Research Paper

Kathleen McKendrick
International Security Department | August 2019

Artificial Intelligence Prediction and Counterterrorism

CHATHAM HOUSE
The Royal Institute of International Affairs

Contents

Summary

- The use of predictive artificial intelligence (AI) in countering terrorism is often assumed to have a deleterious effect on human rights, generating spectres of 'pre-crime' punishment and surveillance states. However, the well-regulated use of new capabilities may enhance states' abilities to protect citizens' right to life, while at the same time improving adherence to principles intended to protect other human rights, such as transparency, proportionality and freedom from unfair discrimination. The same regulatory framework could also contribute to safeguarding against broader misuse of related technologies.

- Most states focus on preventing terrorist attacks, rather than reacting to them. As such, prediction is already central to effective counterterrorism. AI allows higher volumes of data to be analysed, and may perceive patterns in those data that would, for reasons of both volume and dimensionality, otherwise be beyond the capacity of human interpretation. The impact of this is that traditional methods of investigation that work outwards from known suspects may be supplemented by methods that analyse the activity of a broad section of an entire population to identify previously unknown threats.

- Developments in AI have amplified the ability to conduct surveillance without being constrained by resources. Facial recognition technology, for instance, may enable the complete automation of surveillance using CCTV in public places in the near future.

- The current way predictive AI capabilities are used presents a number of interrelated problems from both a human rights and a practical perspective. Where limitations and regulations do exist, they may have the effect of curtailing the utility of approaches that apply AI, while not necessarily safeguarding human rights to an adequate extent.

- The infringement of privacy associated with the automated analysis of certain types of public data is not wrong in principle, but the analysis must be conducted within a robust legal and policy framework that places sensible limitations on interventions based on its results.

- In future, broader access to less intrusive aspects of public data, direct regulation of how those data are used – including oversight of activities by private-sector actors – and the imposition of technical as well as regulatory safeguards may improve both operational performance and compliance with human rights legislation. It is important that any such measures proceed in a manner that is sensitive to the impact on other rights such as freedom of expression, and freedom of association and assembly.

1. Introduction

Counterterrorism policy strikes a balance between maintaining the security of a population and respecting individual rights to privacy and freedoms such as those of expression, association and religion.[1] Novel technological developments can challenge the implementation of these policies by forcing authorities to re-examine how counterterrorism actions are conducted.[2] Artificial intelligence (AI) is one of these technologies.

This paper describes how AI can theoretically contribute to counterterrorism operations, as well as some areas where it is already used. It presents some of the problems with the current use of AI for these purposes, examining both practical limitations and implications for human rights. It also considers the opportunities and risks posed by increased use of AI as part of counterterrorism activity by governments. It makes the case that use of AI in counterterrorism is not inherently wrong, and suggests some necessary conditions for legitimate use of AI as part of a predictive approach to counterterrorism on the part of liberal democratic states.

Uses of AI in counterterrorism centre on generating accurate predictions that help direct resources for countering terrorism more effectively. Predictive AI might also minimize unnecessary intrusion on the majority of the population and mitigate human bias in decision-making. In an era when it is becoming easier to collect, store and analyse on a large scale revelatory data about people's individual activities and associations, predictive methods may prove increasingly important where they can accurately direct attention to the areas or individuals most likely to pose threats, and reduce the number of citizens who are subject to more invasive monitoring.

There are known risks associated with AI. Algorithms can exhibit inherent or learned biases of their own,[3] may be vulnerable in adversarial environments,[4] and are inherently limited by the data to which they have access.[5] Public oversight, already limited by security considerations, could be jeopardized further by the incorporation of automated analysis systems that are inherently unintelligible to humans.[6] A lack of adequate safeguards on the use of AI, and on the vast reservoirs of data on which it relies, could lead not only to its misuse by autocratic states in order to impose totalitarian control over their citizens, but also to excessive infringement on rights such as of privacy and freedom of expression or association in democratic states.

[1] Brokenshire, J. (2013), 'National Security and Civil Liberties – Getting the balance right', speech delivered at National Security Summit at Queen Elizabeth Conference Centre, 3 July 2013, https://www.gov.uk/government/speeches/national-security-and-civil-liberties-getting-the-balance-right (accessed 1 Oct. 2018).
[2] Cornish, P. (2010), 'Technology, Strategy and Counterterrorism', *International Affairs*, 86(4), p. 888, https://www.jstor.org/stable/40865000 (accessed 1 October 2018).
[3] Osoba, O. and Welser, W. (2017), *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*, Santa Monica, CA: RAND Corporation, 2017, https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf (accessed 1 Oct. 2018).
[4] Goodfellow, I., Papernot, N., Huang, S., Duan, Y., Abbeel, P. and Clark, J. (2017), 'Attacking Machine Learning with Adversarial Examples', Open AI blog, 24 February 2017, https://blog.openai.com/adversarial-example-research/ (accessed 1 Oct. 2018).
[5] Waters, R. (2018), 'Why we are in danger of overestimating AI', *Financial Times*, 5 February 2018, https://www.ft.com/content/4367e34e-db72-11e7-9504-59efdb70e12f (accessed 1 Oct. 2018).
[6] Knight, W. (2017), 'The Dark Secret at the Heart of AI', *MIT Technology Review*, 11 April 2017, https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ (accessed 1 Oct. 2018).

The notion of using accurate predictive technologies underpinned by AI for counterterrorism purposes is speculative but feasible. It is useful to consider in advance the possibilities and the costs of such an approach, and how this nascent field might be regulated.

2. Prediction in Counterterrorism and Relevant AI Technologies

This section describes the role of prediction in counterterrorism, and introduces how certain types of AI are used for predictive purposes. It presents a definition of AI and distinguishes it from 'big data' analytics in general. It argues that good prediction is needed to mount a counterterrorism strategy that is effective at preventing terrorism without undue encroachment on citizens' rights. Further, it suggests that predictive AI is likely to be helpful in contributing to this.

Terrorist attacks in liberal democracies seek to undermine public support in the governments they target.[7] Unpredictably striking civilian targets disturbs the illusion of a secure environment underwritten by the nation state. Even if the aftermath of an attack is effectively managed, it is invariably more desirable to prevent such an attack entirely. As such, prevention is often the central focus of counterterrorism strategies.[8]

There are two means to prevent terrorist attacks. One is deterrence: through the protection of infrastructure, the application of security checks and the promise of punishment. Another is the denial of the ability to conduct attacks: by apprehending terrorists before their plots come to fruition, countering recruitment and radicalization of future terrorists, and placing restrictions on the movement and freedom of individuals.

Many of these methods share common characteristics. Specifically, they place demands on scarce resources and, to some extent, infringe on the rights of the people to whom they are applied. Excessive or arbitrary infringement of rights betrays the principles of liberal democracy, and in so doing concedes a victory to the terrorists.[9] Hence, on the basis of pragmatism as well as principle, effective counterterrorism might be characterized as a problem of optimization that aims to provide security for the majority of citizens with minimum infringement of rights and freedoms. To reach this optimal point, effective terrorism prevention would benefit from methods that improve the prediction of attacks or behaviours associated with them.

Prediction allows discretion in the application of preventive measures, minimizing the effect on the population as a whole. Effective prediction might, for example, have the effect that only violent terrorists are met with coercive force or restrictions, while conciliatory measures are directed towards individuals vulnerable to radicalization. In the case of deterrence through physical protection, prediction can serve as a means to improve the allocation of resources to locations most likely to be targets.

[7] Kurth Cronin, A. (2004), 'Sources of Contemporary Terrorism', in Kurth Cronin, A. and Lundes, J. (eds) (2004), *Attacking Terrorism: Elements of a Grand Strategy*, Washington, DC: Georgetown University Press, 2004, p.33.
[8] Monaco, L (2017) 'Preventing the Next Attack; A Strategy for the War on Terrorism' *Foreign Affairs* 96(6), pp.23–29.
[9] Richardson, L. (2006) *What Terrorists Want: Understanding the Enemy, Containing the Threat*, New York: Random House, p. 250.

According to the Engineering and Physical Science and Research Council:

> Artificial Intelligence technologies aim to reproduce or surpass abilities (in computational systems) that would require 'intelligence' if humans were to perform them. These include: learning and adaptation; sensory understanding and interaction; reasoning and planning; optimisation of procedures and parameters; autonomy; creativity; and extracting knowledge and predictions from large, diverse digital data.[10]

For the purpose of making predictions, the type of AI that enables the extraction of knowledge and predictions from large diverse digital data is most relevant. This is related to, but distinct from, the field of 'big data' analytics. 'Big data' refers to datasets whose size is beyond the capability of typical database software tools to capture, store manage and analyse. A popular characterization defines 'big data' in terms of the so-called 'four Vs': velocity, veracity, variety and volume. Notably, characterizations of 'big data' are technical, based on the nature of the data.

This paper chooses to focus on AI, not just 'big data' analytics. The overlapping methods of interest are mainly those that use machine learning to build models based on data, and then make inferences from those models. These methods – the type of AI referred to above – are in some ways a narrower area of interest than 'big data', which can also include use of large datasets for searching and analysis by human operators structuring specific queries. In other ways, specifically the adaptation of traditionally 'big data' methodologies to smaller datasets, they are broader, and not necessarily exclusive to data characterized by the four Vs.

The data analysed, and the type of analyses, are to some extent inextricably linked; and this often causes confusion between AI and 'big data' analytics. The algorithms supporting predictive models are self-programmed based on exposure to data. In many cases, it would be impossible to analyse the data without such an approach, much as it would be impossible to build the models without the data. 'Big data' needs AI to interpret it: AI often needs 'big data' to build and improve itself.

Developments in other areas of AI mean that improving accuracy of predictions might become increasingly important, both for effective counterterrorism and for halting the erosion of privacy. As observed by the UN Special Rapporteur on privacy in 2014: 'Declining costs of technology and data storage have eradicated financial or practical disincentives to conducting surveillance.'[11] Developments in AI have amplified the ability to conduct surveillance without being constrained by resources. Facial recognition technology, for instance, may enable the complete automation of surveillance using CCTV in public places in the near future.[12] AI-driven text analysis could be used to 'understand' the content of private messages without requiring the attention of a human analyst.

[10] Engineering and Physical Sciences Research Council (2018), 'Research Areas: Artificial Intelligence Technologies', *Engineering and Physical Sciences Research Council Website*, https://epsrc.ukri.org/research/ourportfolio/researchareas/ait/ (accessed 1 Oct. 2018).
[11] OHCHR (2014), *The right to privacy in the digital age: Report of the Office of the United Nations High Commissioner for Human Rights, Advance Edited Version*, p. 3, http://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session27/Documents/A.HRC.27.37_en.pdf (accessed 1 Oct. 2018).
[12] The regulation of which remains an open issue. Finding an effective way of doing this is a top priority, for example, in the UK national surveillance camera strategy. See Surveillance Camera Commissioner (2017), *A National Surveillance Camera Strategy for England and Wales*, March 2017, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/608818/NSCS_Strategy_post_consultation.pdf (accessed 31 Jul. 2019).

In principle, the level of justification required to authorize surveillance activity should stay the same, irrespective of the practical potential or technical possibilities.[13] Despite this, it is incorrect not to recognize that the potential to expand surveillance may result in an irresistible draw to do so. The dissolution of practical constraints on the conduct of surveillance means that the question of how surveillance is directed at people of legitimate intelligence interest, rather than being applied indiscriminately, may become increasingly pertinent. In the near future, therefore, making good predictions about who or what should be watched could be central to limiting wholesale misuse of technical means of surveillance.

Terrorism itself is unpredictable by design. The instigation of terror is dependent on the notion that an attack will be conducted by an unknown person at a supposedly random place and time. Despite this, effective counterterrorism strategy is underwritten by the possibility of prediction. The question is not whether there is a place for prediction in countering terrorism, but whether and how the particular approach of using AI can make prediction better.

