Global Network
on Extremism & Technology

Artificial Intelligence and Countering Violent Extremism: A Primer

Marie Schroeter

GNET is a special project delivered by the International Centre for the Study of Radicalisation, King's College London.

The author of this report is Marie Schroeter, Mercator Fellow on New Technology in International Relations: Potentials and Limitations of Artifical Intelligence to Prevent Violent Extremism Online

The Global Network on Extremism and Technology (GNET) is an academic research initiative backed by the Global Internet Forum to Counter Terrorism (GIFCT), an independent but industry-funded initiative for better understanding, and counteracting, terrorist use of technology. GNET is convened and led by the International Centre for the Study of Radicalisation (ICSR), an academic research centre based within the Department of War Studies at King's College London. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing those, either expressed or implied, of GIFCT, GNET or ICSR.

We would like to thank Tech Against Terrorism for their support with this report.

CONTACT DETAILS

For questions, queries and additional copies of this report, please contact:

ICSR
King's College London
Strand
London WC2R 2LS
United Kingdom

T. **+44 20 7848 2098**
E. **mail@gnet-research.org**

Twitter: **@GNET_research**

Like all other GNET publications, this report can be downloaded free of charge from the GNET website at www.gnet-research.org.

Executive Summary

Radicalisation can take place offline as well as online. To what extent the internet plays a role remains contested. Doubtless, there are radical and extreme communities online. This report looked into the ability of artificial intelligence (AI) applications to contribute to countering radicalisation. Mapping the possibilities and limitations of this technology in its various forms, the report aims to support decision-makers and experts navigate the noise, leading to informed decisions unswayed by the current hype. The following findings were the most striking:

1. Manipulated search engines and recommendation systems can contribute to counter-radicalisation by pointing to moderate content

Search engines and recommendation systems have great potential for helping to make online spaces safer by contributing to the prevention of violent extremism through lowering the chances of encountering radicalising content. Search engines help to navigate the jungle of information online, including extremist content. Manipulated algorithms could point to moderate rather than extremist content. Equally, recommendation systems that suggest the next video, song or movie based on browsing history can potentially reinforce extreme viewpoints by recommending confirmatory content. A balanced recommendation system would counter malicious narratives with opposing content or spread information about projects and contact points on the prevention and countering of violent extremism.

2. Natural language processing can help to translate minority languages for better content moderation and support content moderation of niche websites in the long run

Natural language processing (NLP) offers potential for content moderation online, especially with respect to languages spoken only by small groups of people. Often moderation of content in minority languages does not seem profitable enough for investment. Smaller platforms do not always have the technical expertise or resources for content moderation systems, as even employing existing models requires significant time and effort. Other support what might be considered an extreme interpretation of the value of free speech and hence do not want to limit users. Improved NLP can help to translate content into languages in which large numbers of experienced and trained moderators operate. NLP can also detect unusual semantic patterns on websites. This could be helpful to support detecting critical messages on platforms that do not want to or cannot invest in content moderation. However, such measures must respect privacy standards and human rights at all times.

3. Tackling disinformation and manipulated content online lacks automated solutions

To date there are no compelling automated tools that identify and tackle disinformation and manipulated content, which is harmful but legal. The drastically improved digital literacy of users to create digital sovereignty seems to be a more helpful approach in the short-term.

4. Superhuman AI is not going to 'sound an alarm' if individual radicalises online

A general AI that monitors, with superhuman intelligence, the content and behaviour of individuals online and 'sounds an alarm' when indicators for radicalisation are coming together is not feasible and will stay in the realms of science fiction for two reasons. First, there is not enough data to feed an algorithm with definite information on radicalisation and when a radicalised individual turns to violence. Unless there is a technical innovation that allows the creation of reliable systems with far less data, there is no incentive to use technical help as they would not be able to make reliable predictions without enough data on previous cases. Radicalisation and terrorism is – luckily – too rare and diverse to produce enough information for an algorithm. Secondly, predicting the behaviour of individuals would require clearly assignable data on individuals, which would give away every aspect of privacy and potentially result in surveillance on an unprecedented scale. The above described scenario is not compatible with liberal democracies having the right to privacy at their core.

Contents

1 Introduction

According to popular belief, artificial intelligence (AI) will revolutionise everything, including national security. To what effect the internet facilitates radicalisation remains an unanswered question, but the latest terror attacks, in Halle in eastern Germany, Christchurch in New Zealand and at Poway synagogue in California, are just three recent examples of the online sphere playing a significant role in radicalisation today.

How can AI help to counter radicalisation online? Expertise on the matter is divided into different disciplines but can be found among researchers and experts from security and counterterrorism backgrounds, as well as policymakers and tech-experts, who increasingly come together to investigate this domain. Currently, the existing landscape of information makes it difficult for decision-makers to filter real information from the noise. This report wants to shed light on the latest developments in AI and put them in the context of counter-radicalisation efforts in liberal democracies.

This publication contributes to the topic by highlighting some limits and possibilities of AI in counter-radicalisation online. The second chapter briefly explains the key concepts and ideas behind AI. In a 'Deep Dive' at the end of the chapter, special attention is given to the quality of data and bias and manipulation in datasets. The third chapter discusses the potential provided by and limitations of AI-based technological innovations for a 'healthy' online space, free from terrorist content, propaganda material and fake engagement. The assumption is that this healthy online environment contributes to the prevention of radicalisation. The chapter assesses a range of popular AI-based concepts, ranging from Deepfakes to bot armies spreading fake news, and explains why search engines, recommendation systems and, in particular, natural language processing (NLP) have the potential to contribute to this objective in one way or another. The fourth chapter looks solely at a hypothetical 'general AI', the omniscient system that identifies individuals undergoing radicalisation and can consequently help law enforcement to prevent crime before it happens. This chapter also argues, however, that such AI technology will remain solely in the realm of science fiction for the foreseeable future. This leads to a discussion of the reasons behind such a position. Big data debates, especially regarding traditional security, cannot take place in liberal democracies without safeguarding and prioritising privacy. Another 'Deep Dive' in chapter four provides more information for the interested reader. The fifth chapter concludes the report.

The report is based on semi-structured interviews with researchers, policymakers and advisers as well as private sector representatives. Additionally, findings from desk research and media monitoring influenced the positions of this report. I talked to different stakeholders to gain a multi-disciplinary perspective, which considers the fragmented information landscape. However, there are clear limitations to the research as a consequence of information on the use of machine learning that has been either restricted by intelligence services or limited by private sector companies.

2 What is Artificial Intelligence?

Although AI is a very fashionable term today, there is no standard definition shared around the world. This is in part due to the fact that the study of AI is a fast-paced and highly popular subject, constantly producing new findings and blurring the borders between computing, statistics and robotics. Although there is no consensus on definition, it does affect most of our lives. This type of technology recommends our next shopping items online, manages diaries and operates driverless cars.

Alexa, Google's voice assistant, shows how sophisticated automated decision making systems can be: Alexa can conveniently plan an entire date night including booking tickets for a show, reserving a table at a restaurant, ordering an Uber and informing your date about your estimated time of arrival.[1] More generally, the term AI describes a discipline that is concerned with automated and adaptive technological systems. AI performs tasks without constant guidance and is able to improve performance through learning from previous experiences.[2]

It was during a conference in Dartmouth in 1959 that AI was first given its name. Following some initial success, researchers were optimistic that using computer-based algorithms would provide a rapid advancement. In these early stages they were able to write code that could solve problems; programs included elements that improved performance through learning. Limited by the capacity of memory and processors, however, an AI 'winter' followed as investment in such research in the 1960s was frozen and interest waned.

The current AI hype in the 21st century has been made possible by technical developments mainly driven by the private sector. Increasingly cheap solutions for mass storage and software, combined with subject-matter experts and better access to data, allowed the area to flourish. What makes AI so special? First, it makes the analysis of bulk data easier; it is quicker and more efficient than an analysis performed by human operators. Second, such technology is able to work with uncertainty and based on that ability make predictions for the future. Whether these predictions are reliable or not is of secondary concern. It is precisely the ability of algorithms to predict that is their strength. By comparison, the human brain is unable to make decisions based on large datasets, multiple conditions and uncertainty. The power of prediction can be viewed as the signature ability of algorithms.

1 Hao, K. (2019a), 'Inside Amazon's plan for Alexa to run your entire life', MIT Technology Review. Accessed at: https://www.technologyreview.com/s/614676/amazon-alexa-will-run-your-life-data-privacy/
2 Reaktor & University of Helsinki (2018), 'How should we define AI?'. Accessed at: https://course.elementsofai.com/1/1

Focussing on the term 'AI' is in itself misleading. It suggests that there is a similarity to human intelligence or human learning processes. Deep neuronal systems, a special machine-learning technique with several layers of processing units, are indeed inspired by the architecture of the human brain. However, the capabilities of such systems are very different from human neurons. Even in complex cases, the algorithm is able to fill in missing data using prediction models, but it cannot give meaning to its findings. The differences between human and machine intelligence become visible when looking at what is and is not achievable for a neural network. For example, an algorithm can identify breast cancer in its early stages more reliably than humans can. This is done by analysing mammography images with a lower error rate than radiologists.[3] On the other hand, the algorithm cannot understand – as in give meaning to – the emotions of the patient and react accordingly. Empathy requires years of observing as well as emotional intelligence, which is lacking in algorithms. Moreover, the word 'intelligence' in the term AI implies that the system is able to produce original thought, which is definitely far-fetched. Google's AI AlphaGo can easily calculate the most promising moves in the highly complex game Go but that does not mean that AlphaGo understands the game itself.[4] AlphaGo is 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. Attributing meaning and context is a very human thing to do; most children are able to say why playing games is fun. Instead, while the system cannot explain why it does what is does, it is nonetheless capable of identifying the best possible move for a given situation according to the objective its program needs to achieve. It analyses all options and decides mathematically how to minimise risk and therefore deal with uncertainty.

In highlighting the strengths and weaknesses of AI, two categories unfold. AlphaGo might be considered a 'narrow' AI as it handles one task, whereas a 'general' AI would be able to deal with any intellectual task (such an AI currently only exists in science fiction). The intelligence of a system can be weak or strong, which translates into narrow or general AI respectively. The term 'narrow AI' applies to a system that pretends to be intelligent by producing the desired results. The intelligence can be shallow and often based on false structures: for example, algorithms trained to identify trains in pictures would not be able to point out a train itself, but rather recognise the often-occurring parallel tracks that feature on pictures of trains. The algorithm was able to rely on false structures in its neuronal networks because they produced the desired outcome.[5] This obviously carries a risk: it is not yet fully clear what the unintended consequences could be. A known consequence for example is that face recognition systems are weak at recognising people of colour.[6] A general AI would have a genuine mind, consciousness or super intelligence, through which popular media might refer to it. Again, superintelligent systems are so far only present in science fiction.

3 Hao, K. (2020), 'Google's AI breast cancer screening tool is learning to generalize across countries', MIT Technology Review. Accessed at: https://www.technologyreview.com/f/615004/googles-ai-breast-cancer-screening-tool-is-learning-to-generalize-across-countries/

4 Gibney, E. (2017), 'Self-taught AI is best yet at strategy game Go', Nature. Accessed at: https://www.nature.com/news/self-taught-ai-is-best-yet-at-strategy-game-go-1.22858

5 Thesing, L. et al. (2019), 'What do AI algorithms actually learn? – On false structures in deep learning', Arxiv. Accessed at: https://arxiv.org/abs/1906.01478

6 Simonite, T. (2019), 'The best Algorithms Struggle to Recognize Black Faces Equally', Wired. Accessed at: https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/

The meaning of the term AI changed over time. Nowadays, AI and machine learning (ML) are used interchangeably by many news outlets; this report does so also. In general there are two popular areas of ML: supervised and unsupervised. Supervised ML means the algorithm is trained to analyse data based on a training dataset with previously labelled data. Labelling can be understood as the decision between 'data fits condition' or 'data does not fit condition'. Labelled data therefore gives meaning to data points, which requires human input. For example, a picture might show an apple. If it does, it would be labelled 'apple'. To train an algorithm, it is necessary to have access to a large quantity of clearly labelled data. With the help of a test dataset, the performance of the algorithm can be evaluated and, if it is deemed successful, the algorithm can then apply labels onto new data. The advantage of supervised learning is that the algorithm organises the data exactly as programmed. Labelling data by hand is a labour-intense and expensive undertaking. Many internet users have labelled data themselves, as some websites pose 'I am not a robot' security questions. Such questions might, for example, require the user to mark all images within a set of pictures that show cars. At this very point, the user labels data. Google's reCaptcha, which offers such a validation service, uses these results to train ML datasets.[7] For instance, the labelled data could be useful for driverless cars.

