Royal United Services Institute
for Defence and Security Studies

Occasional Paper

Artificial Intelligence and UK National Security

Policy Considerations

Alexander Babuta, Marion Oswald and Ardi Janjeva

Artificial Intelligence and UK National Security

Policy Considerations

Alexander Babuta, Marion Oswald and Ardi Janjeva

RUSI Occasional Paper, April 2020

Royal United Services Institute
for Defence and Security Studies

189 years of independent thinking on defence and security

The Royal United Services Institute (RUSI) is the world's oldest and the UK's leading defence and security think tank. Its mission is to inform, influence and enhance public debate on a safer and more stable world. RUSI is a research-led institute, producing independent, practical and innovative analysis to address today's complex challenges.

Since its foundation in 1831, RUSI has relied on its members to support its activities. Together with revenue from research, publications and conferences, RUSI has sustained its political independence for 189 years.

Published in 2020 by the Royal United Services Institute for Defence and Security Studies.

RUSI Occasional Paper, April 2020. ISSN 2397-0286 (Online).

Royal United Services Institute
for Defence and Security Studies
Whitehall
London SW1A 2ET
United Kingdom
+44 (0)20 7747 2600
www.rusi.org
RUSI is a registered charity (No. 210639)

Contents

Acknowledgements

THE AUTHORS ARE very grateful to all research participants who gave up their valuable time to contribute to this study. The authors would also like to thank a number of individuals who provided helpful feedback on an earlier version of this paper, particularly Nick Jennings, Eric Kind, Sam Sherman, Amy Ertan, the project Steering Committee and RUSI colleagues (past and present) Malcolm Chalmers, Keith Ditcham and Andrew Glazzard.

Executive Summary

RUSI WAS COMMISSIONED by GCHQ to conduct an independent research study into the use of artificial intelligence (AI) for national security purposes. The aim of this project is to establish an independent evidence base to inform future policy development regarding national security uses of AI. The findings are based on in-depth consultation with stakeholders from across the UK national security community, law enforcement agencies, private sector companies, academic and legal experts, and civil society representatives. This was complemented by a targeted review of existing literature on the topic of AI and national security.

The research has found that **AI offers numerous opportunities** for the UK national security community to improve efficiency and effectiveness of existing processes. AI methods can **rapidly derive insights** from large, disparate datasets and identify connections that would otherwise go unnoticed by human operators. However, in the context of national security and the powers given to UK intelligence agencies, use of AI could give rise to **additional privacy and human rights considerations** which would need to be assessed within the existing legal and regulatory framework. For this reason, **enhanced policy and guidance** is needed to ensure the privacy and human rights implications of national security uses of AI are reviewed on an ongoing basis as new analysis methods are applied to data.

The research highlights three ways in which intelligence agencies could seek to deploy AI:

1. The **automation of administrative organisational processes** could offer significant efficiency savings, for instance to assist with routine data management tasks, or improve efficiency of compliance and oversight processes.
2. For **cybersecurity** purposes, AI could proactively identify abnormal network traffic or malicious software and respond to anomalous behaviour in real time.
3. For **intelligence analysis**, 'Augmented Intelligence' (AuI) systems could be used to support a range of human analysis processes, including:
    a. **Natural language processing and audiovisual analysis**, such as machine translation, speaker identification, object recognition and video summarisation.
    b. **Filtering and triage** of material gathered through bulk collection.
    c. **Behavioural analytics** to derive insights at the individual subject level.

None of the AI use cases identified in the research could replace human judgement. Systems that attempt to 'predict' human behaviour at the individual level are likely to be of limited value for threat assessment purposes. Nevertheless, the use of AuI systems to collate information from multiple sources and flag significant data items for human review is likely to improve the efficiency of analysis tasks focused on individual subjects.

The requirement for AI is all the more pressing when considering the need to counter AI-enabled threats to UK national security. **Malicious actors** will undoubtedly seek to use AI to attack the UK, and it is likely that the most capable hostile state actors, which are not bound by an equivalent legal framework, are developing or have developed offensive AI-enabled capabilities. In time, other threat actors, including cybercriminal groups, will also be able to take advantage of these same AI innovations.

- Threats to **digital security** include the use of polymorphic malware that frequently changes its identifiable characteristics to evade detection, or the automation of social engineering attacks to target individual victims.
- Threats to **political security** include the use of 'deepfake' technology to generate synthetic media and disinformation, with the objective of manipulating public opinion or interfering with electoral processes.
- Threats to **physical security** are a less immediate concern. However, increased adoption of Internet of Things (IoT) technology, autonomous vehicles, 'smart cities' and interconnected critical national infrastructure will create numerous vulnerabilities which could be exploited to cause damage or disruption.

There are opportunities and risks relating to privacy intrusion. AI arguably has the potential to **reduce intrusion**, by minimising the volume of personal data that is subject to human review. However, it has also been argued that **the degree of intrusion is equivalent** regardless of whether data is processed by an algorithm or a human operator. Furthermore, use of AI could result in additional material being processed which may not have previously been possible for technical or capacity-related reasons. This would need to be taken into account when assessing proportionality of any potential intrusion, balanced against the increase in effectiveness of analysis that may result.

'Algorithmic profiling' could be considered more intrusive than manual analysis and would raise further human rights concerns if it was perceived to be unfairly biased or discriminatory. Safeguarding against machine bias will require internal processes for ongoing tracking and mitigation of discrimination risk at all stages of an AI project, as well as ensuring demographic diversity in AI development teams.

Much commentary has raised concern regarding the **'black box'** nature of certain AI methods, which may lead to a loss of accountability of the overall decision-making process. In order to ensure that human operators retain ultimate accountability for the decision-making process informed by analysis, it will be essential to design systems in such a way that non-technically skilled users can interpret and critically assess key technical information such as the margins of error and uncertainty associated with a calculation.

Despite a proliferation of 'ethical principles' for AI, it remains uncertain how these should be operationalised in practice, suggesting the need for **additional sector-specific guidance** for national security uses of AI. An agile approach within the existing oversight regime to anticipating and understanding the opportunities and risks presented by new AI capabilities will be essential to ensure the UK intelligence community can adapt in response to the rapidly evolving technological environment and threat landscape.

Note on Sources

THE FINDINGS PRESENTED in this paper are based on a combination of open and closed-source research. The content is primarily derived from confidential interviews and focus groups with respondents from across the UK national security community. Although open-source references are included throughout, it is not always possible to provide a specific source for research findings and conclusions.

Introduction

RUSI WAS COMMISSIONED by GCHQ to conduct an independent research study into the use of artificial intelligence (AI) for national security purposes. The overall aim of the project is to establish an independent evidence base to inform future policy development and strategic thinking regarding national security uses of AI.

The research examined the use of AI within the UK Intelligence Community (referred to throughout as 'UKIC' or 'the agencies'). The findings presented in this paper are based on in-depth consultation with practitioners and policymakers from across UKIC, other government departments, law enforcement agencies, military organisations, private sector companies, academic and legal experts, and civil society representatives. This was complemented by a targeted review of existing academic literature, research reports and government documents on the topic of AI and national security.

The findings presented in this paper are the product of the authors' independent research and analysis. Due to subject-matter sensitivities, certain content has been omitted or sanitised in consultation with project partners. These revisions in no way influence the overall findings or conclusions of the research.

This paper is structured as follows. The introduction provides a brief overview of the context of the project and the issues under consideration. Chapter I examines potential uses of AI in the national security context, as identified in the research. Chapter II summarises the legal framework governing UKIC's use of data, before assessing specific legal and ethical considerations arising from national security uses of AI. Finally, Chapter III provides a summary of existing AI guidance, regulation and oversight frameworks, before considering what additional sector-specific guidance and oversight mechanisms may be needed in the national security context.

The Context

The UK continues to face serious national security threats from a range of sources.[1] There is a high expectation that UKIC will protect citizens from threats to their safety and adopt new methods that may allow them to do this more effectively. At the same time, the public expects the agencies to adapt and innovate in a way that provides reassurances that citizens' rights and freedoms are respected. Achieving this balance is a major challenge for those in the national security community, particularly at a time of such considerable technological change. At the same time, public discourse is increasingly focused on the governance and regulation of data

1. See, for example, Centre for the Protection of National Infrastructure, 'National Security Threats', <https://www.cpni.gov.uk/national-security-threats>, accessed 8 April 2020.

analytics, and there appears to be increasing concern that existing structures are not fit for purpose in terms of the governance and oversight of AI.

The modern-day 'information overload' is perhaps the greatest technical challenge facing the UK's national security community.[2] The ongoing, exponential increase in digital data necessitates the use of more sophisticated analytical tools to effectively manage risk and proactively respond to emerging security threats. For UKIC, this 'obligation to innovate' is even more pressing when considering hostile uses of AI that already pose a tangible threat to UK national security, such as the use of machine learning (ML) algorithms to facilitate cyber attacks, generate malware or automate disinformation campaigns. Against this backdrop, there is a clear driver for UKIC to implement advanced data science techniques to effectively respond to future threats to the UK's hyperconnected digital ecosystem.

While AI offers numerous opportunities for UKIC to improve the efficiency and effectiveness of existing processes, these new capabilities raise additional privacy and human rights considerations which would need to be assessed within the existing legal and regulatory framework. Recent commentary has highlighted potential risks regarding the implementation of AI and advanced analytics for surveillance purposes, particularly relating to the potential impact on individual rights.[3] As summarised by Jonathan H King and Neil M Richards:

> The problem is that our ability to reveal patterns and new knowledge from previously unexamined troves of data is moving faster than our current legal and ethical guidelines can manage. We can now do things that were impossible a few years ago, and we've driven off the existing ethical and legal maps. If we fail to preserve the values we care about in our new digital society, then our big data capabilities risk abandoning these values for the sake of innovation and expediency.[4]

Addressing these concerns is a high priority for the national security community. According to GCHQ, 'it is absolutely essential that we have the debates around AI and machine learning in the national security space that will deliver the answers and approaches that will give us public consent'.[5] GCHQ further notes that 'it is essential that AI is used ethically and is subject

2. See, for example, 'Address by the Director General of the Security Service, Andrew Parker, to RUSI at Thames House', 8 January 2015, <https://www.mi5.gov.uk/news/director-general-speaks-on-terrorism-technology-and-oversight>, accessed 8 April 2020.
3. See, for example, Steven Feldstein, 'The Global Expansion of AI Surveillance', Carnegie Endowment for International Peace Working Paper, September 2019; Ronja Kniep, 'Another Layer of Opacity: How Spies Use AI and Why We Should Talk About It', about:intel, 20 December 2019, <https://aboutintel.eu/how-spies-use-ai/>, accessed 8 April 2020; Andrew Guthrie Ferguson, *The Rise Of Big Data Policing: Surveillance, Race, And The Future Of Law Enforcement* (New York, NY: NYU Press, 2019).
4. Jonathan H King and Neil M Richards, 'What's Up with Big Data Ethics?', *Forbes*, 28 March 2014.
5. Paul Killworth cited in Alexander Babuta, 'A New Generation of Intelligence: National Security and Surveillance in the Age of AI', *RUSI Commentary*, 19 February 2019.

to effective oversight'.[6] Outgoing MI5 Director General Sir Andrew Parker has likewise stated that he is particularly interested in AI 'because of our need to be able to make sense of the data lives of thousands of people in as near to real time as we can get to', but recognises that '[technology] will never replace our need to also have human insight, because technology and data will never tell us what is going on in people's heads'.[7]

Most AI methods under consideration are rapidly becoming more prevalent throughout the commercial sector.[8] However, UKIC is subject to additional levels of scrutiny regarding the acquisition and use of data – scrutiny and oversight to which the private sector is not subject.[9] Furthermore, national security uses of AI will require a higher degree of robustness and resilience than many commercial applications, and many capabilities will not be readily transferable from other sectors.

Clear and evidence-based policy is needed to ensure that the UK national security community can take full advantage of the opportunities offered by these new technologies, without compromising societal and ethical values or undermining public trust.

What is AI?

