorist content. This is particularly important within the context of the one-hour removal limit mandated by the TCO regulation.

The TCAP inclusion policy encompasses the official content of designated terrorist organisations (using an aggregated designation list comprised of those maintained by the United Nations, European Union, UK, US, Australia, Canada and New Zealand), as well as unofficial content that expresses support for these organisations, or which glorifies acts of terrorism or provides instruction for terrorist purposes.[123] All content is verified by a terrorism specialist prior to submission to the TCAP. Subscribed platforms receive alerts via email or can log into the TCAP to see a dashboard of URLs relating to their own services.

From November 2020 to November 2023, the TCAP identified and verified more than 49,000 unique URLs containing terrorist content and sent 29,500 alerts to 125 platforms. This content related to 34 separate terrorist entities.[124] More than 80% of the alerts issued by the TCAP have resulted in the removal of the identified content.[125]

5.2.2 The Global Internet Forum to Counter Terrorism (GIFCT)

Founded by Facebook, Twitter, YouTube, and Microsoft in 2017, GIFCT is an NGO with a current total of 26 members.[126] Its activities include the development of cross-platform technical solutions. Its leading initiative is its hash-sharing database. When a GIFCT member company removes an item of terrorist content (video, image or PDF), it can create a perceptual hash and add it to the shared database.[127] In the event that a user attempts to upload that same item to the platform of another GIFCT member company, the item will automatically be flagged for review. This prevents terrorists jumping from one platform to another and does so without user data being shared

121 See <https://tate.techagainstterrorism.org> accessed 20 November 2023.

122 More information on the TCAP is available at: https://www.terrorismanalytics.org/

123 'Inclusion policy' (Terrorist Content Analytics Platform) <https://terrorismanalytics.org/policies/inclusion-policy> accessed 20 November 2023.

124 'Home' (Terrorist Content Analytics Platform) <https://www.terrorismanalytics.org/> accessed 20 November 2023.

125 Transparency Report: Terrorist Content Analytics Platform – Year Two: 1 December 2021 – 30 November 2022 (Tech Against Terrorism, 2023) <https://techagainstterrorism.org/hubfs/Tech-Against-Terrorism-TCAP-Transparency-Report-2021-2022.pdf> accessed 30 October 2023.

126 'Membership' (Global Internet Forum to Counter Terrorism) <https://gifct.org/membership> accessed 20 October 2023.

127 'GIFCT's Hash-Sharing Database' (Global Internet Forum to Counter Terrorism) <https://gifct.org/hsdb> accessed 20 October 2023.

between companies. There are currently 2.1 million hashes in the database, relating to approximately 370,000 unique items of content.[128]

To access the database, platforms must fulfil the criteria for GIFCT membership.[129] The mentorship programme offered by Tech Against Terrorism provides support for companies seeking to fulfil these criteria. The programme assists with the development of the necessary processes, policies, and enforcement mechanisms, and with ensuring that these are adequately future-proofed.[130] GIFCT is also currently investing in making integration with the hash-sharing database easier.[131]

5.3 Future development

Although this report has highlighted the limits of machine learning content moderation algorithms, it is important to emphasise the possibility of developing AI tools to help ease some of these limitations. For example, there are indications that divergences in how extremists and other users use specific terms and expressions can be used to improve the detection of terrorist content.[132] More generally, generative AI approaches (which will be subject to transparency requirements under the EU's Artificial Intelligence Act) can be used to create new items of content in order to supplement existing items when compiling a training dataset. Techniques such as style transfer may also have value in correcting bias in datasets by generating content of an under-represented minority.[133] OpenAI has also announced that it is now offering GPT-4 users the ability to create their own 'AI-assisted moderation system', which they claim can identify inconsistencies in content moderation policies more quickly than equivalent, conventional processes.[134] Platforms such as Discord and Snap have already announced that they are integrating GPT into their content moderation efforts.[135] There are also ways in which AI can be used to reduce the harmful effects on human moderators, such as varying the level and type of harmful content that they are exposed to; identifying and blurring out areas of images, so that the moderator only views them

128 Global Internet Forum to Counter Terrorism, 2022 GIFCT Transparency Report (GIFCT, 2022).

129 'Resource Guide' (Global Internet Forum to Counter Terrorism) <https://gifct.org/resource-guide/> accessed 20 October 2023.

130 'Tech Against Terrorism Mentorship Programme' (Tech Against Terrorism Knowledge-Sharing Platform) <https://ksp.techagainstterrorism.org/knowledgebase/tech-against-terrorism-mentorship-programme> accessed 20 October 2023.

131 Tom Thorley, 'Advances in Hashing for Counterterrorism' (Global Internet Forum to Counter Terrorism, 29 March 2023) <https://gifct.org/2023/03/29/advances-in-hashing-for-counterterrorism> accessed 20 October 2023.

132 Fernandez and Alani (n 10).

133 Cambridge Consultants (n 4).

134 'Using GPT-4 for content moderation' (Open AI, 15 August 2023) <https://openai.com/blog/using-gpt-4-for-content-moderation> accessed 22 October 2023.

