Centre for Emerging Technology and Security

RESEARCH REPORT

Human-Machine Teaming in Intelligence Analysis

Requirements for developing trust in machine learning systems

Anna Knack, Richard J. Carter and Alexander Babuta

December 2022

About CETaS

The Centre for Emerging Technology and Security (CETaS) is a policy research centre based at The Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to inform UK security policy through evidence-based, interdisciplinary research on emerging technology issues. Connect with CETaS at cetas.turing.ac.uk.

This research was supported by The Alan Turing Institute's Defence and Security Programme. All views expressed in this report are those of the authors, and do not necessarily represent the views of The Alan Turing Institute or any other organisation.

Acknowledgements

The authors would like to thank all those who contributed their time to participate in an interview or focus group for this project, without whom the research would not have been possible. The authors are also very grateful to Professor Tristram Riley-Smith, Emelia Probasco, Brian, James and Ellie for their valuable feedback on an earlier draft of this report.

Executive Summary

This report presents the findings of a CETaS research project examining the use of machine learning (ML) for intelligence analysis within the UK national security context. The findings are based on in-depth interviews and focus groups with national security practitioners, policymakers, academics and legal experts.

The aim of the research was to understand the technical and policy considerations arising from the use of ML within an intelligence analysis context. Specifically, the research explored how to calibrate the appropriate level of trust users should have in machine-generated insights, and best practice for integrating ML capabilities into the decision-making process of an analyst.

Intelligence analysts working in national security face a major challenge in coping with massive volumes of data that may yield crucial insights to current and future events. The ongoing global expansion of data presents both risks (that a crucial 'needle in the haystack' is missed), and also opportunities (more 'haystacks' to look for new 'needles' to gain deeper insights). The use of ML offers real potential to simultaneously reduce such risks and to pursue such opportunities.

There are important considerations to make when deploying ML to support a human decision-making process, including (i) the challenge of explaining and understanding why, and how, the model is functioning the way it does, and (ii) the risk of harm to society and citizens if ML capabilities are used inappropriately. It is recognised that clear guidance on the safe and effective use of ML is required prior to its widescale adoption in high-stakes contexts such as national security.

ML explainability is multifaceted and can refer either to technical properties of model performance, such as expected precision and recall rates at different thresholds (sometimes described as 'global explanations'); or to the specific factors the model took into account to arrive at a particular prediction (sometimes described as 'local explanations'). This study sought to examine intelligence analysts' requirements and priorities regarding both global and local model explanations.

Research conducted for this study involved examining the decision-making process and analytic workflow of intelligence analysts, to understand the technical, behavioural and policy considerations that must be taken into account when integrating ML capabilities into this process. The report's key findings and recommendations are as follows:

1. **ML is most valuable in characterising, discovering and triaging information from large volumes of disparate data.** This offers the best return on investment for ML in intelligence analysis in the short-term as it addresses some of the most pressing needs of the intelligence community. These applications also present a more manageable risk of using ML, as key decisions (such as those pertaining directly to individuals) are still taken by the analyst.

2. **How an analyst treats an output from a ML model is highly context-specific.** The meaning and confidence that an analyst assigns to machine-generated information is shaped by the current context (urgency of decision-making, the priority of the operation, and the perceived impact of subsequent decisions on resources and outcomes). Identifying and understanding the different contexts in which an analyst may use a ML model should therefore be central to the process of developing, testing and verifying a ML model.

3. **The lack of technical explainability of many ML systems is widely acknowledged.** 'Explainable AI' is a growing sub-discipline of data science research, and technical approaches are gaining traction to help explain the behaviour of sophisticated ML models. Whilst the use of such techniques is likely to remain of interest to data scientists and ML engineers, mathematical explainability methods may be of limited utility in improving the analyst's real-world understanding of the behaviour and performance of a model.

4. **Increasing the analyst's trust in ML capabilities involves both trust in the ML output and trust in the whole system, of which the ML is one part.** Analysts do not respond and commit effort to understand an output from a ML model alone, but also take into consideration other factors such as their experience of the model's prior performance, whether the model has been formally approved for operational use, and the nature of the task that the analyst is performing.

5. **ML should be designed from the outset to be integrated into the intelligence analyst's toolset and workflow.** The most effective application of ML should come from understanding the current work environment of the analyst. This requires a deeper understanding of human factors, usability requirements, and the psychology of decision-making to be integrated into the processes of developing the model (data science) and the tools for interacting with the model (software engineering).

6. **The type and amount of technical information about a ML model that is provided to analysts should be context-specific, user-specific and interactive.** The level of information provided to an analyst should improve the transparency and

interpretability of the model and should consist of two parts: mandatory information (context-specific) and custom information which is selected by the analyst (user-specific). The presentation of both types of information should be unambiguous and make it easy for the analyst to traverse different *layers* of explanation (for instance through the use of click-through interfaces).

7. **The complexity of explanation should be determined by the complexity of the problem**. If a problem is cognitively straightforward for a human an explanation of an ML solution to that problem is unnecessary and does not help decision-making. However, for more complex tasks which cannot be easily completed by a human, it is more important for the model to provide some reasoning as to how it arrived at a certain output. There will also be circumstances where a local explanation is neither helpful nor appropriate.

8. **Analysts should be included in the prototyping and testing of a ML model and associated graphical user interface (GUI).** This should elicit the appropriate level of explanation required to support the decisions of the analyst. Tuning the performance of a model (for instance, setting acceptable limits on false positive / false negative thresholds) should be done with diverse representation from the analyst community. Mandatory involvement of analysts in testing of ML models should increase their overall confidence and adoption. The results of these tests should be routinely shared with partner organisations deploying similar ML-enabled systems.

9. **Different thresholds may be required for different uses of the same model and are key to analyst confidence and need to be continuously reviewed.** In some circumstances analysts can tolerate a higher rate of false positives (e.g. in high-priority operations where the risks of missing something important may be catastrophic) that, in other circumstances, would not be acceptable. False negatives are generally more problematic within intelligence analysis, due to the risk of potentially important information 'slipping through the net'.

10. **Language for discussing and explaining ML models should be standardised across the national security community.** Values such as the confidence level of a classifier should be presented to a recognised standard such as the PHIA (Professional Head of Intelligence Assessment) Probability Yardstick. This information should be presented both linguistically and numerically where possible.

11. **Data science should be offered as a support service to analysts.** For example, a small team of data scientists who are dedicated to helping analysts who are using ML models interpret results and investigate concerns. Close support to analysts

should increase their level of understanding of the behaviour and performance of ML models. This should mitigate the risk of inappropriate use of a ML model whilst simultaneously improving the overall aptitude and awareness of the analyst community in the use of ML models.

12. **Effective adoption of ML requires a system-level approach.** The design of a ML model should consider its effect on existing policies and practices including any necessary legal authorisations, the criticality of feedback from analysts on model performance, and consideration of the whole life costs of deploying and maintaining the ML model. Organisational policies and processes may need to be updated to account for these additional requirements.

13. **Additional training and learning materials should be made available to enable those using or overseeing the use of ML systems to acquire a minimum level of data science and ML literacy.** The ability to understand technical properties such as precision, recall and accuracy was cited as the minimum level of literacy that an analyst should have to ensure they have a sufficient understanding of the performance, and therefore the utility, of a ML model.

Further research should aim to:

i. Identify technical and policy considerations for more advanced use of ML in human-machine teaming (such as non-classification use cases).

ii. Understand the explainability requirements for ML in fully automated decision-making applications.

iii. Develop methodologies for understanding the analyst workflow to guide ML application development, and embed behavioural and decision science into software engineering practices.

iv. Systematically assess how the explainability requirements of different users varies across background, their work context and demographic.

v. Develop a standardised lexicon of terminology for communicating the confidence associated with ML-supported analysis based on the PHIA probability yardstick.

1. Introduction

Intelligence analysis in the national security context seeks to gather information and answer questions about a target or group in support of current operations and future intent.[1] Key to an intelligence analyst's ability to do this effectively is acquiring and examining relevant data to identify threats and opportunities more quickly, making the best use of all available data sources.