There is no universally accepted definition of AI. However, a distinction is often made between 'General AI' (machine intelligence with the agency, reasoning and adaptability of a human brain) and 'Narrow AI' (machine intelligence trained to perform well in a narrowly defined cognitive task, such as playing chess, driving a car or translating documents). All existing AI can be characterised as Narrow AI. It is widely accepted that General AI – if it is indeed achievable – is many decades away.

Narrow AI can be understood as 'a set of advanced general-purpose digital technologies that enable machines to perform highly complex tasks effectively'.[10] AI is usually defined in terms of the ability 'to perform tasks that would usually require human intelligence',[11] and can be

6. Jo Cavan and Paul Killworth, 'GCHQ Embraces AI, but not as a Black Box', about:intel, October 2019, <https://aboutintel.eu/gchq-embraces-ai/>, accessed 8 April 2020.
7. Lionel Barber and Helen Warrell, 'MI5 Chief Sees Tech as Biggest Challenge and Opportunity', *Financial Times*, 12 January 2020.
8. Wendy Hall and Jerome Pesenti, 'Growing the Artificial Intelligence Industry in the UK', 15 October 2017, <https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk>, accessed 8 April 2020.
9. The powers given to the UK Intelligence Community (UKIC) are subject to a specific oversight regime set out in intelligence and surveillance legislation, while the private sector's use of data remains governed primarily by data protection frameworks.
10. Paul Martin, *The Rules of Security: Staying Safe in a Risky World* (Oxford: Oxford University Press, 2019), p. 217.
11. Oxford Reference, 'Artificial Intelligence', <https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095426960>, accessed 8 April 2020.

understood as comprising six sub-disciplines: automated reasoning; natural language processing (NLP); knowledge representation; computer vision; robotics; and machine learning (ML).[12]

Recent progress in Narrow AI has been driven primarily by advances in the sub-field of ML. ML enables computer systems to learn and improve through experience, and is characterised by the use of statistical algorithms to find patterns, derive insights or make predictions. An algorithm can be defined as 'a set of mathematical instructions or rules that, especially if given to a computer, will help to calculate an answer to a problem'.[13] ML is a specific category of algorithm that is able to improve its performance at a certain task after being exposed to new data. There are three main types of learning: supervised; unsupervised; and reinforcement learning.

- In **supervised learning**, the agent 'observes some example input–output pairs and learns a function that maps from input to output'.[14] For example, for object classification, training data could include many photographs of different types of fruit, and labels defining which fruit is depicted in each photo. The trained model is considered to 'generalise' well if it is able to correctly identify the type of fruit when presented with new, unfamiliar photos.
- In **unsupervised learning**, 'the agent learns patterns in the input even though no explicit feedback is supplied'.[15] For example, for image recognition, training data could include thousands of individual photographs of five types of animal but no labels identifying the animals. The model is considered to perform well if it is able to correctly divide the photographs into five piles, each containing the photos of one type of animal.
- **Reinforcement learning** is a goal-oriented form of learning, where the agent improves at a task over time based on exposure to positive and negative feedback. For personalised recommender systems, a human listener may be recommended music based on their previous listening habits. The user provides feedback indicating whether they like the computer-recommended track. This feedback helps the algorithm to learn the user's listening preferences, meaning that the recommendations become more accurate over time.
- **Semi-supervised learning** is a fourth category of ML, involving datasets where some input–output pairs are labelled but a large proportion are unlabelled. Returning to the fruit classification example, the model can be pre-trained on the entire training set (using unsupervised methods), before it is fine-tuned using the labelled subset.[16]

12. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd Edition (Upper Saddle River, NJ: Pearson Education Limited, 2016).
13. Cambridge Dictionary, 'Algorithm', <https://dictionary.cambridge.org/dictionary/english/algorithm>, accessed 7 April 2020.
14. Russell and Norvig, *Artificial Intelligence*, pp. 706–08.
15. *Ibid.*
16. See *Ibid.* for all four definitions.

The use of ML has grown considerably in recent years, driven by an exponential growth in computing power coupled with an increased availability of large datasets. In healthcare, ML-based image recognition is used for complex tasks, such as predicting the risk of autism in babies or detecting skin cancer.[17] Local councils are deploying ML algorithms to assist social workers' case prioritisation and identify families most in need of government support.[18] In policing, ML algorithms are used to forecast demand in control centres, predict re-offending and prioritise crimes according to their 'solvability'.[19] With the growth of 'smart cities', ML algorithms are increasingly being used to streamline tasks, such as waste removal, traffic management and sewerage systems.[20] These trends are likely to continue in the coming years, with the UK government's Office for AI estimating that AI could add £232 billion to the UK's economy by 2030.[21]

It is important to note, however, that most AI advancements have been made either in the private sector or academia.[22] The UK government is yet to take full advantage of these opportunities. As summarised by the Committee on Standards in Public Life, 'despite generating much interest and commentary, our evidence shows that the adoption of AI in the UK public sector remains limited. Most examples the Committee saw of AI in the public sector were still under development or at a proof-of-concept stage'.[23] In the coming years, taking full advantage of the opportunities presented by these technologies will be a high priority for the UK government.[24]

17. Heather Cody Hazlett et al., 'Early Brain Development in Infants at High Risk for Autism Spectrum Disorder', *Nature* (Vol. 542, No. 7641, February 2017), pp. 348–51; Matt Reynolds, 'AI Rivals Dermatologists at Spotting Early Signs of Skin Cancer', *New Scientist*, 25 January 2017.
18. Vicky Clayton, 'Why is the What Works Centre Researching Machine Learning?', What Works for Children's Social Care, 8 February 2019, <https://whatworks-csc.org.uk/blog/why-is-the-what-works-centre-researching-machine-learning/>, accessed 8 April 2020.
19. Alexander Babuta and Marion Oswald, 'Data Analytics and Algorithms in Policing in England and Wales: Towards a New Policy Framework', *RUSI Occasional Papers* (February 2020).
20. Nick Huber, 'Internet of Things: Smart Cities Pick up the Pace', *Financial Times*, 29 January 2020.
21. Office for Artificial Intelligence, 'AI Sector Deal – One Year On', 2019, <https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal-one-year-on>, accessed 17 April 2020.
22. Committee on Standards in Public Life, 'Artificial Intelligence and Public Standards: A Review by the Committee on Standards in Public Life', February 2020.
23. *Ibid.*, p. 15.
24. See, for example, Department for Digital, Culture, Media and Sport et al., 'Leading Experts Appointed to AI Council to Supercharge the UK's Artificial Intelligence Sector', 16 May 2019, <https://www.gov.uk/government/news/leading-experts-appointed-to-ai-council-to-supercharge-the-uks-artificial-intelligence-sector>, accessed 8 April 2020.

I. National Security Uses of AI

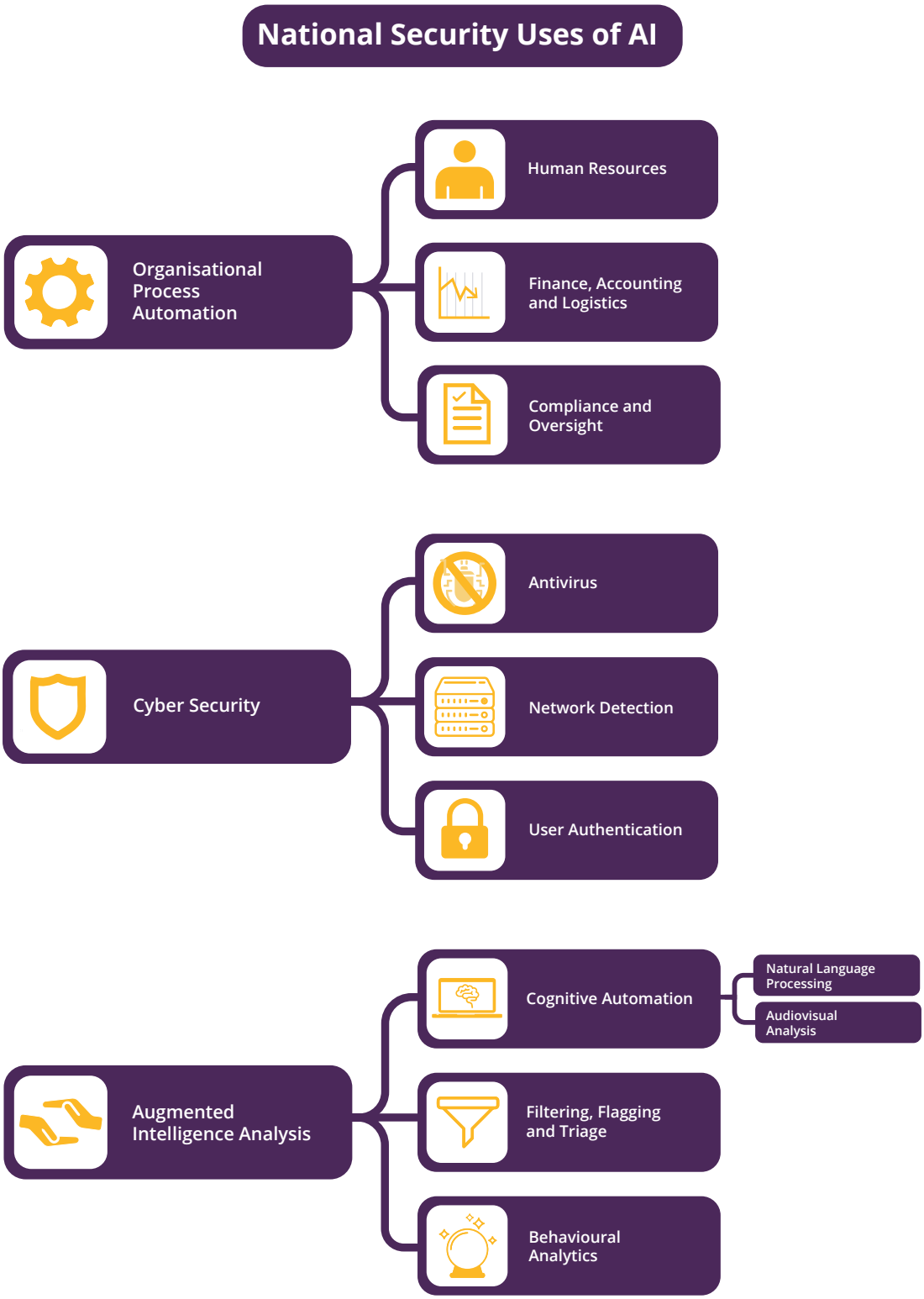
RECENT COMMENTARY HAS highlighted the acute challenges posed to intelligence agencies as a result of the modern-day 'information overload'.[25] As summarised by Greg Allen and Taniel Chan, 'there is more data to analyse and draw useful conclusions from, but finding the needle in so much hay is tougher'.[26]

But the challenge is more than just one of volume. In his 2015 report of the Investigatory Powers Review, David Anderson described how changing methods of communication, the fragmentation of service providers, difficulties in attributing communications, ubiquitous encryption and the emergence of new sources of data have all contributed to a growing 'capability gap' within intelligence agencies.[27] These challenges call for the development of more sophisticated analytical tools, and AI is likely to form an important component of this new toolkit. GCHQ have stated publicly that 'within an organisation like GCHQ, there is a potential to use machine learning and AI to improve our operational outcomes. We can tackle these large problems and potentially deliver intelligence and security solutions to help keep the UK safe, in ways which we couldn't do before'.[28]

There are numerous ways in which UKIC could apply AI to improve the efficiency and effectiveness of existing processes. Potential use cases identified in this research are discussed in turn below, and can be broadly categorised as illustrated in Figure 1.

25. See, for example, Cavan and Killworth, 'GCHQ Embraces AI, but not as a Black Box'; Barber and Warrell, 'MI5 Chief Sees Tech as Biggest Challenge and Opportunity'.
26. Greg Allen and Taniel Chan, 'Artificial Intelligence and National Security', Belfer Center Study, July 2017, p. 27.
27. David Anderson, *A Question of Trust: Report of the Investigatory Powers Review* (London: Stationery Office, 2015), p. 49.
28. Babuta, 'A New Generation of Intelligence'.

Figure 1: National Security Uses of AI



Source: Authors' research.

