The Alan Turing Institute

Centre for Emerging Technology and Security

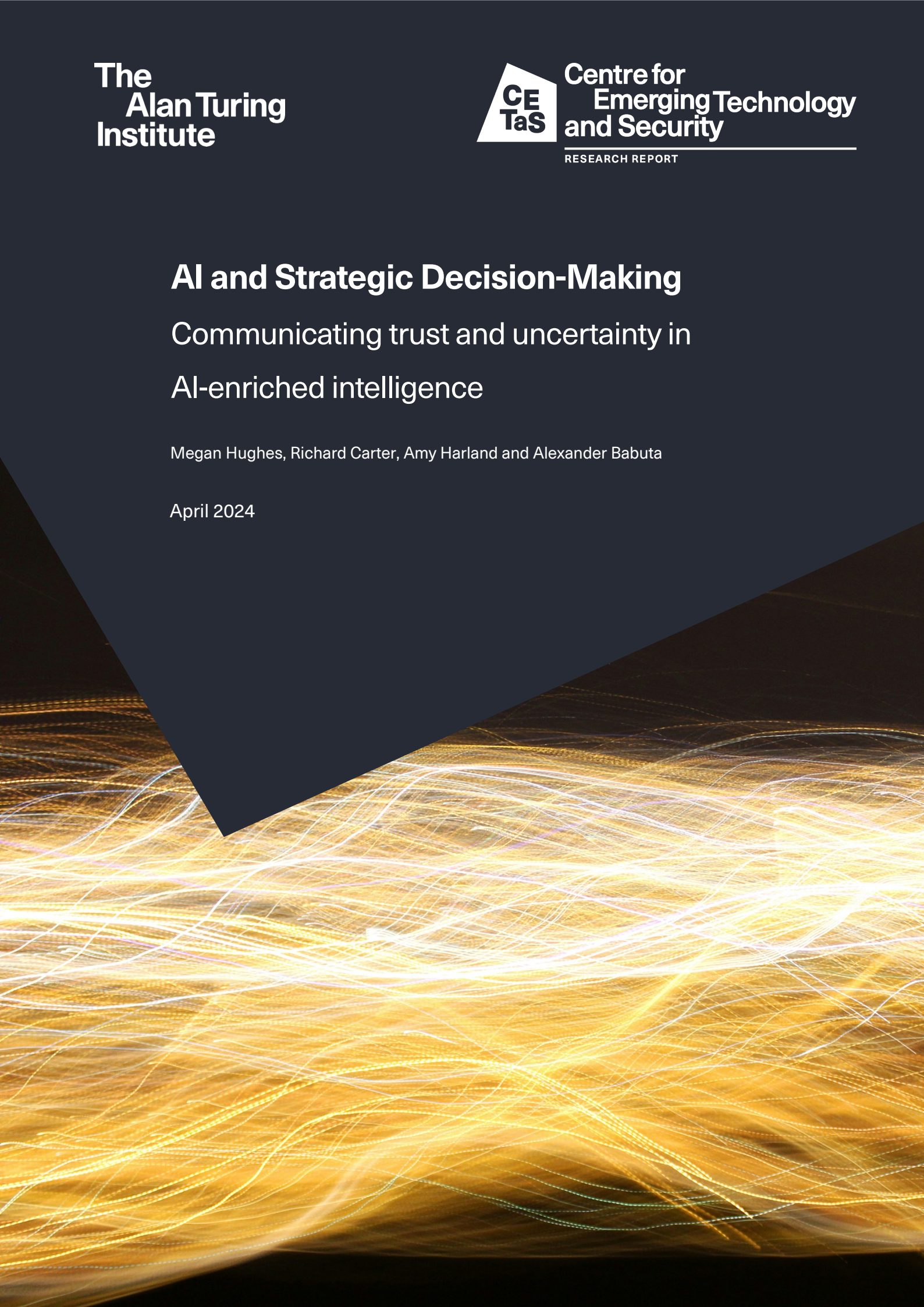
RESEARCH REPORT

AI and Strategic Decision-Making

Communicating trust and uncertainty in AI-enriched intelligence

Megan Hughes, Richard Carter, Amy Harland and Alexander Babuta

April 2024



Foreword

Advances in artificial intelligence (AI) bring new opportunities and hold exciting potential for both intelligence production and assessment, helping to surface new intelligence insights and boosting productivity. AI is not new to GCHQ or the intelligence assessment community. But the accelerating pace of change is. In an increasingly contested and volatile world, we need to continue to exploit AI to identify threats and emerging risks, alongside our important contribution to ensuring AI safety and security.

Across intelligence production and all-source assessment, AI can help to surface new insights and ensure that our analysts can access, at speed, a far greater range of data and information. We must harness the potential of AI to make sense of the ever-expanding volume of material which can inform our assessments. If we don't, we risk drowning in data and failing to spot emerging risks or trends as a result.

At the same time, advances in AI bring some new challenges for intelligence production and assessment. Questions of bias, robustness, and source validation apply just as much to AI systems as they do to the more traditional sources of insight.

This welcome, groundbreaking report explores some of the ways in which we may need to adapt our intelligence system to successfully integrate AI tools into our work. And it seeks to answer the difficult question of what needs to be in place for AI-enriched insights to be used effectively and wisely in the assessments which inform National Security decisions.

We are grateful to the Alan Turing Institute's Centre for Emerging Technology and Security (CETaS) for helping us explore this important issue, and to the large number of people across Government who have contributed to this research.

Madeleine Alessandri CMG
Chair of the Joint Intelligence Committee

Anne Keast-Butler
Director GCHQ

About CETaS

The Centre for Emerging Technology and Security (CETaS) is a 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 are grateful to all those who took part in a research interview, focus group or exercise for this project, without whom the research would not have been possible. The authors are especially grateful to Sam for his contributions and insights throughout the research, and to Claire and Ann for supporting the project and facilitating stakeholder engagement. The authors would also like to thank Sir David Omand, Rupert Barrett-Taylor, Vivien, Rosie, Tom and Emily for their valuable feedback on an earlier version of this report. Design for this report was led by Michelle Wronski.