There is also unsupervised ML. Here the ML algorithm has to find patterns in unlabelled, unorganised data without being trained how input data corresponds to output data. Unsupervised algorithms find patterns within datasets without having any further instructions or classifications. The challenge is that it is not clear whether the patterns that the algorithm will find are in any way useful to the algorithm's programmers. Findings might not be relevant at all. Nevertheless, it saves labour-intense labelling of data and makes use of the vast amount of unlabelled data that is openly available. Unsupervised learning can be used as a first step before working with a supervised learning algorithm.

Reports and news often speak about deep learning. This is a special ML technique where several processing units are connected in a network; the scale of the network allows for the analysis of more complex problems. Likewise, neural networks often make headlines. These are inspired by the human brain structure and allow for storing and processing information simultaneously. Such networks have played a large role in the breakthrough of Big Data, as they enable the processing of large amounts of data.

7 Google ReCaptcha (n.d.). Accessed at: https://www.google.com/recaptcha/intro/v3.html

Deep Dive

Data – Quality, Bias and Manipulation

There is hardly a meeting or conference on AI where somebody doesn't employ the catchphrase 'Data is the new oil'. But is it actually true? Certainly, ML needs a lot of data and data is expensive to acquire. Also, the more algorithms improve, the more data they have. However, not every data point is of the same value, as the rule of diminishing returns applies. An algorithm learns more from early data than from its millionth repetition. Equally, algorithms work especially well if their training data prepared them for lots of eventualities – a dataset that includes both usual and unusual elements. So, data is oil's equal in terms of high prices, high demand (many new business models depend upon data as they use ML) and, at the moment, a relatively small number of holders of the commodity. However, whereas every drop of oil contributes to total oil output, data points have differing influences on the value of data as a whole.

Algorithms are vulnerable to bias, a systematic distortion of reality through the employment of data samples. A bias in input will inevitably lead to a bias in output and, if reinforced, multiply in effect. Bias in current datasets has been proven numerous times, especially against women and people of colour. For example, Amazon had to remove its automated human resources tool, as it discriminated against women. The algorithm rated men better than women as it had been trained with application documents from the last decade. It is not a secret that the tech sector is heavily male dominated, a bias that was reproduced in the algorithm's decisions.[8]

Good, unbiased data is an imperative for an algorithm. Following a consultation process, the United Nations founded the organisation Tech Against Terrorism, which set up the Terrorist Content Analytics Platform (TCAP) to create such a dataset, which could then be consulted by the private sector, academia and civil society. Originally, Tech Against Terrorism intended only to include content from al-Qaeda and Islamic State (IS), but it later announced that it would extend TCAP to include far-right terrorism as well. Other areas deserve attention as well, such as terrorism fuelled by a misogynist ideology, as seen in the Incel movement. Obviously, such datasets have to comply with data security standards and take into account the potential mental health implications for reviewers.

However, data can be manipulated. A dataset that has been tampered with is hard to detect from the outside, especially for non-technical experts. Data does not necessarily need a lot of changes to manipulate an algorithm, as Chinese researchers proved. Their experiments resulted in a self-driving car opting to drive on the opposite side of the road. This illustrates the vulnerability of the systems.[9] For AI applications designed to prevent online radicalisation, a manipulation could create the opposite outcome. One can imagine, for instance, a tricked content moderation system that decides to take down profiles of the opposing political party in the run-up to an election. Another weakness stems from the adversarial training of ML systems: two systems compete with each other, thereby improving mutual performance. For example, one ML system creates fake faces and the other has to filter them out from a set of real ones. Even as the filtering system improves, the face-faking system also develops in skill too. It is not sure yet what consequences this may have.

8 Dastin, J. (2018), 'Amazon scraps secret AI recruiting tool that showed bias against women', Reuters Technology News. Accessed at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

9 Knight, W. (2019), 'Military artificial intelligence can be easily and dangerously fooled', MIT Technology Review. Accessed at: https://www-technologyreview-com.cdn.ampproject.org/c/s/www.technologyreview.com/s/614497/military-artificial-intelligence-can-be-easily-and-dangerously-fooled/amp/

3 AI Countering Radicalisation Online – the Devil is in the Details

How can statistics and automated decision-making help to counter radicalisation online? As always, the devil is in the details and the noise around AI is deafening. That makes it difficult for non-technical experts to understand how much substance there actually is. This chapter looks in depth at the possibilities and limitations of popular technological innovations that are based on ML. It focuses on elements that dominate public conversation. From deepfakes to automated content moderation, search engines and natural language processing, this chapter helps to assess the technology regarding its uses for counter-radicalisation. There are overlaps between different elements, which are unavoidable in such a rapidly changing environment with frequent new developments.

3.1 Shaping the User's Online Experience – What is Visible and Easy to Find

Machine learning can have a significant impact on the user experience online, shaping what users easily find and see. The algorithms in their various forms have great potential to counter radicalisation by contributing to a healthier online space, preventing malicious content. This report looks at search engines, recommendation systems and automated content moderation.

Search engines ultimately help users to navigate the millions of websites and find relevant online content. Much like a 21st-century telephone book, search engines point in the right direction within the mass of information and data online. The algorithms underlying search engines are the linchpin for success. Ten years ago the landscape of search engines was much more diverse, but ultimately Google's secret recipe made it the frontrunner, building trust in its tools by delivering relevant content. Today it serves billions of search requests each day with 15% of its daily requests being new queries. Their user-friendly algorithms not only find the sought-after information but also recognise spelling mistakes and auto-suggest the next word in the search bar. Ultimately, the programming of the algorithm decides what information to present. The accessibility of bomb manuals online reportedly led directly to terrorist activities, as illustrated by the case of Noelle Velentzas and Asia Siddiqui who used, an FBI agent observed, IS's *Inspire* magazine, blog posts about homemade explosives and *The Anarchist Cookbook* to create homemade explosives.[10] Britain carried out Operation Cupcake,

10 United States District Court, Eastern District of New York (2015), United States of America vs. Noelle Velentzas and Asia Siddiqui. Complaint and affidavit in support of arrest warrant, 2014R00196. Accessed at: https://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/04/02/velentzas-siddiqui-complaint.pdf

in which MI6 and GCHQ took down a guide for making a homemade bomb from *Inspire* and replaced it with 'The best cupcakes in America' recipes. In summary: whether it is possible to find manuals for makeshift bombs or whether algorithms have been manipulated to make it harder to find extremist content online can be of the utmost importance. This will not completely stop tech-savvy users, but it nonetheless increases the barrier to reaching such content.

Recommendation systems are a convenient tool to find the next clip, song, article or shopping item based on previously consumed or bought items. Recommendation systems can lead to the discovery of new song or make it easy to find the appropriate bedding for a newly bought mattress. However, the algorithms that suggest the next things to look at or listen to can also create **filter bubbles**, which can reinforce assumptions by suggesting similar material. Consequently, this can also lead to the reinforcement of extremist attitudes. The opacity of how algorithms suggest new items on social media and music, video or film websites gives the consumer no preferences on how things might be recommended: more of the same, opposing views or some other combination of both. As research on autosuggested content on bigger social media platforms from 2019 found out, YouTube's algorithms in particular contributed to a reinforcement of extremist views. Once a video with extremist or fringe content has been watched, similar content was recommended.[11] This is especially alarming given that YouTube is the most popular social media website among adults in the USA and it is not unlikely that many users receive their news from the website.[12]

Mainstream social media came repeatedly under fire for not acting decisively enough against terrorists' exploitation of their platforms. The claim was made that intermediaries should take on the responsibility for dealing with such content given social media's role as a quasi-public space where people meet, exchange arguments and conduct business. However, today innumerable **niche services** offer a diversified landscape of online services across the entire spectrum of the internet infrastructure, ranging from messengers with a high degree of anonymity, niche platforms with no capabilities or inclination to monitor content, to hosting services that allow the dissemination of manifestos and live videos. Recently websites and services related to the gaming industry came under fire for not preventing malicious use.[13] Swansea University showed in recent research how IS's network used a range of hosting services for its online newspaper Rumiyah, thereby decentralising the content and increasing the challenges for an efficient and quick removal.[14] More research and meta-data on how niche services are utilised by terrorists is urgently needed.

11 Reed et al (2019), 'Radical Filter Bubbles', in the 2019 GRNTT Series, an Anthology, RUSI, London.
12 Perrin, A. & Anderson, M. (2019), 'Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018', Pew Research Centre. Accessed at: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/
13 Schlegel, L. (2020), 'Points, Ratings and Raiding the Sorcerer's Dungeon: Top-Down and Bottom-Up Gamification of Radicalisation and Extremist Violence'. Accessed at: https://gnet-research.org/2020/02/17/points-rankings-raiding-the-sorcerers-dungeon-top-down-and-bottom-up-gamification-of-radicalization-and-extremist-violence/
14 Macdonald, S. et al (2019), 'Daesh, Twitter and the Social Media Ecosystem', The RUSI Journal, vol. 164, no. 4, pp.60–72.

3.2 Managing User-created Content

Web 2.0 revolutionised online interactivity. It allowed for a move away from static websites to real-time interactions by large numbers of users worldwide. While this global connectivity on an unprecedented scale enhanced and supported many communities, it also created new challenges for counter-radicalisation efforts.

Automated content moderation on social media platforms aims to prevent terrorist content from being spread. Already 98% of the malicious content on Facebook is filtered out by ML algorithms, as stated in the latest report for the EU's self-assessment reports on the practice of disinformation.[15] Users flag the remaining 2%. Twitter reports that it challenges ten accounts per second;[16] Google, which owns YouTube, removes 80% of inappropriate videos before they receive any views, according to its own information.[17] On the face of it, this all sounds like successful moderation and thus it is fair to say that content moderation has improved over recent years. However, the loopholes remain enormous, especially when leaving standard language zones. At the moment only 14 out of Europe's 26 official languages are covered in Facebook's fact-checking language repertoire. Thanks to third-party contracting, 15 African countries are now being monitored, but that is still less than a third of the continent's countries.[18] It remains unclear whether that applies only to official languages or if it is inclusive of dialects as well. Omitting content in languages spoken by minorities for moderation is a known phenomenon from other highly populated and diverse regions and countries, such as India.[19]

As previously discussed, algorithms cannot give meaning to data, which means algorithms fail to understand the context in which malign behaviour takes place. Examples from Sri Lanka show how Facebook's algorithms were not able to assess multilayered cultural context. Prior to the Easter bombings in Colombo in April 2019, posts slipped through because algorithms were not able to understand the complexity of the hate speech they contained. Even after multiple efforts to report the hate speech, which fed into a polarised, anti-Muslim sentiment, Facebook failed to take the content down or respond with adequate content moderation.[20] To classify the slang language used as hate speech, any algorithms involved in content moderation would have had to have been able to understand the ethnicities of the parties. The problem goes even further: often languages that do not use the Latin alphabet are 'translated' into it for convenience. Some languages have no grammatical explicit future tense, so what would a threat, which implies the future, even look like? If automated filtering is to be scaled up, it has to face these design failures.