Automation of Organisational Processes

As for all large organisations, the most immediate benefit for UKIC in the use of AI will most likely be the ability to automate organisational, administrative and data management processes – repetitive tasks which comprise a significant proportion of overall workload. As summarised in the 2016 White House report on AI, 'AI's central economic effect in the short term will be the automation of tasks that could not be automated before'.[29] This could include assisting with tasks such as human resources and personnel management, logistics optimisation, finance and accounting.

Examples of commercial uses of AI demonstrate its potential benefits in administrative processes. These can be divided into front office and back office uses. In the front office, a combination of computer vision and NLP can be used in processes such as handling insurance claim forms and accompanying information like photographs, carrying out query resolutions more quickly and efficiently by guiding users through repositories of information, and making chatbots act as the first point of contact for enquiries on e-commerce websites.[30] Back office functions include the automation of data capture when scanning images for invoice processing, cross-referencing data between application forms and supplementary documents when servicing loans, and collating swathes of information, such as industry-wide announcements or companies' annual financial data.[31] Similarly, the effective use of AI could significantly reduce administrative workloads across the UK goverment, from improving the efficiency of room booking and diary management systems, to managing job applications or conducting routine background checks.

For UKIC, significant efficiency gains could also be made in the automation of compliance and oversight processes. A recent report by the German think tank, Stiftung Neue Verantwortung (SNV), identified seven tools for 'data-driven intelligence oversight', to enable oversight bodies 'to conduct unannounced checks as well as (semi-)automated audits on intelligence agencies' data processing'.[32] AI could conceivably be applied to any one of these processes. For example, the authors propose a 'hidden pattern detector' to identify inappropriate database use and 'activities that may not be legally compliant', suggesting that 'options for analyzing log files range from simple descriptive methods … to sophisticated machine learning or statistical analysis techniques'.[33] Automating aspects of authorisation and oversight processes could not only help to ensure compliance with relevant legislative

29. National Science and Technology Council, 'Preparing for the Future of Artificial Intelligence', October 2016, <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>, accessed 8 April 2020.
30. Sarah Burnett, 'Automating Content-Centric Processes with AI', Everest Group, 8 December 2017, <https://www.everestgrp.com/automating-content-centric-processes-ai-technology/>, accessed 8 April 2020.
31. *Ibid.*
32. Kilian Vieth and Thorsten Wetzling, 'Data-Driven Intelligence Oversight: Recommendations for a System Update', Stiftung Neue Verantwortung, November 2019, p. 2.
33. *Ibid.*, p. 25.

requirements, but would also free up staff time within oversight bodies to provide scrutiny and advice regarding more complex technical issues.

Cyber Security

Modern-day cyber security threats require a speed of response far greater than human decision-making allows. Given the rapid increase in the volume and frequency of malware attacks, AI cyber defence systems are increasingly being implemented to proactively detect and mitigate threats. While traditional antivirus methods rely on 'blacklisting' historic threats based on virus signatures, AI-based antivirus can recognise aspects of software that may be malicious without the need to rely on a pre-defined list. As summarised in a recent report from Darktrace, an AI cyber security company:

> While legacy security tools can often identify known threats that have already been discovered 'in the wild', artificial intelligence can uniquely spot the weak and subtle signals of a never-before-seen cyber-threat. This capability has become necessary in recent years, as advanced cyber-criminals continue to develop novel tactics, techniques, and procedures specifically designed to evade controls that have been pre-programmed with signatures of past attacks.[34]

Similarly, AI-based network detection systems could be trained to learn what constitutes 'normal' activity on an organisation's network, identify abnormal network traffic based on analysis of log data and respond in real time. Relatedly, these techniques could be used to identify and flag abnormal system activity that may suggest an insider threat. A 2018 report by Cybersecurity Insiders found that '86% of organisations already have or are building an insider threat programme', mainly based around 'Intrusion Detection and Prevention (IDS), log management and SIEM [security information and event management] platforms'.[35]

User authentication is another area of potential value to UKIC. Recent research has focused on the use of so-called 'behavioural biometrics' to identify users based on unique aspects of their digital activity, such as how they handle their mouse or compose sentences in a document.[36] Such active authentication systems could enhance cyber security by ensuring ongoing user authentication following an initial session login.

34. Darktrace, 'Autonomous Response: Threat Report 2019', p. 3, <https://customers.darktrace.com/en/request-resources?pp=wp-cyber-ai-response-threat-report-2019&utm_source=darktrace&utm_medium=mudwall>, accessed 8 April 2020.
35. Cybersecurity Insiders, 'Insider Threat: 2018 Report', 2018, p. 4, <https://crowdresearchpartners.com/wp-content/uploads/2017/07/Insider-Threat-Report-2018.pdf>, accessed 8 April 2020.
36. See, for example, Defense Advanced Research Projects Agency, 'Active Authentication (Archived)', <https://www.darpa.mil/program/active-authentication>, accessed 8 April 2020.

Augmented Intelligence Analysis

AI-assisted intelligence analysis could offer significant benefits in deriving insights from unstructured and disparate datasets, thereby improving the efficiency of the intelligence workflow and potentially reducing collateral intrusion by minimising the volume of content that is subject to human review.

Potential examples of AI-assisted intelligence analysis fall broadly into three categories:

1. **Cognitive automation** of human sensory processing (particularly NLP and audiovisual analysis).
2. **Filtering, flagging and triage** of data gathered through bulk collection, as part of an interactive human–machine analysis workflow.
3. **Behavioural analytics** to derive insights at the individual subject level.

Cognitive Automation

One area where AI offers clear potential benefits could be described as 'cognitive automation', meaning the machine replication of human sensory processing (particularly NLP and audiovisual analysis). Automation of this kind would significantly reduce the time needed for human operators to interpret large volumes of data, while also potentially reducing intrusion by minimising the volume of content that is subject to human review.

Effective use of speech-to-text transcription could dramatically reduce the human resources required to process audio data (such as intercept material). Machine translation also presents clear benefits, either applied to transcribed text or directly to audio data. In addition, speaker identification could make large quantities of voice data searchable in a more efficient way. Rapid progress has been made in language analytics in recent years. In February 2019, OpenAI announced that they had trained a large-scale unsupervised language model – named GPT-2 – that generated coherent paragraphs of text, achieved state-of-the-art performance on many language modelling benchmarks and performed basic reading comprehension, machine translation, question answering and summarisation.[37] Recent research has also demonstrated the potential uses of ML techniques for authorship attribution based on linguistic analysis of stylometric features.[38]

37. Alec Radford et al., 'Better Language Models and their Implications', OpenAI, 14 February 2019, <https://openai.com/blog/better-language-models/>, accessed 8 April 2020.
38. Hoshiladevi Ramnial, Shireen Panchoo and Sameerchand Pudaruth, 'Authorship Attribution Using Stylometry and Machine Learning Techniques', in Stefano Berretti, Sabu M Thampi and Praveen Ranjan Srivastava (eds), *Intelligent Systems Technologies and Applications,* Vol. 1 (Cham, Switzerland: Springer, 2016), pp. 113–25.

AI could also improve the efficiency of video data processing. Object classification and facial matching could substantially reduce the amount of time analysts spend manually trawling through video footage. Another benefit is the ability to classify material in order to shield analysts and investigators from harmful content, such as material depicting violence or sexual abuse. Video summarisation is a further area of interest. An example is the use of ML algorithms to generate a unique summary of a video by selecting key frames which accurately capture the content and context of the original video. This can be used to identify a change that has happened over time and create a video highlighting that change to an analyst. In the military context, the Software Engineering Institute (SEI) has been developing a multi-stage video summarisation pipeline for US military organisations that aims to notify operators of significant events, such as the planting of an explosive device on a road. According to the SEI, the long-term goal would be to 'recognize and search for patterns of life across multiple videos, with the ultimate goal of predicting future activities and events'.[39]

Filtering, Flagging and Triage

It is publicly reported that bulk data gathered by UKIC is processed using a series of automated volume reduction systems to filter, query and select material for examination. Incorporating AI into these systems could improve the efficiency of filtering processes, ensuring that human operators have access only to the information that is most relevant to the analytical task at hand, thereby minimising collateral intrusion. As summarised in a 2015 report from the US National Research Council on the bulk collection of signals intelligence:

> No software-based technique can fully replace the bulk collection of signals intelligence, but methods can be developed to more effectively conduct targeted collection and to control the usage of collected data ... Automated systems for isolating collected data, restricting queries that can be made against those data, and auditing usage of the data can help to enforce privacy protections and allay some civil liberty concerns.[40]

The report of the Bulk Powers Review conducted by Lord Anderson in 2016 provides a detailed account of how GCHQ carried out bulk interception at the time of writing.[41] The report describes three stages of bulk interception, which can be understood as 'collection', 'filtering' and 'selection for examination'. First, bearers are selected on the basis of an assessment of the potential intelligence value of their communications. A degree of filtering is then applied to the traffic of selected bearers, which is 'designed to select communications of potential intelligence

39. Kevin Pitstick, 'Video Summarization: Using Machine Learning to Process Video from Unmanned Aircraft Systems', Carnegie Mellon University, Software Engineering Institute, 22 January 2018, <https://insights.sei.cmu.edu/sei_blog/2018/01/video-summarization-using-machine-learning-to-process-video-from-unmanned-aircraft-systems.html>, accessed 8 April 2020.
40. National Research Council, *Bulk Collection of Signals Intelligence: Technical Options* (Washington, DC: National Academies Press, 2015).
41. David Anderson, *Report of the Bulk Powers Review*, Cm 9326 (London: Stationery Office, 2016), p. 23.

value while discarding those least likely to be of intelligence value'. Finally, 'the remaining communications are then subjected to the application of queries, both simple [relating to an individual target] and complex [combining several criteria], to draw out communications of intelligence value'. Due to the volume of collected data, even when communications relate to specific targets of interest, a triage process is applied to determine which items are most useful. 'Analysts use their experience and judgement to decide which of the results returned by their queries are most likely to be of intelligence value and will examine only these'.[42] In their most recent annual report, the Investigatory Powers Commissioner's Office (IPCO) noted that 'we are confident that the majority of data gathered by way of bulk collection is not reviewed by analysts, although it will be automatically screened against specific criteria to enable the agencies to extract intelligence relevant to clearly identified operational purposes'.[43]

If deployed effectively, AI could identify connections and correlations within and between multiple bulk datasets more efficiently than human operators, improving the accuracy of this screening and filtering process. However, a crucial distinction must be drawn between using AI to identify content of interest to flag to a human operator, and applying behavioural analytics methods to detect or 'predict' suspicious activity (this is discussed further below). For extracting intelligence from bulk data, AI is likely to be most useful when deployed as part of an interactive 'human–machine team' analysis workflow.[44]

It could be argued that AI has the potential to reduce collateral intrusion when searching or filtering data gathered through bulk collection, by minimising the volume of content that is subject to human review. However, it has also been argued that machine analysis is not necessarily inherently less intrusive than human review. This issue is discussed further in Chapter II.

Behavioural Analytics

'Behavioural analytics' can be understood as the application of complex algorithms to individual-level data to derive insights, generate forecasts or make predictions about future human behaviour. There are various ways in which intelligence agencies could hypothetically implement AI to make predictions about future behaviour. These include insider threat detection, predicting threats to individuals in public life, identifying potential intelligence sources who may be susceptible to persuasion and predicting potential terrorist activity before it occurs.

The use of behavioural analytics for counterterrorism purposes has attracted significant public attention in recent years. Following the 2017 attacks in London and Manchester, a joint Operational Improvement Review conducted by MI5 and Counterterrorism Policing proposed a

42. *Ibid.*
43. Investigatory Powers Commissioner's Office (IPCO), *Annual Report 2018*, HC 67, SG/2020/8 (London: Stationery Office, 2018), p. 29.
44. For further discussion on human–machine teaming, see Ministry of Defence, 'Joint Concept Note 1/18: Human–Machine Teaming', May 2018, p. 36, <https://www.gov.uk/government/publications/human-machine-teaming-jcn-118>, accessed 8 April 2020.