Cite this work as: Megan Hughes, Richard Carter, Amy Harland and Alexander Babuta, "AI and Strategic Decision-Making: Communicating trust and uncertainty in AI-enriched intelligence," *CETaS Research Reports* (April 2024).

Executive Summary

This report presents the findings of a CETaS research project commissioned by the Joint Intelligence Organisation (JIO) and GCHQ, on the topic of artificial intelligence (AI) and strategic decision-making. The report assesses how AI-enriched intelligence should be communicated to strategic decision-makers in government, to ensure the principles of analytical rigour, transparency, and reliability of intelligence reporting and assessment are upheld. The findings are based on extensive primary research across UK assessment bodies, intelligence agencies, and other government departments, conducted over a seven-month period throughout 2023-24.

'AI-enriched intelligence' in this context refers to intelligence insights that have been derived in part or in whole from the use of machine learning analysis or generative AI systems such as large language models.

The research considered:

1. Whether national security decision-makers are sufficiently equipped to assess the limitations and uncertainty inherent in assessments informed by AI-enriched intelligence.

2. When and how the limitations of AI-enriched intelligence should be communicated to national security decision-makers to ensure a balance is struck between accessibility and technical detail.

3. Whether further governance, guidelines, or upskilling may be required to enable national security decision-makers to make high-stakes decisions based on AI-enriched insights.

Key findings from the research are as follows:

1. **AI is a valuable analytical tool** for all-source intelligence analysts. AI systems can process volumes of data far beyond the capacity of human analysts, identifying trends and anomalies that may otherwise go unnoticed. Choosing *not* to make use of AI for intelligence purposes **therefore risks contravening the principle of comprehensive coverage in intelligence assessment,** set out in the Professional Head of Intelligence Assessment Common Analytical Standards. Further, if key patterns and connections are missed, the **failure to adopt AI tools could undermine the authority and value of all-source intelligence assessments** to government.

2. However, **the use of AI exacerbates dimensions of uncertainty** inherent in intelligence assessment and decision-making processes. The outputs of AI systems are probabilistic calculations (not certainties) and are currently prone to inaccuracies when presented with incomplete or skewed data. The opaque nature of many AI systems also makes it difficult to understand how AI-derived conclusions have been reached.

3. There is a critical need for **careful design, continuous monitoring, and regular adjustment of AI systems used in intelligence analysis and assessment** to mitigate the risk of amplifying bias and errors.

4. **The intelligence function producing the assessment product remains ultimately responsible for evaluating relevant technical metrics** (such as accuracy and error rates) in AI methods used for intelligence analysis and assessment, and all-source intelligence analysts must take into account any limitations and uncertainties when producing their conclusions and judgements.

5. National security decision-makers currently require a **high level of assurance relating to AI system performance and security** to make decisions based on AI-enriched intelligence.

6. In the absence of a robust assurance process for AI systems, national security decision-makers generally exhibited **greater confidence in the ability of AI to identify events and occurrences** than the ability of AI to determine causality. Decision-makers were more prepared to trust AI-enriched intelligence insights when they were corroborated by non-AI, interpretable intelligence sources.

7. **Technical knowledge of AI systems varied greatly among decision-makers.** Research participants repeatedly suggested that a baseline understanding of the fundamentals of AI, current capabilities, and corresponding assurance processes, would be necessary for decision-makers to make load-bearing decisions based on AI-enriched intelligence.

This report recommends the following actions to embed best practice when communicating AI-enriched intelligence to strategic decision-makers.

1. The Professional Head of Intelligence Assessment (PHIA) should develop guidance for **communicating uncertainty within AI-enriched intelligence in all-source assessment.** This guidance should outline standardised terminology to be used if articulating AI-related limitations and caveats to decision-makers. Guidance should also be provided on the threshold at which assessments should communicate the use of AI-enriched intelligence to decision-makers.

2. **A layered approach should be taken by the assessment community when presenting technical information to strategic decision-makers.** Assessments in a final intelligence product presented to decision-makers should always remain interpretable to non-technical audiences. However, additional information on system performance and limitations should be available on request for those with more technical expertise.

3. The UK Intelligence Assessment Academy should complete a **Training Needs Analysis on behalf of the all-source assessment community** to identify the requirement for training for new and existing analysts. The Academy should work with all-source assessment organisations to develop appropriate training in response to the Analysis.

4. **Training should be offered to national security decision-makers** (and their staff) to build their trust in assessments informed by AI-enriched intelligence**.** Decision-makers should be given basic briefings on the fundamentals of AI and corresponding assurance processes.

5. **Short, optional expert briefings should be offered immediately prior to high-stakes national security decision-making sessions** where AI-enriched intelligence underpins load-bearing decisions. These sessions should brief decision-makers on key technical details and limitations, and ensure they are given advanced opportunity to consider confidence ratings. These briefings should be jointly coordinated by the JIO and National Security Secretariat and should draw from cross-governmental expertise from the network of Chief Scientific Advisers and relevant Scientific Advisory Councils. Guidance on when to offer briefings should be produced, and the need for briefings should be continuously assessed; as decision-makers become more comfortable with consuming AI-enriched intelligence, the level of desired assurance may reduce, and briefings may eventually become unnecessary.

6. **A formal accreditation programme should be developed for AI systems used in intelligence analysis and assessment** to ensure models meet minimum policy requirements of robustness, security, transparency, and a record of inherent bias and mitigation. Technical assurance for the application of a system to a specific problem should be devolved to relevant organisations, and **each organisation's assurance process should be accredited**. This programme will require dedicated resourcing, bringing together understanding of intelligence assessment standards and processes with technical expertise. PHIA should assist in developing principles and requirements, while technical expertise for accreditation and testing should be drawn from technical authorities in the intelligence community and across government.