One of the most pressing challenges for analysts is being able to keep up with the exponentially growing volume and variety of data that is now available. New software techniques such as those incorporating machine learning (ML) have introduced powerful analytical capabilities – all of which hold significant potential to help analysts process more data and information at a speed that is relevant to support decision-making.[2]

Decisions made in national security can involve serious consequences for society and the individual, particularly as intelligence and law enforcement agencies are granted exceptional legal powers to undertake activity that may interfere with individuals' human rights. For this reason, enhanced policy and guidance is required regarding the implementation of ML within a national security decision-making process. Specifically, there is a need to address how decision-making is affected by increased automation of high-stakes information processing tasks, and how ML capabilities should be developed to ensure that analysts can have confidence and skill in using them appropriately.

1.1 The 'black box' problem

ML performance is significantly improving in speed and accuracy,[3] but is also becoming more complex and challenging to understand. The increasing use of advanced ML techniques such as deep neural networks has made it more difficult for human operators to

[1] The UK Government describes intelligence analysis as 'adding value through the process of taking known information about situations and entities of strategic, operational, or tactical importance and characterising the known and the future actions in those situations.' https://www.gov.uk/government/organisations/civil-service-intelligence-analysis-profession/about

[2] Throughout this report, ML and artificial intelligence (AI) are used synonymously, although we recognise that ML is a specific sub-discipline of the wider field of AI.

[3] Lydia P. Gleaves, Reva Schwartz, and David A. Broniatowski, "The role of individual user differences in interpretable and explainable machine learning systems," *ArXiv* (2020).; Jianlong Zhou, Fang Chen and Andreas Holzinger, "Towards explainability for AI fairness," *International workshop on extending explainable AI beyond deep models and classifiers* (2022).; Madalina Busuioc, "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review*, no. 81 (2020): 825-836.

understand how machine learning models arrive at their outputs.[4] This uncertainty inhibits non-technical audiences from understanding the reasoning underpinning ML recommendations in decision-support functions, with outputs seemingly resulting from a 'black box' and earning this phenomenon the designation the 'black box' problem.[5] It will often be possible for the model output to provide a classification (e.g. 'high risk' or 'low risk'; 'civilian vehicle' or 'military vehicle'), but operators may have insufficient information to understand the logic the model followed and which components of the input data the model assessed to be important to predict an outcome.[6]

The 'black box' problem refers to ML capabilities that cannot be directly interpreted nor understood through examination. For example, Google's BERT natural language processing model consists of 110 million parameters.[7] Examining each of those parameters is unfeasible and would not yield an understanding of the logic of the model. As such, large ML models are currently treated as non-interpretable, and eliciting an understanding about the behaviour of such models is often done by modifying the inputs to the model and monitoring the subsequent outputs that the model produces (counterfactual explanations).[8] Hence, such ML models are described in the academic literature as a 'black box' where only the inputs and outputs can be directly interpreted and understood, not the inner workings of the system.

Moreover, important factors such as the training data used to build the model, the confidence level associated with any given prediction, or the variation in accuracy by context are often not apparent to the operator. Accuracy rates can often be misleading, as they may vary considerably when a model is deployed on a new dataset whose distribution varies significantly from the test data. Without the ability to comprehend the reasoning underpinning AI systems (and the limitations and uncertainties inherent in the model), there

[4] Lydia P. Gleaves, Reva Schwartz, and David A. Broniatowski, "The role of individual user differences in interpretable and explainable machine learning systems," *ArXiv* (2020).

[5] Alexander Babuta and Marion Oswald, "Data analytics and algorithms in policing," *RUSI Occasional Papers* (2019).; Andreas Holzinger et al., "Explainable AI methods – a brief overview", *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, (2022).; Michael Veale, Max Van Kleek and Reuben Binns, "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making," *CHI*. (2018).; Jianlong Zhou, Fang Chen and Andreas Holzinger, "Towards explainability for AI fairness," in *xxAI - Beyond Explainable AI,* no. 13200 (2022): 375-386.; Cynthia Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence 1* (2019): 206-215.; Shraddha Mane and Dattaraj Rao, "Explaining network intrusion detection system using explainable AI framework," *ArXiv* (2021).

[6] Madalina Busuioc, "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review*, no. 81 (2020): 825-836. (2020).; Sunil Aryal, "Levels of explainable artificial intelligence for human-aligned conversational explanations," *Artificial intelligence 299* (2021).

[7] Jacob Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," *ArXiv* (2019).

[8] Madalina Busuioc, "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review*, no. 81 (2020): 825-836.

is a risk that human operators may over-trust or under-trust AI systems, particularly when used in high-stakes decision-making contexts such as defence and security.

This potential lack of model explainability is further exacerbated by advances in ML, which are increasing the ability of systems to learn, decide and act autonomously.[9] This challenge is compounded when developing a complex ML system, i.e., multiple models operating in sequence as part of an automated analytic pipeline. Model performance and limitations are not static, but will change when the model is retrained and revalidated with new data.

In general, an inverse correlation between the performance and the interpretability of a model may exist. The higher the performance at a task (e.g., ability to generalise across a wide range of data) the larger and more complex the model is likely to be. Hence, there is a tendency for more powerful ML models – such as deep learning neural networks – to be less interpretable. This presents a conundrum: how to benefit from the predictive power of non-interpretable models, whilst managing the risk of the lack of explanation or predictability of those models and, ultimately, the decision-making accountability if things go wrong.

1.2 Implications of the black box problem for analytic tradecraft

Recent research has discussed how ML systems are increasingly deployed as part of a human-machine team (HMT).[10] Given the challenges outlined above, national security organisations require a number of assurances before deploying a HMT capability to support intelligence analysis and subsequent decision-making.

The challenges of applying ML-supported decision-making are amplified in the national security context given the degree of discretion and professional judgment that an analyst uses to make a high-stakes decision with uncertain and incomplete information, and this may be difficult to quantify or encode accurately in a ML system. Human operators may not be able to fully explain every factor they have taken into account when arriving at a certain decision, but they must be able to provide sufficient explanation to demonstrate that the action taken was *necessary* and *proportionate*.[11] This requires an explanation of their reasoning, hypotheses and conclusions. What constitutes a 'sufficient explanation' in this

[9] David Gunning and David W. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Magazine*, no.40 (2019): 44-58.
[10] Mariarosaria Taddeo et al., "Artificial Intelligence for National Security: The Predictability Problem", *CETaS Research Reports* (September 2022).
[11] 'Investigatory Powers Act 2016' (UK); 'Human Rights Act 1998' (UK).

regard is highly context-specific, and automated systems are not capable of providing the same type of subjective reasoning and rationale, drawing on their own professional judgement in the way that human operators are expected to do. Furthermore, human users' ability to fully explain their reasoning may be harder where an algorithmically-derived insight or prediction has shaped the decision that they are subsequently required to justify or defend.

The lack of interpretability of ML could lead to challenges across the intelligence analysis process, such as:

- **Creating missed opportunities** to transform data to actionable intelligence, and casting doubt on whether the classifications and predictions made are sufficiently accurate.[12]

- **Hindering algorithmic assessments from being challenged** as to whether the decisions and the processes underpinning their assessments are relevant, fair, proportionate and not based on discriminatory inputs.[13]

- **Obfuscating whether a particular type of model serves the aims of a particular context.**[14] ML systems are developed for particular use cases that may not be appropriate for other applications and sometimes these systems are procured off-the-shelf, so their limitations may not be known when they are first deployed.

- **Casting doubt on who should take accountability for decisions.** Opaque ML models may reinforce ambiguity surrounding whether accountability should be assigned to the model itself, its developer, a senior representative within the organisation, or the human-in-the-loop, in circumstances where the human operator is unable to comprehend the overall decision-making process.

---

[12] Ankit Tweari, "Decoding the Black Box: Interpretable methods for post-incident counter-terrorism investigations," *WebSci '20 Companion* (2020).

[13] Jianlong Zhou, Fang Chen and Andreas Holzinger, "Towards explainability for AI fairness," in *xxAI - Beyond Explainable AI,* no. 13200 (2022): 375-386.; Zeynep Akata et al., "Hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible and explainable artificial intelligence," *Computer* (2020).; Alexander Babuta and Marion Oswald, "Data analytics and algorithms in policing," *RUSI Occasional Papers* (2019).; Alexander Babuta and Marion Oswald, "Machine learning predictive algorithms and the policing of future crimes: governance and oversight," in *Policing and Artificial Intelligence* (Oxford: Routledge, 2019).