'step change' in how the organisations use data. This included the need for 'improvements in the ability of MI5 and police to exploit data to detect activity of concern, particularly on the part of closed SOIs [subjects of interest] but in relation also to active SOIs and previously unknown individuals'.[45] Lord Anderson's recently published 'implementation stock-take' describes 'the identification of capabilities and data needed to develop relevant behavioural triggers',[46] which will be achieved by 'increasingly sophisticated use of artificial intelligence and behavioural analytics to extract information from bulk datasets'.[47] The report concludes that:

> Behavioural analytics is here to stay, and its techniques may be effective not just in refining the assessment of risk from existing leads and SOIs but in discovering new leads who would not otherwise have come to the attention of authorities. Some indicators are geared to identifying immediate pre-attack behaviour, such as attempts to obtain firearms or researching attack methodologies. More general indicators – for example, personal frustrations or changes in baseline behaviour – may also have their place when applied to persons who are already under suspicion.[48]

There is a large body of academic research exploring the relative merits of clinical (discretionary) versus statistical (non-discretionary) decision-making, and the debate about which approach is more accurate, justified or informative is intense and ongoing.[49] A number of empirical studies from the 1950s onwards have demonstrated that statistical forecasting typically yields more accurate predictions than unstructured clinical judgement, across many disciplines and in a wide range of decision-making contexts.[50] However, experts argue that aggregated 'predictive accuracy' rates are fundamentally misleading when assessing risk judgements at the individual level, and the evidence shows that violence risk assessment approaches that incorporate a degree of professional judgement yield more successful results than relying purely on statistical methods.[51] Recent research into prediction of life outcomes using a mass collaboration approach

45. David Anderson, *Attacks in London and Manchester, March-June 2017: Independent Assessment of MI5 and Police Internal Reviews* (London: Brick Court Chambers, 2017), p. 32.
46. David Anderson, '2017 Terrorist Attacks MI5 and CTP Reviews: Implementation Stock-Take', 11 June 2019, p. 14.
47. *Ibid.*, p. 18.
48. *Ibid.*, p. 19.
49. For further discussion, see, for example, Caroline Logan and Monica Lloyd, 'Violent Extremism: A Comparison of Approaches to Assessing and Managing Risk', *Legal and Criminological Psychology* (Vol. 24, No. 1), 2019, pp. 14–61.
50. Paul E Meehl, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (Minneapolis, MN: University of Minnesota: 1954); Robyn M Dawes, David Faust and Paul E Meehl, 'Clinical Versus Actuarial Judgment', *Science* (Vol. 243, No. 4899, 1989), pp. 1668–74; William M Grove et al., 'Clinical Versus Mechanical Prediction: A Meta-Analysis', *Psychological Assessment* (Vol. 12, No. 1, 2000), pp. 19–30; Stefanía Ægisdóttir et al., 'The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction', *Counseling Psychologist* (Vol. 34, No. 3, 2006), pp. 341–82.
51. Alan A Sutherland et al., 'Sexual Violence Risk Assessment: An Investigation of the Interrater Reliability of Professional Judgments Made Using the Risk for Sexual Violence Protocol',

concluded that 'despite using a rich dataset and applying machine-learning methods optimized for prediction, the best predictions were not very accurate and were only slightly better than those from a simple benchmark model'.[52]

Moreover, given the relative infrequency of terrorist violence, there is a significantly smaller corpus of historical data to form the basis of a statistical risk model (when compared with other forms of violent offending). Previous analysis of historic cases reported in the academic literature reveals wide variation in perpetrators' backgrounds, behavioural patterns and motivations, and in the precipitatory factors that ultimately lead them to commit an act of extremist violence. As such, there is no consistent 'profile' of a terrorist offender.[53] As summarised by John Monahan, 'existing research has largely failed to find valid nontrivial [statistically significant] risk factors for terrorism. Without the identification of valid risk factors, the individual risk assessment of terrorism is impossible'.[54] Another concern of incorporating statistical methods into terrorism risk assessment processes is the potential loss of relevant contextual information which should be considered when making judgements related to individuals' behavioural patterns or intentions.

Considering these limitations, rather than attempting to 'predict' individual behaviour, efforts should instead focus on developing so-called 'augmented intelligence' (AuI) systems to support human analysis. This is achieved by collating relevant information from multiple sources and flagging significant data items for human review. A degree of human judgement is essential when making assessments regarding behavioural intent or changes in an individual's psychological state. As summarised in a recent article from Palantir's Privacy and Civil Liberties Engineering team:

> AI is overrated. The role of machines is not to replace but facilitate human reasoning. Augmented intelligence (AuI) can help intelligence agencies navigate the data deluge by enabling human analysts to make data-driven decisions in a more transparent and accountable way
>
> …
>
> A computer program can effectively augment human intelligence by providing analysts with a unified data landscape that is moreover presented in a way that makes sense intuitively. For an analyst working

International Journal of Forensic Mental Health (Vol. 11, No. 2, 2012), p. 120; Kevin S Douglas, Melissa Yeomans and Douglas P Boer, 'Comparative Validity Analysis of Multiple Measures of Violence Risk in a Sample of Criminal Offenders', *Criminal Justice and Behavior* (Vol. 32, No. 5, October 2005), pp. 479–510.

52. Matthew J Salganik et al., 'Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration', *Proceedings of the National Academy of Sciences* (Vol. 117, No. 15, 2020), pp. 8398–403.
53. Paul Gill, *Lone-Actor Terrorists: A Behavioural Analysis* (London: Routledge, 2015).
54. John Monahan, 'The Individual Risk Assessment of Terrorism', *Psychology, Public Policy, and Law* (Vol. 18, No. 2, September 2011), p. 19.

> at an intelligence agency, this might mean turning data stored in documents, reports, and tables into persons, objects, and events, and graphically visualizing the relationships between them. AuI does not aim at providing answers but at enabling subject-matter experts to ask the right questions. Asking the right questions in turn enables human analysts to efficiently sift through a morass of data to find the information that actually matters.[55]

This view was emphasised by Metropolitan Police Commissioner Cressida Dick at RUSI's Annual Security Lecture in February 2020:

> So I would talk – in line with many other people – about Augmented Intelligence. I wouldn't put all policing's hopes and fears on what is described as Artificial Intelligence … The term describes better how technology can work to improve human intelligence rather than to replace it. That feels much closer to how we in policing are using technology. I also believe a licence to operate technology in those human terms feels much closer to what the public would expect and accept.[56]

In sum, the evidence reviewed for this paper suggests that it is neither feasible nor desirable to attempt to develop AI systems to 'predict' human behaviour at the individual level – for instance, for counterterrorism risk assessment purposes. Nevertheless, AuI – the use of AI systems to collate relevant information from multiple sources and flag significant data items for human review – has clear potential benefits in this context and is likely to improve the efficiency of analysis tasks focusing on individual subjects. Care will be needed, however, to ensure that relevant case-specific contextual information is not 'screened out' because it is not found to be statistically significant in historic data.

Adversarial AI

Malicious actors will undoubtedly seek to use AI to attack the UK.[57] It is likely that the most capable hostile state actors, which are not bound by an equivalent legal framework, are developing or have already developed offensive AI-enabled capabilities. In time, other threat actors, including cybercriminal groups, will also be able to take advantage of these same innovations. The national security requirement for AI is therefore all the more pressing when considering the need to combat potential future uses of AI by adversaries. This paper divides these threats into three categories: threats to the UK's digital security, political security and physical security.

---

55. Paula Kift, 'Augmentation as Artifice: A Palantir Look at AI', about:intel, 30 October 2019, <https://aboutintel.eu/palantir-augmented-intelligence/>, accessed 8 April 2020.
56. Cressida Dick, 'RUSI Annual Security Lecture', 24 February 2020, <https://rusi.org/event/rusi-annual-security-lecture>, accessed 17 April 2020.
57. See, for example, Centre for the Protection of National Infrastructure, 'National Security Threats'.

Digital Security

The threat from AI-enabled malware is likely to grow and evolve in the coming years. Specifically, polymorphic malware that employs complex obfuscating algorithms and frequently changes its identifiable characteristics could reach a level of adaptability that renders it virtually undetectable to both signature- and behaviour-based antivirus software. AI-based malware could proactively prioritise the most vulnerable targets on a network, iteratively adapt to the target environment and self-propagate via a series of autonomous decisions, potentially eliminating the need for a command-and-control (C2) channel.[58] A further concern is the use of domain-generation algorithms to continuously generate a large number of domain names to be used as rendezvous points between infected devices and C2 servers, which would make it considerably difficult to successfully shut down botnets.[59]

The automation of social engineering attacks is another potential threat. By collating a victim's online information, attackers can automatically generate malicious websites, emails and links that are custom-made for clicks from that victim (sent, for example, from addresses imitating their real contacts). Further developments in this area could see chatbots gaining human trust during longer and more creative online dialogues.[60]

The increased adoption of AI across the UK economy will also create new vulnerabilities which could be exploited by threat actors. Supply-chain attacks on training data ('data poisoning') could cause AI systems to behave in erratic and unpredictable ways, or allow attackers to install a 'backdoor' by which to take control of a system, for instance by training an algorithm to classify a particular malware as benign software.[61]

58. Miles Brundage et al., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', Future of Humanity Institute, February 2018, p. 25, <https://arxiv.org/pdf/1802.07228.pdf>, accessed 8 April 2020. A command-and-control server is a computer controlled by an attacker or cybercriminal which is used to send commands to systems compromised by malware and receive stolen data from a target network. See Trend Micro, 'Command and Control [C&C] Server', <https://www.trendmicro.com/vinfo/us/security/definition/command-and-control-server>, accessed 8 April 2020.
59. For further discussion, see Daniel Plohmann et al., 'A Comprehensive Measurement Study of Domain Generating Malware', in *Proceedings of the 25th USENIX Security Symposium* (Berkeley, CA: USENIX, 2016), pp. 263–78.
60. Brundage et al., 'The Malicious Use of Artificial Intelligence', p. 24.
61. For further discussion, see Nicolas Papernot et al., 'Practical Black-Box Attacks Against Machine Learning', in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (New York, NY: Association for Computing Machinery, 2017), pp. 506–19.

Political Security

The creation of 'deepfake' synthetic media and its impact on democratic processes has recently emerged as a significant concern.[62] Deepfakes involve the use of ML algorithms to combine or superimpose an existing piece of media (such as an image of an individual's face) onto genuine content. In May 2019, researchers at Samsung showcased an AI system that created videos of a person speaking based only on a single photo of that person.[63] The disruptive potential of this technology was showcased in the run-up to the 2019 general election in the UK when the research organisation, Future Advocacy, and artist, Bill Posters, created a deepfake video showing candidates Boris Johnson and Jeremy Corbyn endorsing each other for prime minister.[64] This was intended to warn the public of how AI technology can be used to fuel disinformation, erode trust and compromise democracy. Ahead of the 2020 US presidential election, experts are voicing their concerns about the 'very high likelihood that deepfake technology – video or voice – will be used … to actually compromise the election'.[65]

At present, modified data can be readily identified by media forensic experts. Nevertheless, in the time-sensitive context of an election, the identification of a fake video might simply come too late. Furthermore, given the rapid pace at which digital content spreads online, there is a legitimate concern that individuals in positions of power could take reactive decisions based on false information, with potentially catastrophic consequences.[66]

Physical Security

At present, there are few real use cases of how AI may be weaponised to directly threaten physical security. One area of concern could be the repurposing of commercial AI systems by terrorists – for instance, using drones and autonomous vehicles to carry out explosive attacks or cause serious crashes.[67] These risks may increase as the use of AI becomes increasingly normalised: connected autonomous vehicle uptake has recently been estimated to reach

62. Chris Marsden, Trisha Meyer and Ian Brown, 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?', *Computer Law and Security Review* (Vol. 36, 2020).
63. Egor Zakharov et al., 'Few-Shot Adversarial Learning of Realistic Neural Talking Head Models', Samsung AI Center and Skolkovo Institute of Science and Technology, May 2019, arXiv preprint, 1905.08233v2.
64. Luke O'Reilly, 'Boris Johnson Appears to Endorse Jeremy Corbyn as Prime Minister in Viral Deepfake Video', *Evening Standard*, 12 November 2019.
65. Ragavan Thurairatnam cited in Thor Benson, 'Experts Say Deepfakes Could Swing the 2020 Election', *Inverse*, 11 February 2020, <https://www.inverse.com/innovation/how-deepfakes-could-swing-the-election>, accessed 8 April 2020.
66. See, for example, Russell Goldman, 'Reading Fake News, Pakistani Minister Directs Nuclear Threat at Israel', *New York Times*, 24 December 2016; Ragavan Thurairatnam cited in Thor Benson, 'Experts Say Deepfakes Could Swing the 2020 Election', *Inverse*, 11 February 2020, <https://www.inverse.com/innovation/how-deepfakes-could-swing-the-election>, accessed 8 April 2020.
67. Brundage et al., 'The Malicious Use of Artificial Intelligence', p. 27.