135 Johan Moreno, 'Discord Adds OpenAI-Powered Chatbot, Moderation Features' (Forbes, 10 March 2023) <https://www.forbes.com/sites/johanmoreno/2023/03/10/discord-adds-openai-powered-chatbot-moderation-features/?sh=25f48f33d6bf> accessed 20 November 2023; 'Early Learnings from My AI and New Safety Enhancements' (Snap Privacy and Safety Hub, 4 April 2023) <https://values.snap.com/en-GB/news/early-learnings-from-my-ai-and-new-safety-enhancements> accessed 20 November 2023.

if doing so is necessary to arrive at a moderation decision; and using visual question answering to enable moderators to reach a decision by asking the system questions about the content without viewing it directly.[136]

There are also examples of the largest tech companies developing automated tools for use by other platforms. For example, in 2019 Facebook open-sourced two image-matching technologies,[137] and in 2022 made openly available the Hasher-Matcher-Actioner tool. This built on the earlier software and can be plugged into databases such as GIFCT's hash-sharing database.[138] Microsoft has also announced a collaboration with Tech Against Terrorism, to pilot an AI-powered detection tool. If successful, the tool will be made available to smaller platforms and not-for-profit organisations.[139]

Finally, it is important to reiterate that this report has focused on content-based approaches to identifying terrorist content online. Behaviour-based cues can also be used to detect such content.[140] Moreover, these two types of approach are not mutually exclusive. Early work on combining NLP tools with behavioural signals has yielded some promising results.[141]

6. Conclusion and recommendations

Today, the vast majority of terrorist content on the biggest platforms is identified and removed using automated tools. In broad terms, content-based tools employ either a matching- or a classification-based approach. The second of these categories makes use of different forms of AI, including machine learning. But while these technologies continue to develop – and provide benefits across a variety of different fields – it is important to recognise that automated content moderation 'is not a panacea for the ills of social media'.[142] As important as automation is, given the sheer volume of (terrorist) content posted online, both matching-based and classification-based tools have their limitations. Their use must be supplemented by human input, with appropriate oversight mechanisms in place. This is challenging, in terms of resource, for platforms of all sizes. Accordingly, this report has also considered some of the ways in which these resource issues might be addressed, including ways in which AI might helpfully be further developed.

136 Cambridge Consultants (n 4).

137 Davis and Rosen (n 51).

138 Nick Clegg, 'Meta Launches New Content Moderation Tool as It Takes Chair of Counter-Terrorism NGO' (Meta Newsroom, 13 December 2022) <https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/> accessed 22 November 2023.

139 'Microsoft and Tech Against Terrorism Partner to Leverage Latest AI Technologies to Improve Terrorist and Violent Extremist Content Detection' (Tech Against Terrorism, 10 November 2023) <https://techagainstterrorism.org/news/microsoft-and-tech-against-terrorism-ai-detection-tool> accessed 23 November 2023.

140 Stuart Macdonald, Connor Rees and Joost S., Remove, Impede, Disrupt, Redirect: Understanding & Combating Pro-Islamic State Use of File-Sharing Platforms (Resolve Network, 2022).

141 Thorley and Saltman (n 12).

142 Gillespie (n 3), 2.

On the basis of this discussion, the report offers the following three recommendations.

1. Augmented intelligence in content moderation – that is, 'human-in-the-loop' practices – remain the highest standard of practice for moderating terrorist content whilst protecting freedom of expression. ***(a) We recommend the development of a set of minimum standards for those employing content moderators. This should include examples of best practice, including provision for moderators' wellbeing.*** The growing capacities of AI for a wide variety of purposes can be a benefit in both reducing resource cost and strengthening supports for human-in-the-loop content moderation, particularly moderator mental health and well-being. ***(b) We recommend the promotion of further development of AI tools for safeguarding the wellbeing of content moderators.***

2. Options available for AI content moderation by small to medium sized platforms are limited compared to those available for large platforms. In addition, collaborative initiatives provide an opportunity for small platforms to boost capacity. Product offerings paired with collaborative models across industry actors likely offer the best possible access to tools while making content moderation best practices feasible for most operators. ***We recommend that: (a) small platforms should assess any off-the-shelf offering carefully; (b) such platforms should also explore the opportunity to make use of Tech Against Terrorism Europe's capacity-building programme, the Knowledge-Sharing Platform and Terrorist Content Analytics Platform offered by Tech Against Terrorism, and the hash-sharing database maintained by GIFCT; and (c) where GIFCT membership is not possible, small platforms should seek other potential forms of collaboration to bolster their content moderation resources.***

3. To support broader industry capacity for effective and high standard AI content moderation in relation to terrorist content online, more collaboration is needed across the various industry operators. Knowledge-sharing is essential to boost overall capacity and enable robust responses, especially when world events or terrorist acts spike flows of terrorist content, glorification of violence, and recruitment activities online. ***(a) We recommend that international organisations and governments support the development of openly available automated content moderation tools by NGOs, and that the largest tech platforms develop automated content moderation tools and make them openly available to other platforms.*** There are existing examples of such collaboration. ***(b) Such tools should be accompanied by a good practice guide that explains how the tool works, its limitations, and how it can be integrated into a platform.*** In addition to sharing tools, larger platforms have already demonstrated commitments to collaborative models of engagement across the industry. As such, ***(c) we also recommend that large platforms develop multiple models for collaboration, taking into account both their need to vet partners and protect IP whilst also enabling increased access to tools and collaboration for small to medium platforms.***

7. About Tech Against Terrorism Europe

Tech Against Terrorism Europe (TATE) supports tech companies in complying with the EU's Terrorist Content Online (TCO) regulation and counter the terrorist threat. This project works with hosting service providers to ensure understanding and compliance of the EU's TCO Regulation. TATE is a consortium of partners from academia and civil society. Consortium partners are: Dublin City University (DCU), Ghent University, The JOS Project, LMU Munich, Saher Europe, Swansea University and Tech Against Terrorism. The TATE project is funded by the European Union.

https://tate.techagainstterrorism.org/


tech
against
terrorism
europe