[14] Alexander Babuta, Marion Oswald and Christine Rink, "Machine learning algorithms and police decision-making: legal, ethical and regulatory challenges," *RUSI Whitehall Report* (2018).; Marion Oswald and Alexander Babuta, "Machine learning predictive algorithms and the policing of future crimes: governance and oversight," in *Policing and Artificial Intelligence* (Oxford: Routledge, 2019).

The research found that it is not common practice to design ML-enabled information systems with explainability and interpretability as a formal requirement. However, such systems are increasingly being deployed across the UK public sector, including in criminal justice, local government and healthcare.[15] Similarly, various policy documents in recent years have suggested that ML will become an increasingly integral component of national security analysis and decision-making processes.[16]

Decisions in the national security context are governed by a well-established system of authorisation, audit, oversight and external scrutiny, which has been refined and developed over many decades. The granularity of information required to explain analysts' decisions varies considerably across use cases, and some higher-stakes contexts demand a much higher degree of explanation than others.[17] It would therefore be unreasonable to strive for 'full transparency' of decisions across all contexts, whether ML-supported or otherwise. It remains to be established whether a higher standard of interpretability is required in the outputs and processes governing ML-supported decisions. Moreover, even if a higher standard of interpretability is *not* required, there are unresolved considerations surrounding the types and ranges of error or risk that are more 'acceptable' than others in different security contexts.[18]

To further complicate matters, recent strategic policy documents have emphasised a push towards more synergy across services and allied nations (e.g. through multi-domain integration).[19] However, as ML-enabled information systems become more complex and interconnected, the interpretability challenges mentioned above could become compounded, meaning practitioners may lack the full picture of potential risks and vulnerabilities across a complex integrated system.

---

[15] Lina Dencik et al., "Data scores as governance: investigating uses of citizen scoring in public services," (2018).; Jianlong Zhou, Fang Chen and Andreas Holzinger, "Towards explainability for AI fairness," in *xxAI - Beyond Explainable AI,* no. 13200 (2022): 375-386.; Marion Oswald and Alexander Babuta, "Machine learning predictive algorithms and the policing of future crimes: governance and oversight," in *Policing and Artificial Intelligence* (Oxford: Routledge, 2019); Alexander Babuta and Marion Oswald, "Machine learning algorithms and police decision-making: legal, ethical and regulatory challenges," *RUSI Whitehall Report* (2018).

[16] Office for Artificial Intelligence, Department for Digital, Culture, Media and Sport and Department for Business, Energy and Industrial Strategy, *National AI Strategy* (2021).; Ministry of Defence, *Defence Artificial Intelligence Strategy* (2022).; GCHQ, *Pioneering a New National Security: The Ethics of Artificial Intelligence* (2021).; Alexander Babuta, Marion Oswald and Ardi Janjeva, "Artificial intelligence and UK national security: Policy considerations," *RUSI Occasional Paper* (2020).

[17] Marion Oswald, "Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power," *Philosophical Transactions A* (2018).

[18] Alexander Babuta, Marion Oswald and Christine Rinik, "Machine learning algorithms and police decision-making," *RUSI Whitehall Report* (2018).

[19] Alun Preece et al., "Explainable AI for intelligence augmentation in multi-domain operations," *ArXiv* (2019).

The field of *explainable AI* is an active, global research effort to tackle the problem of non-interpretable ML techniques such as large neural networks, and recent research shows promise in addressing the 'black box' problem.[20] Such research is encouraging but nascent, and will take time to diffuse into standard data science and machine learning practices. As such, the 'black box' problem remains an inherent characteristic of most deep neural networks and large models. With research efforts focussing on developing ever-increasingly complex models trained on larger, more diverse datasets, this challenge is only likely to grow. This must be taken into account when developing future policy and guidance for the use of machine learning systems in higher-stakes contexts, particularly in terms of their potential impact on human decision-making processes.

1.3 Research aims and objectives

This study sought to understand: where ML is viewed as bringing utility in the intelligence analysis workflow, or where it could in future.; the technical and human considerations associated with embedding an ML system within a complex decision-making process; and how the technical behaviour of the model needs to be explained to different users to provide sufficient confidence in the system.

Specifically, the research aimed to address the following three research questions:

> **RQ1:** How does the availability, type and format of an AI explanation influence the level of decision-making confidence of the user?
>
> **RQ2:** What information do intelligence analysts require regarding the performance, robustness and predictability of an AI system, in order to maintain the appropriate degree of decision-making confidence and accountability for the task at hand? How do we avoid automation bias or over-trust in AI systems?
>
> **RQ3:** How does the decision-making confidence and risk appetite of different AI users vary across context and background?

The research focused on two aspects of AI explainability, which can loosely be described as 'local explainability' and 'global explainability'. Local explainability refers to output-level information regarding *why* the model has generated a particular output (for instance, the features taken into account to calculate a particular probability score). Global explainability

[20] Pantelis Linardatos, Vasilis Papastefanapoulos and Sotiris Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy* 23 no. 1 (2020): 18.

refers to model-level technical details related to key values such as precision, recall and classification thresholds.[21]

It soon became apparent in the research that the types of explanations analysts require are not the typical explanations described in the academic literature on explainable AI. Instead, this study takes a broader view of the technical requirements across the full spectrum of humans in-the-loop that may be required to interface with an ML system.

Moreover, the research found that to develop sufficient confidence in the ML system, technical explanations alone are not enough and need to be supported by policy requirements that create a supportive ecosystem for responsible use of ML in intelligence analysis. Explainability is only one method to counteract the black box problem and complexities associated with ML modelling. The requirement for technical explainability needs to be examined alongside the necessary adaptations of organisational structures, policies and training programmes for the use of ML for intelligence analysis. All these issues and more are explored in detail in the following sections.

1.4 Methodology

Research for this study was conducted over a three-month period from July – September 2022. The analysis contained in this report is drawn from a literature review on explainable AI, and interviews with 18 respondents including operational intelligence analysts, legal experts, behavioural scientists, human factors engineers, and defence and security research organisations.

Research participants were identified using a purposive, non-probabilistic sampling strategy. A focus was on identifying individuals with direct experience of using, developing, overseeing or researching ML and related analysis tools within a national security context. A semi-structured interview guide was developed to ensure a broadly consistent line of questioning across interviews, while allowing flexibility to pursue other lines of enquiry identified in the course of discussions. Interviews were conducted on an anonymous, non-attributable basis.

[21] Precision and recall are key metrics used to describe machine learning performance. 'Precision' describes what proportion of positive identifications were correct, and 'recall' describes what proportion of actual positives was identified correctly. The higher the precision, the lower the false positive rate; and the higher the recall, the lower the false negative rate. The 'classification threshold' is the point at which the model classifies a data item into a target category (e.g. a classification threshold set at 0.99 would only create an alert if the model identified a 99% or higher probability that the target variable belongs to a certain class).

Interview data was analysed following a general inductive approach, whereby the focus is on extracting meaning from data and categorising data into relevant themes and sub-categories. The sections of this report broadly correspond to the core themes identified through this analysis process. Throughout this report, an anonymised coding system is used to refer to interview data. The following prefixes are used to indicate the category of research participant to which interview data refers:

G = Government respondent

L = Legal expert

INT = International expert

This research was inevitably limited in scope. The focus of this study is on ML to support analysts' sensemaking of data as part of a decision-support function. The use of ML for *automated decision-making in* intelligence analysis was not examined in detail, because stakeholders stated clearly that decision-support tools have more near-term potential for implementation. Moreover, this report focuses primarily on the use of machine learning classification models ('classifiers') within an intelligence analysis context. Other forms of artificial intelligence – such as the use of ML for data enhancement or data generation – are not the direct focus of this study. Finally, due to the sampling method used, it was not possible to comprehensively evaluate the specific differences in decision-making processes across intelligence analysts working in different fields, organisations and mission areas (see Section 5 for corresponding proposals for further research).