31% of total vehicle sales by 2035 in the UK and wider Europe.[68] Moreover, it is likely that AI will transform what would previously be classed as high-skill attack capabilities into tasks which low-skill individuals can perform with little effort. This may take the form of 'swarming attacks', where 'distributed networks of autonomous robotic systems, cooperating at machine speed, provide ubiquitous surveillance to monitor large areas and groups and execute rapid, coordinated attacks'.[69]

Increased adoption of Internet of Things (IoT) technology,[70] the emergence of 'smart cities' and interconnected critical national infrastructure will create numerous new vulnerabilities which could be exploited by threat actors to cause damage or disruption. While these potential physical threats are yet to materialise, this situation could change rapidly, requiring government agencies to formulate proactive approaches to prevent and disrupt AI-enabled security threats before they develop.

68. Centre for Connected and Autonomous Vehicles, 'Connected and Autonomous Vehicles: Market Forecast', 7 September 2017, <https://www.gov.uk/government/publications/connected-and-autonomous-vehicles-market-forcecast>, accessed 8 April 2020.
69. Brundage et al., 'The Malicious Use of Artificial Intelligence', p. 28.
70. 'Internet of Things' refers to the growth of internet-connected devices (such as home appliances or wearable devices) which have not traditionally been connected to the internet.

II. Legal and Ethical Considerations

THIS CHAPTER SUMMARISES the legal framework regulating UKIC and its use of AI, before considering potential legal and ethical issues that could arise from the use of AI for national security purposes.

Legal Framework

The statutory functions of the UK intelligence agencies are set out in the Security Service Act 1989[71] and Intelligence Services Act 1994.[72] The 1989 Act (in respect of the Security Service) and 1994 Act (in respect of the Secret Intelligence Service and GCHQ) restrict the power to obtain and disclose information to that which is necessary for these agencies' functions. The framework governing most digital investigatory powers – including the interception of communications, equipment interference, obtaining of communications data and the acquisition of bulk datasets – is now laid out in the Investigatory Powers Act 2016 (IPA).[73] The IPA regime subjects the agencies to additional levels of scrutiny regarding their acquisition of data and use of investigatory techniques, scrutiny and oversight to which the private sector is not subject.[74] Section 2 of the IPA introduces several 'general duties in relation to privacy', including a requirement for the public authority to consider 'whether what is sought to be achieved by the warrant, authorisation or notice could reasonably be achieved by other less intrusive means'.[75] Directed and intrusive surveillance and the use of covert human intelligence sources continue to be governed by the Regulation of Investigatory Powers Act 2000.[76]

In addition to primary legislation, the agencies' activities are also governed by statutory codes of practice relating to intrusive powers pursuant to the 2000 and 2016 Acts, together with internal guidance and policies. Part 4 of the Data Protection Act 2018[77] sets out a separate data protection regime for the intelligence services. There are a number of national security

71. 'Security Service Act 1989 (UK)'.
72. 'Intelligence Services Act 1994 (UK)'.
73. 'Investigatory Powers Act 2019 (UK)'.
74. The powers given to UKIC are subject to a specific oversight regime set out in intelligence and surveillance legislation, while the private sector's use of data remains governed primarily by data protection frameworks.
75. 'Investigatory Powers Act 2019 (UK)', s2(2).
76. 'Regulation of Investigatory Powers Act 2000 (UK)'.
77. 'Data Protection Act 2018 (UK)'.

exemption certificates in place pursuant to the national security exemption in the Act,[78] although the agencies continue to be required to ensure that the use of personal data is both lawful and secure.

Several other general legal frameworks are relevant to the agencies' exercise of their functions. In particular, the Human Rights Act 1998 – which transposes into UK law the rights contained within the European Convention on Human Rights (ECHR) – ensures the protection of fundamental human rights and political freedoms, subject to certain restrictions. Some of these rights are absolute, meaning they cannot be limited or infringed under any circumstances (such as Article 3, the prohibition on torture and inhumane or degrading treatment or punishment, and Article 6, the right to a fair trial). Others – including the right to respect for one's private and family life, their home and their correspondence (Article 8), and the rights to freedom of expression (Article 10), assembly and association (Article 11) – are qualified rights, meaning that the state has the power to interfere with these rights provided that such interference is 'in accordance with the law' and 'necessary in a democratic society'. Conversely, the state has positive obligations under Articles 2 and 3 of the ECHR which it would breach if it fails to 'take measures within the scope of their powers which, judged reasonably, might have been expected to avoid' a risk to an individual or society.[79] One could therefore also argue that the agencies have a positive obligation to adopt new technological methods that would improve their ability to protect the public from threats to their safety.

Julia Black and Andrew D Murray argue that:

> [W]hilst it is important that the overall regime for AI regulation is coherent, it does not need to, and indeed should not, operate in isolation from existing regulatory regimes. Where an activity is already regulated under a specific regulatory regime, then the use of AI in the development or deployment of that activity, for example in the development of medical treatments or devices, is captured within the perimeter of an existing regulatory regime. Those regulators need to develop norms for the use of AI, and quickly, but the mechanism is there.[80]

In addition to the 1989, 1994 and 2016 Acts, further protections exist by virtue of the Human Rights Act 1998 such that any analysis of data obtained through intrusive techniques, or otherwise, must be carried out only as necessary for the agencies' statutory functions and subject to the required ongoing human rights proportionality assessment. The UK Supreme Court has developed a four-stage proportionality test for assessing, pursuant to the Human Rights Act 1998, whether a measure that infringes a fundamental human right is proportionate.

78. Information Commissioner's Office, 'National Security Certificates', <https://ico.org.uk/about-the-ico/our-information/national-security-certificates/>, accessed 8 April 2020.
79. Emma Lazarovna Tagayeva and Others v Russia, Application Nos. 26562/07, 49380/08, 21294/11, 37096/11, 14755/08, 49339/08, 51313/08, 13 April 2017, para. 482.
80. Julia Black and Andrew D Murray, 'Regulating AI and Machine Learning: Setting the Regulatory Agenda', *European Journal of Law and Technology* (Vol. 10, No. 3, 2019).

This test was set out in the *Bank Mellat* case as follows:

1. Is the objective of the measure pursued **sufficiently important** to justify the limitation of a fundamental right?
2. Is it **rationally connected** to the objective?
3. Could a **less intrusive measure** have been used without unacceptably compromising the objective?
4. In regard to these matters and to the severity of the consequences, has a **fair balance** been struck between the rights of the individual and the interests of the community?[81]

This human rights proportionality test provides criteria that the agencies can use to assess the legitimacy of new uses of technology, including AI. However, because existing authorisation processes focus on the *collection* of data (rather than subsequent analysis), internal processes will need to continue to re-assess the necessity and proportionality of any potential intrusion if AI is subsequently applied to data previously obtained. This reflects a point made in a report by IPCO's Technology Advisory Panel:

> It will be increasingly difficult with the growth of Artificial Intelligence (AI) to know what analytical work has been done on the data. The intrusion caused by obtaining and retaining the data is not a fixed impact but will vary according to the people whose data it is and what other data is available and may be combined with the original data. *It is essential to reassess the potential for intrusion constantly as analytic processes change and develop*.[82]

In his oral evidence to the House of Commons Public Bill Committee on the Investigatory Powers Bill, surveillance expert Eric Kind (formerly Eric King) described an 'intermediary stage' after data is collected but before it is reviewed by an analyst, where additional safeguards may be needed to account for the analytical processes applied between the point of collection and human analysis.[83]

Much concern over the acquisition of communications data focuses on the insights that can be gleaned about an individual's personal life.[84] The type of analysis applied to a collected dataset has direct implications in this regard, implying the need for an additional assessment of the extent of intrusion into individual rights, and (assuming intrusion exists) necessity and proportionality focused specifically on the *analytical processes* which may be applied to collected data. This is

81. Bank Mellat v HM Treasury (No. 2), UKSC 39, UKSC 2011/0040, 2013.
82. Technical Advisory Panel of the Investigatory Powers Commissioner's Office, 'Metrics of Privacy', 14 November 2018, <https://www.ipco.org.uk/docs/Formal%20report_Metrics%20of%20Privacy%20Conference.pdf>, accessed 8 April 2020. Emphasis in original.
83. House of Commons Public Bill Committee, 'Investigatory Powers Bill', 24 March 2016, <https://publications.parliament.uk/pa/cm201516/cmpublic/investigatorypowers/160324/am/160324s01.htm>, accessed 17 April 2020.
84. UN General Assembly, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue', A/HRC/23/40, 17 April 2013.

particularly important considering the analysis of bulk datasets will include the processing of data about many individuals who are not of intelligence interest.

Any future policy or guidance for national security uses of AI must pay due regard to issues such as necessity and proportionality, transparency and accountability, and collateral intrusion risk. As these issues are highly context-dependent, any future policy should be 'mission-agnostic' and principles-based, establishing standardised processes to ensure that AI projects follow recommended routes for empirical evaluation of algorithms within their operational context, and assess each project against legal requirements and ethical standards.

Machine Intrusion

The question of whether the use of AI represents increased privacy intrusion or a method by which intrusion could be reduced remains a matter of debate. The use of AI arguably has the potential to reduce intrusion, both in terms of minimising the volume of personal data that needs to be reviewed by a human operator, and by resulting in more precise and efficient targeting, thus minimising the risk of collateral intrusion. However, it has also been argued that the degree of intrusion is equivalent regardless of whether data is processed by an algorithm or a human operator. According to this view, the source of intrusion lies in the collection, storage and processing of data. The methods by which this is achieved – whether automated or manual – are immaterial.

European case law has long held that the collection and retention of personal data can constitute an infringement of human rights, regardless of whether it is reviewed by a human or a machine.[85] The Anderson report highlighted a difference in opinion between European and UK courts in this regard, noting that:

> … the ECtHR [European Court of Human Rights] has traditionally been readier than the English courts to find that Article 8 is engaged, or engaged in more than a minor respect. In the context of investigatory powers, it is engaged not only when material is read, analysed and later shared with other authorities, but also when it is collected, stored and filtered, even without human intervention.[86]

In *R (National Council for Civil Liberties) v Secretary of State for the Home Department*,[87] the court highlighted a fundamental difference of opinion between the claimant and the government by stating that:

> It is common ground between the parties that there is an interference with the right to respect for private life at all material stages, including at the stage when data is obtained and retained. However, the Defendants submit that there is no 'meaningful' intrusion into privacy rights until the stage when

85. S and Marper v UK, ECHR 1581, 2008.
86. Anderson, *A Question of Trust*, p. 75.
87. R (National Council for Civil Liberties) v Secretary of State for the Home Department, EWHC 2057 (Admin), 2019.

> the data is selected for examination. The Claimant submits that that is wrong and inconsistent with 'decades' of authority from the European Court of Human Rights. It also submits that this is a proposition which is not only 'startling' but 'dangerous and artificial'.[88]

Furthermore, it has also been suggested that the algorithmic analysis of data could be *more* intrusive than parametric keyword searches. 'If the automatic collection and storage of information already constitutes a violation of privacy — a view that is supported by the European Court of Human Rights — the algorithmic analysis of data that goes beyond a simple keyword search must do so even more', explains Ronja Kniep.[89] Use of AI could result in additional material being processed which may not have previously been possible for technical or capacity-related reasons. This would need to be taken into account when assessing proportionality of any potential intrusion, balanced against the increase in effectiveness of analysis that may result.