1. Introduction

This report presents the findings of a CETaS research project commissioned by the Joint Intelligence Organisation (JIO) and GCHQ on the topic of artificial intelligence (AI) and strategic decision-making. The research sought to examine the question:

> 'How should AI-enriched intelligence be communicated to strategic decision-makers in government, to ensure the principles of analytical rigour, transparency, and reliability of intelligence reporting and assessment are upheld?'

Throughout this report, 'AI' is used to refer to machine learning (ML), and the phrase 'AI-enriched intelligence' refers to intelligence insights that have been derived in part or in whole from the use of ML analysis, or generative AI systems such as large language models (LLMs).

A key function of the UK intelligence analysis profession is to provide timely and accurate insights to support strategic decision-making. All-source intelligence analysts draw together diverse sources of information and contextualise them for strategic decision-makers (SDMs) across government. This involves drawing on intelligence and other information and adding a layer of professional judgement to form all-source intelligence assessments to support decision-making.[1] Analysts draw conclusions from incomplete information whilst highlighting gaps in knowledge and effectively communicating uncertainty.

Assessing and evaluating incomplete and unreliable information is a core responsibility of an intelligence analyst. The decisions taken on the basis of intelligence assessments can be highly consequential and load-bearing – for instance, whether to authorise military activity, diplomatic responses, or domestic public safety measures in the event of national emergencies.

Over the past two decades, there has been a huge growth in the volumes of data potentially available for analysis. Intelligence assessment functions have a significant challenge to identify, process, and analyse these exponentially growing sources and quantities of information. AI has the potential to offer both incremental and transformational improvements to the rigour and speed of intelligence assessments, and has been shown to

[1] HM Government, *About us* (Intelligence Analysis), https://www.gov.uk/government/organisations/civil-service-intelligence-analysis-profession/about.

be a crucial tool for improving productivity and effectiveness in intelligence analysis and assessment.[2]

In 2020, the Royal United Services Institute's independent review of AI and UK National Security identified 'numerous opportunities for the UK national security community' to use AI to improve efficiency and effectiveness of existing processes, concluding that 'AI methods can rapidly derive insights from large, disparate datasets and identify connections that would otherwise go unnoticed by human operators'. The review identified three specific priorities for 'Augmented Intelligence' systems within intelligence analysis:

(i) **Natural language processing and audio visual analysis** (such as machine translation, speaker identification, object recognition or video summarisation);

(ii) **Filtering and triage** of material gathered through bulk collection;

(iii) **Behavioural analytics** to derive insights at the individual subject level.

According to one US-based study, an all-source analyst could save more than 45 days a year with the support of AI-enabled systems completing tasks such as transcription and research.[3] AI has also been identified as key to maintaining strategic intelligence advantage over increasingly sophisticated adversaries.[4] A failure to adopt AI tools could therefore lead to a failure to provide strategic warning.

However, the use of AI-enriched intelligence to inform all-source intelligence assessment is not without risk. AI could both exacerbate known risks in intelligence work such as bias and uncertainty, and make it difficult for analysts to evaluate and communicate the limitations of AI-enriched intelligence. A key challenge for the assessment community will be maximising the opportunities and benefits of AI, while mitigating any risks.

This report considers strategic decision-making in the context of national security and defines strategic decision-making as the process of making key decisions that have a significant impact on national security outcomes. Such decisions typically include

[2] Adam C and Richard Carter, "Large Language Models and Intelligence Analysis," *CETaS Expert Analysis* (July 2023); Anna Knack, Richard Carter and Alexander Babuta, "Human-Machine Teaming in Intelligence Analysis: Requirements for developing trust in machine learning systems," *CETaS Research Reports* (December 2022); Alexander Babuta, Ardi Janjeva and Marion Oswald, "Artificial Intelligence and UK National Security: Policy Considerations," *RUSI Occasional Papers* (April 2020); GCHQ, "Pioneering a New National Security," (2021), https://www.gchq.gov.uk/files/GCHQAIPaper.pdf.

[3] Mitchel et al., "The future of intelligence analysis," *The Deloitte Center for Government Insights,* (2019).

[4] CSIS Technology and Intelligence Task Force, *Maintaining the Intelligence Edge* (Center for Strategic & International Studies: January 2021).

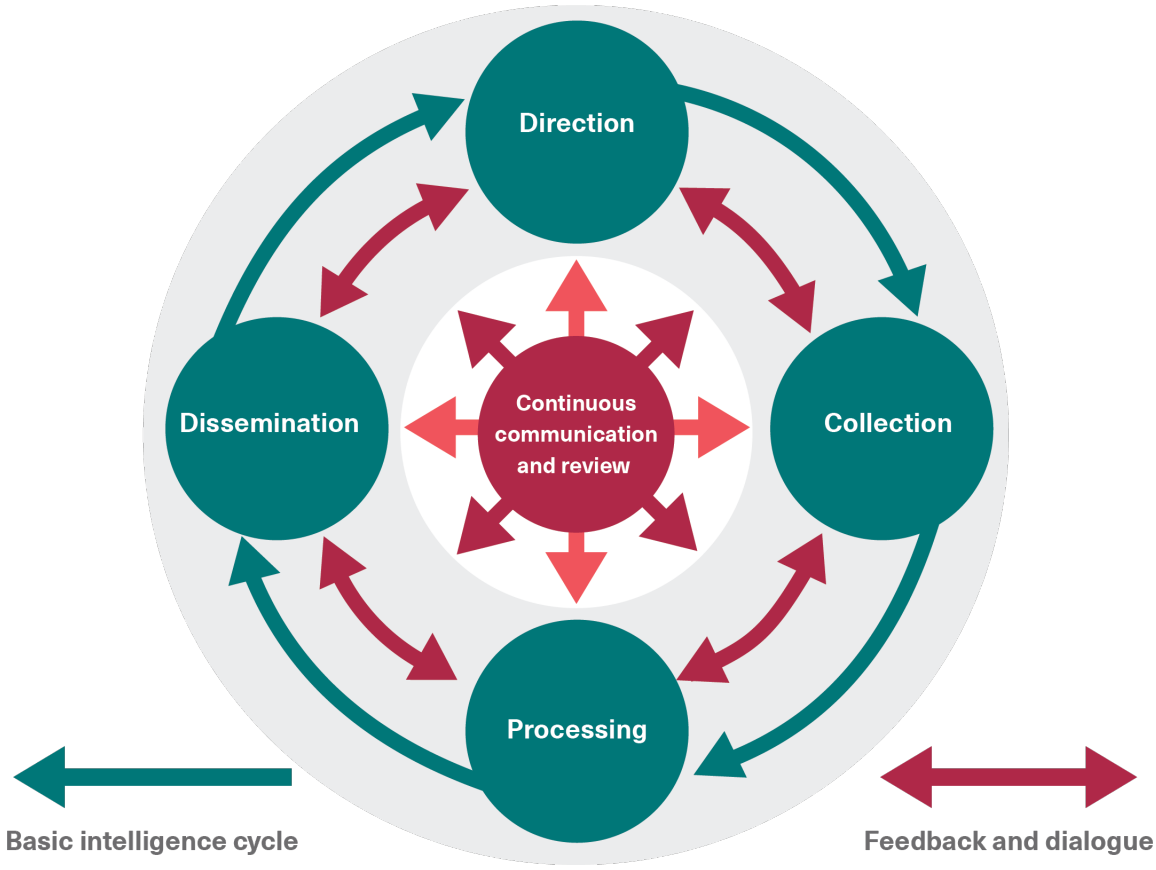
consideration of the potential impact on the safety and prosperity of the public or the country's global standing in the world. A strategic decision-maker is an individual whose contribution to the process has a material bearing on the outcome. Such decision-makers may be government officials such as senior civil servants (e.g. relevant departmental Director Generals or Permanent Secretaries), or ministers and Secretaries of State attending the National Security Council (e.g. the Foreign Secretary, Defence Secretary or Prime Minister).