15 Facebook (2019), Facebook report on the implementation of the code of practice for disinformation. Accessed at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62681

16 Twitter (2019), Twitter Progress Report: Code of Practice against Disinformation. Accessed at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62682

17 Google (2019), EC EU Code of Practice on Disinformation. Accessed at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62680

18 Africa Times (2019), 'Facebook expands fact-checking to 15 African nations'. Accessed at: https://africatimes.com/2019/10/10/facebook-expands-fact-checking-to-15-african-nations/

19 Perrigo, B. (2019), 'Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch', *Time*. Accessed at: https://time.com/5739688/facebook-hate-speech-languages/

20 Wijeratne, Y. (2019a), 'The Social Media Block isn't helping Sri Lanka', *Slate*. Accessed at: https://slate.com/technology/2019/04/sri-lanka-social-media-block-disinformation.html and Wijeratne, Y. (2019b), 'Big Tech is as Monolingual as Americans', Foreign Policy. Accessed at: https://foreignpolicy.com/2019/05/07/big-tech-is-as-monolingual-as-americans/

Many social media companies are keen to prevent their platforms from being exploited by malicious actors. The effectiveness of algorithms to detect propaganda material or terrorist activity also depends upon the quality and availability of data with which the algorithm has been trained. Facebook admitted that a lack of training data was responsible for its failure to identify and filter out livestreams showing shootings, such as the Christchurch attack, which was broadcast live. It now uses footage from British police officers' body cams during terrorist training exercises.[21]

Another ML-based innovation that could help to manage user-created content and scale up content moderation is **natural language processing** (NLP). This describes technical procedures and tools to analyse language and process it. The application of NLP is manifold: from customer support chat bots, dictation software, automated translation to talking to Siri, NLP can be found everywhere. In particular, language translation has shown how much technology has progressed in recent years. In the past it felt very hit-and-miss whether an online translation would be successful, but these days such translations work with a far higher degree of reliability. Nonetheless, automated translation is not flawless yet or ready to take over from interpreters. Among translators, Google Translate was in the news when a screenshot was shared that translated 'I am smart' and 'I am beautiful' from English into Spanish and French in the masculine. For the sentence 'I am beautiful and not smart' it used the feminine.[22] Those flaws must be fixed and research and development is underway. A new technique called masking, created by the Chinese company Baidu, allows a translation program to go beyond word by word translation and actually take into account context, thereby producing more robust results.[23] This could be helpful amid the latest reports on the right-wing extremist Boogaloo movement, which is reported to use coded language online in order to avoid automated take-downs on social media platforms.[24]

The technology has great potential to support content moderation on websites with an extreme understanding of free speech. Well-known examples are 8chan, 4chan and Gab for right-wing extremist ideology and many other forms of group-related misanthropy such as anti-Semitism, xenophobia and white supremacy. The 'no policy' approach can facilitate radical environments as it is allowed to share everything but illegal content, such as child pornography, under United States legislation. The shooters in the synagogue in Poway, the Walmart in El Paso and the mosque in Christchurch in 2019 all posted on 8chan before committing their terrorist attacks. More fine-grained research is necessary, but these final posts stand out in the midst of the usual banter, irony and extremely offensive language on those websites. They all refer to other attacks and share a link to a manifesto, livestream or other writing; the shooters mention that they could potentially die. The posts have an affectionate tone. It is conceivable

21 Manthorpe, R. (2019), 'Police share "shooting" video with Facebook to help identify live-streamed attacks', *SkyNews*. Accessed at: https://news.sky.com/story/police-share-shooting-video-with-facebook-to-help-identify-live-streamed-attacks-11843511

22 'Marta Ziosi', LinkedIn, Accessed at: https://www.linkedin.com/posts/marta-ziosi-3342007a_googletranslate-googletranslate-women-activity-6603598322009808896-MQJX

23 Baidu Research (2019), 'Baidu's Pre-training Model ERNIE Achieves New NLP Benchmark Record'. Accessed at: http://research.baidu.com/Blog/index-view?id=128

24 Owen, T. (2020), 'The Boogaloo Bois are all over Facebook', *Vice*. Accessed at: https://www.vice.com/en_us/article/7kpm4x/the-boogaloo-bois-are-all-over-facebook

that a NLP could identify rising threat levels if certain indicators in a post are fulfilled. Such indicators would need to be adjusted to the specifics of the platform.

NLP could also create resilience in online communities using minority languages by providing better content moderation. Automated content moderation for languages from minorities fails to produce reliable results. Instead of hoping that algorithms will soon perform better even though the economic incentive is too small to compel companies to invest in improvements, the solution could lie with NLP, as better translation into languages spoken by experienced and trained moderators could be the answer. Content moderation could oversee other, less well-covered languages if the automated translation is of an acceptable standard.[25] However, potential applications must always respect privacy standards and comply with human rights.

3.3 Fabricated Content Driven by AI – How to Turn the Tables

Manipulated content has the potential to allow extremist thought to leak into mainstream discourse and can facilitate radicalisation and lead to violence in the real world. Political disinformation is not a new strategy but the possibility to reach unprecedentedly large audiences at the press of a button to steer public debates poses new challenges. This chapter looks at ways to counter trolls, bots, fake news and deep fakes.

Trolls or **bots** are social media accounts that spread specific content or produce artificial engagement on social media platforms. 'These bots can be programmed to perform tasks normally associated with human interaction, including follow users, favour tweets, direct message (DM) other users and, most importantly, they can tweet content, and retweet anything posted by a specific set of users or featuring a specific hashtag.'[26] Programming a bot does not require sophisticated technical knowledge and can be done easily, with the help of readily available manuals online.[27]

Trolls used in large numbers are known as a troll or bot network or army. Manipulated content orchestrated through many messengers can influence public discourse or attitudes according to one's own agenda. A well-known example is the Russian interference in the American election of 2016, where strategically placed fake engagement supported Donald Trump's campaign and attacked Hillary Clinton. Estimates suggest between 5% and 15% of online accounts are fake (these numbers are contested).[28] According to a study from Pew Research, the five hundred most active bots on Twitter are responsible for 22% of tweeted links whereas the five hundred most active humans only account for around 6%. Meanwhile, 66% of the accounts sharing links to the most popular

25 Wijeratne, Y. (2019b).
26 Symantec Security Response (2018), 'How to Spot a Twitter Bot', Symantec Blogs/Election Security. Accessed at: https://www.symantec.com/blogs/election-security/spot-twitter-bot
27 Agarwal, A. (2017), 'How to write a Twitter Bot in 5 Minutes', Digital Inspiration. Accessed at: https://www.labnol.org/internet/write-twitter-bot/27902/
28 Burns, J. (2018), 'How many Social Media Users are Real People?', *Gizmodo*. Accessed at: https://gizmodo.com/how-many-social-media-users-are-real-people-1826447042

websites are bots and only 34% are human.[29] The need for troll farm regulation has been shown lately by the investigative research by Katarzyna Pruszkiewicz,[30] who worked for six months in a Polish troll company. She and her colleagues steered online conversations in favour of paying customers, among them public broadcasters. It is unclear how far online engagement really translates into votes,[31] but it is unacceptable for politicians and state institutions in democracies to win arguments with money.

Mark Zuckerberg, CEO of Facebook, named AI as the solution to content moderation, which includes the identification and removal of fake engagement of every kind. Only automated systems can process the content of million of users in different languages and with diverse cultural backgrounds, according to Zuckerberg.[32] However, details remain unclear. Zuckerberg also admitted in his hearing in the US Senate in 2018 that AI might be ready in five to ten years to detect nuances in language, but the technical developments are not there yet.[33] Existing technological solutions claim the identification of bots is possible. The assumption is that a bot which is set up for a specific purpose would create and engage on single-issue or very narrow thematic content, as opposed to a human, who would be interested in a broader range of topics. Additional information for analysis includes the date and time of creation of the account.[34] The promise of such technology contrasts with findings from Riga's NATO Centre of Excellence. Its recent investigation, which included Facebook, Twitter, Instagram and YouTube, proved that identification and take-down of fabricated engagement is insufficient.[35] For just €300 researchers were able to buy 3,530 comments, 25,750 likes, 20,000 views and 5,100 followers. The platforms failed to classify inauthentic behaviour or accounts: four weeks after the purchase, four in five items were still online. Even after reporting a sample, 95% of the content was still online three weeks after the websites were notified. Given the determination of actors to spread malicious content via fake accounts or troll armies, platforms have to have a proactive approach to identifying fake accounts to avoid content moderation becoming a game of whack-a-mole.

Fake news or **junk news** contains fabricated content, straightforwardly false information or conspiracy theories, which are not necessarily illegal but definitely harmful. A more differentiated terminology distinguishes disinformation from misinformation. The former is spread with intent, whereas the latter is distributed unintentionally. Both have the potential to leak extremist thought into mainstream discourse, which can facilitate radicalisation and lead to real-world violence. Disinformation can be part of a political strategy and, if

29 Wojcik, S. et al. (2017), 'Bots in the Twittersphere', Pew Research Centre. Accessed at: https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/

30 Davies, C. (2019), 'Undercover reporter reveals life in a Polish troll farm', *The Guardian*. Accessed at: https://www.theguardian.com/world/2019/nov/01/undercover-reporter-reveals-life-in-a-polish-troll-farm

31 Eckert, S. et al. (2019), 'Die Like Fabrik', *Sueddeutsche Zeitung*. Accessed at: https://www.sueddeutsche.de/digital/paidlikes-gekaufte-likes-facebook-instagram-youtube-1.4728833

32 Cao, S. (2019), 'Facebook's AI Chief Explains How Algorithms Are Policing Content – And Whether It Works', *The Observer*. Accessed at: https://observer.com/2019/12/facebook-artificial-intelligence-chief-explain-content-moderation-policy-limitation/

33 Harwell, D. (2018), 'AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how', *The Washington Post*. Accessed at: https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/

34 Gupta, S. (2017), 'A Quick Guide to Identify Twitterbots Using AI', Hackernoon. Accessed at: https://hackernoon.com/a-quick-guide-to-identify-twitterbots-using-ai-c3dc3a7b817f

35 Bay, S. & Fredheim R. (2019), 'Falling Behind: How social media companies are failing to combat inauthentic behaviour online', NATO STRATCOM COE. Accessed at: https://www.stratcomcoe.org/how-social-media-companies-are-failing-combat-inauthentic-behaviour-online

spread effectively (perhaps via the aforementioned fake accounts and bot armies), influence public discourse. For instance, 'Pizzagate' was a conspiracy theory from the 2016 American election campaign. Following the leak of John Podesta's private emails, then campaign chairman of Democratic presidential candidate Hillary Clinton, opponents spread the news that in the flood of emails one could find a code that connected senior leaders from the Democrats with human trafficking and child sex. Far-right supporters in particular spread the theory on the image boards 4chan and 8chan as well as on Reddit and Twitter during the electoral campaign. A number of restaurants were mentioned to have facilitated the machinations of the alleged paedophiles. The owners and employees of the restaurants received threats, among them death threats. Eventually Edgar Maddison Welch, inspired by the online posts, decided to go to one of the restaurants. He fired three shots. No one was hurt. In interrogation after the shooting he denied that the information was fake news.

Although the content is not illegal, platform providers can penalise violations of self-set community standards. Yet, filtering out fake news can run counter to the business model of social media companies. Polarising, spectacular content increases user engagement and makes people spend more time on their websites, which allows companies to gather more data on customers. This data is the backbone of their financial model as it improves targeted advertising and increases the returns. Nonetheless there are efforts from different actors to tackle fake news. Researchers from the Canadian University of Waterloo developed an AI-based tool that can support fact-checking to an unprecedented extent. By cross-checking statements of an article with other sources, the system offers an indication of whether it is likely to be fake news or not. According to the researchers, the system is correct nine times out of ten.[36] This could be an important step forward in the fight against fake news.