[13] Office of Surveillance Commissioners (2016), *OSC Annual Report for 2015-2016*, HC 214 SG/2016/66, printed 7 July 2016: House of Commons, p. 7, para 2.8.

3. Practical Applications of AI in Counterterrorism

This section describes how the contribution of predictive AI to countering terrorism has already been recognized and, to a limited extent, put to use. It introduces some of the approaches and actors involved.

Automated data analytics are used to support the activities of the intelligence and security services, particularly through data visualization.[14] Algorithms prioritize terrorist suspects,[15] and routinely assess the risk of air-travel passengers.[16] Information can be collected and stored by default, to be analysed at a later time with a view to revealing patterns and links that expose terrorist networks or suspicious activities.[17] Machine learning approaches allow the interpretation and analysis of otherwise inaccessible patterns in large amounts of data.[18] These approaches may involve filtering, analysis of relationships between entities, or more sophisticated image- or voice- recognition tools.[19]

Intelligence agencies and security services are not the only ones to both recognize and attempt to realize the predictive value of data. The result has been a devolution of methods and capabilities to civilian authorities (e.g. the police).[20] Social network analysis of urban gangs,[21] citywide alert systems,[22] crime-spot prediction,[23] and custody decision-making aids[24] are all examples of predictive tools underpinned by AI that are applicable to counterterrorism and that are already in use by law enforcement agencies.

Beyond this, commercial actors are also involved, sometimes as a result of governments demanding action from communication service providers to monitor and exclude terrorist activity on their own

[14] Van Puyvelde, D., Coulthart, S. and Hossain, M. S. (2017), 'Beyond the Buzzword: big data and national security decision-making', *International Affairs*, 93(26), pp. 1397–1416.
[15] In the case of the Manchester suicide attack, for example, an experimental algorithmic method categorized the bomber as a subject of interest worthy of closer investigation prior to the attack. Anderson, D. (2017), *Attacks in London and Manchester March–June 2017*, independent assessment of M15 and police internal reviews, December 2017, para 2.38, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/664682/Attacks_in_London_and_Manchester_Open_Report.pdf (accessed 1 Oct. 2018).
[16] Elias, B. (2014), *Risk-Based Approaches to Airline Passenger Screening*, Congressional Research Service Report, 31 March 2014, https://www.hsdl.org/?view&did=752251 (accessed 1 Oct. 2018); and US Department of Homeland Security (2013), *Passenger Name Record (PNR) Privacy Policy*, 21 June 2013.
[17] Akhgar, B., Saathoff, G. B., Arabnia, H., Hill, R., Staniforth, A. and Bayerl, P. (2015), *Application of Big Data for National Security*, Oxford: Butterworth-Heinemann.
[18] Van Puyvelde et al. (2017), 'Beyond the Buzzword: big data and national security decision-making', p. 1398.
[19] Weaver, M. (2016), 'Search for UK jihadi in Isis video to use voice and vein recognition software', *Guardian*, 4 January 2016, https://www.theguardian.com/world/2016/jan/04/isis-video-uk-jihadi-voice-vein-recognition-software (accessed 1 Oct. 2018).
[20] Galdon Clavell, G. (2016), 'Policing, Big Data and the Commodification of Security', in van der Sloot, B., Broeders, D. and Schrijvers, E. (eds) (2016), *Exploring the Boundaries of Big Data*, Netherlands Scientific Council for Government Policy (WRR) Report, Amsterdam: Amsterdam University Press, p. 91.
[21] Gunnell, D., Hillier, J. and Blakeborough, L. (2016), 'Social Network Analysis of an Urban Street Gang using Police Intelligence Data', Home Office Research Report 89, January 2016, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/491578/horr89.pdf (accessed 1 Oct. 2018).
[22] Levine, E. S., Tisch, J., Tasso, A. and Joy, M. (2017), 'The New York City Police Department's Domain Awareness System', *Interfaces* 47(1), pp. 70–84.
[23] For example, PREDPOL, in use by various police departments across the UK and US. See PREDPOL (2018), http://www.predpol.com/.
[24] Baraniuk, C. (2017), 'Durham Police AI to help with custody decisions', BBC Technology, 10 May 2017, https://www.bbc.com/news/technology-39857645 (accessed 1 Oct. 2018).

platforms.[25] Some technology companies employ a mix of human expertise and increasingly sophisticated predictive measures to monitor and disrupt terrorist activity on their platforms.[26] The future could see an increasing involvement by these kinds of technology companies in attempting to address terrorism themselves, rather than simply closing down unacceptable sites or user profiles.[27] The development and use of AI in the financial services sector has been spurred by mandatory reporting of suspicious activity in financial transactions.[28] Private-sector companies are also heavily involved in developing software and making datasets available for use by the public sector in the area of general law enforcement.

In the case of counterterrorism, investigative approaches applied prior to an attack taking place are traditionally undertaken by the intelligence and security services. Broadly speaking, these approaches focus on working outwards from partially discovered plots or known suspects in order to find other involved parties or to identify links leading deeper into terrorist organizations.[29] The granting of access to data related to a particular individual is contingent on their having a link to one or more existing investigations.

Distinct from this, and recently made plausible by developments in AI, is the analysis of all individuals' routine activity to predict terrorist events, or to identify terrorists by distinguishing what is distinct in the activity of a specific subgroup. The vast amount of digital information now generated by the average individual means that more of this routine activity could be understood through analysis. Sources include communications metadata and internet connection records, but also extend to location and activity tracking, purchases and social media activity. Much of this information is not in the hands of the intelligence and security services, meaning that the actors involved in exploiting it are diverse. Some narrow cases of use are described below.

Timing and location of attacks

Large amounts of effort, particularly from the academic community, have been devoted to developing models that predict the location and timing of terrorist attacks.[30] Basic approaches have incorporated the 'aftershock effect', whereby the chance of another event is increased in the wake of an attack (a phenomenon also observed with crimes such as burglary) to make surprisingly accurate predictions about terrorist attacks.[31] Other approaches have predicted the impact of external

[25] Stewart, H. (2017), 'May calls on internet firms to remove extremist content within two hours', *Guardian*, 20 September 2017, https://www.theguardian.com/uk-news/2017/sep/19/theresa-may-will-tell-internet-firms-to-tackle-extremist-content (accessed 1 Oct. 2018); Titcomb, J. (2017), 'Internet giants insist they are tackling terrorism, but it is right to demand more', *Telegraph*, 17 October 2017; Chazan, G. (2018), 'Twitter suspends top AfD MP under new German hate speech law', *Financial Times*, 2 January 2018, https://www.ft.com/content/19f89fb2-efc7-11e7-b220-857e26d1aca4 (accessed 1 Oct. 2018).
[26] Bickert, M. (2017), 'Hard Questions: How We Counter Terrorism', *Facebook Newsroom*, 15 June 2017, https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/ (accessed 1 Oct. 2018).
[27] Murphi, M. (2018), 'Facebook pays terror victims to talk down extremists on Messenger', *Telegraph*, 27 February 2018, https://www.telegraph.co.uk/technology/2018/02/27/facebook-funds-terror-victims-talk-extremists-messenger/ (accessed 1 Oct. 2018).
[28] Sadwick, R. (2018), 'Your Money Helps Fight Crime: Using AI To Fight Terrorism, Trafficking And Money Laundering', *Forbes*, 9 January 2018, https://www.forbes.com/sites/rebeccasadwick/2018/01/09/ai-money-laundering/#73fd22d22aec (accessed 1 Oct. 2018).
[29] Intelligence and Security Committee of Parliament (2014), *Report on the intelligence relating to the murder of Fusilier Lee Rigby*, House of Commons, 25 November 2014.
[30] Subrahmanian, V. S. (ed.) (2013), *Handbook of Computational Approaches to Counterterrorism*, New York: Springer.
[31] Dickerson, J. P., Simari, G. I. and Subrahmanian, V. S. (2013), 'Using Temporal Probabilistic Rules to Learn Group Behaviour', in Subrahmanian, V. S. (ed.) (2013), *Handbook of Computational Approaches to Counterterrorism*, New York: Springer.

factors – such as political conditions – on the incidence of attacks.[32] In 2015, for instance, the technology start-up PredictifyMe claimed that its model, computing more than 170 data points, was able to predict suicide attacks with an accuracy of 72 per cent.[33] It is not possible to validate this claim, and should be noted that the start-up in question subsequently collapsed. In other instances, however, sophisticated models based on open-source information that aim to predict various other types of events have achieved prescient results. Increasingly these models incorporate open-source data generated by individuals using social media and applications on their mobile phones. One such example is the Early Model-Based Event Recognition using Surrogates (EMBERS) system, which incorporates the results of various separate predictive models in order to forecast events such as disease outbreaks and civil unrest events. The project is a collaboration between the academic and business communities and is funded by the US Intelligence Advanced Research Projects Activity (IARPA)'s Open Source Indicators Program. Among other things, inputs include RSS feeds from news websites and blogs, Twitter feeds, events pages on social networking sites and restaurant booking applications.[34]

Vulnerability to radicalization

As technology companies increasingly assume duties related to safeguarding their users, tools that identify suicide risk or vulnerability to mental health issues have possibilities for repurposing as tools that could assess susceptibility to violent extremist ideologies.[35] Considering radicalization specifically, the Jigsaw subsidiary of Alphabet Inc. (formerly Google Ideas) has the stated vision of being 'an incubator [...] that builds technology to tackle some of the toughest global security challenges today',[36] for example by disrupting online radicalization and propaganda. Among Jigsaw's projects is the 'Redirect Method', which targets users of video-sharing sites who may be susceptible to propaganda from terrorist groups such as Islamic State of Iraq and Syria (ISIS), and redirects them to videos espousing a credible counter-narrative.[37]

[32] Choi, K., Asal, V., Wilkenfeld, J. and Pattipati, K. R. (2013), 'Forecasting the Use of Violence by Ethno-Political Organizations: Middle Eastern Minorities and the Choice of Violence', in Subrahmanian, V. S. (ed.) (2013), *Handbook of Computational Approaches to Counterterrorism*, New York: Springer.
[33] Lo, C. (2015), 'Safer with data: protecting Pakistan's schools with predictive analytics', *Army Technology*, 8 November 2015, http://www.army-technology.com/features/featuresafer-with-data-protecting-pakistans-schools-with-predictive-analytics-4713601/ (accessed 1 Oct. 2018).
[34] Ramakrishnan, N. et al. (2014), '*Beating the News' with EMBERS: Forecasting Civil Unrest using Open Source Indicators*, New York: KDD, 14 August 2014, http://people.cs.vt.edu/naren/papers/kddindg1572-ramakrishnan.pdf (accessed 1 Oct. 2018).
[35] Bhui, K., Everitt, B. and Jones, E. (2014), 'Might Depression, Psychosocial Adversity, and Limited Social Assets Explain Vulnerability to and Resistance against Violent Radicalisation?', *PLoS ONE* 9(9), https://www.kcl.ac.uk/kcmhr/publications/assetfiles/2014/Bhui2014.pdf (accessed 1 Oct. 2018); and the use by Facebook of AI to detect suicidal posts, reported by Constine, J. (2017), 'Facebook rolls out AI to detect suicidal posts before they're reported', *TechCrunch* blog, 27 November 2017, https://techcrunch.com/2017/11/27/facebook-ai-suicide-prevention/ (accessed 1 Oct. 2018).
[36] Jigsaw (2018), 'How can technology make people in the world safer?', https://jigsaw.google.com/vision/ (accessed 1 Oct. 2018).
[37] Jigsaw (2016). '*The Redirect Method*', www.redirectmethod.org (accessed 1 October 2018).