The remainder of this report is structured as follows. Section 2 summarises cross-cutting findings emerging from the research regarding the importance of decision-making context and the stage in the analytic workflow where ML is deployed. Section 3 discusses specific technical requirements identified in the research as important to consider when developing ML systems for use in intelligence analysis. Section 4 explores the organisational policy considerations associated with deploying ML systems into the intelligence workflow. Finally, Section 5 concludes by highlighting priority areas for further research.

2. Cross-cutting Findings

The research first investigated the sequence of steps required to consider how ML might be integrated into an intelligence analysis workflow. The literature review suggested that, in the first instance, it would be important to engage with users to understand where in the analytic workflow ML would bring the most utility and could help address challenges users encounter, as well as how best to manage the new risks that ML could pose. This investigation highlighted that user acceptance of ML in intelligence analysis was highest in the information filtering, categorisation and triage stages of analytic tradecraft (rather than deriving insights at the level of individual data points).

This section explores the context of an intelligence analyst's workflow, and the points in this process where ML systems could provide the most value.

2.1 The importance of context

Engagement with potential users of ML-enabled tools highlighted that it would be important not just to explain the model, but to design the overall experience of using the model from the analyst or decision-maker's perspective, and incorporate all the assurances required to have confidence in the ML output. With this in mind, the *context* of the decision process the ML is used to inform was repeatedly mentioned as the single most important consideration in the development of a ML system for use in the intelligence context.

Interviewees agreed unanimously that ML explanations need to be designed with a specific context in mind. The same ML capability could be deployed for numerous purposes, and a system that has relatively low-risk consequences in a commercial context may be associated with a much higher level of risk when deployed in an intelligence context. For example, recommender algorithms prioritise films or songs based on consumers' past choices and are relatively innocuous with low-risk consequences, so do not require significant explanations. This is in stark contrast to a recommender system that may be used to inform intelligence analysts' conclusions, which may subsequently result in direct action being taken towards an individual. Furthermore, the same model could be used to support longer-term and lower-priority analysis tasks, or very high-priority and time-sensitive operations. For this reason, it is neither feasible nor desirable to prescribe 'model-level explainability requirements' for a single model that may be used in multiple settings.

Intelligence analysts are typically seeking, extracting and identifying 'needles from haystacks,' or useful details and patterns from large amounts of data in multiple stages,

involving highly manual data analysis that entails collecting and tagging enormous volumes of text, visual and audio information from various sources alongside structured data.[22] There are a limited number of individuals who can review and translate material, creating an enormous burden of effort that could lead to cognitive overload for analysts.[23]

These combined activities help the analyst to piece together an understanding of the 'big picture' which can support decision-making. Various consumers and decision-makers may exploit the intelligence products that analysts produce, which could include senior decision-makers outside the analyst's organisation such as other government departments or other nations' intelligence agencies.[24]

Operational tempo sometimes requires near real-time inputs from analysts potentially within minutes,[25] but can also involve retrospective analysis that allows analysts to take time to develop an understanding of a threat or risk. Analysts' intelligence outputs can feed into relatively low-stakes tasks, but can also feed into high-stakes decisions such as military action or arrests, so there is very little tolerance for error.

Intelligence analysts are trained to make context-specific judgements in conditions of uncertainty when presented with incomplete, imperfect and fragmented information.[26] For this reason, increasing the analyst's trust in ML capabilities requires a trust model that involves both trust in the *output* and trust in the *system.* To illustrate with a human analogy, a team leader must have trust to delegate tasks to team members, building a mental model of who in the team can be trusted with specific types of work. Without buy-in from both decision-makers and analysts, ML-enabled tools that may have utility in intelligence analysis could fail to achieve their potential, if users do not have the confidence to deploy or use them operationally. Moreover, adoption of these tools in analysts' day-to-day work is only likely to yield benefits if ML is embedded in the right stages of intelligence analysis, where users could conceivably imagine developing sufficient confidence in the ML tool. There is a clear risk to deploying an ML capability without a clear understanding of the overall decision-making process which it is used to support, as this could damage analysts' trust and confidence in ML more generally, creating challenges for wider adoption of ML for national security.

[22] G1 & G2; L2; L3; INT5
[23] G7; G8 & G9; INT2
[24] L1
[25] G1 & G2; INT1
[26] G7

2.2 Embedding ML into the intelligence analysis pipeline

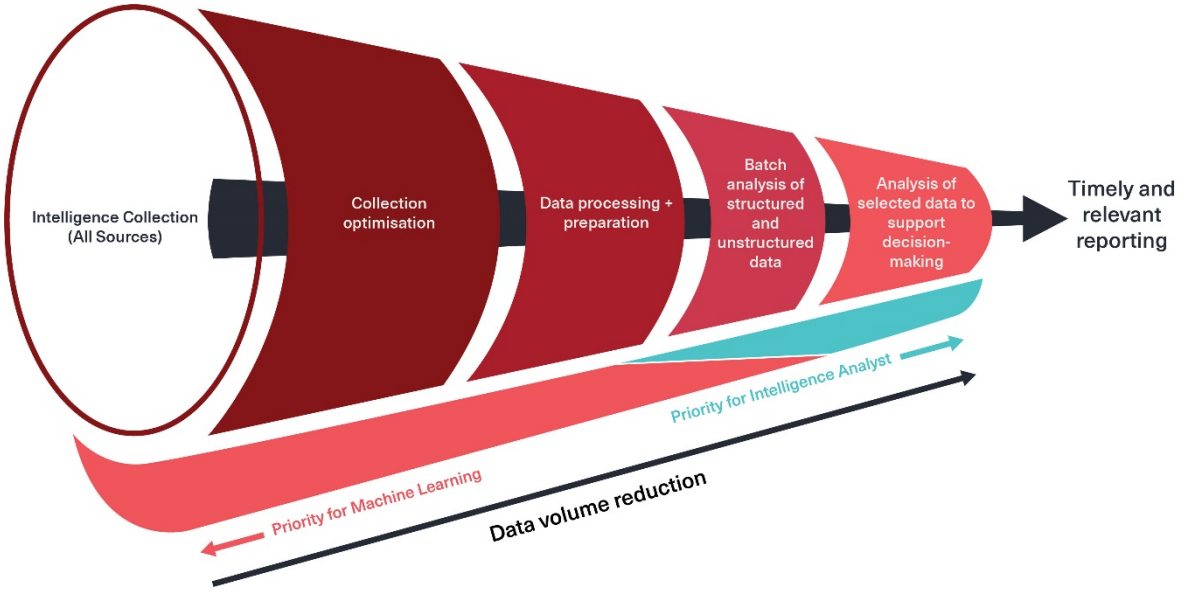
In the near-term, the most frequently cited stage of intelligence analysis where ML was perceived to hold potential is in information filtering and prioritisation, to make discovery more efficient or, as one interviewee described it, 'to reduce the signal-to-noise ratio'.[27] As summarised by one interviewee:

> 'I wouldn't rule out [ML-enabled intelligence analysis tools] as part of the chain for anything, but it's more what are the follow on steps, such as kinetic action or advising something is in a particular location or police arrests and knocking doors down. It would not be appropriate for [an ML tool] to spit out taking those kinds of direct action. You'd need parallel evidence. I would be happy to use and trial ML output as part of process like in filtering and collection decision, which is so far away from an actual outcome.'[28]

Figure 1 presents an overview of the typical intelligence analysis pipeline and corresponding system functions, and illustrates where interviewees believed ML would provide the most value. In general, the 'collect' and 'process' stages were identified as being the stages in the pipeline where ML would provide the most utility for analysts. The perceived value added by ML decreases as we move down the pipeline towards the decision-making and reporting that ultimately results from the analysis process.

[27] G1 & G2; INT1; INT5; G3; G7; L2
[28] G7

*Figure 1. Overview of intelligence analysis pipeline and priority areas for incorporating machine learning**



*This image is illustrative of where in the process stakeholders believed that ML or human involvement is most useful. The study team understands these stages of analytic tradecraft may differ in different contexts.