Newsguard, a private sector effort from the tech giant Microsoft, provides an example of ineptitude. Newsguard is an add-on to web browsers that provides an assessment of the credibility of media outlets in the form of ratings. On social media it shows a small label to indicate the trustworthiness of information. This is a bulky solution: the user is required proactively to download the add-on, which then does not assist in assessing specific articles but only offers a general rating of the outlet. This would not have helped in the aforementioned Pizzagate scandal, which was spread through private accounts. Breitbart, a medium spreading white supremacy and far-right ideology, or the Russian propaganda channel RT are both rated with an overall green label, but Newsguard then highlights in the text below its label that these sites have 'significant limitations'. Breitbart also brought Facebook under attack: Facebook launched a 'news' section, displaying stories from verified media outlets. Those media outlets have been identified in collaboration with journalists and adhere to Facebook's internal guidelines against hate speech and click-bait content. Including Breitbart among the outlets sparked protests. Until now, Facebook has defended its decision, using the

36 Grant, M. (2019), 'New tool uses AI to flag fake news for media fact-checkers', *Waterloo News*. Accessed at: https://uwaterloo.ca/news/news/new-tool-uses-ai-flag-fake-news-media-fact-checkers

argument of free speech. Meanwhile, Twitter announced, in the wake of the British general election in 2019, a ban on all political advertisement.[37]

In general, banning extremist content plays into the hypothesis of creating a healthier online space by limiting the chances to encounter potentially radicalising content online. Nevertheless, it can always only be part of the answer to harmful content online, as it does not look at the root causes of the expressed opinion.

Deepfakes are an extreme version of synthetic and manipulated data and give a fundamentally new meaning to 'putting words in someone's mouth'. The latest technological developments allow users to create videos with people's facial expressions and voices. This can result in videos of politicians that look and sound very realistic, but the person might have never said the recorded words. An Indian politician used a deepfake video to cater to the multilingual environment during a recent electoral campaign, which prompted mixed reactions.[38] Organisations warning about deepfakes also circulated their findings, such as the video in which Boris Johnson and his opponent Jeremy Corbyn endorse each other for the December 2019 election. The majority of fabricated data exists in pornography, making women the biggest victims of this new technology. It means women may star in pornographic videos without their knowledge or consent. This technology seems especially harmful in combination with **de-identification** tools. These tools alter pictures and videos in a way that makes it impossible for algorithms to identify the new, slightly altered version as the original face. It would rather identify it as a new item altogether. Users however, would still recognise the original face in the new, altered version. This could challenge a quick and effective take-down. The industry-led Global Internet Forum to Counter Terrorism (GIFCT) created a Hash-Sharing Consortium – a database of digital 'fingerprints' of malicious content, so called hashes.[39] Through collaboration between different companies, the forum wants to increase effectiveness.[40] It is unclear whether the database could withstand the systematic use of de-identification software, particularly as extremist content is dispersed strategically through a range of actors and across platforms.

Realistic deepfakes require expert knowledge – especially when the results are intended to deceive people. The necessary technical knowledge still prevents a quick scale-up in the technology, particularly when there are other, less demanding methods that could achieve the same objective. Also, as Hwang reports, there is the threat of exposure through de-identification tools. Prohibition policies and the danger of public exposure can make deepfakes less

37 The differentiation between political and issue advertisement remains contested and there are no universally accepted defintions. The arising difficulties produced calls to treat all advertisements with one strict standard to increase transparency and allow scrutiny on their impact. For more information see: Frederik J. Zuiderveen Borgesius et al. (2018): Online Political Microtargeting: Promises and Threats for Democracy. Utrecht Law Review, 14 (1). 82-96. Accessed at: https://www.ivir.nl/publicaties/download/UtrechtLawReview.pdf; and Call Universal Advertising Transparency by default (2020). Accessed at: https://epd.eu/wp-content/uploads/2020/09/joint-call-for-universal-ads-transparency.pdf

38 Christopher, N. (2020), 'We've just seen the First Use of Deepfakes in an Indian Election Campaign', *Vice*. Accessed at: https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

39 GIFCT, 'Joint Tech Innovation'. Accessed at: https://www.gifct.org/joint-tech-innovation/

40 Llansó, E. (2019), 'Platforms want centralised Censorship. That should scare you', *Wired*. Accessed at: https://www.wired.com/story/platforms-centralized-censorship/; and Windwehr, S. and York, Jillian (2020), 'One Database to rule them all: The invisible Content Cartel that undermines the freedom of expression online', EFF. Accessed at: https://www.eff.org/deeplinks/2020/08/one-database-rule-them-all-invisible-content-cartel-undermines-freedom-1.

attractive tools to influence campaigns.[41] Twitter used its new labels on manipulated content for the first time on content created by the White House social media director.[42] The policy states that it would flag manipulated videos or photos but not remove them, unless such content threatened someone's physical security.

Getting disinformation and artificial engagement under control requires a stronger digital literacy among users. As research on the dissemination of fake news on Twitter shows, falsehoods spread more quickly and broadly than the truth and this is because of human interaction. Bots add to virality but do not cause the wide distribution of falsehoods. Researchers accounted the emotional response and relative newness of the content for such distribution.[43] The finding clearly shows that there is no alternative to an adequate education that enables users to navigate online spaces with increased resilience.

41 Hwang, T. (2020), 'Deepfakes – A grounded threat assesment', Centre for Security and Emerging Technology.
42 Dent, S. (2020), 'Twitter labels video retweeted by Trump as "manipulated data"', Engadget Online. Accessed at: https://www.engadget.com/2020/03/09/twitter-labels-trump-retweet-manipulated-media/
43 Vosoughi, S. et al. (2018), 'The spread of true and false news online', *Science* vol. 359, no. 6380, pp.1146–51. Accessed at: https://science.sciencemag.org/content/359/6380/1146

4 Predicting Radicalisation Before It Happens – General AI for Law Enforcement

It is easy to imagine the room where the magic happens; let's borrow the image from the science fiction classic *Minority Report*: a big blue touch screen, showing the results of a super-intelligent machine that is supposed to help law enforcement. The result displayed is based on available data on individuals as well as online real-time behaviour. The system's sirens go off in warning as soon as it spots critically increasing risk factors indicating a dangerous level of radicalisation. According to the observed behaviour the system will then send appropriate units to the location. The system allows the police to strike before something happens thanks to the predictive power of the strong AI-based system. This scenario may seem tempting, even if it is exaggerated and reminiscent more of science fiction than reality. Nonetheless, in the nexus between new technologies and security there is appetite for such a developments. This chapter solely focuses on the myth of a super-intelligent general AI to surveil content and the behaviour of individuals online to counter radicalisation.

Predictive policing projects are exploring how AI can help law enforcement agencies in their work. Such projects are ML applications that forecast on future crimes based on statistical correlations to support law enforcement.[44] The effectiveness of such systems is a topic of strong controversy. Kent Police in the UK, for example, discontinued using the American software PredPol to forecast crimes, as the added value was not convincing.[45] The civil rights organisation Big Brother Watch reported an overpolicing of minorities and the re-creation of bias against certain areas in predictive policing projects, since increased patrolling in areas historically more prone to crime also leads to more reporting of crimes and creates a cycle of reinforcing structural bias.[46] The shift towards using indicators for crime prediction involves interpretative reasoning and non-causal forms of risk thinking. The change indicates a move towards emphasising the importance of context in risk analysis and can be seen as a step away from profiling, which has been perceived as unjust and discriminatory in many societies.[47] Racial or ethnic profiling stressed the relationship between law enforcement and pluralistic societies.[48]

44 Moses, B. L. & Chan, J. (2018), 'Algorithmic prediction in policing: assumptions, evaluation, and accountability', *Policing and Society*, vol. 28, no. 7, pp.806–22.

45 Big Brother Watch Submission to the Centre for Data Ethics and Innovations (2019), 'Bias in Algorithmic Decision Making (Crime and Justice)', Big Brother Watch. Accessed at: https://bigbrotherwatch.org.uk/wp-content/uploads/2019/06/Big-Brother-Watch-submission-to-the-Centre-for-Data-Ethics-and-Innovation-Bias-in-Algorithmic-Decision-Making-Crime-and-Justice-June-2019.pdf

46 Ibid.

47 Monaghan, J. & Molnar, A. (2016), 'Radicalisation theories, policing practices, and "the future of terrorism?"', *Critical Studies on Terrorism*, vol. 9, no. 3, pp.393–413.

48 Open Society Foundations (2019), 'Ethnic Profiling: What it is and Why it must end', Accessed at: https://www.opensocietyfoundations.org/explainers/ethnic-profiling-what-it-and-why-it-must-end

The indicator-based approach, however, could seem tempting for counter-radicalisation efforts online. A list of indicators based on online behaviour as well as content consumed could serve as a basis to support the search for individuals currently radicalising. However, a transfer from current predictive policing systems to counter-radicalisation work is hard to imagine, mainly for three reasons.

The first obstacle is that there is a lack of clarity or precise understanding of radicalisation processes.[49] It is also unclear, when precisely a radicalised individual moves on to commit a crime, which would justify intervention. Radicalisation and terrorism luckily do not occur often enough to provide a reliable dataset. Radicalisation is a highly complex and individualised process and although researchers have identified certain elements reoccurring in radicalisation processes,[50] there is not enough information to train an algorithm. Current AI systems need a vast amount of data to develop their predictive power. Unless a technological breakthrough allows AI technology to work with significantly less data, it does not look promising. Today there are no indicators for such a shift.

Current predictive policing systems are running on group-based assumptions in combination with information on location and time of crime-prone areas. This means a mixture of open source information, governmental data and data provided by private companies feed the algorithms to make informed predictions. The underlying assumptions are much more economical and rational-choice-based. Consider a burglary, for example: the criminal, once successful at a certain time and location, is likely to strike again using a similar location and time for the next crime in order to repeat the success. The idea is, that criminals want to minimise risk and maximise success as much as possible. Those assumptions do not necessarily work for radicalisation and terrorism. This is not to say that there is no rational reasoning in choosing terrorism, but it does not play out in the same way as in procurement crime. Contrarily, dying for the right purpose has been an explicit pull-factor in IS's 'You only die once – why not make it martyrdom' propaganda campaign to join the caliphate.[51]

The third reason is the limitation in liberal democracies stemming from the idea that an individual is protected by the state and from the state. As a thought experiment: what would be needed to predict individual behaviour? An algorithm predicting individual behaviour would need more differentiated data than the available group-based information. Namely, it would require non-anonymised data on the behaviour of individuals, the more the merrier, to ensure predictions are reliable. This would require a surveillance of individual behaviour in an unprecedented way: live-coverage of an entire society would be necessary. This is not compatible with existing privacy rights. Nor is it desirable in a free society, for ethical and moral reasons. It would have repercussions on fundamental rights, such as freedom of speech, of the press, of association, of telecommunications secrecy and so on.[52] Essentially, it would create nothing less than a dystopia.

49 see Monaghan & Molnar.
50 see Neumann.
51 Kingsley, P. (2014), 'Who is behind Isis's terrifying online propaganda operation?', *The Guardian*. Accessed at: https://www.theguardian.com/world/2014/jun/23/who-behind-isis-propaganda-operation-iraq
52 Ganor, B. (2019), 'Artifical or Human: A New Era of Counterterrorism Intelligence?', *Studies in Conflict and Terrorism*.

Deep Dive

Democratic AI by Design

Algorithms run on data and their thirst for more data is never ending. They first have to be trained with a huge dataset, then tested with more data just to continue working their way constantly through more information. Not surprisingly, privacy and data protection come to mind, especially in combination with AI and security issues.

Instead of trying to regulate automated decision-making systems to fit the standards of democratic societies, democratic values should be woven into the design of the technology in the first place. Technological development should be 'private by default', meaning to treat user-data according to the highest privacy standards, unless users agree to give away information. This has implications for the content curation on social media platforms and the mass tracking of personal behavioural data. Furthermore, systems have to produce transparent results or provide explanations, which allow human operators to evaluate the assessments of the algorithm and decide upon the credibility of the results. This would be a clear alternative to current 'Blackbox AI', the results of which cannot be explained. Increased transparency for example in the realm of recommendation systems or content curation would also enable public interest and research, thus leading to a better understanding of online radicalisation. Accountability in decision-making processes can only be reached with a transparent and trustworthy AI. Audits of automated decision-making applications would ensure lawful implementation and set incentives for a fair and democratic AI.