This report examines whether, in today's context of data proliferation and fast-developing AI technology, current practices are sufficient to maintain the rigour, transparency, and reliability demanded by intelligence assessment standards. Uncertainty is not new or unique to AI – it is inherent in all intelligence analysis and assessment. However, AI has the potential to *exacerbate* uncertainty. The research investigated when and how the limitations of AI-enriched intelligence should be communicated by all-source intelligence analysts to national security SDMs, while ensuring a balance is struck between accessibility and technical detail. Additionally, the research explored whether further governance, guidance, or upskilling may be required – both to enable the effective communication of AI-enriched intelligence within the assessment community, and to enable SDMs to make load-bearing decisions based on judgements informed by AI-enriched insights.

1.1 The intelligence cycle

This section presents a simplified overview of the UK intelligence process to outline the stages at which AI-enriched intelligence may become relevant. The simplified cycle presented here has four core functions: tasking (or direction, whereby requirements for information are set), collection (conducted by the intelligence agencies), all-source analysis and assessment (or processing, conducted by assessment bodies including the Joint Intelligence Organisation), and dissemination of finished products to decision-makers. While this is presented as a four-stage process, all activities may be conducted concurrently, and there is continuous communication and review between each stage. This is illustrated below.

Figure 1: Joint Doctrine Publication 2-00, Intelligence, Counter-intelligence and Security Support to Joint Operations, Ministry of Defence, 2023



AI-enriched intelligence could enter the intelligence cycle either at the collection or processing stage. In either instance, it would be the responsibility of the all-source analysis and assessment function to contextualise the AI-enriched intelligence (alongside all other available information held on the same requirement) and ensure that any limitations in the evidence base are communicated appropriately to SDMs. This report is therefore focused on the analysis and assessment and dissemination stages of the intelligence cycle.

1.2 Research methodology

1.2.1 Aims and research questions

The main research aim was to gather new insight on the factors that shape the degree of confidence SDMs feel when making load-bearing decisions on the basis of AI-enriched intelligence assessment. This report addresses the following research questions:

- **RQ1:** In what circumstances (if any) is it necessary to communicate and distinguish the use of AI to strategic decision-makers, and at what stage in the reporting chain does the use of AI become unnecessary to communicate?
- **RQ2:** How should AI-enriched information be communicated to strategic decision-makers to ensure they understand the reliability, confidence and limitations of the intelligence product – and how does this vary across intelligence contexts and types of AI system?
- **RQ3:** How do we effectively educate strategic decision-makers to make high-stakes decisions based on AI-enriched reporting, and achieve the appropriate level of understanding, trust and confidence in AI systems and their outputs?
- **RQ4:** What additional governance, oversight and upskilling is required to provide assurances that AI-generated insights are being used appropriately to support senior decision-making in this context?

1.2.2 Methodology

The primary data sources for this study comprised semi-structured interviews and focus groups with stakeholders from assessment bodies across government and the UK intelligence community (UKIC).[5] A tabletop exercise was also conducted with a group of senior government officials, to test SDMs' responses to AI-enriched intelligence in a simulated scenario. This study was conducted over a seven-month period from June 2023 – January 2024. Data collection involved the following core research activities:

- **Systematic literature review** of academic and grey literature to establish the state-of-the-art in current methodologies, challenges, and perspectives regarding trust in AI. A small number of experts from academia and industry also provided their viewpoints on approaches to developing and implementing trustworthy AI systems in high-stakes environments.
- **Semi-structured interviews and focus groups** with intelligence analysts, assessment staff, and other government officials. A total of 30 research participants engaged in this phase of the research.
- **Tabletop exercise (TTX)** with 16 senior officials from numerous UK Government departments and agencies. The purpose of the TTX was to examine the decision-making process of SDMs when presented with assessments that were notionally

[5] The UKIC is defined here as the Security Service (MI5), the Secret Intelligence Service (MI6) and the Government Communication Headquarters (GCHQ).

based on AI-enriched intelligence in a simulated high-stakes scenario. The scenario used for the TTX centred on the theme of election security, and discussions were framed around fictitious outputs from a notional (but technically plausible) ML classification system.