Identifying terrorists

Leaked details of the US National Security Agency's SKYNET, which was purportedly used in Pakistan in 2007, are useful in illustrating how quantitative methods might predict involvements in terrorism. As reported, the algorithm was used to analyse metadata from 55 million domestic Pakistani mobile phone users.[38] This was a machine learning model built by exposure to those data; it classified the phone users into two separate groups, one of which exhibited a usage pattern matching that of a small group of persons known to be terrorist couriers, the other comprising the remainder of the mobile phone users. The model was able to narrow the large population size down, and was reported to have falsely identified individuals as potential couriers in only 0.008 per cent of cases. It is important to note, however, that the scale of the initial dataset in relation to the total population of Pakistan (at that time approaching 200 million) implies that a false positive rate of 0.008 per cent would result in the wrongful identification of some 15,000 individuals as of interest. Furthermore, the 0.008 per cent false positive rate could only be achieved with a 50 per cent accuracy rate for identifying known couriers, meaning that half of the known couriers could be identified using the model. From these figures, it is obvious that the model used was not effective in its own right, but it shows how seemingly non-sensitive data may have predictive value when identifying close links with terrorism or likely intelligence value.

These limited examples of cases of the use of predictive AI in countering terrorism hint at the possibilities, rather than providing any credible proof of concept. It is not realistic to expect AI to provide immediate solutions to complex questions. US Immigration and Customs Enforcement discovered this when attempting to use machine learning models in data mining across various internet sources to assist with the vetting of visa applicants. The pursuit of a technical solution to this task was abandoned after it became clear that no such capability was available for immediate procurement.[39]

In summary, there are already numerous examples of AI that predict terrorism, or aspects of terrorism. Often, the ability to develop AI tools for this purpose rests with those who have access to data, or who are custodians of it by virtue of the service they provide. Where predictive AI is useful to police forces and other authorities (such as border enforcement agencies), its development is often outsourced to the software industry. Assuming that the trend of digitization continues, and that the performance of AI improves, there will be more scope to derive accurate predictions about terrorism from AI in future, and its uptake for counterterrorism use is likely to increase.

[38] Robbins, M. (2016), 'Has a rampaging AI algorithm really killed thousands in Pakistan?', *Guardian*, 18 February 2018, https://www.theguardian.com/science/the-lay-scientist/2016/feb/18/has-a-rampaging-ai-algorithm-really-killed-thousands-in-pakistan (accessed 1 Oct. 2018); The Intercept (2015), 'SKYNET: Courier Detection by Machine Learning', 8 May 2015, https://theintercept.com/document/2015/05/08/skynet-courier/ (accessed 1 Oct. 2018).
[39] Harwell, D. and Miroff, N. (2018), 'ICE just abandoned its dream of "extreme vetting" software that could predict whether a foreign visitor would become a terrorist', *Washington Post*, 17 May 2018, https://www.washingtonpost.com/news/the-switch/wp/2018/05/17/ice-just-abandoned-its-dream-of-extreme-vetting-software-that-could-predict-whether-a-foreign-visitor-would-become-a-terrorist/ (accessed 1 Oct. 2018).

4. Challenges Associated With the Current Use of Predictive AI in Counterterrorism

AI can be used to make predictions about terrorism based on communications metadata, financial transaction information, travel patterns and internet browsing activity, as well as publicly available information such as social media activity. The potential to prevent terrorist attacks may drive an impetus to increase the use of this technology. In 2017, for example, a UK Operational Improvement Review, completed in the wake of four terrorist attacks, called for a step change in the ability to 'exploit data to detect activity of concern' with regard to both subjects of interest previously known to security services and those unknown to them.[40]

A number of challenges exist or are arising regarding the use of predictive AI in countering terrorism. Its application should address two types of concerns: human rights problems; and practical implications.

Human rights problems

Lack of well-established norms for the use of AI technology

In 2014, the report of the Office of the United Nations High Commissioner for Human Rights (OHCHR) on the right to privacy in the digital age observed that 'examples of overt and covert digital surveillance in jurisdictions around the world have proliferated, with governmental mass surveillance emerging as a dangerous habit rather than an exceptional measure'.[41] AI could be used to facilitate mass surveillance; recent developments in video analysis and natural language processing indicate that future technologies are likely to make this even more possible with potentially lower entry costs.[42]

AI is widely becoming recognized as a technology that might deliver great benefits while at the same time posing significant risks, with numerous organizations devoting resources to exploring these.[43] One recent, multi-author report on the malicious use of AI has considered its impact on political freedom, including a possible future scenario of a predictive 'civil disruption' system designed to pre-empt threats before they come to fruition.[44] As regards sector-specific regulation on the use of

[40] Anderson (2017), *Attacks in London and Manchester March–June 2017*, p. 33, para 3.38.
[41] OHCHR (2014), *The right to privacy in the digital age*, http://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session27/Documents/A.HRC.27.37_en.pdf, Advance Edited Version, p. 3 (accessed 1 Oct. 2018).
[42] Amazon's Rekognition Application Program Interface is an example of this, making facial recognition software widely available at very low prices. See Amazon.com Inc. (2018), Amazon Rekognition website, Amazon WorkSpaces, https://aws.amazon.com/rekognition/ (accessed 1 Oct. 2018). This particular technology has been the focus of criticism after falsely matching the faces of members of Congress to those of people who had been arrested. Snow, J. (2018), 'Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots', 26 July 2018, https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28 (accessed 1 Oct. 2018).
[43] For example, the Future of Life Institute, the Leverhulme Centre for the Future of Intelligence, and the Centre for the Study of Existential Risk. Most directly related to the topic of this paper in particular is the Human Rights, Big Data and Technology Project, based at the University of Essex: see https://hrbdt.ac.uk/.
[44] Brundage, M. et al (2018), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, p. 28, https://maliciousaireport.com/ (accessed 1 Oct. 2018).

AI for public safety, consensus is yet to be found over where the line should be drawn on the use of these technologies, whether it is by intelligence and security services, law enforcement or the private sector.

Some of the rights and freedoms at stake are elaborated in the Universal Declaration of Human Rights, and consolidated in treaties such as the International Covenant on Civil and Political Rights (ICCPR). Adopted in 1966 and in force since 1976, the ICCPR has global reach, having 173 state parties by July 2019. It protects rights to privacy, freedom of thought and religion, expression and association.[45] Most pertinent to the subject of this paper is the UN General Assembly Resolution on the right to privacy in the digital age (68/167, adopted in December 2014). This resolution places the onus on states to ensure their activities comply with international law,[46] but different interpretations of concepts such as 'arbitrary' or even 'unlawful' interference with an individual's privacy mean that the resolution is relatively weak as a binding norm. Regional rights protection treaties[47] serve to strengthen this in some areas, particularly where they underwrite the authority of supranational courts. Aside from areas covered by the jurisdiction of courts with supranational authority, controls against the abuse of AI and the vast repositories of data on which it relies exist at state level. There are wide variations across the world in what is considered acceptable use of AI, in terms of both methods and purpose. China, for example, has been reported as embracing facial recognition and video behavioural analysis technology at an unprecedented scale, for a range of purposes that include identifying jaywalkers and apprehending wanted criminals at public events.[48]

The mass collection and retention of domestic data alone, on some form of which the use of predictive AI would likely be contingent, remains a hotly contested point. At the European Court of Human Rights (ECtHR), for instance, there are various pending cases that assert that states including Germany, Sweden and the UK violate the European Convention on Human Rights through their interception, retention and use of bulk data.[49] Various states have also had powers of data retention stripped back in rulings by the Court of Justice of the European Union (CJEU).[50] For example, in December 2016 the CJEU ruled against the UK's Data Retention and Investigatory Powers Act of 2014, stating that it failed to meet fundamental rights guaranteed by the EU Charter.[51] Legal challenges to the replacement legislation – the Investigatory Powers Act 2016 – are

[45] At articles 17, 18, 19 and 22. United Nations (1966). *International covenant on civil and political rights*, https://www.ohchr.org/EN/ProfessionalInterest/Pages/CCPR.aspx (accessed 10 Feb. 2019).
[46] United Nations General Assembly (2014), 'Resolution adopted by the General Assembly on 18 December 2013, 68/167. The right to privacy in the digital age', http://undocs.org/A/RES/68/167 (accessed 1 Oct. 2018). Co-sponsored by 54 nations, the resolution was adopted without a vote.
[47] For example, the European Convention on Human Rights, adopted in 1953 by 47 parties, the American Convention on Human Rights, adopted in 1978 by 24 parties (only 11 of whom have ratified the treaty), and the African Charter on Human and Peoples' Rights, adopted in 1986 by 54 state parties.
[48] Mozur, P. (2018), 'Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras', *New York Times*, 8 July 2018, https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html (accessed 1 Oct. 2018).
[49] See European Court of Human Rights cases: *Breyer v. Germany (on telecommunications companies having to retain data)*; *Centrum för Rättvisa v. Sweden*; *Big Brother Watch and others v. the United Kingdom* (on electronic surveillance); *Catt v. the United Kingdom* (on data retention); and *Association Confraternelle de La Presse Judiciaire v. France* (on French Intelligence Act 2015), https://www.echr.coe.int (accessed 1 Oct. 2018).
[50] Bulgaria, Czech Republic, Cyprus, Germany and Romania. See http://europeanlawblog.eu/2017/01/12/tele2-sverige-ab-and-watson-et-al-continuity-and-radical-change/ (accessed 1 Oct. 2018).
[51] Court of Justice of the European Union (2016), *Tele2 Sverige AB v. Post- och telestyrelsen and Secretary of State for the Home Department (United Kingdom of Great Britain and Northern Ireland) v. Tom Watson and Others*, Judgement of the Court (Grand Chamber), 21 December 2016 In Joined Cases C-203/15 and C-698/15, https://eur-lex.europa.eu/legal-content/EN/SUM/?uri=CELEX%3A62015CJ0203 (accessed 1 Oct. 2018).

already under way, and initial objections have been upheld by the domestic High Court leading to modification of the act.[52]

As with the question of limits on data collection, retention and analysis in general, there is no agreed position on the relative proportionality of automated analysis and human attention. Sometimes, the use of automated analysis has tended to exacerbate rather than assuage concerns. For example, use of automated analysis was cited as part of the reason why the transfer of Passenger Name Record (PNR) data to Canada by the EU was blocked by a CJEU opinion in July 2017.[53] This reflects the previous stipulation by that court that derogation and limitations in the protection of private data must only exist as far as strictly necessary, and that the need for safeguards to ensure this 'is all the greater where [...] personal data are subjected to automatic processing'.[54] Aside from this, there is scant international jurisprudence explaining the impact of automated analysis on proportionality.

The power to access, collect and store citizens' data brought about by the information age could represent a change in the relationship between states and citizens, and demands revision of the measures designed to safeguard not just privacy, but other freedoms critical to democratic functioning, such as those of expression and association.[55] The further developments of the digital age, such as the ability to automate large parts of the analysis of that data, reiterate this requirement.

In the use and limitation of these powers, it should be possible to draw a distinction between democratic and authoritarian states.[56] Democratic institutions have essential roles to play in setting and maintaining limits and in creating a clear, legislated context for any data privacy interference.[57] In the UK, the revision of the legislation governing the investigatory powers of the police and intelligence services in the UK – from the Regulation of Investigatory Powers Act 2000, to the Data Retention and Investigatory Powers Act 2014, and subsequently to the Investigatory Powers Act 2016 – is an attempt to achieve this. The April 2018 ruling by the domestic High Court upholding a challenge to parts of the most recent version of the act, and requiring them to be rewritten if they are to be legal, shows that adjustment is ongoing rather than resolved.