5 Conclusion

The aim of the report was to discuss how AI-enabled technologies can support counter-radicalisation efforts online.

AI offers new opportunities for analysing big data and making predictions for the future. There is room for a narrow application of the technology to support the prevention of violent extremism and reduce the risk of encountering radicalising content online. Especially promising AI-based tools are search engines, recommendation systems and NLP. NLP offers the potential for improved content moderation online, especially for languages spoken only by small groups of people. Often, the assumed financial returns for bigger platforms to invest in content moderation of minority languages – especially human content moderators – are not great enough. Smaller platforms do not always have the technical expertise or resources for content moderation systems, as even employing existing models requires significant time and effort. Other platforms supporting an extreme interpretation of free speech and claim that they do not want to limit users. Improved NLP can help to translate content into languages spoken by more trained moderators. Furthermore, it can also detect unusual semantic patterns on websites that do not want to invest in content moderation. However, such measures must always respect privacy standards and human rights.

Content moderation for large social media platforms remains a challenge. The vast number of languages in combination with a colourful range of cultural context is still too overwhelming for algorithms to filter out malicious content. A broad public discourse on so called 'grey-area material' is needed – content that is harmful but legal. A society should agree on a common ground where free speech finds its limitations online. This decision should not be left solely to private companies. AI technology is likewise under-developed and inexpedient to fight artificial engagement online, including trolls, bots and fake news. Such devices are still insufficiently detected. Users must be brought into the loop and educated to ensure responsible behaviour online leading to digital sovereignty; the design of platforms should allow for a transparent 'Notice and Action' system. However, the weight must not only be on the customers' or users' shoulders. The quest for a safer internet must be supported by policies that disincentivise malicious or fabricated content online and forbid business models which prioritise harmful content, as it increases online engagement and hence benefits advertisement revenues. The crux will be in platform operators avoiding taking down too much content and encroaching on free speech while still providing adequate measures to prevent malicious content.

It may have become obvious that a general AI, a system with super-intelligence, is not an option to forecast online radicalisation of individuals for two reasons. The first is technical: given the current status of AI technology, algorithms need vast amounts of data to make useful predictions on the future.

Luckily, radicalisation and terrorism do not occur often enough to produce enough data for a general AI forecasting the behaviour of individuals regarding online radicalisation. The rate of false positives and false negatives would be intolerable. Investment into human resources would be more beneficial. The second reason is around privacy: a system observing real-time online behaviour of individuals, storing the data and analysing it, would not comply with privacy standards in liberal democracies. It could potentially lead to the mass surveillance of society.

In the long term the application of AI-based technologies should follow clear standards. Those standards should protect users from unfair automated decision-making tools, for example through biased datasets or discriminatory curation of content based on gender, sex, religion or other characteristics protected under the Human Rights Act. Outcomes of algorithms need to be transparently reached to allow for accountability of decisions based on the algorithmic calculations. The development of an AI that respects privacy rights and is free of discrimination, as well as comprehensible for the operator, must be the way forward for the technology.

Policy Landscape

This section is authored by Armida van Rij and Lucy Thomas, both Research Associates at the Policy Institute based at King's College London. It provides an overview of the relevant policy landscape for this report.

Introduction

Preventing the glorification of violence and terrorism, the spread of disinformation and other forms extremist content online are challenges shared by policy actors and tech platforms across the globe. *Artificial Intelligence (AI) and Countering Violent Extremism (CVE): A Primer*, a report for the Global Internet Forum to Counter Terrorism (GIFCT), gives a comprehensive overview of the opportunities and challenges posed by AI in the CVE space.

National and international policymakers, as well as technology companies, are under mounting pressure to moderate and remove extremist content more quickly and effectively. In part, this is because real-world harms, tragic in their frequency and scope, have occurred as a result of malicious content online, such as the Christchurch mosque shooting in March 2019, the 2015 Charleston church shooting in the USA, and Canada's Quebec City mosque shooting in 2017. Subsequent efforts have been made by the tech industry and by national and multinational policymakers to remove a range of extremist content produced by Islamic State and violent jihadist groups, as well as white supremacist, misogynistic, anti-Semitic and Islamophobic content.

AI and CVE: A Primer describes existing AI technologies designed to aid, accelerate and make accurate online content moderation and removal. These range from tools that are 'trained' in language to identify and flag harmful content, to technology that detects deepfake videos, as well as the processes of machine learning to build algorithmic identification tools. The report also finds several challenges and barriers to the effective deployment of these technologies. First, recommendation systems based on the same AI technology can lead users down 'rabbit holes' of increasingly harmful content. Secondly, an emphasis on European language content moderation leaves non-dominant content under-recognised and under-moderated. Thirdly, there are at present no effective measures to counter disinformation online, from both technical and ethical points of views. Lastly, a general AI system to monitor real-time exchanges online relies on impossible levels of non-anonymised data, which would endanger privacy and free speech rights.

In this report, we explore how these opportunities and challenges have been addressed in nine key national and supranational policy actors: Canada, France, Japan, Ghana, New Zealand, the UK, the USA, the European Commission and the UN Counter-Terrorism Executive Directorate. We present an overview of these efforts in a case-by-case basis and conclude by presenting policy recommendations.

AI and CVE: Addressing the Challenges and Assessing New Developments

Canada

The Canadian government has cultivated a robust counter-terrorism strategy and its online CVE efforts and initiatives form one part of a broader, holistic CVE policy. Like many governments, its investment in and attention to countering violent extremism online was unfortunately triggered by real-world harm.

In late January 2017, Québécois Alexandre Bissonnette opened fire on the Islamic Cultural Centre of Quebec City, killing six and injuring five. Subsequent investigation found that Bissonnette had been active in far-right and racist online circles prior to the shooting, as well as regularly checking Twitter accounts of conspiracy theorists, white nationalists and alt-right online personalities like Ben Shapiro and Alex Jones of InfoWars.[53]

Unlike the perpetrators of many other high-profile, online-fuelled terrorist attacks, Bissonnette did not post a manifesto or statement of intent online.[54] Nevertheless, the increasing use of terrorist manifestos is a wider trend that AI could help to combat. Far-right manifestos often refer to one another; for instance, admiration for previous or recent attacks or repetition of memes or internet short-hands. AI could help to identify uploads of harmful content, such as far-right manifestos, in order to intervene before offline, real-world attacks occur.

Canada's response to online violent extremism, as laid out in its National Strategy on Countering Radicalization to Violence,[55] is threefold: to develop counter-messaging with civil society, to support CVE research and to partner with international initiatives and tech companies. The third, in particular, has been the space in which Canada has invested in the AI-CVE nexus.

Most significantly, in 2019 Canada commissioned Tech Against Terrorism, an international UN-sponsored initiative that works with the global tech industry, to develop the Terrorist Content Analytics Platform (TCAP),[56] a database that hosts verified terrorist material and content from existing datasets and open sources. It is hoped that the TCAP will act as a real-time alert service for terrorist and violent extremist content on smaller internet platforms: verified malicious content on these platforms will be rapidly shared with and acted upon by their content moderation teams. In the medium and long term,

53 Riga, A. (17 April 2018), 'Quebec Mosque Killer Confided He Wished He Had Shot More People, Court Told', *Montreal Gazette*. Accessed: https://montrealgazette.com/news/local-news/quebec-mosque-shooter-alexandre-bissonnette-trawled-trumps-twitter-feed/. See also: Mahrouse, G. (2018), 'Minimizing and denying racial violence: Insights from the Quebec Mosque shooting', *Canadian Journal of Women and the Law*, vol. 30, no. 3, pp.471–93.

54 For example, Robert Bowers (perpetrator of the 2018 Pittsburgh synagogue shooting), Dylann Roof (2015 Charleston church shooting), Brenton Tarrant (2019 Christchurch mosque shooting), Patrick Crusius (2019 El Paso shooting), Anders Breivik (2011 Utøya massacre), and many others have posted manifestos on a variety of online platforms shortly before their attacks. See: Ware, J. (2020), *Testament to Murder: The Violent Far-Right's Increasing Use of Terrorist Manifestos* – Policy Brief, International Centre for Counter-Terrorism – The Hague. Accessed: https://icct.nl/publication/testament-to-murder-the-violent-far-rights-increasing-use-of-terrorist-manifestos/

55 'National Strategy on Countering Radicalization to Violence', Public Safety Canada. Accessed: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-strtg-cntrng-rdclztn-vlnc/index-en.aspx#s7

56 The TCAP has also been covered in the Policy Landscape section of a GNET report on 'Decoding Hate: Using Experimental Text Analysis to Classify Terrorist Content.' Accessed: https://gnet-research.org/wp-content/uploads/2020/09/GNET-Report-Decoding-Hate-Using-Experimental-Text-Analysis-to-Classify-Terrorist-Content.pdf

the TCAP will function as a historical archive for improved quantitative and qualitative academic analysis.[57]

In the AI space specifically, one of TCAP's stated aims is to support an ecosystem of algorithmic content classifiers.[58] As GNET's primer report, *AI and CVE*, demonstrates, 'algorithms need vast amounts of data to make useful predictions on the future.'[59] Automated content moderation mechanisms based on machine learning and natural language processing rely on Big Data analysis to train AI to recognise data for what it is, learning to recognise the elements of IS propaganda videos (logos, flags, etc) to identify and mark future videos with the same or similar elements. The TCAP, as the first unified platform for online terrorist content, is a veritable goldmine of data and information for developers designing machine learning algorithms to identify and classify terrorist material.

In providing verified terrorist content from across internet platforms as a historical archive, the TCAP could provide a significant technological leap forward in countering violent extremism online. The Canadian government, as co-sponsor of the platform, has demonstrated how targeted and smart investment in cross-sectoral initiatives can provide opportunities for academia, industry and civil society to collaborate to drive AI forward in a CVE context.

European Commission

In its AI White Paper from February 2020, the European Commission stated that 'AI tools can provide an opportunity for better protecting EU citizens from crime and acts of terrorism'.[60] The EU wants to take a dual approach to using AI: regulatory and investment-focused. It is particularly interested in enabling 'trustworthy AI' by building a sound regulatory framework to help protect European citizens as well as 'create a frictionless internal market' for the development of AI.[61] 'Trustworthy AI, in this case, means technically robust and accurate applications'.[62] The EU also intends to increase investment on AI to at least €20 billion per year by 2030.[63]

The European Commission appointed a High-Level Expert Group on Artificial Intelligence (AI HLEG) in 2019. This group established seven criteria to ensure trustworthy AI. These seven principles are: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability.[64] On this basis, the Commission's White Paper calls for an ecosystem of trust to ensure the protection of fundamental rights.[65]

57 'Press Release: Tech Against Terrorism Participates in UN General Assembly Week in New York', Tech Against Terrorism. Accessed: https://www.techagainstterrorism.org/2019/10/08/press-release-tech-against-terrorism-participates-in-un-general-assembly-week-in-new-york/
58 Ibid.
59 Schroeter, M. (2020), 'AI and CVE: A Primer', Global Network on Extremism and Technology, p.25.
60 European Commission, (19 February 2020), 'White Paper on Artificial Intelligence – A European Approach to Excellence and Trust', p.2. Available from: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
61 Ibid., p.10.
62 Ibid., p.20.
63 Government of France, Ministry of Europe and Foreign Affairs, 'Transparency and accountability: The challenges of artificial intelligence'. Available from: https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/transparency-and-accountability-the-challenges-of-artificial-intelligence/
64 Ibid.
65 Ibid.