It is possible to anticipate future disputes arising about the interpretation of intrusion in circumstances where AI has been deployed to analyse data. European case law suggests that it should not be assumed that the use of automated data processing methods is inherently less intrusive than human review. In some cases, automated processing may lead to the examination of data which would not otherwise have been flagged for human review. This would need to be considered when assessing the necessity and proportionality of potential intrusion. Even a low level of intrusion by a machine could potentially lead to a higher level of mistaken intrusion – if improper machine analysis leads to the wrong course of action being taken. Conversely, AI analysis could be viewed as a more proportionate alternative to human review if it resulted in more effective use of data to identify and mitigate threats. The potential consequences of analysis will also need to be considered when making these judgements, as this will have significant implications for potential intrusion into individual rights.

It is important to note, however, that standardised processes already exist to assess the necessity and proportionality of any potential intrusion when accessing previously collected data. As summarised in IPCO's most recent annual report (with regards to GCHQ):

> Where there is an operational requirement to access data, which will include bulk communications data (BCD) and/or bulk personal data (BPD), an analyst must justify why the access and examination of the data are necessary and proportionate and must record the specific intelligence requirement and priority for each search
>
> …
>
> Whenever GCHQ analysts conduct a query of bulk data, they are required to draft a statement explaining why their query is necessary and proportionate. Overall, we concluded that these justifications were

88. *Ibid.*, para. 6.
89. Kniep, 'Another Layer of Opacity'.

> meeting the required standard and analysts were accounting for the proportionality of their queries of bulk data in sufficient detail.[90]

Considering the potential privacy implications as new analysis methods are applied to previously collected datasets, such internal processes will need to continue to assess the necessity and proportionality of any potential intrusion if AI is subsequently applied to previously collected data.

Finally, there is also a concern that the use of multiple AI systems in conjunction with each other could result in a 'cumulative intrusion risk'. The risks highlighted above could be compounded when automated systems interact with each other, resulting in an interconnected network of systems that results in significantly greater levels of intrusion than in the case of each system in isolation. As discussed in the Anderson report, 'intrusions into privacy have been compared, compellingly, to environmental damage: individually their impact may be hard to detect, but their cumulative effect can be very significant'.[91] This suggests the need for internal processes to monitor the overall cumulative effects of automated data processing systems and any resulting compound intrusion risk, as well as the extent to which this is judged to be both necessary and proportionate.

Collection and Retention

Collection

Large datasets may be needed to train ML algorithms, and much of the information contained therein may not be of national security concern. Training data could come from a number of different sources. For instance, text data could be used to train a machine translation system, or image databases to train an object classifier. An agency's operational data could be used to train a system to identify potential targets or relationships between entities within bulk datasets. The privacy and human rights implications will vary considerably depending on the source of training data used and how it is acquired.

The justification for collection of bulk datasets has been discussed at length elsewhere. For example, Anderson's Bulk Powers Review concluded that 'bulk powers play an important part in identifying, understanding and averting threats in Great Britain, Northern Ireland and further afield. Where alternative methods exist, they are often less effective, more dangerous, more resource-intensive, more intrusive or slower'.[92] In *Big Brother Watch v UK*, the court concluded that 'it is clear that bulk interception is a valuable means to achieve the legitimate aims pursued, particularly given the current threat level from both global terrorism and serious crime'.[93] In

90. IPCO, *Annual Report 2018*, pp. 49–52.
91. *Ibid.*, p. 27.
92. Anderson, *Report of the Bulk Powers Review*, p. 1. The bulk powers defined in the Investigatory Powers Act 2016 are bulk interception, bulk acquisition, bulk equipment interference and bulk personal datasets.
93. Big Brother Watch v UK, Nos. 58170/13, 62322/14 and 24960/15, 13 September 2018.

relation to bulk personal datasets (BPDs),[94] a recent court ruling concluded that 'in some areas, particularly pattern analysis and anomaly detection, no practicable alternative to the use of BPDs exists. Where an agency does not have the "seed" of intelligence usually needed to begin an investigation, these techniques enable it to spot hostile activity or actors'.[95] This conclusion was based on operational examples provided to the court, reflecting the importance of such evidence to any future determination of necessity.

The ongoing challenges by Privacy International and Liberty to the IPA bulk powers should be noted. In particular, there is pending reference at the European Court of Justice (CJEU) questioning whether the activities of intelligence agencies relating to bulk communications data are governed by EU law.[96] Challenges to the use of bulk data are likely to continue with implications that would need to be considered carefully for the potential use of such data within AI systems.

Data Retention and 'Model Leakage'

There is an ongoing academic debate over whether the retention of a trained ML model could be viewed as equivalent to the retention of the underlying training data. For UKIC, this will have implications for the retention, security classification and handling requirements of trained ML models.

Recent academic research has demonstrated that ML methods can be vulnerable to a range of cyber security attacks that may lead to breaches of confidentiality.[97] Of concern in this regard are 'model inversion' and 'membership inference' attacks. As summarised by Michael Veale, Reuben Binns and Lilian Edwards, model inversion 'turns the journey from training data into a machine-learned model from a one-way one to a two-way one, permitting the training data to be estimated with varying degrees of accuracy', while membership inference 'does not recover the training data, but instead recovers information about whether or not a particular individual was in the training set'.[98] The authors conclude that 'where models are vulnerable to such attacks, they gain an additional dimension – not only an analytic product potentially protected

94. 'Bulk personal datasets' are datasets retained by the intelligence services about individuals, the majority of whom are not likely to be of intelligence interest. See 'Investigatory Powers Act 2016 (UK)', Part 7.
95. R (National Council for Civil Liberties) v Secretary of State for the Home Department, EWHC 2057 (Admin), 2019, para. 222.
96. See Investigatory Powers Tribunal, 'IPT/15/110/CH', <https://www.ipt-uk.com/judgments.asp?id=40>, accessed 8 April 2020; InfoCuria, 'Case C-623/17', <https://tinyurl.com/t87ubts>, accessed 8 April 2020. It is also worth noting that the UK's compliance with EU fundamental rights will have implications for any data protection adequacy decision after the Brexit transition period.
97. For further discussion, see Liyang Xie et al., 'Differentially Private Generative Adversarial Network', arXiv preprint, arXiv:1802.06739, 2018.
98. Michael Veale, Reuben Binns and Lilian Edwards, 'Algorithms That Remember: Model Inversion Attacks and Data Protection Law', *Philosophical Transactions of the Royal Society* (Vol. 376, 2018).

by intellectual property rights, but also a set of personal data, conceptually close to the idea of "pseudonymization" in the GDPR [General Data Protection Regulation]'.[99]

It is important to note, however, that this is only one academic interpretation of the legal question as to whether a model should be treated as 'personal data'. Mark Leiser and Francien Dechesne, on the other hand, argue that:

> ... as opposed to *databases*, inversion and membership inference models can only ever contain unstructured, anonymous data. While an attack might 'leak' data, this does not make the model personal … the model does not constitute the crucial information referring to a natural person. Rather, the model, containing correlation of numeric parameters of training data, is the tool in the process.[100]

The authors conclude that there is no personal data contained within a model and a reidentification would not be done by a model but by a 'skilled human'.[101] The extent to which these concerns are relevant to UKIC is also unclear, as the models in question would be deployed in a secure environment and are therefore less likely to be vulnerable to confidentiality attacks from adversarial actors.

Where possible, differential privacy methods may protect models from potential confidentiality attacks (although this may lead to a reduction in the model's accuracy). Pseudonymisation could also be used to store data in such a way that it is possible to conduct analysis without inferring properties about individuals, while homomorphic encryption could make it possible to perform operations on a dataset without needing to decrypt the data. It is likely that the protections required for a trained model will depend largely on the type of ML used, the extent to which these methods are vulnerable to model inversion or membership inference attacks, and the context and environment in which the models are used.

Testing and Deployment

Does it Work?

A potential risk to individuals is the reliability of AI systems used to process personal data. In a national security context, the consequences of errors can be very high, particularly if an AI system is integrated into a decision-making process which may result in direct action being taken against an individual. For such analysis to be justified, agencies must be sufficiently confident that the capability they are seeking to deploy will deliver the desired outcomes while balancing the potential benefits against the level of intrusion arising from data collection and analysis. Ensuring the validity and reliability of statistical algorithms used by the agencies will require establishing context-specific evaluation processes that assess the real-world effectiveness of a

99. *Ibid.*
100. Mark Leiser and Francien Dechesne, 'Governing Machine Learning Models: Challenging the Personal Data Presumption', *International Data Privacy Law* (forthcoming 2020).
101. *Ibid.*

tool when deployed in a live operational context. As well as evaluating reliability and statistical accuracy, this process should also include developing standardised terminology for how error rates and other relevant technical information should be communicated to human operators.

Behavioural Profiling, Bias and Discrimination

Concerns have been raised regarding the ability of ML algorithms to build comprehensive 'profiles' of individuals in a way that traditional methods do not.[102] Such 'algorithmic profiling' could be considered inherently more intrusive than manual analysis of collected data, and would raise further human rights concerns if it were perceived to be unfairly biased or discriminatory. A report from Cardiff University's Data Justice Lab highlighted particular risks regarding the 'possibility for targeting, stigma and stereotyping of particular groups with the labelling of "risk"' and the 'lack of transparency, public knowledge, consent, and oversight in how data systems are being implemented and used'.[103] 'Predictive policing' tools have received much criticism in this regard, with claims that they over-predict individuals from certain racial groups, or particular neighbourhoods where postcodes function as a 'proxy variable' for race.[104] In an intelligence context, there is also a risk that biases in historic data may result in important case-specific information being overlooked, and that the reliance on historic data may only reveal insights related to threats which appear similar to data items that have been encountered previously.

However, while much commentary has focused on the ability of AI systems to replicate or amplify biases inherent in collected data, it is often argued that these systems are likely to be no more biased than existing human decision-making processes. There are over 180 known cognitive biases, and although some of these are more trivial than others, research has consistently shown that human decision-makers do not have the insight into their own decisions that is often assumed.[105] More importantly, the use of AI could potentially reveal underlying biases in datasets which would otherwise go unnoticed. As summarised by Helen Margetts, 'some of our existing systems are designed in a way that makes it impossible to measure bias … One of the

102. See, for example, Dimitra Kamarinou, Christopher Millard and Jatinder Singh, 'Machine Learning with Personal Data', Queen Mary School of Law Legal Studies Research Paper 247, 2016.
103. Lina Dencik et al., 'Data Scores as Governance: Investigating Uses of Citizen Scoring in Public Services', Cardiff University Data Justice Lab, December 2018, p. 4.
104. See, for example, Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact', *California Law Review* (Vol. 104, No. 3, June 2016), p. 671; Rashida Richardson, Jason M Schultz and Kate Crawford, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice', *New York University Law Review*, Online Feature, <https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/>, accessed 8 April 2020; Danielle Ensign et al., 'Runaway Feedback Loops in Predictive Policing', *Proceedings of Machine Learning Research* (Vol. 81, No. 1, 2018), pp. 1–12.
105. See Daniel Kahneman, *Thinking, Fast and Slow* (New York, NY: Macmillan, 2011); Ben Yagoda, 'The Cognitive Biases Tricking Your Brain', *The Atlantic*, September 2018.

good things about machine learning technologies is that they have exposed some bias which has always been there'.[106]

Nevertheless, law and regulation has developed over time to govern such human frailties and to safeguard against bias in human decision-making, but the same safeguards do not yet exist in the context of algorithmic decision-making. For this reason, internal processes are needed to ensure fairness in algorithm-assisted decision-making. Throughout all stages of an AI project, attention should be given to the representativeness of the model's outputs, and whether they display any evidence of unfair discrimination. Processes are needed for ongoing tracking and mitigation of discrimination risk (alongside ongoing re-evaluation of a model's precision and recall). As discussed by the Committee on Standards in Public Life, this will require ensuring diversity in AI project teams, as 'a workforce composed of a single demographic is less likely to check for and notice discrimination than diverse teams'.[107] Workforce diversity is not only important for identifying the risk of bias within datasets, but also for identifying operational impacts that may be more detrimental for certain demographic groups.