Inherent disproportionality

Predictive AI would rely on analysis of data belonging to the general public to distinguish suspicious from normal behaviour, or to discern trends that might help predict attacks. The vast majority of data under analysis would be generated by people who are not of interest to intelligence services.

[52] National Council for Civil Liberties (Liberty) (2017), 'Liberty gets go-ahead to challenge Snoopers' Charter in the High Court', press release, 30 June 2017, https://www.liberty-human-rights.org.uk/news/press-releases-and-statements/liberty-gets-go-ahead-challenge-snoopers%E2%80%99-charter-high-court (accessed 1 Oct. 2018) and Cobain, I. (2018), 'UK has six months to rewrite snooper's charter, high court rules', *Guardian*, 27 April 2018, https://www.theguardian.com/technology/2018/apr/27/snoopers-charter-investigatory-powers-act-rewrite-high-court-rules (accessed 1 Oct. 2018).
[53] Court of Justice of the European Union (2017), Press Release No 84/17, 26 July 2017, https://curia.europa.eu/jcms/upload/docs/application/pdf/2017-07/cp170084en.pdf (accessed 1 Oct. 2018).
[54] Court of Justice of the European Union (2014), Judgement of the Court, Grand Chamber 8 April 2014, Digital Rights Ireland vs Others, C-293/12 and C-594/12, para 55.
[55] Balkin, J. M. (2008), 'The Constitution in the National Surveillance State', Faculty Scholarship Series, Paper 225, http://digitalcommons.law.yale.edu/fss_papers/225 (accessed 1 Oct. 2018).
[56] Ibid., p. 17.
[57] Ignatieff, M. (2004), *The Lesser Evil: Political Ethics in the Age of Terror*, Edinburgh: Edinburgh University Press, p. 81.

Because of this, one of the specific areas of concern associated with predictive AI technology would be that it constitutes a surveillance measure applied to the whole population, and that this would render it indiscriminate and therefore inherently disproportionate. This has also been one focus of objections to states' bulk data collection and analysis programmes. While such programmes are not directly linked to the use of AI, the use of predictive AI for countering terrorism would be likely to rely on the ability to collect and analyse data in bulk.

Even the collection and analysis of seemingly non-sensitive data can still create a significant interference with privacy. UN General Assembly Resolution 68/167 on the right to privacy in the digital age describes the 'unlawful or arbitrary collection of personal data' as a highly intrusive act that could violate 'the rights to privacy and to freedom of expression and may contradict the tenets of a democratic society'.[58] This is because of the long-standing recognition that significant personal information can be inferred from this kind of data,[59] and that the aggregation and analysis of data constitutes an invasion of privacy even if actions are conducted in public space.[60] The controversy remains therefore, even if the data concerned are collected from open sources.

Some statutes of international law give guidelines for assessing whether any data privacy infringement is justifiable. The right to data privacy is protected most specifically under Article 8 of the EU Charter, and is also covered by the right to privacy in general at Article 8 of the European Convention on Human Rights. Because of this, the CJEU and the ECtHR have both passed relevant judgments. For both institutions, interference must meet a three-fold criterion of fulfilling a legitimate aim, being undertaken within a legal framework and satisfying conditions of necessity and proportionality.[61] Necessity in this context means that the measure is suitable, in that it has a causal relationship with the policy objective, and that it does not curtail more rights than is necessary given alternative options.[62] The CJEU has tightened this to a criterion of strict necessity, demanding that the surveillance is targeted; that it is limited or subject to differentiation based on the objective pursued;[63] that it is subject to precise and objective guidelines; and that adequate and effective safeguards against abuse are in place.[64]

Where state powers of data retention and analysis have been rescinded by the CJEU, which arbitrates on the application of the Charter of the European Union, it is because they have fallen short of these criteria. The CJEU made clear that its judgment against the UK on data retention can be interpreted as 'precluding national legislation which, for the purpose of fighting crime, provides for general and indiscriminate retention of all traffic and location data of all subscribers and

[58] United Nations General Assembly (2014), 'Resolution adopted by the General Assembly on 18 December 2013, 68/167. The right to privacy in the digital age'.
[59] European Court of Human Rights (1984), *Judgement in the case of Malone v. United Kingdom*, 2 August 1984, para 83–88.
[60] European Court of Human Rights (2001), *Judgement in the case of P.G. and J.H. v. the United Kingdom*, 25 September 2001.
[61] The ECtHR evaluates whether a measure of mass surveillance is necessary in a democratic society by considering 'whether the interference was in accordance with the law, pursued a legitimate aim or aims and was proportionate to the aim(s) pursued'. European Court of Human Rights Press Unit (2018), *Mass Surveillance – Fact Sheet*, https://www.echr.coe.int/Documents/FS_Mass_surveillance_ENG.pdf, p. 1. (accessed 10 Feb. 2019).
[62] This definition is taken from an influential judicial analysis of proportionality: UK Supreme Court (2013), *Judgment in Bank Mellat vs HM Treasury*, para 74, http://www.bailii.org/uk/cases/UKSC/2013/39.html (accessed 10 Feb. 2019). Necessity is a precondition of proportionality, and the first three steps of the fourfold process represent conditions of necessity. The first step is the presence of a legitimate aim, and the fourth is that the measure is proportionate, in that the benefits outweigh the costs of the infringement of the right.
[63] Court of Justice of the European Union (2015), *Judgement in the case of Schrems vs Data Protection Commissioner*, C-362/14 para 93.
[64] Court of Justice of the European Union (2014), *Judgement in the case of Digital Rights Ireland vs Others*, paras 54–55. Strict necessity is also described this way in the ECtHR case law: see European Court of Human Rights, Research Division (2013), *National security and European case-law*, paras 31–33, https://rm.coe.int/168067d214 (accessed 10 Feb. 2019).

registered users relating to all means of electronic communication'.[65] The UK's powers of data retention fell short of the criteria of necessity. Methods that collect and analyse all data might be considered disproportionate not because blanket retention is wrong in principle, but because it cannot be linked to a specific legitimate objective with a clear causal relationship to the policy.

A screening operation conducted in Germany in an effort to identify terrorist 'sleeper cells' provides an example of where a method involving the analysis of bulk data falls short of scrutiny regarding its ability to deliver a legitimate objective. The screening operation involved cross-referencing a group of individuals identified using static characteristics – such as religious affiliation – with other databases. It was deemed to be unconstitutional by Germany's Federal Constitutional Court on the grounds that it was interfering with the 'informational self-determination of people against whom no suspicion was present' in the absence of a specific threat.[66] The characteristics that were used to separate out these individuals were selected based on impressions from trends in previous attacks, which was not a sufficiently rigorous evidence base.

More recent judgments by the ECtHR shed further light on the legitimacy of bulk data collection and retention regimes. In September 2018 the court found the UK to have been in violation of the right to privacy due to the inadequate safeguards and oversight on its bulk data collection regime, but also held that 'In reaching this conclusion, the Court found that the operation of a bulk interception regime did not in and of itself violate the Convention'.[67] In June 2018 the ECtHR had held that Sweden's bulk interception of electronic signals did not excessively interfere with citizens' right to privacy, partly because it was limited to communication that crossed the Swedish border.[68] In this case, a relatively broad condition was accepted as being sufficiently limiting in the presence of independent oversight bodies and clear definition in law.

Legislation and governance frameworks for the use of bulk data tend to focus mainly on the powers of access to data, with limited description of how the data might be used, or the corollary considerations of how they are properly safeguarded from misuse. Without giving more detail on the use of data, it is difficult to prove necessity by clearly linking the retention of data to something that can be shown to be a legitimate objective. If AI is used to a greater extent, the framework under which it is used must prove – or otherwise guarantee – its contribution to a legitimate objective.

An expanding but weakly regulated private sector role

In countries where governments hold themselves to comparatively strict legal obligations, private companies have exercised significantly more discretion over how they use customer data. The way predictive AI is currently used in countering terrorism sees responsibility fall on a wide range of actors, primarily on the basis of their role as custodians of data.[69] This relative freedom is part of

[65] Court of Justice of the European Union (2016), *Tele2 Sverige AB v. Post- och telestyrelsen and Secretary of State for the Home Department (United Kingdom of Great Britain and Northern Ireland) v. Tom Watson and Others*.
[66] De Hert, P. and Lammerant, H. (2017), 'Predictive Profiling and its Legal Limits: Effectiveness Gone Forever?', in van der Sloot, B., Broeders, D. and Schrijvers, E. (eds) (2017), *Exploring the Boundaries of Big Data*, Netherlands Scientific Council for Government Policy (WRR) Report, p. 157.
[67] European Court of Human Rights, *Press Unit (2018) Mass Surveillance – Fact Sheet*, p. 4.
[68] Ibid.
[69] Characterized by one NGO as a 'delegation of law enforcement and quasi-judicial responsibilities to Internet intermediaries under the guise of "self-regulation" or "cooperation"'. McNamee, J. (2011), *The Slide from 'Self-Regulation' to Corporate Censorship*, European Digital Rights Discussion Paper, 24 January 2011, https://www.edri.org/files/EDRI_selfreg_final_20110124.pdf (accessed 1 Oct. 2018).

the reason why, to date, a significant proportion of publicized efforts to use AI in predicting terrorism has been initiated by technology companies.[70]

Certain companies, working within legal restrictions on the collection and sale of data, have made information about the public into a commodity, which the same companies can sell to law enforcement and security agencies. Governments might buy data that help them maintain security if they are unable to obtain them legally by other means, such as telecommunications interception.[71] Data are a valuable commodity, meaning that commitments by private companies to protect customer privacy are not always as principled as they may appear. While some technology companies have explicitly blocked governments from using data from their platforms,[72] these same data remain available to third-party companies; hence, such stipulations are predominately rhetorical in value. Pragmatism about access to large markets leads technology companies to renege on their stated principles in some of the states where user privacy and freedom are most at risk.[73] Moreover, there is limited transparency as to where and when they do this. At one extreme, in countries where there is greater government control of private industry, unlimited opportunities exist for governments to collect and analyse personal or sensitive information on their subjects.[74]

While the comparative flexibility in the private sector might afford opportunities, vesting too much responsibility in the hands of commercial companies could result in a growing deficit of accountability. Although some private companies make efforts at transparency, these are partial and are conducted at their own discretion.[75] Despite the fact that private companies do have an increasing role by virtue of their responsibilities to their users, it is questionable whether they should be relied on to counteract excessive infringement of privacy and freedom of expression by autocratic states, or to balance security from terrorism with right to privacy on behalf of democracies.[76]

Lack of redress

The inherent opacity of methods of using predictive AI in countering terrorism, or even in the use of retained bulk data in general, make it difficult to trigger legal guarantees of redress. The EU's General Data Protection Regulation (GDPR), one of the world's most far-reaching data protection regimes, attempts to balance this by securing a right of redress on any fully automated decision that

[70] Bickert, M. (2017), 'Hard Questions: Are We Winning the War on Terrorism Online?', *Facebook Newsroom*, 28 November 2017, https://newsroom.fb.com/news/2017/11/hard-questions-are-we-winning-the-war-on-terrorism-online/ (accessed 1 Oct. 2018).
[71] Cagle, M. (2016), 'Facebook, Instagram, and Twitter Provided Data Access for a Surveillance Product Marketed to Target Activists of Color', ACLU Northern California website, https://www.aclunc.org/blog/facebook-instagram-and-twitter-provided-data-access-surveillance-product-marketed-target (accessed 1 Oct. 2018).
[72] Twitter, for example, has blocked government agencies from accessing certain commercial social media analysis tools. See Levin, S. (2016), 'Twitter blocks government 'spy centers' from accessing user data', *Guardian*, 15 December 2016, https://www.theguardian.com/technology/2016/dec/15/twitter-dataminr-user-data-aclu (accessed 1 Oct. 2018).
[73] See, for example, Cook, J. and Archer, J. (2018), 'How Google reversed its course on opposition to censorship in China', *Telegraph*, 1 August 2018, https://www.telegraph.co.uk/technology/2018/08/01/google-reversed-course-opposition-censorship-china/ (accessed 1 Oct. 2018).
[74] In China, for example, data-generating companies China Rapid Finance and Ant Financial Services Group are reported to be among eight companies competing to generate the best algorithms with which to allocate scores to citizens as part of the country's social credit system. See Botsman, R. (2017), 'Big data meets Big Brother as China moves to rate its citizens', *Wired*, 21 October 2017, https://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion (accessed 1 Oct. 2018).
[75] Google, for example, produces an annual transparency report. See Google (2018), Google Transparency Report web page, https://transparencyreport.google.com/ (accessed 1 Oct. 2018).
[76] A European digital rights initiative paper observes: 'The dangers of extra-judicial policing and punishment by private companies have not been assessed with regard to fundamental rights. Furthermore, they have not been assessed with regard to their effectiveness for fighting crime.' McNamee (2011), *The Slide from 'Self-Regulation' to Corporate Censorship*, p. 3.

is made. However, it is unclear whether this criterion could be met by decisions that involve an arbitrary level of human oversight. Further, it contains an exemption for government agencies.