The response from major technology companies on the AI White Paper was mixed. Google called on the EU to use existing regulation and regulatory frameworks, rather than build new regulatory frameworks to which technology companies need to adhere. In parallel, Google, Facebook and other technology platforms will need to prepare for the Digital Services Act, expected later this year, which will seek to 'regulate the online ecosystem across a range of areas including … offensive content'.[66] The EU is also expected to follow its AI White Paper with legislation on AI and safety, liability, fundamental rights and data later in 2020.[67]

France

In France, there are a number of key stakeholders whose remit involves AI. The Ministerial Coordinator on Artificial Intelligence has been tasked with analysing and developing proposals on changes related to digital innovation applicable to security.[68] Within the Ministry of Defence, there is also a Defence Artificial Intelligence Coordination Unit, as part of the Defence Innovation agency.

France has been adapting its legal framework to allow for safe and more efficient use of AI-enabled technologies to safeguard France's population. In terms of policy developments, France published its AI strategy in March 2018. Its main objectives are: to improve the AI education and training ecosystem to develop and attract the best AI talent; to establish an open data policy for the implementation of AI applications and the pooling of assets together; and to develop an ethical framework for transparent and fair use of AI applications.[69] In accordance with the EU Directive on Network and Information Systems Security, France developed its own cybersecurity law.[70] France is currently synthesising its thinking on using AI-enabled technologies for military use.[71]

France has launched a series of initiatives focusing on AI. At the G7 Multistakeholder Conference on AI in 2018, France and Canada announced the launch of an International Panel on Artificial Intelligence, which would support the responsible adoption of AI.[72] They also jointly spearheaded the development on a new Global Partnership on Artificial Intelligence (GPAI), joined by a number of other countries. The purpose of the initiative is to guide the responsible development and use of AI, taking into account human rights, inclusion, diversity, innovation and economic growth.[73] Specifically, it will serve to bridge the gap between theory and practice on AI by supporting research on AI-related activities.

66 Stolton, S. (23 June 2020), 'Platform clamp down on hate speech in run up to Digital Services Act', *EURACTIF*. Available from: https://www.euractiv.com/section/digital/news/platforms-clamp-down-on-hate-speech-in-run-up-to-digital-services-act/

67 Kayali, L., Heikkila, M. and Delcker, J. (19 February 2020), 'Europe's digital vision, explained', *Politico*. Accessed: https://www.politico.eu/article/europes-digital-vision-explained/

68 Government of France, Prime Minister's Office, (13 July 2018), 'Action plan against terrorism', p.20. Accessed: http://www.sgdsn.gouv.fr/uploads/2018/10/20181004-plan-d-action-contre-le-terrorisme-anglais.pdf

69 European Commission, *France AI strategy report*. Accessed: https://ec.europa.eu/knowledge4policy/ai-watch/france-ai-strategy-report_en

70 Government of France, National Cyber Security Agency, *Directive network and information system security (NIS)*. Accessed: https://www.ssi.gouv.fr/entreprise/reglementation/directive-nis/

71 Pannier, A. and Schmitt, O. (2019), 'To fight another day: France between the fight against terrorism and future warfare', *International Affairs* vol. 95, no. 4. Accessed: https://academic.oup.com/ia/article/95/4/897/5492774

72 Government of Canada, Prime Minister's Office (6 December 2018), *Mandate for the International Panel of Artificial Intelligence*. Accessed: https://pm.gc.ca/en/news/backgrounders/2018/12/06/mandate-international-panel-artificial-intelligence

73 Government of France, Ministry of Europe and Foreign Affairs (15 June 2020), *Launch of the Global Partnership on Artificial Intelligence by 15 founding members*. Accessed: https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding

In France, the GPAI is supported by a Centre of Excellence, a sister-institution to the GPAI Centre of Excellent in Montreal. The GPAI is also supported by the OECD.

Based on the AI strategy, France may envisage launching the Digital Republic Act, which would serve to 'open up public data, strengthen the protection of users' rights and data privacy and to ensure that the opportunities due to digitalisation benefit all'.[74]

Ghana

Efforts to combat violent extremism online in Ghana are limited, since political violence in the country has not been fuelled by terrorist activities, unlike its neighbouring states of Nigeria and Chad.[75] The Global Terrorism Database, a database of global terror attacks since 1970, lists only 21 incidents with 23 fatalities in 50 years in Ghana.[76]

Ghana does not face the same issues as some of its neighbouring countries around government shutdowns of the internet or governmental use of social media to suppress political dissent.[77] Such governments have exploited a legacy of colonial laws, historically used to violate freedoms, to 'legitimise many … attempts to make extra-legal demands of the private sector'.[78] The 'Ranking Digital Rights' 2019 report shows that social media platforms and internet service providers have had to respond to extra-legal government shutdown demands, raising concerns about excessive surveillance and censorship.[79]

Although the Ghanaian government has not made such illegal demands as yet, civil society groups and journalists have expressed concern about the future.[80] Prior to the 2016 elections, the Ghanaian police chief announced a possible social media shutdown.[81] Although the president resisted such plans, anxieties around digital rights in Ghana are growing.

Liberal freedom of expression laws in Ghana leave digital spaces open to abuses, such as hate speech and cyberbullying (particularly of women).[82] Calls for tighter regulation of social media platforms, therefore, are growing. An expert from the Freedom of Expression Media Foundation for West Africa has noted that 'If there isn't regulation then other pieces of legislation will be used to prosecute people in ways that might be excessive', similar to the governmental demands described above.

---

74 European Commission, *France AI strategy report*.
75 Credit to Tomiwa Ilori, researcher at the Expression, Information and Digital Rights Unit at the Centre for Human Rights, University of Pretoria, for these insights via e-mail communication.
76 Global Terrorism Database, START. Accessed: https://www.start.umd.edu/gtd/
77 Ilori, T. (2020), 'Content Moderation Is Particularly Hard in African Countries', Information Society Project at Yale Law School. Accessed: https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/moderate-globally-impact-locally-content-moderation-particularly-hard-african-countries
78 Ilori, T. (2020), 'Stemming digital colonialism through reform of cybercrime laws in Africa', Information Society Project at Yale Law School. Accessed:https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/stemming-digital-colonialism-through-reform-cybercrime-laws-africa
79 Ranking Digital Rights, '2019 RDR Corporate Accountability Index'. Accessed: https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf
80 Majama, K. (2019), 'Africa in urgent need of a homegrown online rights strategy', Association for Progressive Communications. Accessed: https://www.apc.org/en/news/africa-urgent-need-homegrown-online-rights-strategy
81 Olukotun, D. (16 August 2019), 'President of Ghana says no to internet shutdowns during coming elections,' *AccessNow*. Accessed: https://www.accessnow.org/president-ghana-says-no-internet-shutdown-elections-social-media/
82 Endert, J. (2018), 'Digital backlash threatens media freedom in Ghana', DW Akademie. Accessed: https://www.dw.com/en/digital-backlash-threatens-media-freedom-in-ghana/a-46602904

Government regulations of social media must, however, find a balance between protecting users from harm and protecting users' free speech. A prominent civil society group that campaigns against internet shutdowns, has warned against government regulation of social media: 'Once you allow government to regulate the internet – and you have examples from other countries – then you will end up with that government telling you how to use the internet.'[83]

It is not yet clear if there are plans to develop AI-based tools to aid in online content regulation in Ghana. Nevertheless, regional threats to freedom of expression via colonial-era laws have shown that the country must place its citizens' digital rights at the forefront of any technological tools or legislative effort to monitor harmful content online. In a welcome move, Ghana passed a Right to Information Bill in 2019 that guarantees access to information held by public institutions.[84] The bill signals that the Ghanaian government wants to handle digital rights with transparency and accountability. Any future developments in online content moderation should follow these commitments and standards in order to protect privacy and freedom of expression rights.

Japan

Japan channels most of its CVE efforts through the Association of Southeast Asian Nations (ASEAN).[85] As early as 2004, the ASEAN member nations, in partnership with Japan, issued a set of declaratory statements of cooperation to combat international terrorism. As well as signalling political intentions, the declaration committed signatories to 'preventing, disrupting and combatting international terrorism through information exchange, intelligence sharing and capacity building,' establishing a precedent for regional cooperation to counter violent extremism and terrorism.[86]

Japan reaffirmed its commitment to multinational collaboration in Southeast Asia in 2015 in combatting violent extremism and terrorism and to cooperate in implementing the ASEAN Plan of Action to Prevent and Counter the Rise of Radicalisation and Violent Extremism (2018–2025).[87] The plan of action prioritises partnership 'with the business community and technology sector in promoting moderation and enhancing dialogue to prevent radicalisation and violent extremism', as well as strengthening 'strategic communications' to prevent the misuse of social media by VE and terrorist actors.[88]

83 Ibid.
84 Yahya Jafru, M. (26 March 2019), 'Right to information – RTI bill passed into law', *Graphic Online*. Accessed: https://www.graphic.com.gh/news/politics/ghana-news-rti-bill-passed.html
85 'Japan: Extremism & Counter Extremism', Counter-Extremism Project. Accessed: https://www.counterextremism.com/countries/japan
86 'ASEAN-Japan Joint Declaration for Cooperation to Combat International Terrorism', ASEAN. Accessed: https://asean.org/?static_post=asean-japan-joint-declaration-for-cooperation-to-combat-international-terrorism-2
87 'Chairman's Statement of the 18th ASEAN-Japan Summit, Kuala Lumpur, November 22 2015', ASEAN. Accessed: https://www.asean.org/wp-content/uploads/2015/12/6.-Chairmans-Statement-of-the-18th-ASEAN-Japan-Summit-Final-Final.pdf; 'Japan's cooperation with ASEAN 2025 (Counter-Terrorism)', Mission of Japan to ASEAN. Accessed: https://www.asean.emb-japan.go.jp/asean2025/jpasean-ps03.html
88 '2018 ASEAN Plan of Action to Prevent and Counter the Rise of Radicalisation and Violent Extremism (2018–2025), adopted in Myanmar, October 31 2018', ASEAN. Accessed: https://cil.nus.edu.sg/wp-content/uploads/2019/10/2018%20ASEAN%20Plan%20of%20Action%20to%20Prevent%20and%20Counter%20the%20Rise%20of%20Radicalisation%20and%20Violent%20Extremism%20(2018-2025).pdf

The 2020 Olympic and Paralympic Games (delayed until 2021 due to the coronavirus crisis) will be held in Tokyo. Hosting the Olympics has traditionally been viewed as a 'test' of a nation's security capabilities and the games are an opportunity to pilot innovations in AI, security and law enforcement.[89]

In one such pilot ahead of the games in 2018, the Kanagawa Prefectural Police announced the launch of a predictive policing system in order to predict crimes and attacks based on a deep machine-learning algorithm.[90] Since then, leading technology corporations have confirmed the provision of a large-scale facial recognition, biometric authentication and behaviour detection systems at the games and at ports and airports.[91] These systems will have the ability to scan faces for particular emotions and to confirm identity against facial features and personal information.

It is not yet clear whether these AI-security capabilities will extend to social media and online activity. The systems trialled in Kanagawa could have reportedly included monitoring social media content to combat crime, which could be understood as an embrace by Japanese law enforcement of AI-based social media monitoring to combat malicious or potentially dangerous content online. Such a move would be risky, given the outcry in 2017 in response to the government's controversial counter-terrorism bill that critics saw as imperilling civil liberties.[92] Japan will need to give serious attention to protecting citizens' rights to privacy and freedom of expression as it develops AI technologies for countering violent extremism online and offline.

New Zealand

Governance of countering violent extremism online in New Zealand involves the coordination of manifold agencies and bodies. These include the Cabinet External Relations and Security Committee; police, intelligence and security communications agencies; and foreign affairs, trade, defence, transport, innovation and development agencies. New Zealand's overarching counter-terrorism strategy is outlined in their National Strategy Overview, released in February 2020.[93]

In the aftermath of the Christchurch mosque shootings in March 2019, the governments of New Zealand and France brought together a coalition of heads of state with social media and technology companies under the Christchurch Call to Eliminate Terrorist and Violence Extremist Content Online.[94] The call commits the supporting countries to enforce laws that prohibit

89 See, for example: Soria, V. (2011), 'Beyond London 2012: The Quest for a Security Legacy,' *The RUSI Journal*, vol. 156, no. 2, pp.36–43.