Transparency and Accountability

Much commentary has raised concerns regarding the 'black box' nature of certain ML methods, which may lead to a loss of accountability of the overall decision-making process.[108] Deep learning methods are generally inscrutable to human users, meaning it is not possible to assess the factors that were taken into account during computation. In some cases, the use of black-box methods may not introduce any additional risks to the data subject(s) or human operators. In other cases, particularly when AI systems are deriving insights at the individual subject level, it may be unacceptable for human users to have no knowledge of the factors that were considered during computation. There is also a related risk that operators may become over-reliant on AI systems or 'defer' to algorithmic insights at the expense of their own professional judgement, rendering the resultant decision a de facto automated one.[109]

---

106. Helen Margetts cited in Committee on Standards in Public Life, 'Artificial Intelligence and Public Standards', p. 27.
107. *Ibid.*
108. See, for example, Cavan and Killworth, 'GCHQ Embraces AI, but not as a Black Box'; Roger Levy, 'The Black Box Problem', *RUSI Journal* (Vol. 164, No. 5–6, July/August 2019), pp. 82–87; Davide Castelvecchi, 'Can We Open the Black Box of AI?', *Nature* (Vol. 538, No. 7623, 5 October 2016); Wojciech Samek, Thomas Wiegand and Klaus-Robert Müller, 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models', arXiv preprint, arXiv:1708.08296v1, 28 August 2017.
109. For further discussion, see Hannah Couchman and Alessandra Prezepiorski Lemos, 'Policing by Machine: Predictive Policing and the Threat to Our Rights', Liberty, February 2019; Keith Dear, 'Artificial Intelligence and Decision-Making', *RUSI Journal* (Vol. 164, No. 5–6, July/August 2019), pp. 18–25; Ben Koppelman, 'How Would Future Autonomous Weapon Systems Challenge Current Governance Norms?', *RUSI Journal* (Vol. 164, No. 5–6, July/August 2019), pp. 98–109; Michael A Froomkin, Ian Kerr and Joelle Pineau, 'When AIs Outperform Doctors: Confronting the Challenges

There is unlikely to be a one-size-fits-all solution to the concerns expressed about transparency and accountability. The extent to which it is necessary to explain the factors which were considered when arriving at a certain output will depend largely on the context in which the algorithm is applied and the overall decision-making process that it informs. In order to ensure that human operators retain ultimate accountability for the overall decision-making process informed by analysis, it will be essential to design systems in such a way that non-technically skilled users can interpret key technical information, such as the margins of error and uncertainty associated with a calculation. Intelligence professionals are trained to make decisions in conditions of uncertainty. The output of an AI system should be treated as another source of information for the user to consider in conjunction with their own professional judgement. Context-sensitive internal oversight processes are needed to ensure AI tools are used to support (rather than replace) human judgement, considering the reality that each deployment will give rise to different transparency and accountability challenges.

It will also be important to maintain senior organisational accountability for the development and deployment of AI systems, ensuring those with management, monitoring and approval responsibilities fully understand the limitations and risks associated with different methods. Achieving this will require developers and technical experts to be able to translate complex information (such as error rates and confidence intervals) in such a way that senior decision-makers can assume overall accountability for the tool and how it is deployed operationally.

---

of a Tort-Induced Over-Reliance on Machine Learning', *Arizona Law Review* (Vol. 61, No. 1, 2019), p. 33.

III. Regulation, Guidance and Oversight

THIS CHAPTER SUMMARISES existing guidance and professional standards relating to the development and deployment of AI, and considers additional sector-specific guidance that may be needed in the context of national security. This is followed by a review of the roles and responsibilities regarding monitoring and oversight.

Existing Guidance

Although discussions on the ethical use of AI are now well-established, these are largely yet to translate into operationally relevant guidance that stakeholders can implement in practice. Without establishing clear boundaries regarding permissible and unacceptable uses of AI, the fear of falling on the wrong side of the ethical divide may impede the potential of better results from newer, often experimental methods.

The paper's annex provides 19 examples of AI-related guidance. In the UK public sector, the most relevant guidance has been provided by the Department for Digital, Culture, Media and Sport[110] and the Alan Turing Institute.[111] Other guidelines – such as the National Cyber Security Centre guidance on 'Intelligent Security Tools'[112] and the HM Treasury *Aqua Book*[113] – are also relevant. These focus on maximising the potential of data analytics projects in a responsible, proportionate way which is mindful of the potential limitations at each stage of a project lifecycle.

110. Department for Digital, Culture, Media and Sport, 'Data Ethics Framework', 13 June 2018, <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737137/Data_Ethics_Framework.pdf>, accessed 8 April 2020.
111. Government Digital Service and Office for Artificial Intelligence, 'Understanding Artificial Intelligence Ethics and Safety', 10 June 2019, <https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety>, accessed 8 April 2020.
112. National Cyber Security Centre, 'Intelligent Security Tools: Assessing Intelligent Tools for Cyber Security', 2019, <https://www.ncsc.gov.uk/collection/intelligent-security-tools>, accessed 8 April 2020.
113. HM Treasury, *The Aqua Book: Guidance on Producing Quality Analysis for Government* (London: Stationery Office, 2015), p. 6.

Guidelines produced by the OECD,[114] European Commission[115] and the UN[116] focus on aspects of 'trustworthiness' in AI systems, and on identifying ways of managing the intersections between AI and longstanding legal principles. In the private sector, Google,[117] Microsoft[118] and IBM[119] have been proactive in setting out their recommendations for how companies should approach AI projects in a responsible and ethical way. These recommendations focus on issues of unfair bias, unintended consequences, and transparency and accountability. The World Economic Forum has also provided 10 guidelines for AI procurement processes,[120] and the AI Now Institute has published its version of an 'Algorithmic Impact Assessment' for 'public agency accountability'.[121]

In the area of predictive policing, the National Police Chiefs' Council has adopted the 'Algocare' model, and together with additional explanatory documentation, recommends its use to chief constables.[122] Algocare aims to translate key public law and human rights principles into practical considerations and guidance that can be addressed by public sector bodies when implementing AI, and could also be a useful starting point for national security-specific AI guidance.

114. OECD, 'OECD Principles on AI', May 2019, <http://www.oecd.org/going-digital/ai/principles/>, accessed 8 April 2020.
115. European Commission, 'Ethics Guidelines for Trustworthy AI', 8 April 2019, <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>, accessed 8 April 2020.
116. UN, 'Secretary-General's High-Level Panel on Digital Cooperation', 12 July 2018, <https://www.un.org/sg/en/content/sg/personnel-appointments/2018-07-12/secretary-generals-high-level-panel-digital-cooperation>, accessed 8 April 2020; UN Interregional Crime and Justice Research Institute, 'Centre on Artificial Intelligence and Robotics', <http://www.unicri.it/topics/ai_robotics/>, accessed 8 April 2020.
117. Google, 'Artificial Intelligence at Google: Our Principles', 7 June 2018, <https://ai.google/principles>, accessed 8 April 2020.
118. Saleema Amershi et al., 'Guidelines for Human–AI Interaction', in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper No. 3 (New York, NY: Association for Computing Machinery, 2019).
119. Ryan Hagemann and Jean-Marc Leclerc, 'Precision Regulation for Artificial Intelligence', IBM Policy Lab, 21 January 2020, <https://www.ibm.com/blogs/policy/ai-precision-regulation/>, accessed 8 April 2020.
120. World Economic Forum, 'AI Government Procurement Guidelines', September 2019, <https://www.weforum.org/whitepapers/ai-government-procurement-guidelines>, accessed 17 April 2020.
121. Dillon Reisman et al., 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability', AI Now Institute, April 2018, <https://ainowinstitute.org/aiareport2018.html>, accessed 8 April 2020.
122. See Marion Oswald et al., 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental" Proportionality', *Information and Communications Technology Law* (Vol. 27, No. 2, 2018), pp. 223–50.

Various bodies are also engaged in advisory activities relating to the responsible and ethical use of AI, including the Centre for Data Ethics and Innovation,[123] the Information Commissioner's Office,[124] the Office for AI,[125] parliamentary and independent committees,[126] bodies with sector expertise or policymaking functions,[127] and campaigning organisations.[128] The roles and responsibilities of these stakeholders, as well as their regulatory remit, will need to be more clearly defined to ensure that work is not duplicated and they are able to provide meaningful oversight of government AI projects.

National Security-Specific Guidance

UKIC operates within a highly specific regulatory framework. The agencies may wish to implement AI systems in very different ways and for different purposes, and will therefore need to consider a range of factors which may not be relevant for other sectors. This, in turn, demands a more sector-specific approach to guidance and oversight. Future guidance should establish standardised processes to continuously assess the risks and benefits of national security AI deployments on an ongoing basis. An agile approach within the existing oversight regime to anticipate and understand the opportunities and risks presented by new AI capabilities appears essential to avoid creating excessive layers of oversight. Without finding this balance, there is a risk of stifling innovation and undermining the agencies' ability to adapt in response to the rapidly evolving technological environment and threat landscape.

Moreover, discussions regarding the potential risks of AI are often focused on extreme examples of theoretical future uses, which are typically detached from the reality of how the technology is currently being used. As a result, valid concerns regarding the ethical implications of AI may be overshadowed by speculation over unrealistic worst-case-scenario outcomes. It may therefore be difficult for organisations to develop clearer, operationally relevant guidance for the legitimate use of AI if discussions do not consider the likely and realistic applications of AI as an incremental development of existing capabilities and processes.

123. See UK Government, 'Centre for Data Ethics and Innovation', <https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation>, accessed 8 April 2020.
124. See ICO, <https://ico.org.uk/>, accessed 8 April 2020.
125. See GOV.UK, 'Office for Artificial Intelligence', <https://www.gov.uk/government/organisations/office-for-artificial-intelligence>, accessed 8 April 2020.
126. See, for example, UK Parliament, 'Select Committee on Artificial Intelligence', <https://www.parliament.uk/ai-committee>, accessed 8 April 2020; All-Party Parliamentary Group on Data Analytics, <https://www.policyconnect.org.uk/appgda/home>, accessed 8 April 2020.
127. See, for example, Partnership on AI, <https://www.partnershiponai.org/about/>, accessed 8 April 2020; AI Now, <https://ainowinstitute.org/>, accessed 8 April 2020; IEEE (Institute of Electrical and Electronics Engineers), <https://www.ieee.org/>, accessed 8 April 2020.
128. See, for example, Liberty, <https://www.libertyhumanrights.org.uk/>, accessed 8 April 2020; Privacy International, <https://privacyinternational.org/>, accessed 8 April 2020; Big Brother Watch, <https://bigbrotherwatch.org.uk/>, accessed 8 April 2020.

In developing a clearer policy framework for national security uses of AI, there is an opportunity for UKIC to take a more active role in government AI policymaking more broadly. The current approach to AI development across the UK government has been characterised as disjointed and uncoordinated, for instance by the Committee on Standards in Public Life, which found that '[p]ublic sector organisations are not sufficiently transparent about their use of AI and it is too difficult to find out where machine learning is currently being used in government'.[129] Developing a more coherent cross-government approach will require drawing on diverse, multi-disciplinary expertise from across the public sector, and there would be considerable value in leveraging the deep technological expertise within the agencies for the benefit of wider government policy development.

But policy and guidance can only go so far. The legitimate use of AI will also require complex and context-specific judgements to be made by individuals on a case-by-case basis. In a context where the regulatory framework is not yet fully established, this gives rise to the risk of increasing 'responsibilisation' of the individual user to determine what the ethical position is in any given context. This raises further questions about the distribution of responsibility if mistakes were made in the operationalisation of AI. In light of these issues, it is important to foster a culture where users and decision-makers feel empowered to make informed ethical judgements, supported by a collaborative environment in which open communication is actively encouraged.