Systems such as the Investigatory Powers Tribunal (IPT) in the UK attempt to provide redress over the actions of government agencies, but are limited in so doing, with some procedures being conducted in secret.[77] Although the IPT has ruled against the government, indicating its independence, the system has been criticized for offering only limited redress, relying on trust in the institutions of government and being difficult for the general public to access.[78] However, because of the need to make a trade-off between transparency and operational security that is typical of counterterrorism methods in general, the IPT represents a compromise that is considered to be fair by many. Despite the inherent limits to transparency in countering terrorism, an important aspect of successful counterterrorism policy is maximizing transparency wherever possible. The use of AI is problematic if it poses – or is seen to pose – a threat to this.

Practical concerns

The ability of AI to achieve adequate predictive value

The low incidence of terrorism, and the tendency of terrorist tactics to evolve fairly rapidly, makes it difficult to build good predictive models. There are multiple different roles in terrorism, and multiple pathways to fulfilling those roles. This means that it is impossible to list definitive indicators of involvement or exclusion from terrorist activities.[79] While there appear to be some discernible trends in characteristics common to terrorists, the tiny number of terrorists within the general population renders broad characteristics based on profiling of no predictive value.[80] For example, the likelihood of an individual in the subgroup of male converts to Islam being convicted of terrorism is less than three-hundredths of a per cent, whereas more than 90 per cent of convicted terrorists are from outside that subgroup.[81] Machine learning models function well across large populations, but are poor predictors of individual behaviour.

[77] Investigatory Powers Tribunal (2018), website, http://www.ipt-uk.com/content.asp?id=12 (accessed 1 Oct. 2018).
[78] Keenan, B. (2015), 'Going "below the waterline", the paradoxical regulation of secret surveillance in the UK', LSE Law Policy Briefing, February 2015, http://eprints.lse.ac.uk/64056/1/Policy%20Briefing%209_2015.pdf (accessed 1 Oct. 2018).
[79] Borum, R. (2015), 'Assessing Risk for Terrorism Involvement', *Journal of Threat Assessment and Management*, 2(2), pp. 63–87.
[80] A study that presents a statistical analysis of UK Islamist terrorists between 1998 and 2015 notes some trends: for example, a higher proportion of converts being implicated in terrorist activity, compared with the overall proportion of converts to Islam in the UK Muslim population. Stuart, H. (2017), *Islamist Terrorism: Analysis of Offences and Attacks in the UK (1998-2015)*, Henry Jackson Society Report, http://henryjacksonsociety.org/wp-content/uploads/2017/03/Islamist-Terrorism-preview-1.pdf (accessed 1 Oct. 2018).
[81] Author's calculation, based on Stuart (2017), *Islamist Terrorism: Analysis of Offences and Attacks in the UK (1998-2015)*, p. 945.

The promise of more sophisticated AI is that it will be able to make better predictions based on behaviours in a range of different areas, rather than by crude profiling. As terrorist organizations conduct more activity online, research has shown that it might be possible to use AI to analyse communications and discern characteristics such as degree of radicalization,[82] aggressive intent,[83] or the genesis of terrorist movements,[84] and even to predict the incidence of violent attacks.[85] The implication is that whereas being able to predict terrorist involvement was previously impossible, this may no longer be the case. Accuracy of predictive models based on one source of data can be progressively improved by integrating the results from other feeds. The possibility to do this, however, is restricted most immediately by access to data.

Access to data

More data builds better models. The quality of models achievable is limited by restrictions on what types of data can be accessed and how those data can be used. For government agencies, limits are imposed based on national regulations, international human rights laws or physical access and technical capabilities. For this reason, the use examples given earlier in this paper relate to intelligence agencies operating outside their home nation, academia, or private sector actors with access to data by virtue of their roles as service providers. This section argues that some of the limits on access, as they stand, do not necessarily provide a guarantee of legitimacy, whereas limits on how data are used might do so better.

A common restriction on access to data centres on the differentiation between domestic and foreign data. Legal powers to analyse foreign communications are broader, and encompass 'discovery' capabilities to identify previously unknown threats.[86] In the past, where a distinction was made based on the existence of a foreign element within communications, this was more meaningful. As the academic Philip Bobbitt observed in 2008:

> Signals intelligence in the twentieth century meant intercepting analogue signals carried along dedicated voice channels. In the twenty first century, communications are mostly digital: they carry billions of bits of data, are dynamically routed in packets and are globally networked.[87]

Domestic communications may be routed through a server in a foreign country solely on the grounds of expediency. Moreover, while the antecedents of legal distinctions between foreign and domestic data lie in essential protection against politically motivated surveillance, changes to the nature of communications means that distinction based on the origin and destination of those communications is not necessarily able to provide those protections.

[82] Prentice, S., Rayson, P. and Taylor, P. J. (2012), 'The language of Islamic extremism: Towards an automated identification of beliefs, motivations and justifications', *International Journal of Corpus Linguistics*, 17(2), pp. 259–86, https://benjamins.com/#catalog/journals/ijcl.17.2.05pre/details (accessed 1 Oct. 2018).
[83] Pennebaker, J. W. (2011), 'Using computer analyses to identify language style and aggressive intent: The secret life of function words', *Dynamics of Asymmetric Conflict*, 4(2), pp. 92–102, doi: 10.1080/17467586.2011.627932 (accessed 1 Oct. 2018).
[84] Magdy, W., Darwish, K. and Weber, I. (2015), *#FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support*, Association for the Advancement of Artificial Intelligence, https://arxiv.org/pdf/1503.02401.pdf (accessed 1 Oct. 2018).
[85] Johnson, N. F. et al. (2016), 'New online ecology of adversarial aggregates: ISIS and beyond', *Science* 352(6292), 17 June 2016, http://science.sciencemag.org/content/352/6292/1459 (accessed 1 Oct. 2018).
[86] Intelligence and Security Committee of Parliament (2015), *Privacy and Security: A modern and transparent legal framework*, report presented to Parliament, printed on 12 March 2015, http://isc.independent.gov.uk/committee-reports/special-reports (accessed 1 Oct. 2018).
[87] Bobbitt, P. (2008), *Terror and Consent: The Wars for the Twenty-first Century*, London: Allen Lane, p. 307.

The US hosts a significant amount of global internet infrastructure on its territory.[88] This fact gives rise to the opportunity to have wide-ranging foreign interception capability. To comply with national commitments to protect citizens' privacy, automated minimization procedures exist to filter out domestic citizens' communications that are incidentally intercepted, without subjecting them to any further analysis.[89] Such activity implicitly accepts the automated analysis of domestic citizens' data (with a view to filtering them out), and allows even wider freedom over how foreign data are analysed.

The difference in treatment of foreign communications could cause problems with information sharing. For example, if the disparity in treatment of foreign and domestic citizens is interpreted as a discrepancy in respect for the privacy of those two groups, it seems understandable that communications service providers might be reluctant to uphold these unequal standards against citizens of their home countries. Low compliance rates by communications service providers to law enforcement access requests from other countries might be explained by this.[90] At the national level, close intelligence relationships between states could either be inhibited by these restrictions, or undermine them altogether by allowing the sharing of information derived from infringements of the other states' citizens' privacy.

Government agencies are limited by their technical capability to intercept communications. The UK Government Communications Headquarters (GCHQ), for example, is believed by many to have sweeping intercept capability,[91] but it is still only able to access some of the international communications that enter or leave the UK. Subsequently, GCHQ can only analyse a small percentage of these.[92] If data are obtained through these methods to contribute to predictive models, technical capability is an arbitrary limitation rather than a safeguard of legitimacy. Better access arrangements with communications service providers, predicated on legitimacy of practices and methods involved, would be preferable. As it stands, compliance with such requests from companies based in other countries is unreliable, leaving coverage incomplete.[93]

Given the dependence on data access in order to make good predictive models, restrictions on the ability to access and exploit data in turn restrict the extent to which predictive AI can be used in counterterrorism. The restrictions can be technical or legal, or can be due to a lack of trust between actors involved which thus prevents the sharing of information. Legal restrictions are necessary to prevent abuse, but restrictions that focus solely on access to data have a detrimental impact on the quality of any models that might be developed, while failing to address the need for specific standards governing how data can legitimately be used. Both access to data and the way those data are used must be lawful. Because legality of access to data is dependent on conditions of necessity, these are interlinked. Access to data is justified depending on how it is used.

[88] For example, submarine cables – see https://www.weforum.org/agenda/2016/11/this-map-shows-how-undersea-cables-move-internet-traffic-around-the-world/, internet exchange points – see http://www.datacentermap.com/ixps.html, and data centres – see https://www.cbronline.com/data-centre/top-10-biggest-data-centres-from-around-the-world-4545356/ (accessed 1 Oct. 2018).
[89] National Security Agency (2007), *Minimization Procedures used by the NSA in connection with acquisitions of foreign intelligence*, declassified 18 November 2013.
[90] A problem highlighted in the report to the Intelligence and Security Committee of Parliament (2015), *Privacy and Security: A modern and transparent legal framework*, p. 158.
[91] OHCHR (2014), *The right to privacy in the digital age*, p. 3.
[92] Intelligence and Security Committee of Parliament (2015), *Privacy and Security: A modern and transparent legal framework*, p. 4.
[93] Intelligence and Security Committee of Parliament (2014), *Report on the intelligence relating to the murder of Fusilier Lee Rigby*, p. 135.

Lack of common validation standards

While access to data is important, it cannot in itself guarantee success in building accurate predictive models. Validation and testing are essential to measuring the predictive accuracy of models and assessing the proportionality of their use.

One salient lesson from experience in the related field of predictive policing is to reiterate the requirement for systems to be validated and tested. The fairness of various commercially developed algorithms for predictive policing – or of those that inform legal decisions – has been called into question, which could have been avoided by employing a more thorough validation process prior to their use.[94] One study, published in 2013, of different types of offender risk-assessment software used to inform bail decisions in the US yielded the unexpected finding that:

> There were very few U.S. evaluations examining the predictive validity of assessments completed using instruments commonly used in U.S. correctional agencies. In most cases, validity had only been examined in one or two studies conducted in the United States, and frequently those investigations were completed by the same people who developed the instrument.[95]

This indicates an absence of standards that are systematically applied, and a tendency to trust the assessments of commercial software providers in evaluating their own products.