This report is narrowly focused on the use of AI in intelligence analysis and assessment to inform strategic decision-making for national security. The following themes are out of scope of this project and are recommended as topics for future research:

- The use of AI to inform operational and tactical decision-making (as opposed to strategic decision-making).
- Communicating uncertainty in AI-enriched intelligence shared by allies and partners outside the UKIC.
- The use of AI-enriched intelligence to justify investigative activity or warrant applications.
- The vulnerabilities of AI systems used within national security to adversarial attacks or tampering.

This report tackles a sensitive and under-researched topic and therefore heavily relies upon primary research. Participants during the TTX may have been subject to the Hawthorne effect, whereby subjects may change their behaviour in response to their awareness of being observed.

The remainder of this report is structured as follows. Section 2 outlines challenges relating to introducing AI into current analysis and assessment practices. Section 3 presents opportunities for AI in intelligence analysis and assessment. Section 4 explores enabling factors for communicating AI-enriched intelligence to strategic decision-makers. Section 5 concludes with a set of recommendations for best practice when communicating AI-enriched intelligence to strategic decision-makers.

2. AI-enriched Intelligence and Uncertainty

This section provides an overview of the Professional Head of Intelligence Assessment (PHIA) Common Analytical Standards for best practice across the UK intelligence assessment community, and the two key reviews which informed the development of contemporary UK intelligence assessment standards: Lord Butler's 2004 Review of Intelligence on Weapons of Mass Destruction in Iraq;[6] and Sir John Chilcot's subsequent Report of the Iraq Inquiry, published in 2016.[7] It also considers how AI-enriched intelligence may pose challenges to existing intelligence assessment standards, and outlines strategies for building trust in AI systems used to inform intelligence assessment.

2.1 UK intelligence assessment principles

2.1.1 Interpreting the Butler and Chilcot principles

The Butler Review and Chilcot Inquiry are landmark evaluations of the intelligence processes and decision-making procedures that led the UK into conflict in Iraq in 2003. The reports sought to understand how and why the strategic decision-making system faltered, and proposed recommendations to avoid future missteps.

The Butler Review found that several key judgements in the Joint Intelligence Committee's (JIC) assessments in the lead-up to the Iraq conflict did not appropriately reflect the limitations of the underlying intelligence.[8] The Butler Review emphasised several key principles for effective and robust intelligence analysis, including:

- **Access to information:**[9] the need for rigorous, evidence-based intelligence assessments based on access to a wide range of information.
- **Transparency of sources:**[10] the importance of clearly communicating the reliability of sources in intelligence assessments. Assessments should clearly delineate between confirmed facts, interpretations, and speculation.

[6] Robin Butler, *Review of Intelligence on Weapons of Mass Destruction* (Committee of Privy Counsellors: 2004).
[7] John Chilcot, *The Report of the Iraq Inquiry* (Committee of Privy Counsellors: 2016), https://www.gov.uk/government/publications/the-report-of-the-iraq-inquiry.
[8] Robin Butler, *Review of Intelligence on Weapons of Mass Destruction* (Committee of Privy Counsellors: 2004).
[9] Robin Butler, *Review of Intelligence on Weapons of Mass Destruction* (Committee of Privy Counsellors: 2004), 153.
[10] Robin Butler, *Review of Intelligence on Weapons of Mass Destruction* (Committee of Privy Counsellors: 2004), 159.

- **Effective challenge:**[11] the promotion of a culture that values and encourages challenge.

The Chilcot Inquiry was a comprehensive investigation into the UK's involvement in Iraq.[12] While its remit was wider than the Butler Review, its findings regarding intelligence assessment and decision-making echoed and expanded on many of Butler's recommendations. The inquiry emphasised the importance of:

- **Measured, collective decision-making**:[13] decisions of significant consequence must be based on comprehensive and robust debates, considering a wide array of perspectives.

- **Critical examination of intelligence:**[14] decision-makers must fully understand the confidence and robustness of the evidence base.

The guidance to the assessment community and SDMs from Butler and Chilcot emphasised the importance of rigorous decision-making and the necessity for evidence-based assessments and careful consideration of intelligence limitations. These principles are formalised across the UK intelligence assessment community in the form of the PHIA Common Analytical Standards (CAS).

2.1.2 Common Analytical Standards

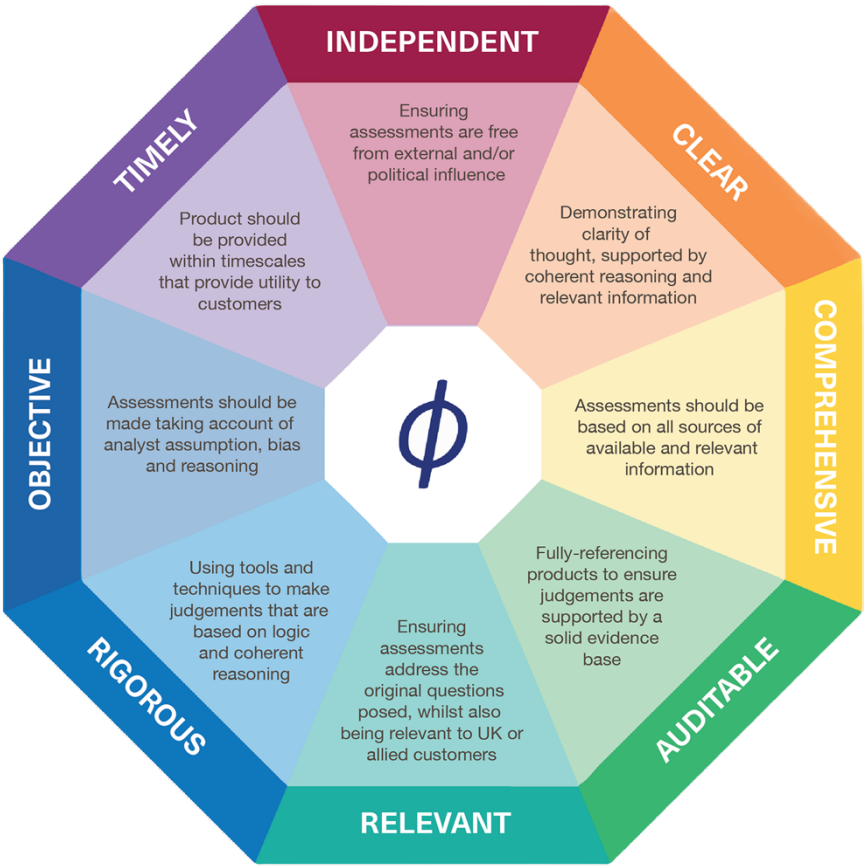
The PHIA was established in response to the Butler Review and leads on "the development of the UK intelligence analysis community's analytical capability providing training, standards and products".[15] The PHIA CAS are designed to standardise rigour, integrity, language, and best practice across the intelligence assessment community.