90 'Kanagawa police to launch AI-based predictive policy system before Olympics', *Japan Times* (29 January 2018). Accessed [paywall]: https://www.japantimes.co.jp/news/2018/01/29/national/crime-legal/kanagawa-police-launch-ai-based-predictive-policing-system-olympics/

91 The Government of Japan (2019), 'All is Ready for a Safe and Secure Tokyo Games'. Accessed: https://www.japan.go.jp/tomodachi/2019/autumn-winter2019/tokyo2020.html; 'NEC Becomes a Gold Partner for the Tokyo 2020 Olympic and Paralympic Games', NEC Corporation (2015). Accessed: https://www.nec.com/en/press/201502/global_20150219_01.html

92 'Japan passes controversial anti-terror conspiracy law', *BBC* (15 June 2017). Accessed: https://www.bbc.co.uk/news/world-asia-40283730

93 Government of New Zealand, Officials' Committee for Domestic and External Security Coordination, Counter-Terrorism Coordination Committee (February 2020), 'Countering terrorism and violent extremism national strategy overview'. https://dpmc.govt.nz/sites/default/files/2020-02/2019-20 CT Strategy-all-final.pdf

94 See https://www.christchurchcall.com/

the dissemination of terrorist and violent extremist content online while respecting the international human rights law including freedom of expression. The countries also work to support frameworks, capacity-building and awareness-raising activities in order to prevent the use of online services to disseminate terrorist and violent extremist content.

The Christchurch call also commits companies, including Amazon, Facebook, Google, Twitter and YouTube, to greater industry standards of accountability and transparency. The companies must enforce their community standards and terms of service by prioritising content moderation and removal actions, and identifying content in real-time for review and assessment. Collectively, the countries and companies are developing efforts with civil society to promote community-led activities in order to intervene in the processes of online radicalisation.

Following the March 2019 attack, a Royal Commission of Inquiry was established to assess agencies' response to the shootings and to determine what additional measures could be taken to prevent future attacks.[95] The Commission's report will shed light on current CVE strategy and future direction in New Zealand and will be a useful resource to understand the extent to which AI forms part of this future strategy. Due to the coronavirus crisis, the report's publication has been delayed until the winter of 2020.

The New Zealand government is also holding itself to more rigorous standards of transparency and accountability in its use of algorithms for governance. As *AI and CVE: A Primer* remarks, algorithmic use can exacerbate existing biases.[96] In July 2020, the government published the Algorithm Charter for Aotearoa New Zealand, a comprehensive review of state use of algorithms in sectors ranging from transport to justice, and a commitment for greater transparency, stakeholder engagement, safeguards for privacy and human oversight of algorithm use.[97] The charter – the first of its kind globally – has, at the time of writing, twenty-five state agencies confirmed as signatories.

Crucially missing, however, are the agencies and bodies responsible for CVE online. To that end, it remains unclear the extent to which policymakers in New Zealand are considering the development of AI and algorithmic tools for countering malicious content online and the standards to which such tools would be held. The charter signals a step in a positive direction and applying such standards to AI-based CVE actions would be a welcome development in policymaking.

95 The Royal Commission of Inquiry into the Attack on Christchurch mosques. See: https://christchurchattack.royalcommission.nz/

96 See also Babuta, A. and Oswald, M. (2019), 'Briefing Paper: Data Analytics and Algorithmic Bias in Policing,' RUSI. Accessed: https://www.gov.uk/government/publications/report-commissioned-by-cdei-calls-for-measures-to-address-bias-in-police-use-of-data-analytics; Benjamin, R. (2019), *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity); Benjamin, R., 'A New Jim Code?', Berkman Klein Center for Internet & Society at Harvard University. Recording accessed: https://cyber.harvard.edu/events/new-jim-code

97 'Algorithm charter for Aotearoa New Zealand', data.govt.nz. Accessed: https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter

United Kingdom

In February 2018, the UK announced the development of a machine learning-based algorithmic tool to detect IS terrorist content online. The software was trained to identify and flag recognisable audio-visual elements in IS propagandistic content – flags, logos, formatting, structures and soundtracks. Giant tech platforms like YouTube and Facebook have invested heavily in developing their own automated content moderation tools over the years. The tool was designed to be unspecific to any particular platform and therefore open source for smaller internet and social media platforms such as Vimeo to use.

Although promising, the tool is severely limited in its efficacy and reception. First, as our colleague Charlie Winter highlighted, IS online content ranges from videos to photographs, written pieces and radio bulletins. Although combatting video content is a positive step, 'at best it's going to mitigate [the problem] somewhat, but it is very far from a solution.'[98] Secondly, the Home Office commissioned the tool to recognise the most explicit and shocking IS video content. The software development company who designed the tool explained that 'it's less about volume and more about how impactful they [the Home Office] think it is to tackle particular sets of videos.'[99] However, many academic studies have described the wide-ranging impact of 'softer' propagandistic content and its radicalising potential over long periods of time.[100] Limiting AI to narrow functions in three ways – IS, video content and extreme content – hampers the technical efficacy of the tool. The tool was made available for free to smaller tech firms, although as of April 2019, no companies are yet to adopt it.[101]

The UK government's approach to using AI to combat violent extremism online also demonstrates the potential for conflicts of interest to arise during collaboration between governments and industry. The government's Online Harms White Paper, published in April 2019, set out a comprehensive case for greater national regulation of social media.[102] Under this new regulatory framework, social media and technology companies will bear a new statutory duty of care to its users, enforceable via Ofcom, the UK's regulatory body for communications. Ofcom will subject platforms to financial and technical penalties – websites could be blocked at ISP level and fined up to 4% of their global turnover – for non-compliance with the framework and violations of the statutory duty of care.[103] While announcing the algorithmic tool in February 2018, then-Home Secretary Amber Rudd signalled that companies may be obliged via legislation to adopt the tool.

Such regulatory moves are not concerning in and of themselves. However, the data analytic and software development company that developed the tool, formerly ASI Data Science (now known as

98 Temperton, J. (13 February 2018), 'ISIS could easily dodge the UK's AI-powered propaganda blockade', *Wired*. Accessed: https://www.wired.co.uk/article/isis-propaganda-home-office-algorithm-asi
99 Ibid.
100 'Hashtag Terror: How ISIS Manipulates Social Media', Anti-Defamation League (21 August 2014). Accessed: https://www.adl.org/education/resources/reports/isis-islamic-state-social-media
101 Murgia, M. and Bond, D. (6 April 2019), 'Businesses show no appetite for anti-terror AI tool', *Financial Times*. Accessed [paywall]: https://www.ft.com/content/fda2d218-56fb-11e9-91f9-b6515a54c5b1
102 HM Government (April 2019), 'Online Harms White Paper'. Accessed: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf
103 Crawford, A. (29 June 2020), 'Online Harms bill: Warning over "unacceptable" delay', *BBC*. Accessed: https://www.bbc.co.uk/news/technology-53222665

Faculty), was commissioned for data modelling in the Vote Leave and Leave.EU campaigns and as such implicated in the Cambridge Analytica scandal.[104] Additionally, as of May 2020, the company has been awarded at least seven publicly funded government contracts in eighteen months, and has noteworthy personal and commercial links to Dominic Cummings, chief adviser to the prime minister.[105]

These facts raise concerns about a conflict of interest. Public and business trust in the tool is undermined by awarding the tool's development contract to a firm with close links to the government's inner circle, which has been embroiled in public scandal, as well as by championing legislation that would oblige social media platforms to use it in order to guarantee ongoing business for the company.

A development company independent from government could have developed a more technically effective tool responsive to broader demands (able to recognise more than a subset of IS video content, for example), and increased both trust in and uptake of the tool. Transparency and accountability are 'not mere slogans to which lip service must be paid: they are crucial to the success of problem-solving efforts' in countering violent extremism online using AI.[106] The UK has missed a significant policy opportunity to develop and provide cutting-edge AI technology to moderate harmful content online by undermining trust in the tool and compromising its technical effectiveness.

UN Counter-Terrorism Committee Executive Directorate

The UN Counter-Terrorism Committee Executive Directorate (UN CTED) was established by UN Security Council Resolution 1535 (2004) as an expert body in support of the Security Council's Counter-Terrorism Committee.[107] Its initial aim was to assess UN Member States' implementation of Security Council resolutions on counterterrorism and support their efforts through dialogue. The UN CTED works closely with the Security Council, major technology companies and civil society organisations through the GIFCT.

There currently exist several UN council resolutions relating to abuse of the internet for terrorist purposes and UN CTED is looking to develop more coherence and streamline the intersection between UN Security Council resolutions and the role of IT. Security Council resolution 2129 (2013) recognises the increasing relationship between terrorism and IT and the use of technologies such as the internet to commit and facilitate terrorist acts, by allowing the incitement,

104 Cadwalladr, C. (7 May 2017), 'The great British Brexit robbery: how our democracy was hijacked', *The Guardian*. Accessed: https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy

105 Evans, R. and Pegg, D. (4 May 2020), 'Vote Leave AI firm wins seven government contracts in 18 months', *The Guardian*. Accessed: https://www.theguardian.com/world/2020/may/04/vote-leave-ai-firm-wins-seven-government-contracts-in-18-months; Pegg, D., Evans, R. and Lewis, P. (12 July 2020), 'Revealed: Dominic Cummings firm paid Vote Leave's AI firm £260,000', *The Guardian*. Accessed: https://www.theguardian.com/politics/2020/jul/12/revealed-dominic-cummings-firm-paid-vote-leaves-ai-firm-260000; Pegg, D. and Evans, R. (2 June 2020), 'AI firm that worked with Vote Leave given new coronavirus contract', *The Guardian*. Accessed: https://www.theguardian.com/technology/2020/jun/02/ai-firm-that-worked-with-vote-leave-wins-new-coronavirus-contract

106 'Tackling the Information Crisis: A Policy Framework for Media System Resilience', The Report of the LSE Commission on Truth Trust & Democracy, p.32. Accessed: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

107 Chowdhury Fink, N. (2012), 'Meeting the challenge: A guide to United Nations counterterrorism activities', *International Peace Institute*, p. 45. https://www.ipinst.org/wp-content/uploads/publications/ebook_guide_to_un_counterterrorism.pdf

recruitment, fundraising or planning of terrorist acts.[108] This resolution also re-enforces the mandate of UN CTED. Resolutions 2354 (2017), 2395 (2017) and 2396 (2017) implores UN Member States to cooperate to prevent terrorist organisations from exploiting ICT and to work with the private sector and civil society to develop effective measures to prevent the abuse of the internet for terrorist purposes.[109] Security Council Resolution 1373 calls on UN Member States to develop and accelerate the 'exchange of operational information' on the use of IT by terrorist organisations and halt terrorist recruitment.[110]

The UN Secretary General's High-level Panel on Digital Cooperation seeks solutions to mitigate AI risks.[111] The panel's recommendation 3C states:

> *'We believe that autonomous intelligent systems should be designed in ways that enable their decisions to be explained and humans to be accountable for their use. Audits and certification schemes should monitor compliance of AI systems with engineering and ethical standards, which should be developed using multi-stakeholder and multilateral approaches. Life and death decisions should not be delegated to machines. We call for enhanced digital cooperation with multiple stakeholders to think through the design and application of these standards and principles such as transparency and non-bias in autonomous intelligent systems in different social settings.'*[112]

One of the areas of focus is regarding the protection of human rights in the digital era.[113]

United States

The United States' Counter-Terrorism Strategy identifies CVE online as a priority area and commits itself to working with business and industry to combat terrorist recruitment, fundraising and radicalisation processes online. In terms of cross-national initiatives, the USA works with initiatives such as Tech Against Terrorism and the Global Counterterrorism Forum, which relies on partnership with other signatories, civil society and the tech sector to craft approaches to countering violent extremism online in the medium and long term.