A recent body of research is concerned with the vulnerability of AI to bias, and debunks the notion that the results of predictive AI might be considered scientifically objective.[96] Avenues of bias are multiple; they may be inherent in the data on which the AI is trained,[97] or in aspects of the design, such as the presence of hidden feedback loops,[98] in the use of culturally mediated factors, such as assumptions based on travel to certain countries, or in differential treatment of subgroups.[99] An algorithm that appears to be performing effectively may be exhibiting bias, and even exacerbating that bias in the real world. To mitigate this, in-depth evaluation of predictive models must include assessments of the quality of the model across different subgroups.[100]

Given the wide range of actors involved, common standards become increasingly important. The uptake of techniques more traditionally used by intelligence agencies in law enforcement[101] and the use of AI by the private sector mean that there are an increasing number of actors involved in the development and use of predictive AI for countering terrorism.

---

[94] For example, see Larson, J., Mattu, S., Kirchner, L. and Angwin, J. (2016), *How We Analyzed the COMPAS Recidivism Algorithm*, 23 May 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed 1 Oct. 2018), which showed differing accuracy rates for black and white defendants rated as high or low risk of reoffending. Black defendants were twice as likely to be wrongly classified as high risk, whereas white defendants were twice as likely to be wrongly classified as low risk. Also see Ensign, D., Friedler, S., Neville, S., Scheidegger, C., Venkatasubramanian, S. (2017), 'Runaway Feedback Loops in Predictive Policing', *Proceedings of Machine Learning Research*, 81, pp. 1–12, 2018, https://arxiv.org/pdf/1706.09847.pdf (accessed 1 Oct. 2018).
[95] Desmarias, S. and Singh, J. (2013), *Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States*, Council of State Governments Justice Centre Report, 27 March 2013, p. 2.
[96] O'Neil, C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, London: Penguin; Osoba and Welser (2017), *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*.
[97] Sweeney, L. (2013), *Discrimination in Online Ad Delivery*, 28 January 2013, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240 (accessed 1 Oct. 2018).
[98] The widely used PREDPOL algorithm, which identifies areas where crimes are likely to take place, features a hidden feedback loop, which renders an area more likely to be a crime hotspot by directing police attention to it. See Ensign, D. et al. (2017), 'Runaway Feedback Loops in Predictive Policing'.
[99] Larson, J. et al. (2016), *How We Analyzed the COMPAS Recidivism Algorithm*.
[100] Some methods of doing so are explored in Hardt, M., Price, E. and Srebro, N. (2016), 'Equality of Opportunity in Supervised Learning', *Computing Research Repository*, 11 October 2016, https://arxiv.org/pdf/1610.02413.pdf (accessed 1 Oct. 2018).
[101] Galdon Clavell (2016), 'Policing, Big Data and the Commodification of Security', p. 93.

Currently, the range of different approaches leads to some unintuitive outcomes, specifically the deletion of data on individuals who are of genuine intelligence interest alongside the blanket retention of data on individuals who are not.[102] Private-sector actors policing content on their own platforms take down sites or close accounts, thereby also deleting potentially valuable intelligence information that is genuinely likely to be linked to criminal or terrorist activity. In the case of the murder of Fusilier Lee Rigby in the UK in May 2013, a social media platform that uses AI in identifying accounts associated with terrorism (*inter alia*) had closed 11 accounts linked to one of the attackers. One of these accounts contained an exchange described by the parliamentary Intelligence Select Committee as the single piece of evidence that could have prevented the attack.[103] This is something against which the existence of common validation standards could mitigate, by improving information sharing based on trust.

There are efforts at the national and international level to harmonize overall standards and encourage the development of beneficial counterterrorism tools. Examples include the UN-sponsored Tech Against Terrorism project,[104] and the extremism blocking tool developed by the UK government for use by tech companies that do not have the resources to develop their own tool.[105] Thus far, these measures are advisory or limited in scope. In all sectors, the increasing use of predictive AI should go hand in hand with the implementation of appropriate validation and testing regimes. As use of predictive AI in countering terrorism is nascent, the absence of attempts to give assurance of the existence and enforcement of validation standards may be understandable. If predictive AI is used more widely in countering terrorism in the future, failing to prove that predictive models are validated could continue to undermine transparency and information sharing.

Performance in adversarial environments

Recent research in image recognition has shown that small changes to images can completely alter how a machine perceives those images, while a human may continue to perceive their original content.[106] Furthermore, there are well-documented examples of certain images having been wrongly classified at high confidence levels.[107] If an adversary is deliberately trying to deceive an image recognition application, there are effective ways to do this.[108]

[102] McNamee, J. (2018), *Europol: Delete criminals' data, but keep watch on the innocent*, EDRi blog, 27 March 2018, https://edri.org/europol-delete-criminals-data-but-keep-watch-on-the-innocent/ (accessed 1 Oct. 2018).
[103] Intelligence and Security Committee of Parliament (2014), *Report on the intelligence relating to the murder of Fusilier Lee Rigby*, pp. 131–32.
[104] See https://www.techagainstterrorism.org/.
[105] Lee, D. (2018), 'UK unveils extremism blocking tool', BBC Technology, 13 February 2018, http://www.bbc.co.uk/news/technology-43037899 (accessed 1 Oct. 2018).
[106] Hosseini, H., Xiao, B. and Poovendran, R. (2017), 'Google's Cloud Vision API Is Not Robust To Noise', Cornell Research Repository, 20 July 2017, https://arxiv.org/pdf/1704.05051.pdf (accessed 1 Oct. 2018).
[107] Examples include an image recognition AI being 'tricked' into classifying a panda as a gibbon in Goodfellow, I., Schlens, J. and Szegedy, C. (2015), 'Explaining and Harnessing Adversarial Examples', Cornell Research Repository, 20 March 2015, https://arxiv.org/abs/1412.6572 (accessed 28 July 2019); and a turtle as a rifle in Athalye, A., Engstrom, L., Ilyas, A. and Kwok, K. (2018),'Synthesizing Robust Adversarial Examples', *Proceedings of Machine Learning Research*, 80, pp. 284–293, http://proceedings.mlr.press/v80/athalye18b.html (accessed 28 July 2019). Also, for examples of patterns as various different animals or objects at high confidence: Nguyen, A., Yosinski, J. and Clune, J. (2015), 'Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images', Cornell Research Repository, 12 April 2015, https://arxiv.org/abs/1412.1897 (accessed 1 Oct. 2018).
[108] Goodfellow et al. (2017), 'Attacking Machine Learning with Adversarial Examples'. See also Sharif, M., Bhagavatula, S., Bauer, L. and Reiter, M. K. (2017), 'Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition', Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf (accessed 1 Oct. 2018).

This vulnerability to adversarial action is an ongoing and unsolved problem in AI research. It is of particular concern when using AI in a security context. Adequate human oversight may go some way towards mitigating vulnerabilities, but in circumstances in which AI is being used to conduct a task that humans cannot do – such as match a behaviour pattern in perhaps millions of examples – it is impossible to back up the machine's performance with human oversight simply because of the limitations on human time. Even in areas where it is feasible to implement direct human oversight, the quality of this oversight depends on factors ranging from the degree of operator training to how results are presented. How human oversight is delivered requires in-depth, system-specific consideration. In most cases it will mean that AI will never be a full solution to a security problem: merely an additional tool that can augment, rather than replace, the role of humans in tackling that problem.

5. Can Predictive AI Be Used Legitimately as a Tool for Detecting Terrorism?

The challenges described above have resulted from the unsystematic development of AI as a means of predicting terrorism by various agencies operating under different regulatory frameworks, none of which have been specifically designed with that purpose in mind.

This section considers whether AI can be used legitimately as a tool of the state for predicting terrorism. This would mean giving government agencies broader access to public data – including data belonging to domestic citizens – and accompanying that access with clearer regulation on how the data are used. Within such a framework, private-sector activity to counter terrorism might be a statutory requirement, of which the government would have oversight. Any duty to report associated with this and other powers of bulk data access would be directly linked to the use of predictive AI.

This kind of approach would require the acceptance of wholesale collection and analysis of portions of domestic citizens' data, something that liberal democracies do not currently entertain for the purposes of countering terrorism. The opportunities opened by such an approach, and the risks posed by it, are explored in more detail below.

Opportunities

Automation changes proportionality assessments

Thus far, the question of how automated analysis (versus human observation) changes proportionality assessments has not been definitively addressed. Broader rights of access to data by law enforcement and security services could be granted for the purpose of automated analysis, based on the fact that this is less intrusive than human observation.

Returning to the criteria mentioned in the section on human rights concerns in the previous chapter, the necessity and proportionality of using a predictive model to automate analysis of public data would require proof of effectiveness (securing a causal relationship with the policy objective), as well as the absence of alternative options, and proof that the benefits outweigh the costs in terms of rights infringement.

Addressing the latter criterion requires an evaluation of the scale of rights intrusion associated with analysis of data. The idea that automated analysis, when compared with human analysis, might fall short of infringing the right to privacy is an argument that is tacitly accepted by many in the private sector.[109] It should be noted that this has not yet been actively put forward to meet concerns over

[109] Roettgers, J. (2017), 'Google Will Keep Reading Your Emails, Just Not for Ads', *Variety*, 23 June 2017, http://variety.com/2017/digital/news/google-gmail-ads-emails-1202477321/ (accessed 1 Oct. 2018).

the proportionality of the use of automated analysis by governments, but it is semi-implicit in the acceptance of bulk communications interception.[110]

For the majority of the population (whose information is filtered out before further analysis), meeting a criterion of proportionality requires acceptance that an automated process of collecting, analysing and deleting the majority of data is non-intrusive; or that it at least constitutes less intrusion than would the storage of that data for later analysis by a human. This is not a straightforward question, in that the type of data and the nature of analysis make a difference in such judgments. Accepting arguments about the proportionality of automated methods depends on standards being in place for their use; it also depends on the ability to share assessments of confidence in the application of those standards.

At present, the default approach is blanket data retention, which might be used as a comparator in assessing proportionality. Early automated analysis could remove the requirement to retain data that is not likely to be of intelligence interest, and improve compliance with the principle of data minimization. Access to redress might apply to those whose data were retained for no reason; but in comparison with having their data retained anyway, further infraction of privacy would be minimized.

The impact on individuals falsely identified as of interest by a given predictive model is another important factor in assessing proportionality. This is inextricably linked to the quality and accuracy of any model used. High false positive rates are likely to render an automated model disproportionate if they necessitate human scrutiny for an unreasonably high number of these falsely identified individuals. The false positive problem may be intractable, or it may improve. An important factor, compared with current approaches, is that it can be measured and compared. This brings the possibility of basing proportionality assessments on quantitative tools as well as qualitative judgments.

Alongside the number of times predictive AI is wrong, the nature of the intervention mounted on the basis of any prediction has a bearing on proportionality. A highly critical evaluation of the efficacy of a predictive programme implemented by the Chicago Police Department in 2013 to identify vulnerable individuals focused on how poor the related interventions were, rather than the quality of predictions.[111]

The output of predictive AI should never constitute a standard of proof in its own right; rather, it should be a cue for human attention. That attention might then authorize and direct activity that might or might not generate an evidence base for any coercive or restrictive measures subsequently imposed. While not having direct human oversight at the earlier stages of automated analysis would be an essential condition for securing proportionality, this subsequent level of human oversight would be equally essential to safeguarding proportionality later on in the process. This is reflected

[110] The proportionality of which may be considered contingent on the automatic deletion of a significant percentage of the communications. Intelligence and Security Committee of Parliament (2015), *Privacy and Security: A modern and transparent legal framework*, paras 65–73, p. 3; and National Security Agency (2007), *Minimization Procedures used by the NSA in connection with acquisitions of foreign intelligence.*
[111] Saunders, J. (2016), 'Pitfalls of Predictive Policing', RAND Corporation blog, https://www.rand.org/blog/2016/10/pitfalls-of-predictive-policing.html (accessed 1 Oct. 2018); Saunders, J., Hunt, P. and Hollywood, J. S. (2016), 'Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot', *Journal of Experimental Criminology* 12(3), pp. 347–371, https://link.springer.com/article/10.1007%2Fs11292-016-9272-0 (accessed 1 Oct. 2018).

in the guarantees of redress over entirely automated decisions provided by recent regulation such as the EU's GDPR. Under that regulation, automated decisions must have legal or similarly significant effect in order to qualify for any such guarantees. The implication is that significant interventions require human oversight, while less consequential actions do not.[112]

Automated analysis could assuage concerns about proportionality, provided the right controls on use – such as adequate human oversight at the appropriate time, and standards on the quality of model used – are in place.