These standards state that all intelligence analysis work should be independent, clear, comprehensive, auditable, relevant, rigorous, objective, and timely.[16]

---

[11] Robin Butler, *Review of Intelligence on Weapons of Mass Destruction* (Committee of Privy Counsellors: 2004), 146.
[12] John Chilcot, *The Report of the Iraq Inquiry* (Committee of Privy Counsellors: 2016), https://www.gov.uk/government/publications/the-report-of-the-iraq-inquiry.
[13] John Chilcot, *The Report of the Iraq Inquiry* (Committee of Privy Counsellors: 2016), 129, https://www.gov.uk/government/publications/the-report-of-the-iraq-inquiry.
[14] John Chilcot, *The Report of the Iraq Inquiry* (Committee of Privy Counsellors: 2016), 131, https://www.gov.uk/government/publications/the-report-of-the-iraq-inquiry.
[15] HM Government, *About us* (Intelligence Analysis), https://www.gov.uk/government/organisations/civil-service-intelligence-analysis-profession/about.
[16] HM Government, *Professional Development Framework for all-source intelligence assessment* (Intelligence Analysis), https://www.gov.uk/government/publications/intelligence-analysis-professional-development-framework/professional-development-framework-for-all-source-intelligence-assessment-html.

Figure 2: Professional Development Framework for all-source intelligence assessment, HM Government



Since the establishment of the PHIA, the UK intelligence context has changed significantly. The volumes of data potentially available for analysis have rapidly increased, and the analytic tooling available to exploit this data has evolved. There is now a need to consider how all-source analysis and assessment should adapt to this context, while maintaining the high standards and requirements established by the CAS.

2.2 Potential risks associated with AI in intelligence analysis

All intelligence work carries an inherent degree of uncertainty, which in turn introduces risk in decision-making. The first Principle of the College of Policing's Authorised Professional Practice (APP) on Risk is that 'The willingness to make decisions in conditions of uncertainty (that is, risk taking) is a core professional requirement of all members of the police service'.[17]

[17] College of Policing, *Risk,* (October 2013), https://www.college.police.uk/app/risk/risk.

This principle is equally applicable to the intelligence analysis profession. The APP notes that 'By definition, decisions involve uncertainty, that is, the likelihood and impact of possible outcomes cannot be totally predicted, and no particular outcome can be guaranteed.'[18]

All-source intelligence analysts working within UK national security are trained to manage risk by evaluating uncertainty in intelligence underpinning judgements and conveying this uncertainty to SDMs using structured communication frameworks such as the Probability Yardstick and the Analytical Confidence Rating (AnCR) framework.[19]

AI could potentially amplify existing uncertainties inherent in intelligence and introduce additional challenges that are difficult for intelligence analysts to evaluate and communicate.

At the sociotechnical level, ethical and societal considerations – such as the replication of social biases in the outputs of AI systems – add layers of complexity and unpredictability in AI-enriched intelligence. Whilst progress has been made in improving the quality of the data used to train AI models, there is a trade-off between the volume of data used to train a model and the subsequent performance of that model. Improving performance requires additional training data, which is costly to maintain to a high quality.

At the technical level, AI is a probabilistic statistical method – meaning all AI outputs are associated with a degree of inherent mathematical uncertainty. Moreover, reliance on biased, inaccurate, or incomplete training data can skew AI decisions, making them unpredictable, unreliable, and inconsistent.[20] Furthermore, the complex and opaque nature of many AI algorithms makes it difficult to understand how AI-derived conclusions have been reached.[21]

AI systems can behave unpredictably. Models trained for specific purposes may not perform as expected on new, unseen data. ML models have also been shown to degrade over time in 91 per cent of cases, as the data on which they are deployed increasingly differs from that on which they were trained.[22] Furthermore, when considering complex AI systems comprising multiple underlying models, the compound effect of ML models interpreting and

[18] College of Policing, *Risk* (October 2013), https://www.college.police.uk/app/risk/risk.

[19] HM Government, *Professional Development Framework for all-source intelligence assessment* (Intelligence Analysis), https://www.gov.uk/government/publications/intelligence-analysis-professional-development-framework/professional-development-framework-for-all-source-intelligence-assessment-html.

[20] Alexander Babuta and Marion Oswald, "Data Analytics and Algorithmic Bias in Policing," *Royal United Services Institute for Defence and Security Studies* (2019).

[21] Lockey, Gillespie, Holm, and Someh, "A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions," in *Proceedings of the 54th Hawaii International Conference on System Sciences* (2021).

[22] Vela et al., "Temporal quality degradation in AI models," *Scientific Reports* 12 (2022): 11654.

acting on data generated by different ML models can lead to biases and errors accumulating or interacting in unforeseen ways, and this could lead to distorted outcomes or decisions that significantly deviate from their original intent.

The limitations and unpredictability inherent in AI systems may interact with existing cognitive biases and heuristics in decision-making, potentially amplifying the effect of human decision-making biases. Subsequently, there is a critical need for careful design, continuous monitoring, and regular adjustment of AI systems to mitigate the risk of amplifying bias and errors in intelligence assessment. The following table illustrates how AI could amplify or perpetuate three of the most common and well-documented cognitive biases.