Returning to the 'needle in the haystack' metaphor, interviewees suggested that ML is useful in helping to produce 'piles of hay,' so analysts need not review 'the entire field', but analysts do not wish for the ML to decide what the 'needles' are.[29] Put differently, one interviewee suggested that 'complex ML-enabled capabilities are probably too far off, so the emphasis is on procuring faster horses rather than cars.'[30] In this sense, 'faster horses' are systems that get through more data and automatically identify information of interest to the analyst, for instance keyword searching. Ideally, models would generate outputs that contain tranches of relevant data with some peripheral information or 'hay' to give the analyst some confidence that nothing is missed.[31] The analyst remains responsible for any decision-making that then results on the basis of analysing that data.

Moreover, interviewees suggested that it is important to distinguish between an ML-enabled *tool*, and ML-enabled systems that adopt a 'teammate' role. For example, if the system is limited to triage or filtering of bulk data, the system is merely a tool performing a narrow task.[32] An ML-enabled teammate would do more than this and be involved in:

> '...joint problem solving by helping reinstate memory and helping the analyst recall strands of "the big picture" after the weekend. A teammate may also support the analyst

[29] G1& G2; INT3; INT 5; L1; L2; INT2
[30] G1 & G2
[31] L2; INT3; INT4
[32] INT4; G8 & G9

> by taking notes, evolving beside them and "backing them up" when analysts are fatigued or helping them generate audit trails of data underpinning their analyses.'[33]

Such a system could also learn and adapt to how much information the user wants and when the user wants to receive alerts, or have an understanding of the legality of situations and rules of engagement.[34] One behavioural scientist cited previous research that discovered that intelligence analysts demonstrated optimal performance at their peak cognitive load before a given saturation point,[35] so a machine 'teammate' might tailor itself accordingly to analysts' peak cognitive load.[36] Another human factors engineer suggested that future systems might be able to listen to conversations between teammates and linkages made by the analyst to point out links that the human analysts may be blind to.[37] The study team found no evidence of systems of this kind currently in use or planned in the near future, but if existing ML *tools* are accepted and trusted now, then user acceptance of more complex systems and 'teammates' will be more likely in the future.

Finally, it is also important to note that the binary distinction often made between 'human-machine teaming' and 'automated decision-making' may be an over-simplification of how ML decision-support tools are deployed in practice. In the future, three different types of interaction between human analysts and ML are conceivable and warrant further consideration, including an ML model that:

- Triages a result that the user can validate manually;
- Produces a result, where the user themselves cannot validate the result manually, but an expert could;
- Produces a result which cannot be validated by human review.

Future efforts to develop ML-enabled information systems for national security should start with a clear understanding of which of these human-machine interaction models best characterises the intended use of the system, as this will have direct implications for the design decisions made at the development stage of the system.

[33] INT4
[34] INT4
[35] G8 & G9
[36] Richard J. Carter, *Cognitive advantage: How artificial intelligence is changing the rules for winning in business and government* (London: Mayhill Publishing, 2021).
[37] INT2

3. Technical Considerations

This section explores the technical considerations identified in the research as important to consider in the development, design and deployment of an ML system for intelligence analysis.

As mentioned previously, the research initially focused on the technical concept of 'AI explainability', meaning tools and techniques that can be used to describe how a model arrived at a particular output or prediction. However, the types of explanation that interviewees focused on were much broader than the strict technical definition of ML explainability, and participants were more concerned with ways that ML outputs and performance could be communicated to a range of stakeholders throughout the lifecycle and application of a ML capability. This is the focus of the following section.

3.1 Determining thresholds: false positives and false negatives

For ML classification systems, it is necessary for the developer to determine a classification threshold above which a data item is categorised into a certain class. For example, a classification threshold set at 0.99 would only create an alert if the model identified a 99% or higher probability that the target variable belongs to a certain class. The same model could instead be set to a 0.95 classification threshold, meaning a 95% or higher probability would trigger an alert. A lower classification threshold inevitably entails a higher number of false positives, while a higher classification threshold increases the risk of false negatives. This is a crucial factor to consider in the development of any ML model, and attention must be paid to the relative 'costliness' of different types of error (i.e., whether it would be a worse outcome to create more false positives for a human to review, or to risk more false negatives meaning potentially important data items are missed).

The research found that false negatives are generally considered to be the costliest type of error in an intelligence context. Across all decision-making settings (whether ML-assisted or otherwise) analysts' risk appetite for *any* false negatives is very low. Furthermore, interviews revealed that the risk tolerance for false positives is likely to increase in high-stress, time-constrained or high-stakes decision-making contexts. This is because the consequences of *not* taking action in an urgent or high-stakes situation could be significantly worse than the consequences of incorrectly taking action on the basis of a false positive.

For example, a ML model that is predicting an event or connection of relevance to a high priority investigation is likely to be acted upon immediately to avoid any potential risk of harm, whereas for lower priority investigations (where there is no immediate risk of harm) an analyst is more likely to need convincing of the correctness of the model. As such, in particularly high-priority and time-sensitive circumstances, an analyst may not require a detailed explanation of why the model produced the prediction and nor would they have the time to review such information in an urgent operational situation. However, interviewees raised the cautionary note that analysts may lose confidence in a system the more false positives they encounter,[38] so an analyst's tolerance for false positives is a finite resource and context-specific.

These observations suggest that a careful balance must be struck to minimise false negatives to the lowest possible threshold, while not creating so many false positives that the analyst becomes frustrated with the system. Determining this threshold is again likely to be highly context-specific, emphasising the crucial importance of user testing of ML classifier models with a group of target users prior to their operational deployment, to set an acceptable upper limit on false positives and false negatives for any given user group.

3.2 Different explanations for different audiences

ML systems used for national security may need to meet a higher level of explainability than those used in less sensitive contexts such as the commercial sector. To interpret and explain the output from a ML model to the requisite standard for it to be used confidently for intelligence analysis, it is insufficient to treat explainability requirements as isolated to a single stakeholder (e.g. the analyst). Information on the performance and behaviour of a ML capability should be provided for multiple stakeholders, at multiple levels of granularity, and throughout the lifecycle of the ML capability.

Our research identified the types of information required to provide sufficient explanations to satisfy the needs of three different categories of stakeholders involved in the use of ML to support intelligence analysis. These are each considered in turn.

3.2.1. Explanations for the senior responsible owner

Interviewees emphasised key differences between the explanations required by data scientists or those developing policy or approving the deployment of a ML system (the

[38] G8 & G9; INT4; Patricia McDermott et al., *Human-Machine Teaming Systems Engineering Guide*, (2018).

senior responsible owner), and the explanations required by analyst end-users.[39] While analysts also require key information regarding system performance and known limitations, this information needs to be presented in a different way for the end user than for those developing or approving the system.

Crucial to any trust model is an understanding of the system's limitations and the calibration points where performance may differ from the user's expectation. Some stakeholders discussed the utility of a 'model card,'[40] a reference point that explicitly sets out ML system limitations and the intended use case for the system. Examples of contextual information that could be contained within a model card include:

- **Highlighting the data used and assumptions made in the evaluation procedure** documented, such as the geographic and time limitations of the data.[41]
- **Providing quantitative information on the base rate,** or prior probability, of an object belonging to a target category.[42]
- **Insights on the historical precision and recall** performance of the model at different classification thresholds.
- **Clearly stating the contexts in which the model can be expected to work less well**.[43] This could include prompts on whether a system is better at translating a written letter in contrast to a technical report.

For those authorising or governing the development and deployment of ML, the most important information to convey relates to the performance of the system and any known limitations. Those governing the deployment of ML systems may not require output-level explanations ('local explainability'), but rather model-level technical details related to key metrics such as precision, recall and classification thresholds (discussed further in subsequent sections). Such technical information should be readily accessible in relevant supporting documentation, for instance via the organisation's intranet. Training

[39] L1; L3
[40] G8 and G9; INT5
[41] Margaret Mitchell et al., "Model cards for model reporting," *ArXiv* (2019).; David Lonsdale and Maria dos Santos Lonsdale, "Handling and communicating intelligence information: a conceptual, historical and information design analysis," *Intelligence and National Security*, no. 34 (2019): 703-726.; G3; G4, G5 & G6; L2
[42] Rodgers, R. Scott, "Improving analysis: Dealing with information processing errors," *Air Force Research Laboratory Human Effectiveness Directorate Warfighter Interface Division* (2006).
[43] G3

requirements for users are also a crucial consideration and are discussed further in Section 4.5.