Quantitative evaluation of methods to prove adequacy

The use of automated methods gives the opportunity to assess the performance of models with a view to quantitatively proving their effectiveness. Metrics such as recall (i.e. the percentage of times a model correctly identifies an event or an individual) and precision, or the number of false alarms, could be used to contribute to proportionality assessments – specifically to prove or disprove the link to a legitimate aim, described above as the second criterion of proportionality. Other metrics include ways of modifying models so that performance across different subgroups mitigates discrimination.[113]

Quantitative validation as a prerequisite for use delineates between a working predictive model and a 'fishing expedition'.[114] Approaches that give a detailed analysis of the quality of a particular predictive model might provide a more reassuring justification for its use than that currently available for the wholesale retention of data.[115]

Another area for analysis that is relevant to proportionality is the impact of different factors on the performance of a model. This allows a clearer decision to be made about whether any privacy intrusion in the collection and retention of data comprising those factors is justified. Discussing predictive policing, RAND Corporation analyst John Hollywood comments that 'increases in predictive power have tended to show diminishing returns'[116] because the relationships underpinning crime, especially those linked to location and timing, are relatively simple and do not feature complex non-linear relationships which AI can best be used to exploit.[117] There is a limit to the amount of useful prediction that can be extracted from data, and including more data does not necessarily improve the quality of a prediction. Validation and testing extend to analysing whether extra data or factors are actually increasing that predictive value.

Although validation is essential, it remains a developing field. There is no unified or agreed way of ensuring that a model is fair. Continued work on objective validation is of paramount importance.

[112] Information Commissioner's Office (2018), 'What does the GDPR say about automated decision-making and profiling?', https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-does-the-gdpr-say-about-automated-decision-making-and-profiling/ (accessed 1 Oct. 2018)
[113] Hardt et al. (2016), 'Equality of Opportunity in Supervised Learning'.
[114] A phrase used by a German court in handing down its judgment on a data screening operation. See De Hert and Lammerant (2017), 'Predictive Profiling and its Legal Limits', p. 157.
[115] The former UK Information Commissioner Christopher Graham lamented the lack of such an approach at a select committee hearing on the draft Investigatory Powers Bill in January 2016. See https://parliamentlive.tv/Event/Index/ac0e244b-5748-4348-97c2-0e7f2bcdfd6c (accessed 1 Oct. 2018).
[116] Hvistendahl, M. (2016), 'Can Predictive Policing prevent crime Before it happens?', *Science*, 28 September 2016, https://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens (accessed 1 Oct. 2018).
[117] Ibid.

Quantitative measures of performance will be required to contribute to proof of necessity. While they will never be able to make proportionality judgments entirely, they may mean that those judgments are better informed.

A single system

Regulation of new capabilities tends to lag behind their adoption, and the decentralization of capabilities exacerbates this.[118] Centralizing analysis, capabilities and source data, or alternatively the results of analysis conducted by different agencies under the same regulatory regime, could make their activities easier to regulate coherently and improve the ability to direct appropriate resources to individuals either at risk or of concern.

For good reasons, the idea of treating people who are at risk of radicalization and those who are actual terrorists within the same system causes discomfiture. If that system is sufficiently discerning, a meaningful distinction between these subgroups could not only be maintained but could also be proved. If we imagine that, in the future, predictive systems based on the analysis of digitally generated or stored data could be seen as neutral, genuinely trustable and effective, they could be used as a better way of directing earlier, non-coercive interventions (such as those undertaken to deter young people from being attracted to violent extremist ideologies) than can be achieved by current systems, that place reporting requirements that can jeopardize trust between individuals. In order to decide whether this kind of surveillance could ever be palatable, it is critical to understand whether the problems arise from the very idea of preventing terrorism through targeting individuals, rather than from the methods involved in identifying and managing these individuals.

While centralization may have benefits in terms of improving overall accuracy, it also has drawbacks, and may miss some protections that naturally come with compartmentalization and the development of assured separate sources to cross-validate results. Notably however, barriers to intelligence sharing have already undermined much of this potential. Moreover, having one system or set of standards does not necessarily mean having one vast reservoir of data. Models could instead be developed and applied to different private datasets, and the results of those models could be tested and integrated centrally.

Barriers to information sharing between intelligence and law enforcement agencies exist to prevent the misuse of powers that should be confined to the security services. It is worth considering at what point these and other control measures might be replaced with technical safeguards – such as audit and access records – as well as limitations on the types of access that different users have.

[118] The exploitation of mobile phone capabilities by police forces provides an example. See Williams, O. (2018), 'Police forces slammed for harvesting people's mobile data', *NS Tech*, 27 March 2018, http://tech.newstatesman.com/news/police-forces-mobile-data (accessed 1 Oct. 2018).

The potential for transparency

It might be possible to release quantitative measures of the performance of models without jeopardizing operational security. Making this information available would improve oversight and could go some way to addressing grievances over methods that are otherwise seen as discriminatory. For example, if individuals can be identified for in-depth 'stop and search' procedures at airports based on a model that integrates suspicious travel patterns, unusual payment methods and physiological signs of uneasiness, and this has demonstrably led to a 50 per cent reduction in the total number of people who are subjected to such searches, travellers could perhaps be informed of this fact to allay concerns about the possibility that such procedures are solely based on profiling with a racial or ethnic bias.[119] The same might apply for other measures, such as the placement of automatic number-plate recognition (ANPR) cameras. Clearly, any such transparency measures would have to be mediated with security concerns, but if the standards to be shared evaluate overall performance, rather than describing a method in detail, this might be achievable.

Sharing quantitative standards could also facilitate trust and set the conditions for better information sharing between international actors and agencies. For example, foreign communications service providers might be more amenable to requests for access if they could be assured that standards for these requests are sufficiently high. Such assurance measures might take the form of proving that access is linked to a legitimate aim: that of showing how models are audited, and that interventions based on their results are proportionate.

Quantitative standards provide a measure of technical transparency. Also important is the transparency of the legal framework within which predictive analysis is undertaken. Centralizing responsibility for overseeing and acting on this analysis with government agencies improves the potential to achieve a clear and coherent understanding about what analysis is undertaken and why.

Risks

Technical possibility

Irrespective of how sophisticated AI becomes, or of how much data is made available, predicting involvement in terrorism, the incidence of attacks or vulnerability to radicalization with sufficient accuracy may prove to be beyond the bounds of possibility.

Methods would have to be proposed, tried and assessed in advance – and retired if sufficient predictive value was not obtainable. Many examples of machine learning AI that have been so successfully used in industry work by improving their predictive capabilities as they are used. To some extent, the benefits of machine learning-based approaches are linked to their ability to dynamically adapt to the system they are analysing, including adapting to changes to background data within that system. Imposing standards of validation before use, and incrementally updating

[119] Noting that there are obvious difficulties in expunging bias derived from the inclusion of culturally mediated factors in models that are inherent to the data itself, such as training on a majority ethnic background, as described above in the section on the lack of common validation standards.

models based on new data, rather than allowing them to dynamically adjust, may limit their utility. This limitation would arguably have to be accepted as a guarantor of fair use.

The requirement to develop and test models would also mean that historical datasets would have to be available for experimentation, constituting a compromise of data privacy with no immediate causal link to a legitimate aim.

This might be acceptable, but only if there was a reasonable chance that a useful predictive model could be developed. A decision about when and whether to accept experimentation of this form would be a prerequisite for concerted use of predictive AI for counterterrorism purposes.

The core problem of identifying terrorists or predicting terrorism in other ways cannot be solved by using computational methods alone. Coupled with some of the inherent technical limitations identified earlier, such as vulnerability to adversarial action and lack of contextual understanding, this places definitive limits on how AI should be used in predicting terrorism. Specifically, the results of any predictive system provide possibility, not proof, and careful attention should be given to where that system needs to supplement, rather than replace, existing methods.

The societal effects of mass surveillance

Proportionality, as discussed so far in this paper, has considered privacy intrusion at the individual level. However, the right to privacy at this level underpins other rights, such as those of freedom of expression and association, and has beneficial social effects.[120]

Many academics have examined the way widespread digital surveillance could cause a 'chilling effect' on engagement with sensitive political issues or activities, dissenting opinion and critical thought, as individuals conform to perceived collective norms.[121] The issue has been a concern of civil society groups, particularly following revelations (commonly referred to as the Snowden revelations) in 2013 about the scale of national surveillance programmes.[122] Psychologists have observed that the mere perception of being watched – even one based on a simple trigger like a picture of a pair of human eyes – can lead to discernible changes in behaviour.[123] While automated analysis might reduce privacy intrusion at the individual level, it also gives rise to the possibility that everybody might feel as if they are being watched at all times. Because of this, automated analysis could have higher costs to other rights, such as rights of expression and association, which would be felt most acutely when individual behavioural changes are aggregated at the societal level.

[120] Solove, D. J. (2007), '"I've Got Nothing to Hide" and Other Misunderstandings of Privacy', *San Diego Law Review*, 44, p. 745.
[121] Kaminski, M. E. and Witnov, S. (2015), 'The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech', *University of Richmond Law Review*, 49; Ohio State Public Law Working Paper No. 288, https://ssrn.com/abstract=2550385 (accessed 10 Feb. 2019).
[122] Human Rights Watch (2014), *With Liberty to Monitor All: How Large-Scale US Surveillance is Harming Journalism, Law and American Democracy*, https://www.hrw.org/sites/default/files/reports/usnsa0714_ForUPload_0.pdf (accessed 10 Feb. 2019).
[123] Goldman, J. G. (2014), 'How being watched changes you – without you knowing', BBC Future, http://www.bbc.com/future/story/20140209-being-watched-why-thats-good; Nettle, D., Nott, K. and Bateson, M. (2012), '"Cycle Thieves, We Are Watching You": Impact of a Simple Signage Intervention against Bicycle Theft', *PLoS ONE* 7(12), doi: https://doi.org/10.1371/journal.pone.0051738 (accessed 10 Feb. 2019).