Cognitive bias	Risk	How AI may amplify bias	Illustrative example
Confirmation bias[23]	Seeking out, interpreting, and remembering information that confirms pre-existing beliefs or hypotheses, while giving disproportionately less consideration to alternative possibilities.	• Lack of attention paid to examining alternative sources of information, as an expected and convenient answer could be returned far quicker by an AI tool.[24] • Training data might reflect confirmation biases, leading to skewed outputs that reinforce pre-existing beliefs. • Human feedback on the perceived performance of AI models may create a self-reinforcing feedback loop thus perpetuating confirmation bias.	An AI system trained on past military intelligence data might tend towards repeating historical assessments rather than objectively analysing present circumstances, leading to over- or under-estimation of current threat levels.
Anchoring bias[25]	Depending too heavily on one initial piece of information, known as the 'anchor', when making decisions.	• Disproportionate weighting given to an initial AI-derived insight, regardless of subsequent human analysis.	Decision-makers' threat perception being influenced by an initial AI-enriched report predicting an imminent attack, despite subsequent human analysis suggesting a lower risk.
Availability bias[26]	Placing greater weight on information which easily comes to mind.	• Trends in public discourse and media reporting regarding developments in AI technology may influence individuals' level of (mis)trust in AI systems.	Decision-makers choosing to disregard the output from an AI system, because of a recent high-profile case of a different AI system proving unreliable.

[23] Raymond Nickerson, "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* 2, no. 2 (1998): 175–220.

[24] Author interview with government participant, 21 August 2023.

[25] Tversky and Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, no. 4157 (1974): 1124–1131.

[26] Tversky and Kahneman, "Availability: A heuristic for judging frequency and probability," *Cognitive Psychology* 5, no. 2 (1973): 207-232.

2.3 Challenges to best practice in intelligence assessment

The following examples illustrate the limitations of AI-enriched intelligence in relation to the PHIA CAS. Some limitations are new and specific to AI, while some are known challenges faced by human analysts, which may be mirrored and exacerbated by AI. To ensure the integrity of assessments, intelligence analysts must guard against these risks where possible and clearly communicate the limitations of AI-enriched intelligence to SDMs.

Rigorous and Comprehensive. The use of AI for summarising or triaging intelligence and other information may inadvertently lead to a myopic focus (searching for a needle in the same haystack repeatedly, rather than examining the entire field of haystacks). This underscores the risk that AI, if overly tuned towards specific datasets or patterns, might narrow the scope of search and analysis to familiar territories, overlooking broader, more diverse, or more relevant information. Such a constrained approach could limit the ability to detect threats or opportunities that lie outside expected parameters, essentially missing 'needles' in other 'haystacks'.

Objective, Clear, and Auditable. The 'black box'[27] nature of AI systems could make it challenging for intelligence analysts to fully understand the limitations of AI-enriched intelligence. The output of a model could be a combination of sources of information with varying degrees of reliability. The issue of uncertainty could be further compounded by: (a) outputs from one AI model being used as the input to other models, and/or (b) a feedback loop where a biased AI model perpetuates and amplifies its biases by influencing the collection of similarly biased data. This process leads to the model assigning higher confidence scores to its predictions when applied to the new, biased data.

Independent. AI is only as reliable as the data on which it has been trained. If the training data reflects biases the AI outputs will likely mirror these flaws, resulting in assessments unwittingly influenced by biases. If intelligence analysts are overly reliant on AI systems and perceive them as infallible due to their computational abilities, it may dissuade them from challenging the AI system's outputs. The lack of explainability of powerful AI systems could exacerbate this risk and discourage challenge.

Relevant. AI lacks human judgement and the ability to contextually understand nuanced information. There is a risk that analysts might inappropriately frame questions when

[27] The 'black box' problem refers to an opaque system where calculation processes are invisible to the user.

interacting with AI systems and receive irrelevant outputs. AI systems may not appropriately account for cultural, social, or political complexities that a human analyst might consider in an assessment. Mitigation of this risk is dependent on the analyst having superior subject-matter knowledge to the AI model (to judge the relevance of the AI model's outputs), which may not always be guaranteed.

Timely. Attempts to manually corroborate AI outputs could be highly time consuming, eroding any gains in timeliness.

2.4 Building trust in AI systems

This section outlines risk mitigations for developers and users of AI capabilities for intelligence assessment.

2.4.1 Developers of AI capabilities

Mitigating technical errors and 'black box' problems. Several techniques and strategies can mitigate uncertainty in AI systems and avoid a compounding effect in a chain of AI models, including: model calibration;[28] uncertainty quantification;[29] ensemble methods;[30] probabilistic programming and Bayesian methods;[31] active learning;[32] and meta-learning.[33] More transparent and explainable modelling techniques can help users understand how AI is generating its results, which can help identify and correct for biases.[34] Explainable AI (XAI) is a growing field and some have argued *all* software systems can be made sufficiently interpretable.[35] Techniques such as Local Interpretable Model-Agnostic Explanations (LIME) or Shapley Additive exPlanations (SHAP) can help visualise and understand AI decision processes. When considering LLMs specifically, explainability for intelligence

[28] Kuleshov, Fenner, and Ermon, "Accurate Uncertainties for Deep Learning Using Calibrated Regression," in *Proceedings of the 35th International Conference on Machine Learning* 80 (2018): 2796-2804.

[29] Kendall and Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?," *Advances in Neural Information Processing Systems* (2017): 5574-5584.

[30] Lakshminarayanan, Pritzel, and Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," *Advances in Neural Information Processing Systems*, (2017): 6402-6413.

[31] Ghahramani, "Probabilistic Machine Learning and Artificial Intelligence," *Nature* 521, no. 7553, (2015): 452–459.

[32] Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, no. 1 (2012): 1-114.

[33] Finn, Abbeel, and Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *Proceedings of the 34th International Conference on Machine Learning* 70 (2017): 1126-1135.

[34] Ribeiro, Singh, and Guestrin, ""Why Should I Trust You?" Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016): 1135–1144.