As well as key technical metrics such as precision and recall, it is also important to account for the risk of data drift – when the distribution of the target data on which the model is deployed differs significantly from the distribution of the training or test data.[44] This can lead to significant reductions in performance when a model is deployed on new, unfamiliar data, meaning the model may need to be retired or retrained.[45]

In addition, those authorising or governing the development and deployment of ML should have access to information justifying the necessity and proportionality of access to the data used to train the model, and any access to data the model may require on an ongoing basis.[46] This is important to maintain senior accountability throughout the full development lifecycle.

3.2.2. Explanations for analysts

Interviewees highlighted that it is more important for analysts to trust the organisational governance of ML systems, rather than be provided with detailed information on their technical operation. If internal authorities have tested, accredited and 'signed off' a ML capability for mission use, presenting an explanation of the operation of the model is less important to the analyst user. System-level features, intended use cases and limitations still need to be explained to the analyst for them to maintain trust in the model,[47] but not to the extent that the level of information provided creates an unnecessary cognitive burden.

Furthermore, stakeholders highlighted that complete trust in the output from a ML model should not be the aim of the system since intelligence analysts should never wholly trust an output without verifying it against other information.[48] Instead, the user interface should indicate the overall certainty in the output (in accessible language), so analysts can calibrate their confidence accordingly.[49]

When considering the type and format of explanation provided to analysts, two distinct requirements were identified in the research. First, the need for system-level technical properties of the model (such as precision, recall and thresholds) to be available when

[44] G3
[45] INT2
[46] L1
[47] G8 & G9; INT2; L3; INT4
[48] G8 & G9
[49] INT3

needed, for instance via Wiki pages on the organisational intranet. And second, the need for an interactive and layered user interface, incorporating click-through functions for analysts to query individual model outputs and be presented with more granular information about how certain predictions have been made. The requirement for such a local explanation will vary according to context and application. One interviewee described a potential layered interface as containing 'more general detail required for everyone, then if you want more info, you find more, and the people who need to know the *n*th degree are right down the pointy end. Most people don't need to know that much detail but need the ability to go down to different stages of detail.'[50]

Crucially, there may be variation in analysts' level of understanding of the technical aspects of the system, so explanations of system parameters must be translated into plain English to be accessible to all users.[51] As summarised by one interviewee:

> 'Analysts are not going to be data engineers or data scientists, so they may not need to understand the underpinnings of the model, but knowing the range of what they are designed to output. The analyst needs to know what to focus their mind on… "is this a common average case?" Just a text description of, "this is what's been the output before", and "this is what it's designed to consider"'.[52]

When considering the visual presentation of these explanations, analysts require intuitive interfaces with simple, clear explanations and visual aids that make use of plain English.[53] This should be accompanied by the option to 'click through' for more detailed information on specific outputs, such as the highest scoring features that led to a particular model prediction.[54] Documentation containing system-level performance metrics such as precision and recall at different thresholds, and technical information such as the distribution of the dataset used to test the model, should be available on request but not necessarily embedded into the user interface of the software itself.

3.2.3. Explanations for oversight bodies

The UK national security community operates within a tightly restricted framework of oversight and compliance. For compliance and inspection purposes, it may be necessary to present a log of underpinning information that supported conclusions made by analysts and

[50] L2
[51] INT2; INT4
[52] INT5
[53] L1; INT3
[54] G3; G7; L3

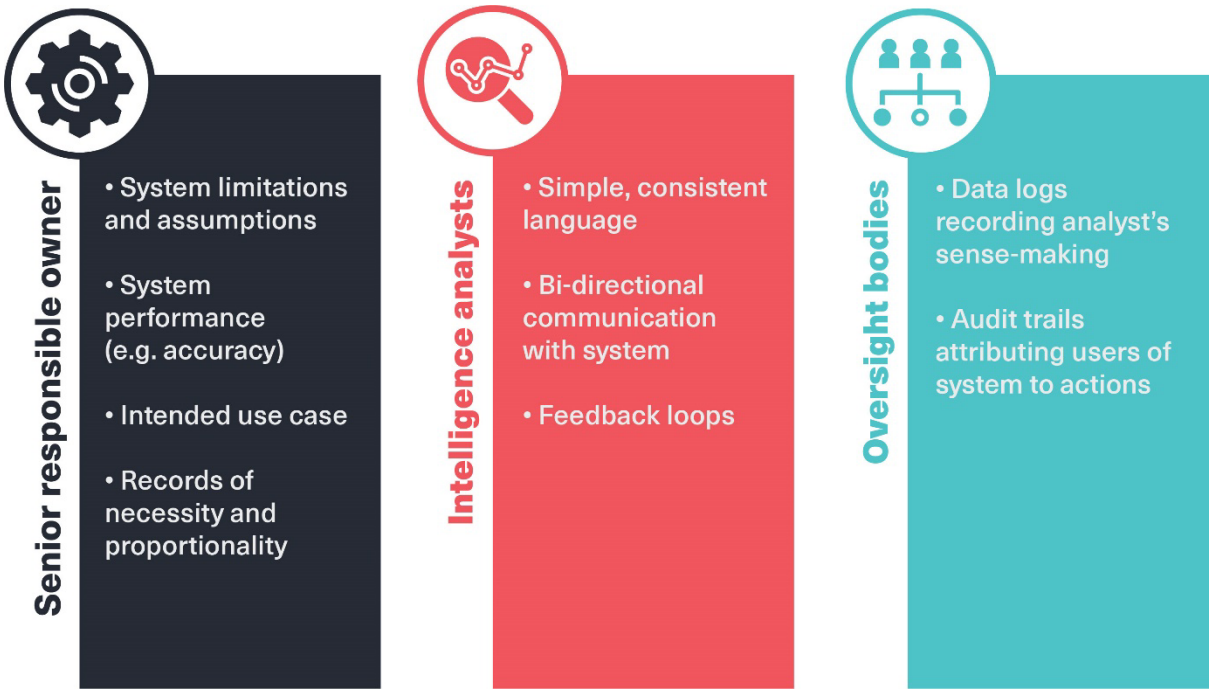
other decision-makers.[55] Decisions may later be challenged as part of a legal process or an operation may have adverse effects, leading to an inquiry or judicial review.[56] As summarised by one interviewee, 'It's the decisions that will be challenged (as part of inquests and so on), and we will need some history and audit trail to explain why certain decisions were made.'[57] Another characterised this requirement as follows:

> 'You need to be able to say there is an audit trail to explain how those conclusions have been reached and why bits of data have been selected [...] without going into masses of detail and tying people in knots. Sometime people try to justify what they do, but it doesn't need to be long paragraphs with clever words. You just need the core explanation – especially where privacy is involved and justifying the necessity and proportionality of a query for information or use of certain data.'[58]

This audit trail for any data-driven intelligence analysis would need to capture the actions of different users involved in the process, as well as the post-hoc rationale for why certain data was accessed and analysed. Specifically, this should capture the justification for why such access was judged to be necessary and proportionate: 'This data was accessed by this analyst on this date for this reason and the reason that was used in court is because...[59]' Another example in the ML context is if a system were leveraging supervised learning (where the features or input variables are pre-labelled by a human), it would be important to know who generated the data labels used to train the model.[60] This audit trail is particularly important because the volume of data that could be processed through ML systems could change the assumptions surrounding the proportionality of the analyst's access to that data.[61]

[55] L1; L2; INT5
[56] G7
[57] L1
[58] G7; L1; L2
[59] L1
[60] G3
[61] L1

Figure 2. Overview of information requirements for different stakeholders to enable effective use and oversight of ML capabilities



These findings highlight the critical importance of ensuring a human-interpretable audit trail of ML-enriched decision-making in the national security context. While there are various existing standards in place to ensure the integrity of analysis processes and subsequent reporting, these will need to be adapted to account for the use of ML decision-support systems – particularly if those systems are informing a decision process which may have a legally significant effect on an individual. It is beyond the scope of this study to elaborate further on the legal considerations raised by the use of machine learning and automated analytics for intelligence analysis. The issue of proportionality of intrusion of automated analytics is the focus of a separate ongoing CETaS research project, with a final report expected in early 2023.