Scepticism still exists as to the chilling effects of surveillance on online activity,[124] although a number of studies have attempted to document these as an empirically observable phenomenon.[125] Historical experience of the behaviour of citizens under totalitarian regimes provides compelling evidence of chilling effects, although it is impossible to disaggregate the impact of surveillance *per se* with that of other measures designed to induce fear and force changes in people's behaviour. More recently, commentators have held up the continuance of political dissent and free speech after the 2013 revelations, and the increase in voluntary information sharing via social media, as evidence of the absence of meaningful chilling effects. Proponents of the chilling effect of surveillance acknowledge that it is not possible to anticipate exactly what the nature of that effect will be, or to measure it.[126] It is also possible that the most drastic effects are not in response to government observation, but to the scrutiny of peers and competitors.[127]

Although it seems logical that the possibility of ubiquitous observation that advances in automation may soon offer could have chilling effects, whether or not they will do so is definitely speculative at this stage, and it is beyond the scope of this paper to try and project those effects. It will be incumbent on governments to pre-empt these possible impacts when they design suitable legal controls on surveillance activity. Addressing this future concern might require the development of models that assess the wider harms of individual privacy infringements based on their societal impact.[128]

Erosion of standards of proof

While dealing with evidentiary standards below the level of definite proof is already a characteristic of counterterrorism operations, concerns exist that 'the growing availability of data might lead to an erosion of existing standards, rendering police and governmental interventions legitimate because the computer "said so"'.[129] The potential offered by the ability to collect and use data may be coupled with a lower tolerance of risk and uncertainty at the societal level, which could lead to a greater focus on prevention – and even on pre-emption – than previously existed.[130] In the extreme, some commentators worry that:

> Deploying new predictive technologies makes it possible to gather specific knowledge of the future. According to this logic, authorities can punish and intervene before a crime happens.[131]

[124] Kendrick, L. (2012), 'Speech, Intent and the Chilling Effect', *William & Mary Law Review*, 54, Virginia Public Law and Legal Theory Research Paper, No. 2012-37, https://ssrn.com/abstract=2094443 (accessed 10 Feb. 2019); Sklansky, D. A. (2014), 'Too Much Information: How Not to Think About Privacy and The Fourth Amendment', *California Law Review*, 102(5), pp. 1097–8.
[125] Penney, J. (2017), 'Internet surveillance, regulation, and chilling effects online: A comparative case study', 27 May 2017, *Internet Policy Review*, 6(2), https://policyreview.info/articles/analysis/internet-surveillance-regulation-and-chilling-effects-online-comparative-case (accessed 10 Feb. 2019); Rainie, L. and Madden, M. (2015), 'Americans' Privacy Strategies Post-Snowden', Pew Research Center Report, http://www.pewinternet.org/2015/03/16/Americans-Privacy-Strategies-Post-Snowden/ (accessed 10 Feb. 2019).
[126] Penney, J. (2018), '(Mis)conceptions About the Impact of Surveillance', https://freedom-to-tinker.com/2018/02/14/misconceptions-about-the-impact-of-surveillance/ (accessed 10 Feb. 2019).
[127] Townsend, J. (2014), 'Online chilling effects in England and Wales', *Internet Policy Review*, 3(2), https://policyreview.info/articles/analysis/online-chilling-effects-england-and-wales (accessed 10 Feb. 2019).
[128] A possible approach to this is explored in Wright, D., Friedewald, M. and Gellert, R. (2015), 'Developing and testing a surveillance impact assessment methodology', *International Data Privacy Law*, 5(1), pp. 40–53, doi:https://doi.org/10.1093/idpl/ipu027 (accessed 10 Feb. 2019).
[129] De Hert and Lammerant (2017), 'Predictive Profiling and its Legal Limits', p. 167.
[130] Galdon Clavell (2017), 'Policing, Big Data and the Commodification of Security', p. 90.
[131] Van Brakel, R. (2017), 'Pre-emptive Big Data Surveillance and Its (Dis)Empowering Consequences: the Case of Predictive Policing', in van der Sloot, B., Broeders, D. and Schrijvers, E. (eds) (2017), *Exploring the Boundaries of Big Data*, Netherlands Scientific Council for Government Policy (WRR) Report, p. 117.

Crucial to banishing this spectre of 'pre-crime' is the recognition that the output of any predictive model always represents possibility rather than proof. Hence, interventions undertaken on the basis of predictive modelling must ultimately be based on a human decision made by someone who is sufficiently informed about the limits and capabilities of the model being used.

'Functional creep'

There are precedents where measures have been adopted for the purposes of countering terrorism, and then appropriated for their general use in fighting crime. The New York City Police Department's Domain Awareness System is one such. Integrating the feeds of 3,000 CCTV and ANPR cameras alongside other sensors, the system was originally installed – and is often touted – for its value in countering terrorism. However, an appraisal of its use indicates that there are very few examples where it has actually been used in this role.[132]

A hoard of data about the general public would be invaluable in identifying start points for criminal investigations, or even for directing any number of beneficial social interventions. The potential for 'functional creep' is massive, and requires safeguarding against. An appropriate measure might be that the dataset concerned was only used for interrogation by certain, pre-authorized models, rather than being available for directed specific queries by analysts. This could render the dataset easier to control, and could also give a level of transparency about how the data are used.

Continued susceptibility to abuse

Even in the presence of regulation and limitations, interpretation by governments will be subjective. The term 'terrorist' is often applied with latitude to suit underlying political purposes. Gathering public data, and allowing those data to be analysed, will carry with it an inherent risk of abuse. Safeguards such as anonymization have been repeatedly undermined in their ability to protect against analysis that links data back to specific personal identities.[133]

Whether any such powers are acceptable comes down in part to the population's trust in governments, and whether they can garner enough public support to allow them to exercise those powers. The limitations of public oversight as a control measure mentioned above are exacerbated by a lack of public understanding of the magnitude of data created, and how the data might be used.

The GDPR has implications for AI, but exemptions for statutory use of data by government agencies and a focus on the use of data for commercial purposes means that the regulation stops short of adequately covering the use of predictive AI in countering terrorism. It is necessary to regulate data use further to formalize assessments of proportionality and incorporate transparency. While this additional legislation might go some way to constraining the use of technology that might otherwise be used less scrupulously, it still does not fully address the risks.

[132] An automated alert was set immediately after the Boston Marathon bombing in 2013 that would have identified the terrorists' vehicle if it had entered New York City. The vehicle was intercepted before this point. In addition, several suspicious packages have been identified, all of which turned out to be false alarms. See Levine et al. (2017), 'The New York City Police Department's Domain Awareness System'.
[133] Bohannon, J. (2015), 'Credit card study blows holes in anonymity', *Science*, 347(6221), 30 January 2015, p. 468.

Maintenance of public support is not a safeguard in itself. Public acceptance of the argument that individuals should have nothing to hide neglects the principle that privacy should be protected as a right, and can lead to creeping societal effects when privacy invasion is not being taken into account.[134] Public willingness to accept exceptional measures in countering terrorism could reduce a population's ability to hold its government to account over the invasion of privacy.[135] This is exacerbated by a lack of awareness by the general public about how much data it generates and how the data might be used. As people's lives become increasingly digitalized, voluntarily or otherwise, the legislative and procedural safeguards that are put in place in order to prevent a surveillance state in the real world need to be robustly transferred to the digital space.

This should not, however, place the use of predictive AI in counterterrorism beyond use. Many aspects of counterterrorism practice are inherently susceptible to misuse. The sound development and regulation of AI capabilities is key to safeguarding against this danger. Such capabilities would demand a continuous process of management and monitoring, similar to those in effect for other counterterrorism practices.

[134] Solove, D. J. (2007), '"I've Got Nothing to Hide" and Other Misunderstandings of Privacy'.
[135] Ignatieff, M. (2004), *The Lesser Evil: Political Ethics in the Age of Terror*, p. 58.

6. Conclusion

The use of AI for predictive purposes in countering terrorism is neither inherently good nor inherently bad. Rather, the way such capabilities are used is critical. The current constructs that regulate the use of predictive AI in countering terrorism seem unlikely to either safeguard against misuse or to enable the most beneficial use of these technologies, both in terms of operational performance and adherence to human rights principles.

Pursuing more concerted efforts to use predictive AI in counterterrorism operations would require commitment in terms of research and experimentation, in order to develop models that are ready to use. If AI technologies for predicting terrorism reach maturity, greater data access and centralization – under strict safeguards – could offer a way of mediating infringement of privacy to proportionate ends.

Currently, specific regulation of the use of AI for the purpose of predicting involvement or risk of terrorism are partial, or entirely absent. Where restrictions do exist, they often focus on access to data and not on how the data are used. Paradoxically, restricting access could limit the ability to develop good models that could otherwise improve compliance with conditions of proportionality and non-discrimination. More centralized access with better regulation could be fairer than incidental access at the discretion of whichever actor or agency is able to obtain it.

Existing assumptions – such as the beliefs that broader access to data is always deleterious for human rights, and that centralization of analysis is inherently bad – should be reviewed in the light of technical possibilities for controlling powers of access and increased transparency. The development and use of predictive capabilities can be a valid justification for wider access to, and use of, public data, provided that models are thoroughly validated before use, that the initial stages of analysis are automated, and that technical measures of control are in place to prevent misuse. Continued access to any data for the purposes of countering terrorism should be contingent on the ability to derive sufficient predictive value from those specific data – meaning that proportionality of access is directly linked to fulfilment of a legitimate aim.

Governments will use new technological means at their disposal when pursuing critical objectives such as public safety. The fact that AI makes invasion of privacy at scale much easier means that the use of those technologies remains a public policy concern. This does not mean that use of AI for predicting terrorism by liberal democracies should be off limits. In fact, good predictive capabilities based on the automated analysis of less intrusive data could be part of sensible restrictions on disproportionate use of measures that present greater threats to privacy and other associated freedoms. A decision-making process, with measurable performance, with regard to who should be subject to more intrusive surveillance may be key to limiting the use of technically enabled surveillance where practical limitations are likely to be eroded. How successfully states manage the powers that new technology brings them will continue to reflect how well established their institutions are, and the strength of their commitment to protecting citizens' rights in general.

About the Author

Major Kathleen (Kitty) McKendrick is a British army officer. Her service history includes overseas operational tours in Iraq and Afghanistan, as well as delivering counterterrorism education and training for NATO personnel at the Centre of Excellence Defence Against Terrorism (COE-DAT) in Ankara, Turkey. During 2017–18 she was the Army Chief of General Staff visiting research fellow in the International Security Department at Chatham House.

Her research areas of interest include defence, security and the application of artificial intelligence in military operations.

She is a chartered engineer, and holds a BEng in aeromechanical systems engineering from Cranfield University, and an MSc in international relations from the London School of Economics and Political Science.

Acknowledgments

Thanks are due to a large number of people for their help and support throughout this project.

Among the staff of the International Security Department of Chatham House, particular thanks go to James de Waal and Hannah Bryce for their input on early drafts, and to Patricia Lewis for her generous oversight, mentoring and guidance. Thank you, too, to Calum Inverarity for shepherding this paper towards publication.

Lorna Woods, of the University of Essex, helped me to get a grip on the legal aspects of the topic of this paper by lending her expertise, insight and time. Lorna again, as well as Andrew Jillions, Lieutenant Colonel Al Brown and anonymous peer reviewers, challenged the arguments in the paper, commented on drafts and helped make it significantly better.

Thanks also go to the editors of this paper, Vera Chapman Browne and Jo Maher.

And thank you to my husband, James Small, for all his support and encouragement.

Independent thinking since 1920

Chatham House, the Royal Institute of International Affairs, is a world-leading policy institute based in London. Our mission is to help governments and societies build a sustainably secure, prosperous and just world.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical including photocopying, recording or any information storage or retrieval system, without the prior written permission of the copyright holder. Please direct all enquiries to the publishers.

Chatham House does not express opinions of its own. The opinions expressed in this publication are the responsibility of the author(s).

Copyright © The Royal Institute of International Affairs, 2019

Cover image: Surveillance cameras manufactured by Hangzhou Hikvision Digital Technology Co. at a testing station near the company's headquarters in Hangzhou, China, on 28 May 2019.

Photo credit: Copyright © Qilai Shen/Bloomberg/Getty

ISBN 978 1 78413 300 9

This publication is printed on FSC-certified paper.

FSC

Typeset by Soapbox, www.soapbox.co.uk

The Royal Institute of International Affairs
Chatham House
10 St James's Square, London SW1Y 4LE
T +44 (0)20 7957 5700 F +44 (0)20 7957 5710
contact@chathamhouse.org www.chathamhouse.org

Charity Registration Number: 208223