In the USA's domestic legislative area, calls for regulation of social media and technology platforms were prompted by reports of Russian interference and information operations online in the 2016 presidential elections. At the same time, social media corporations have continued to grow in terms of user numbers and subsidiary services and products. In late 2019, the US Senate Banking

108 UN, Security Council Counter-terrorism Committee, (14 September 2018), 'Public-private efforts to address terrorist content online: A year of progress – what's next?'. Accessed: https://www.un.org/sc/ctc/news/event/public-private-efforts-address-terrorist-content-online-year-progress-whats-next/; Global Initiative Against Transnational Organised Crime, *Responding to terrorist use of the internet* (21 January 2019). Accessed: https://globalinitiative.net/terrorist_use_internet/

109 UN, Security Council Counter-terrorism Committee, 2018.

110 Global Initiative Against Transnational Organised Crime, 2019.

111 Government of France, Ministry of Europe and Foreign Affairs, 'Transparency and accountability: The challenges of artificial intelligence'.

112 UN (16 December 2019), High-level panel follow-up roundtable 3C – Artificial Intelligence – Meeting note. Accessed: https://www.un.org/en/pdfs/HLP%20Followup%20Roundtable%203C%20Artificial%20Intelligence%20-%201st%20Session%20Summary.pdf

113 UN, Secretary-General's High-level panel on digital cooperation. Accessed: https://www.un.org/en/digital-cooperation-panel/

Committee and Congressional Committee on Energy and Commerce held hearings over Facebook's proposed cryptocurrency service, Libra. The hearings provided an opportunity for legislators in the USA to question Facebook executives over manipulation and misuse of its platform,[114] and to establish big tech regulation as a viable legislative option.[115]

As Congress considers legislative action, the USA's intelligence community has pushed forward in using AI to counter violent extremism online. In spring and summer 2019, the USA was rocked by a spate of mass shootings whose perpetrators has extensive histories of online engagement with violent extremism. For instance, the Poway synagogue shooter, who opened fire on a California synagogue in late April 2019, posted a manifesto to 8chan shortly before the attack. The manifesto refers to other online-fuelled shootings, such as the Christchurch mosque shootings and the Pittsburgh synagogue shooting, as well as typical far-right and white nationalist online personalities and references.

In this context, the Federal Bureau of Investigation (FBI) released a bid proposal for private contractors to develop technology that gives the Bureau 'near real time access to a full range of social media exchanges' to 'detect, disrupt, and investigate an ever growing diverse range of threats to U.S. National interests.'[116] A similar bid was released in January 2020.[117] In June 2020, as #BlackLivesMatter protests swept across the nation, the FBI extended its contracts with Dataminr, a social media monitoring and analytics company, and Venntel, a location data company.[118]

These technologies and access to data would parallel the general AI system as set out in section four of *AI and CVE: A Primer*, a predictive system that allows law enforcement to intervene based on an alerting mechanism. Such tools would present a remarkable ethical threat to users' privacy rights, since real-time surveillance of individual behaviour for law enforcement would rely upon non-anonymised data. The collection of identifiable data would undermine rights to personal security, identity protection and freedom of expression.

The USA's approach to using artificial intelligence to counter violent extremism online demonstrates the legal and ethical challenges inherent in tracking and moderating online material. As Marie Schroeter writes, such an approach 'would create nothing less than a dystopia.'[119]

114 US House of Representatives Committee on Energy and Commerce, 'Facebook: Transparency and Use of Consumer Data,' transcript of 11 April 2018, p.33. Accessed: https://docs.house.gov/meetings/IF/IF00/20180411/108090/HHRG-115-IF00-Transcript-20180411.pdf

115 Molla, R. and Stewart, E. (2019), 'How 2020 Democrats think about breaking up Big Tech', *Vox*. Accessed: https://www.vox.com/policy-and-politics/2019/12/3/20965447/tech-2020-candidate-policies-break-up-big-tech

116 US Government Federal Acquisitions Service, 'Contract Opportunity: Social Media Alerting Subscription.' Accessed: https://beta.sam.gov/opp/b6de554012cf4ab9ab795f52c638467c/view

117 US Government Federal Acquisitions Service, 'Request for Proposal – FBI Social Media Alerting.' Accessed: https://beta.sam.gov/opp/2b3003e9b0b34b639687786e8420013b/view

118 US Government Federal Acquisitions Service, 'Contract Information – Dataminr. Inc.' Accessed: https://beta.sam.gov/awards/90552288%2BAWARD?keywords=15F06720P0000950&sort=-relevance&index=&is_active=true&page=1; Fang, L. (24 June 2020), 'FBI Expands Ability to Collect Cellphone Location Data, Monitor Social Media, Recent Contracts Show', *The Intercept*. Accessed: https://theintercept.com/2020/06/24/fbi-surveillance-social-media-cellphone-dataminr-venntel/

119 Schroeter, M. (2020), 'AI and CVE: A Primer', Global Network on Extremism and Technology, p.22.

Policy Recommendations

Existing initiatives and actions, like those described above, provide insights and recommendations for policymakers across the globe. Based on our findings, we make the following policy recommendations:

Recommendation 1: Establish an independent regulatory body at the cross-national level for oversight on national efforts to counter violent extremism online using artificial intelligence

Governmental legislation that creates penalties for social media companies that fail to moderate harmful content[120] can be highly effective,[121] but risks limiting citizens' right to free speech, since fear of incurring penalties may lead to over-removal of content. As described above, in the cases of the UK, Japan and the USA, there is also potential for content moderation efforts and legislation to run into legal and ethical issues around privacy, trust and accountability.

Social media self-regulation, where companies create and enforce their own standards, codes and policies for removing malicious content online, can be effective, but applicability of the standards can be uneven and non-transparent.[122] Many major companies do publish high-level data on content moderation, but there is no obligation to do so.[123]

Co-regulation between government, civil society and industry, overseen by a cross-national, independent body should be established to ensure adherence to standards in accountability, transparency and ethics. An independent body that is committed to global standards of protecting privacy,[124] that has enforcement mechanisms and that is independent of government would mitigate these issues.

Joint regulation between government, social media platforms and civil society would ensure that users' interests are at the forefront of regulation efforts. Government legislation that authorises the regulatory body protects it from shifting political interests and ensures that its enforcement mechanisms are feasible. In obliging platforms to comply with the mechanisms, moderation efforts are applied more evenly and more fairly. Privacy, free speech and accountability should form the basis of the body's values and inform its governance.

120 For instance, the 2017 German Network Enforcement Act, known as NetzDG, imposes penalties of up to €50 million for social media platforms that do not remove illegal content within twenty-four hours. See: https://www.ceps.eu/system/files/RR%20No2018-09_Germany%27s%20NetzDG.pdf

121 Elhai, W. (2020), 'Regulating Digital Harm Across Borders: Exploring a Content Platform Commission', *SMSociety'20: International Conference on Social Media and Society*, https://doi.org/10.1145/3400806.3400832, pp.223–4.

122 See Matsakis, L. (2 March 2018), 'YouTube Doesn't Know Where Its Own Line Is', *Wired*. Accessed: https://www.wired.com/story/youtube-content-moderation-inconsistent/

123 Ibid.

124 Such as the Global Network Initiative. See https://globalnetworkinitiative.org/

Recommendation 2: Include measures to combat algorithmic bias into software development at point of design

Content moderation and curation policies and practices by online platforms rarely involve public input or accountability.[125] Algorithms, often developed 'behind closed doors' in the tech industry and with huge influence on billions of users' online experience, can be extremely biased. Algorithms 'learn by being fed certain images, often chosen by engineers', who are overwhelmingly and disproportionately white and male.[126] Such bias has led to severe real-world issues, including software that identifies black defendants as more likely to commit a future crime,[127] and a Google Photo app that mistakenly tagged a black user's photos as gorillas.[128]

In terms of CVE, such algorithmic bias means that non-Western extremist content is under-recognised and under-moderated. Since the globe's largest tech companies are headquartered in the West, 'engineers and executives responsible for designing technology products might be unfamiliar with the catalysts of violence and discrimination in cultures other than their own.'[129] The situation in Myanmar illustrates what is at stake when non-Western content moderation and removal is under-developed. Burmese-language hate speech online against the Rohingya community incited wide-scale violence. Yet an escalation of racist hatred of the minority went largely unchecked, as Facebook employed only two Burmese-speaking content reviewers.[130]

To combat such bias, tech platforms can utilise existing expertise in civil society and academia to feed into the software development stage. Companies should undertake extensive and regular audits of their use of algorithms. The results of these audits should be made publicly available to increase accountability, transparency and public trust.[131]

Expanding content moderation capabilities – in linguistic and geographic terms for human teams; and in developing non-Western AI tools – is a costly investment for social media and technology companies. Efforts to advocate for this much-needed expansion can offer the tech industry the opportunity to position themselves as industry leaders for content moderation and removal strategies.

125 'Tackling the Information Crisis: A Policy Framework for Media System Resilience,' The Report of the LSE Commission on Truth Trust & Democracy, p18. Accessed: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

126 Crawford, K. (25 June 1016), 'Artificial Intelligence's White Guy Problem', *New York Times*. Accessed [paywall]: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

127 Angwin, J. et al. (23 May 2016), 'Machine Bias', *ProPublica*. Accessed: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

128 Nieva, R. (1 July 2015), 'Google apologizes for algorithms mistakenly calling black people "gorillas"', *CNET*. Accessed: https://www.cnet.com/news/google-apologizes-for-algorithm-mistakenly-calling-black-people-gorillas/

129 Elhai, p.221.

130 Stecklow, S. (2018), 'Special Report: Why Facebook is losing the war on hate speech in Myanmar', *Reuters*. Accessed: https://www.reuters.com/article/us-myanmar-facebook-hate-specialreport/special-report-why-facebook-is-losing-the-war-on-hate-speech-in-myanmar-idUSKBN1L01JY

131 Turner Lee, N. (2019), 'Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms', Brookings Institute. Accessed: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Recommendation 3: National and multinational stakeholders and initiatives to fund publications that presents the technology, challenges and opportunities afforded by artificial intelligence in clear and accessible language

As *AI and CVE: A Primer* explores, there is considerable hype and noise around AI. Many policymakers misunderstand what AI is and what it can do. Additionally, 'Parliamentary discussions and committee hearings in the UK following the 2018 Cambridge Analytica scandal, in the US Congress and at the European Parliament, revealed shockingly low levels of media literacy and understanding among senior parliamentarians and policy-makers.'[132]

Poor understanding of the digital and media environment, as well as AI's capabilities and limitations, can lead to suboptimal policymaking outcomes. Policymakers need to be educated in this space in order to make evidenced, balanced and informed policy decisions.

National and multinational initiatives should focus on producing and publishing regular guidance on what AI is, as well as its challenges and opportunities in the policy world. Civil society and academic experts in online and offline harms in particular contexts should also produce reactive primers on emerging and future threats. For instance, experts familiar with an upcoming contentious election in a non-Western context would produce a briefing for policymakers on the context and catalysts to harmful content and how this could translate to real-world harms.

Such guidance and briefings should be written in clear and accessible language, cutting through the technical jargon or sensationalist hype around AI. As such, policymakers and laypeople alike would be able to contribute to the public discourse on AI and CVE.

132 'Tackling the Information Crisis: A Policy Framework for Media System Resilience', The Report of the LSE Commission on Truth Trust & Democracy, p.38. Accessed: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

Global Network
on Extremism & Technology

CONTACT DETAILS

For questions, queries and additional copies of this report, please contact:

ICSR
King's College London
Strand
London WC2R 2LS
United Kingdom

T. **+44 20 7848 2098**
E. **mail@gnet-research.org**

Twitter: **@GNET_research**

Like all other GNET publications, this report can be downloaded free of charge from the GNET website at www.gnet-research.org.