3.3 Factors influencing the required granularity of explanation

Finally, interviewees agreed that ML tools for intelligence analysis do not need to be perfect or fully explainable. Numerous factors influence the degree of detail that may be required to explain technical attributes of the model or individual outputs. Key findings in this regard were as follows.

The complexity of explanation should be determined by the complexity of the problem. If a problem is cognitively straightforward for a human, an explanation of an ML solution to that problem is unnecessary and does not help decision-making.[62] For example, if a model suggests that an image looks like it contains a firearm, an explanation of why it produced that result is unlikely to help the analyst – they can simply verify the image in question and make their own judgement as to whether it contains a firearm. However, for more complex tasks which cannot be easily completed by a human (for instance, identifying patterns and connections across large volumes of bulk data), it is more important for the model to provide some reasoning as to how it arrived at a certain output.

The higher the likelihood that a significant decision will be made on the basis of a model output, the more detailed the explanation should be.[63] The model explanation should be proportionate to the severity of the impact of the decision. For example, if a ML model serves the wrong advert to someone on a shopping website, there is little need for a detailed explanation of why that happened. However, if the decision could significantly impact human rights, then there would need to be a more detailed explanation of the factors that the ML used to produce an output.[64]

As the level of detail required regarding technical attributes of the model or the features that led to individual outputs is dependent on a range of factors, it is unfeasible for model developers to prescribe model-level explainability requirements that are appropriate for the full range of potential contexts in which a model may be deployed. This emphasises again the need for a layered approach, whereby key technical attributes about the model are available on demand as required (for instance, in supporting documentation available on an organisational intranet page), and explanations related to individual model predictions are provided with varying levels of granularity as part of a 'click-through' interface.

[62] INT1
[63] G1 & G2
[64] L2

4. Management Processes and Structures

The previous sections detailed the technical requirements that must be considered at the development stage for a ML system used to support intelligence analysis. In addition to ensuring these technical requirements are in place, some adaptations to management structures and processes should also be considered to facilitate the effective and responsible use of ML systems for intelligence analysis. This includes formalised processes for designing and commissioning ML-enabled systems, the standardisation of language and processes, early testing and prototyping of new ML-enabled systems with analysts, and standardised ML operations processes throughout the full software development pipeline.

4.1 Diversity of thought

The research highlighted the critical importance of ensuring that a wide range of relevant stakeholders are involved when developing the requirements for ML-enabled tools to support analysts,[65] including end-users, user researchers, behavioural scientists and software engineers with appropriate support from policy managers. End-users must be involved in requirement setting to ensure that systems aid rather than hinder their work, as well as to avoid the perception that the system is being enforced upon analysts, which could contribute to rejection of the system. As one interviewee said, 'If the analysts understand how the process works, they become part of it rather than it being something that is done to them'.[66]

Analysts interviewed stated that they would endorse the involvement of user researchers and behavioural scientists when developing interfaces for ML systems in order to optimise the overall user experience (e.g. by preventing 'cognitive overload,' when human working memory is oversaturated and starts to reduce the analyst's capacity to take in information or execute tasks). One successful example of this approach is the US Defense Advanced Research Projects Agency (DARPA) Explainable AI programme, which has involved research teams comprising both technical AI/ML specialists, and psychologists who understood the limitations and common failure modes of human decision-making and biases.[67]

[65] INT1; G3

[66] G1 & G2

[67] Further information: https://www.darpa.mil/program/explainable-artificial-intelligence

4.2 Standardisation of terminology and explainability practices

Experts unanimously emphasised the need for standardisation of linguistic terminology when communicating technical information related to ML models and their outputs. This should also be accompanied by establishing universal explainability standards for ML systems used across government, partners and suppliers.[68]

It was noted that there is currently no standard approach to ML explainability, and no gold-standard test that can be run on an explanation to assess the utility of that explanation. The following forms of assurance were suggested to ensure better standardisation of ML explainability practices, as well as the linguistic communication of technical information:

- **Use plain English and consistent terminology** to describe the technical performance and limitations of the model (such as expected false positive and false negative rates, and classification thresholds).
- **Define and publish precision and recall rates at different thresholds,** so analysts can interpret the current performance of the model. Interviewees also highlighted that error rates do not necessarily explain the *value* of the insights generated.[69]
- **Verification and validation procedures** for approving and releasing a model into operational use.[70]

Several ongoing initiatives may be drawn upon to harmonise and streamline standards related to ML-enabled information systems for intelligence analysis, for instance the MITRE 2018 Human-Machine Teaming Systems Engineering Guide.[71] Alongside standardisation of linguistic terminology and explainability practices, further research is also needed to explore the integration of ethical considerations into the development and deployment of ML models used in this context.

[68] G1 & G2; INT3
[69] INT1; INT3
[70] INT3
[71] Further information: https://www.mitre.org/news-insights/publication/human-machine-teaming-systems-engineering-guide

4.3 The importance of user testing

There was strong recognition of the need for user testing of ML systems prior to full operational deployment. Interviewees suggested that the most effective and safe way to integrate ML into analyst workflows is to trial an ML system as a proof-of-concept or sandbox environment that analysts can experiment with. Lessons learned from such initiatives should guide further development and deployment of the capability, for instance to develop a better understanding of the error rate of the system, and users' tolerances for false positives vs. false negatives, which could help developers assess how to optimise the model's thresholds.[72] There was also recognition of the importance of adopting standard software engineering practices into the ML development and testing process.

Controlled trials could also be conducted as part of the user testing and evaluation process. For example, a trial could compare the error rates and speed of using ML tools to complete intelligence analysis tasks, in contrast to conducting the task without ML, to assess whether the ML system has made a positive and beneficial difference.[73] An interviewee described how ML capabilities were initially introduced to support low-risk tasks and then gradually introduced into higher-stakes decision contexts as user acceptance grew.[74] Another interviewee highlighted that 'The gold standard for assessing an explanation is still probably a well-designed human subjects research study to assess in a specific operational context what is the value of an explanation', and specifically to assess the right level of explanation for a particular user community.[75]

There was general agreement that the lessons learnt from controlled experiments and ML user testing provide helpful evidence for future deployments, and should be routinely shared with partner organisations deploying similar ML-enabled information systems.[76]

4.4 Resource requirements

There was universal recognition of the significant resource investment required to develop sustainable and high-quality ML systems. Some interviewees noted the risk of not investing appropriately in the skills, infrastructure or governance of ML-enabled systems, which could lead to poorly performing models that are nevertheless still used by analysts. Consideration

[72] G3; INT4
[73] L3
[74] G4, G5 & G6
[75] INT1
[76] INT1; G4, G5 & G6

should therefore be given to how to sustain ML-enabled systems such that the necessary standard of development, deployment and use is maintained. One interviewee raised the specific example that retraining a particular ML system had proven to be prohibitively expensive.[77] Another raised an example of a system requiring a legal process to change the extraction rules every time it was used, highlighting an issue that needed to be addressed before the system was deployed on a larger scale.[78]

There was concern that some organisations may not have any guidance in place regarding the development and deployment of ML systems, nor the resources or expertise required to develop such guidance. Devoting resource to forecasting such potential bottlenecks in development before fielding a capability would be beneficial to organisations committed to deploying an ML-enabled information system. Furthermore, some interviewees mentioned that knowing when to decommission and retire a ML capability, and having a well-documented and understood process to make such decisions, would also be key to ensuring that scarce resources are freed up and re-deployed efficiently.

In summary, there was recognition that technical explainability alone is not sufficient to ensure responsible and trustworthy use of ML for intelligence analysis. Instead, a range of organisational policies and processes are required to govern the full development and deployment lifecycle, from the stage of when the system is being scoped and conceptualised, through to trialling and evaluation, and ongoing monitoring and evaluation of performance.

4.5 Training and support

Finally, there were mixed responses from interviewees regarding the level of training and support required for users to effectively exploit ML systems. Some interviewees suggested that analysts should not need considerable additional training,[79] although an improvement (or at least a well-defined baseline) in all analysts' data literacy and foundational understanding of ML would be beneficial to overall analytic tradecraft. Others suggested that dedicated training programmes should be established to upskill analysts in the use of ML. It was also suggested that skills development could be geared at training a portion of more technically skilled analysts to build models that support intelligence analysts themselves. The role of data scientists to provide *deep support* to analysts was also cited by

[77] INT1
[78] G1 & G2
[79] G8 & G9

interviewees, and should be considered as part of the work environment for using ML-enabled tools.