Monitoring and Oversight

Finally, for any new policy framework to provide meaningful safeguards, it will be important to reassure the public of the robustness and resourcing of oversight. Recent events have highlighted concerns regarding the oversight of the agencies' use of data capabilities. In June 2019, IPCO reprimanded MI5 for the 'unlawful' handling of personal data, and claimed they would need to 'be satisfied to a greater degree than usual' that the agency's data handling regime was 'fit for purpose'.[130] However, in October 2019, IPCO released a statement following the conclusion of their targeted inspections of MI5. It concluded that 'MI5 has devoted substantial resources both to the programme of work to fix the compliance problems identified and to service this intensive inspection regime'. In addition, it stated that it was 'impressed by MI5's reaction to our criticisms, in the speed, focus, and dedication with which they acted to rectify the situation'.[131]

These recent public disputes illustrate the pressing need to ensure the regulatory apparatus is appropriately equipped to provide robust and comprehensive oversight of complex technical issues. IPCO has a central role to play as the appropriate regulatory body responsible for oversight in this context. In addition to the judicial commissioners' statutory responsibilities for authorisation and inspection of warrants, it will be important to ensure that specific technical

129. Committee on Standards in Public Life, 'Artificial Intelligence and Public Standards', p. 6.
130. David Bond, 'MI5 Under Fire for "Unlawful" Handling of Personal Data', *Financial Times*, 11 June 2019.
131. IPCO, 'Compliance Inspections of MI5 Complete', 22 October 2019, <https://www.ipco.org.uk/Default.aspx?mid=4.32>, accessed 8 April 2020.

issues regarding the development and deployment of AI can be reviewed and discussed on an ongoing basis. The legal and ethical issues discussed above can be highly subjective, and mechanisms are needed to ensure that the national security community considers the perspectives of a diverse range of stakeholders when making internal policy decisions. As noted by Jamie Gaskarth, 'to ensure that there is robust challenge of intelligence policy in the coming age of AI, there will need to be a strong system of vernacular accountability in place, with contributions from individuals from a diverse range of backgrounds, questioning everyday practice and policy assumptions'.[132]

Beyond its statutory oversight and inspection roles, IPCO could also play an important role in convening external experts to discuss these issues in a confidential environment. Its most recent annual report details how it has been involved in various external engagement activities with academics, NGOs and others, suggesting that it intends to expand these engagement activities in the coming years:

> The Inspectorate only reached full strength in January 2019 and our much-needed policy and engagement teams only joined during the summer of 2019. Even with these welcome developments, at the end of 2019, we do not yet have a full team in place. This has undoubtedly had an impact on our ability to take the initiative in a number of areas, especially in terms of our external communications and engagement, both of which are important fields that undoubtedly need development.[133]

In addition to IPCO, there are a number of other stakeholders to consider in the context of monitoring and oversight, including the Intelligence and Security Committee of Parliament and the Independent Reviewer of Terrorism Legislation.

This research has highlighted the importance of drawing on diverse multidisciplinary expertise when understanding the opportunities and challenges posed by AI in the national security context. This will need to be reflected both in the resourcing of oversight and the approach to external stakeholder engagement. In addition, it is crucial to ensure that those responsible for monitoring and oversight have access to sufficient technical expertise and the information needed to make informed and context-specific judgements regarding acceptable uses of new technology, including AI.

132. Jamie Gaskarth, *Secrets and Spies: UK Intelligence Accountability after Iraq and Snowden* (Washington, DC: Brookings Institution Press: 2020), p. 120.
133. IPCO, *Annual Report 2018*, p. 7.

Conclusions

AI HAS THE POTENTIAL to enhance many aspects of intelligence work. Taking full advantage of these opportunities requires establishing standardised processes for developing, testing and evaluating new AI tools in their operational context. The agencies may seek to deploy AI in numerous ways. These vary considerably in terms of their data requirements, potential impact on decision-making and ethical implications. Many uses will be uncontentious, if they simply reduce the time and effort required to work through large volumes of data which would have previously been processed using less efficient manual methods. Other uses may raise complex privacy and human rights concerns, requiring processes for regular review and reassessment of the necessity and proportionality of any potential intrusion, the choice of training data used to build a model and the decision-making process into which an algorithm may be embedded. At the outset of any new AI project, internal processes are needed to assess potential privacy and human rights implications and the level of oversight that will therefore be needed.

UKIC operates within a tightly restricted legal framework. The IPA regime subjects the agencies to additional levels of scrutiny regarding their acquisition of data and use of investigatory techniques – scrutiny and oversight to which the private sector is not subject. Collection capabilities such as equipment interference and bulk powers, whose existence was secret until 2015, are now publicly avowed as essential components of UKIC's technical toolkit. Nevertheless, the use of AI introduces a number of additional considerations, suggesting that enhanced policy and guidance are needed to ensure that AI analysis capabilities are deployed in an ethical and responsible way and with due regard to issues such as necessity and proportionality, transparency and accountability, and collateral intrusion risk.

Concerns regarding the ethical use of AI are highly subjective and context specific. Experts continue to disagree over fundamental questions such as the relative level of intrusion of machine analysis when compared with human review. Despite a proliferation of ethical principles, there is a lack of clarity on how these should be operationalised in different sectors, and who should be responsible for oversight and scrutiny.

Moreover, it is crucial for UKIC to continue to engage with external stakeholders to inform the development of internal policy regarding its use of new technologies, including AI. In addition to engaging with other government departments and those with oversight responsibilities, this should also include incorporating views from civil society organisations and other public interest groups, as well as drawing on lessons learned from other sectors in the development and deployment of AI.

About the Authors

Alexander Babuta is a Research Fellow at RUSI. His research focuses on the use of emerging technologies for national security and policing. He publishes regularly on issues related to surveillance policy, data ethics and artificial intelligence.

Marion Oswald is Vice-Chancellor's Senior Fellow in Law at the University of Northumbria, an Associate Fellow of RUSI and a solicitor (non-practising). She is Chair of the West Midlands Police and Crime Commissioner and West Midlands Police Data Ethics Committee, a member of the National Statistician's Data Ethics Advisory Committee, a member of the Advisory Board to the Ada Lovelace Institute Ryder Review of the Governance of Biometric Technologies and an executive member of the British and Irish Law, Education and Technology Association.

Ardi Janjeva is a Research Analyst at RUSI. His research currently spans numerous areas within Organised Crime and National Security, including the application of emerging technologies for national security and law enforcement, intellectual property crime and counterfeiting, and cyber-enabled fraud.

Annex: Selected AI Guidance and Ethical Principles

Organisation	Publication	Content
Department for Digital, Culture, Media and Sport *UK*	'Data Ethics Framework', 13 June 2018, <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737137/Data_Ethics_Framework.pdf>, accessed 9 April 2020.	Seven principles against which to measure projects, including: clear user need and public benefit; use of data proportionate to user need; understanding of data limitations with robust evaluation plan.
HM Government and Alan Turing Institute *UK*	'Understanding Artificial Intelligence Ethics and Safety', 10 June 2019, <https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety>, accessed 9 April 2020.	Set of recommendations for data projects including: building culture of responsible innovation; establishing actionable principles tailored to the design of AI systems.
National Cyber Security Centre *UK*	'Intelligent Security Tools: Assessing Intelligent Tools for Cyber Security', 2019, <https://www.ncsc.gov.uk/collection/intelligent-security-tools>, accessed 9 April 2020.	Aims to assist organisations in procuring or developing AI security tools. Includes a series of principles focused on: identifying tools which have functionality suited to the organisation's problem and way of working; using data correctly to ensure the AI tool can learn its task well; ensuring necessary skills, expertise and resources are available to support the project; assessing reliability, resilience and limitations of AI systems.
HM Treasury *UK*	*The Aqua Book: Guidance on Producing Quality Analysis for Government* (London: Stationery Office, 2015).	Not specific to AI. A good practice guide for those producing analysis for government. Includes principles of quality assurance for fit-for-purpose analysis: proportionality of response; assurance throughout development; verification and validation; analysis with RIGOUR.

Organisation	Publication	Content
US Office for Science and Technology Policy *US*	'Principles for the Stewardship of AI Applications', 2019, <https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf>, accessed 9 April 2020.	Ten draft principles based on: avoiding a precautionary approach or holding AI systems to impossibly high standards; public trust and participation in AI; consistent risk management/ assessment.
US Department of Defense *US*	'Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance our Security and Prosperity', 2018, <https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF>, accessed 9 April 2020.	Sets out an approach which: advocates experimentation and risk-taking in AI; uses the Joint Artificial Intelligence Center to accelerate delivery of AI-enabled capabilities and synchronise DoD AI activities.
European Commission *EU*	'Ethics Guidelines for Trustworthy AI', April 2019, <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419>, accessed 9 April 2020.	Outlines a focus on areas where legal frameworks intersect with AI; sets out policy options which support a regulatory and investment-oriented approach with twin objectives of promoting uptake of AI and of addressing risks; advocates improving legislative frameworks to ensure effective application of existing EU and national legislation.
UN *International*	'High-Level Panel on Digital Cooperation', 2018, <https://www.un.org/en/digital-cooperation-panel/>, accessed 9 April 2020. 'UNICRI Centre for Artificial Intelligence and Robotics', <http://www.unicri.it/topics/ai_robotics/>, accessed 9 April 2020.	Facilitating cooperation to leverage digital technology while mitigating unintended consequences; raising AI awareness, exchanging education and information, and synchronising stakeholder aims.

Organisation	Publication	Content
OECD *International*	'OECD Principles on AI', May 2019, <http://www.oecd.org/going-digital/ai/principles/>, accessed 9 April 2020.	Identifies five ways to ensure that AI maintains public trust: use AI to drive inclusive growth and sustainable development; include adequate safeguards enabling human intervention; allow for responsible disclosure to give people right of redress; continually revise potential risks of AI systems throughout life cycles; make developers and operators accountable for proper functioning.
Google *Private sector*	'Artificial Intelligence at Google: Our Principles', 7 June 2018, <https://ai.google/principles>, accessed 9 April 2020.	Seven principles to assess AI applications including: avoiding creating or reinforcing unfair bias; being accountable to people; upholding high standards of scientific excellence.
Microsoft *Private sector*	Saleema Amershi et al., 'Guidelines for Human–AI Interaction', in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper No. 3 (New York, NY: Association for Computing Machinery, 2019).	Eighteen human–AI interaction design guidelines including: helping users understand frequency of AI mistakes; ensure AI language and behaviours do not reinforce unfair biases; immediately update how user actions will impact future behaviours of an AI system.
IBM *Private sector*	Ryan Hagemann and Jean-Marc Leclerc, 'Precision Regulation for Artificial Intelligence', IBM Policy Lab, 21 January 2020, <https://www.ibm.com/blogs/policy/ai-precision-regulation/>, accessed 9 April 2020.	Outlines five policy imperatives for companies: designating a lead AI ethics official; setting different rules for different risks; avoid hiding an AI system; explain an AI system; test an AI system for bias.
Institute for Electrical and Electronics Engineers *Academia / NGO / civil society*	'Global Initiative on Ethics of Autonomous and Intelligent Systems', 2016, <https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf>, accessed 9 April 2020.	Prioritising ethical considerations in the design and development phases of new AI systems; ensuring the people behind these systems have the requisite education and training to do this effectively.

Organisation	Publication	Content
Partnership on AI *Academia / NGO / civil society*	'Human–AI Collaboration Framework & Case Studies', September 2019, <https://www.partnershiponai.org/wp-content/uploads/2019/09/CPAIS-Framework-and-Case-Studies-9-23.pdf>, accessed 9 April 2020. 'Explainable Machine Learning in Deployment', 2019, <https://arxiv.org/pdf/1909.06342.pdf>, accessed 9 April 2020.	Applies themes of 'nature of collaboration', 'nature of situation', 'AI system characteristics' and 'human characteristics' to seven different case studies; sets a framework for promoting transparency through thorough consideration of the target audiences of explainable AI.
World Economic Forum *Academia / NGO / civil society*	'Guidelines for AI Procurement', September 2019, <http://www3.weforum.org/docs/WEF_Guidelines_for_AI_Procurement.pdf>, accessed 9 April 2020.	Ten guidelines for AI procurement processes including: define public benefit of using AI while assessing risks; align procurement with governmental strategies; focus on mechanisms of algorithmic accountability and transparency norms; implement process for continued engagement of AI provider with acquiring party.
AI Now Institute *Academia / NGO / civil society*	'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability', April 2018, <https://ainowinstitute.org/aiareport2018.pdf>, accessed 9 April 2020.	Five steps for early phase engagement with AI systems: conduct self-assessment of automated decision systems and evaluate their potential impact; develop external researcher review processes to track impacts over time; provide notice to the public disclosing existing and proposed systems; solicit public comments to clarify outstanding concerns; provide avenues for people to challenge inadequate assessments or inappropriate system uses.