Interviewees generally agreed that there is a need to ensure that data literacy is increased across the community, and that the analyst workforce is 'knowledgeable about our data, where it comes from, and what processes we apply to it. We need this before we get anywhere near ML'.[80]

Interviewees stated that it may not be necessary for analysts to have in-depth expertise on ML. However foundational knowledge on common ML limitations would help users calibrate the level of trust they should attribute to any ML capability.[81] As one interviewee summarised, 'The training I want is teaching analysts to be intelligent consumers of ML outputs'.[82] In high-stakes contexts where users are required to retain a high degree of accountability for individual decisions, it is all the more important to avoid potential over-trust in ML systems, and effective training was seen as crucial in this regard.

Some interviewees suggested that intelligence analysts are trained and experienced in working under conditions of uncertainty and therefore may not need further training in this regard,[83] while others expressed concern that automating some parts of the intelligence process may remove elements of reflective thinking (automation bias).[84] Part of securing analyst buy-in could mean developing some analysts' skills to build ML-enabled tools for other analysts, described as the 'democratisation of data science' – enabling more people with only limited expertise to conduct data science tasks with sufficient data and user-friendly tools – generating confidence in the model and the safeguards embedded within the system.

Finally, it was also suggested that data science should be offered as a 'support service' to analysts using ML-enabled tools operationally. This could benefit both the data scientist and the analyst as each becomes more familiar and aware of the knowledge of the other.

[80] G1 & G2; G3; G4, G5 & G6
[81] INT1; G3
[82] G3
[83] G8 & G9
[84] INT3

5. Further Research

This study was inevitably limited in scope, as it focused specifically on the technical and policy considerations regarding the use of machine learning to support intelligence analysis, with an emphasis on explainability. Several closely related topics were also identified in the research and are recommended for further investigation.

5.1 Human-machine teaming

The difference between ML as a tool and ML as a team-mate for an analyst is significant and important. To treat a ML-enabled system as a 'team-mate', analysts need to be confident in delegating tasks, treating suggestions and outputs from the machine with the same level of consideration as a fellow analyst, and possibly even allowing the machine to set its own goals. This can best be understood as the distinction between a 'decision-support tool' and human-machine teaming (HMT).[85]

The technical, psychological, operational, legal and ethical considerations of HMT are broader and more complex than a ML decision support tool. Future considerations for HMT within intelligence analysis should be investigated further, building on the insights from this research and the broader field of research into human-machine teaming. Some stakeholders suggested that future HMT systems should enable analysts to ask questions, rather than just presenting information.[86] One interviewee proposed that 'an ideal system could be a bit feisty or argumentative',[87] to prompt the analyst to consider countervailing hypotheses and mitigate cognitive biases. At the same time, a behavioural scientist highlighted that 'you don't want something that questions everything you do or people will just turn it off'.[88] There is therefore an important balance to be struck in terms of *how* the outputs and suggestions from the system are presented to users, to enable optimal teaming of both human and machine capabilities.

Dialectic communication could also take the form of regular human feedback loops for the user to relay the usefulness of the ML output, and for the system to learn from the analyst's method of enquiry.[89] A prompt to an analyst might say, 'This classifier has a 90 per cent

[85] Mariarosaria Taddeo et al., "Artificial Intelligence for National Security: The Predictability Problem," *CETaS Research Reports* (September 2022).
[86] G8 & G9; INT1; INT5
[87] G8 & G9
[88] G8 & G9
[89] G1 & G2; G8 & G9

accuracy rate, and by the way, at 10pm, we are going to send you 10 random data inputs, could you please classify them?'[90] These exchanges could be particularly valuable in the case of online learning systems, which can improve their performance and optimise precision and recall over time in response to user feedback.[91] This is a promising avenue for further research and warrants detailed scrutiny.

5.2 Automated decision-making

Automated decision-making – where a machine-generated insight could instigate real-world action in the physical or information domain – was mentioned during interviews, but was largely outside the scope of this research. Some factors were perceived to overlap, such as the importance of context to the perceived level of risk and the need for extensive testing with human experts. Nevertheless, the level of risk, and therefore the level of oversight and explainability required, for using automated decision-making was deemed considerably higher. The realistic use of automated decision-making in the intelligence analysis workflow should be comprehensively investigated.

5.3 Human-centred engineering of ML/AI applications

A major insight from this study is the importance of understanding how intelligence analysts think, behave and act, *and then* developing the ML capability to complement and support the analyst. It is not sufficient to take a dataset, develop a ML model, and yield useful information from it. That insight needs to be of value to the analyst. We have already highlighted the need for behavioural scientists to be involved, along with analysts themselves, in the development of ML capabilities. We recommend taking this a step further to develop a methodology for integrating analyst-centred design in multidisciplinary capability development teams. The emerging field of *decision intelligence*, behavioural science methods such as target audience analysis, software engineering disciplines such as UX (User Experience) and UI (User Interface) design, and data science hypothesis-driven project design, all provide key ingredients. What is now needed is the recipe (methodology) to bring the right mix of skills, experience and tools to comprehensively understand the intelligence analyst. This should be central to any future research efforts to develop and design ML systems for use in an intelligence analysis context.

[90] G3
[91] G3

5.4 Technical requirements for users from different backgrounds

Finally, a key finding of this study was the importance of user context, as different users will require differing levels of technical information regarding a ML system and how it is operating. However, given the resource and data sample limitations of this study, it was not possible to explore every factor relevant to the user context, or to systematically assess how the explainability requirements of users varies across background. This study focused primarily on the factors related to the user's role (an intelligence analyst), but more specific recommendations may be given for different types of intelligence analysts across different backgrounds and contexts. This is an important avenue for future research and warrants further consideration. Finally, given the wide range of stakeholders required to interpret ML-supported analysis, further work should seek to develop a standardised lexicon of terminology for communicating the confidence associated with ML-supported analysis (for instance based on the PHIA probability yardstick).

About the Authors

Anna Knack is a Senior Research Associate in the Turing Defence and Security programme and Lead Researcher at the Centre for Emerging Technology and Security. Anna's recent and ongoing research is focused on human-machine teaming, AI explainability, AI-augmented decision-making and cyber-AI. Prior to joining The Turing, Anna was Deputy Co-ord Lead of the Technology, Disruption & Uncertainty research workstream in RAND, where she coordinated a portfolio of work that leveraged futures methods to help policymakers understand and prepare for the future strategic operating environment.

Dr. Richard J. Carter is a Senior Research Consultant at the Centre for Emerging Technology and Security. He is a computer scientist and strategic advisor to government and industry on emerging technologies and strategic change. Rich advises the UK government on artificial intelligence and undertakes research in AI-human partnering as part of his honorary research position at the University of Bristol. Rich has been part of the UK's Defence & Security sector for most of his 25-year career but has also had stints in the video games industry where he produced award-winning and BAFTA-nominated video games, and in academia where he co-founded the UK's first national nanotechnology centre at the University of Bristol. Rich is a Fellow of the Royal Society of Arts, a Fellow of the British Computer Society, and a Chartered IT Professional. He holds a PhD in Complexity Science, a Master's in Business Administration, a Master's in Nanotechnology, a Bachelor's degree in Computer Science and a Postgraduate Certificate in Law.

Alexander Babuta is Head of the Centre for Emerging Technology and Security. His research interests include the applications of artificial intelligence and data science for UK security and policing, the regulation of investigatory powers, and the psychology of criminal offending. Prior to joining The Alan Turing Institute, he worked within the UK Government as AI Futures Lead at the Centre for Data Ethics and Innovation, and before this as Research Fellow for National Security at the Royal United Services Institute. He is also Chair of the Essex Police Data Ethics Committee. Alexander holds an MSc with Distinction in Crime Science from University College London (UCL), where his research explored the use of data science methods for police risk assessment of missing children. He also holds a first class Bachelor's degree in Linguistics from UCL.

CE TaS

Centre for Emerging Technology and Security

RESEARCH REPORT