[35] Adadi and Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* 6 (2018): 52138-52160; Kroll, "The fallacy of inscrutability," *Philosophical Transactions of the Royal Society A* 376, no. 2133 (2018); Lipton, "The mythos of model interpretability," *Queue* 16, no. 3 (2018): 30-57.

analysts requires the model to be able to cite sources accurately to allow for the verification of information.[36]

Improving data representativeness and quality. Biased datasets can be mitigated by ensuring that the data used to train AI is representative of the phenomenon being modelled (where possible).[37] It is also important to consider whether available data is relevant and appropriate to use. This could involve scrutinising how data was obtained and considering whether any gaps in the data exist.[38] Conducting regular bias audits can also help to identify and mitigate AI bias.[39] This involves assessing the system's outputs for fairness and neutrality, for instance through Reinforcement Learning from Human Feedback.[40]

Model cards. Model cards are formal records that provide standardised metadata on AI models (e.g. training data information, potential limitations, intended use) and are intended to increase transparency on model development and use.[41]

Adversarial training. Adversarial training, whereby AI systems are trained with the addition of intentionally crafted misleading inputs, can make models more robust and less susceptible to bias.[42] This process, akin to stress-testing, can prepare the AI to handle outliers or edge cases. Benchmark tests have now been developed to objectively measure the comparative performance of models and the degree of bias they exhibit, which should help in the design and testing of AI systems.[43]

Experimentation and periodic review. AI systems should be periodically reviewed and updated to ensure their outputs remain valid and trustworthy.[44] Continual learning techniques can be employed to allow the system to evolve over time, correcting biases that may have been introduced due to model or data drift. Less formally, iteration and 'trial and error' will give users the opportunity to experiment and familiarise themselves with new

[36] Author interview with government participant, 18 August 2023.

[37] Buolamwini and Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 1–15.

[38] Author interview with government participant G2.

[39] Friedler et al., "A Comparative Study of Fairness-enhancing Interventions in Machine Learning," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, (2019): 329–338.

[40] Christiano et al., "Deep Reinforcement Learning from Human Preferences," (2017), https://arxiv.org/pdf/1706.03741.pdf.

[41] Hugging Face, "Model Cards," https://huggingface.co/docs/hub/model-cards.

[42] Goodfellow, Shlens and Szegedy, "Explaining and Harnessing Adversarial Examples," *3rd International Conference on Learning Representations*, 7-9 May 2015.

[43] Abbas, Langlais, Rashid and Rezagholizadeh, "Context-aware Adversarial Training for Name Regularity Bias in Named Entity Recognition," *Transactions of the Association for Computational Linguistics* 9 (2021): 586–604.

[44] Chen et al., "Continual Learning for Sentiment Classification in Online Review Platforms," *IEEE Transactions on Knowledge and Data Engineering* 32, no. 6 (2018): 1195–1208.

technology.[45] A track record of historic use cases can also help analysts to gauge the accuracy of a model and build trust.[46]

2.4.2 Users of AI capabilities

Carefully considering when (not) to deploy AI. Intelligence analysts should carefully consider when the use of AI models is appropriate and most valuable. If an explainable method could be used to comprehensively answer a problem, utilising a black box ML model may lead to unnecessary risk. It will be important to critically evaluate the models and the data being used and balance the risk of using AI (e.g. lack of transparency) against the reward (e.g. speed and comprehensive coverage). This concept was termed 'context assurance' by one research participant.[47] Additionally, the risk of *not* utilising AI (e.g. missing a key insight) should be considered.

Critical thinking. Analysts within the assessment community are trained in critical thinking and challenge and encouraged to be sceptical of information.[48] These qualities are key to the responsible and appropriate use of AI.[49] It will be important that the assessment community continues to cultivate a culture of challenge and 'puts meaningful bumps in the road' to allow for humans to interrogate machine outputs.[50] While AI can be a powerful tool, it is not infallible, and model outputs should not be accepted uncritically. Triangulating outputs across multiple models and human review should help to build trustworthy models.[51]

Prompt engineering. An additional consideration when utilising LLMs specifically is the need for effective prompt engineering to achieve the desired outcomes (knowing what question to ask, and how to phrase the question appropriately). Analysts using LLMs would need to learn how to interact with the system to ensure information is being sufficiently interrogated.[52] This is particularly relevant as LLMs are often designed to be conversational in style.[53]

---

[45] Author interview with government participant, 18 August 2023.

[46] Author interview with government participant, 21 August 2023.

[47] Author interview with government participant, 21 August 2023.

[48] Author interview with government participant, 23 August 2023.

[49] Nickerson, "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* 2, no. 2 (1998): 175–220.

[50] Author interview with government participant, 11 August 2023.

[51] Author interview with government participant, 23 August 2023.

[52] Author interview with government participant, 18 August 2023; CETaS workshop, 10 November 2023.

[53] Author interview with government participant, 23 August 2023.

Collaborative human-machine decision-making. Involving both human judgement and AI recommendations in the assessment process can help counteract biases. Research across a wide range of fields has consistently shown that human decision-making is reliably improved through the introduction of statistical support tools.[54] Humans and AI have different strengths and weaknesses, and effective collaboration between the two should simultaneously maximise the strengths while minimising the weaknesses of both human and AI computational abilities.

[54] Meehl, *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence* (University of Minnesota: Oxford University Press, 1954); Dawes, Faust and Meehl, "Clinical versus actuarial judgment," *Science* 243, no. 4899 (1989):1668-1674; Grove et al., "Clinical versus mechanical prediction: a meta-analysis," *Psychological assessment* 12, no. 1 (2012); Ægisdóttir et al., "The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction," *The Counseling Psychologist* 34, no. 3 (2006): 341-382.

3. Integrating AI into Analysis and Assessment Processes

This section outlines potential opportunities and benefits for integrating AI into the intelligence cycle, and considers when it is necessary for SDMs to be notified that AI has been used in the analysis and assessment processes.

3.1 Opportunities and benefits

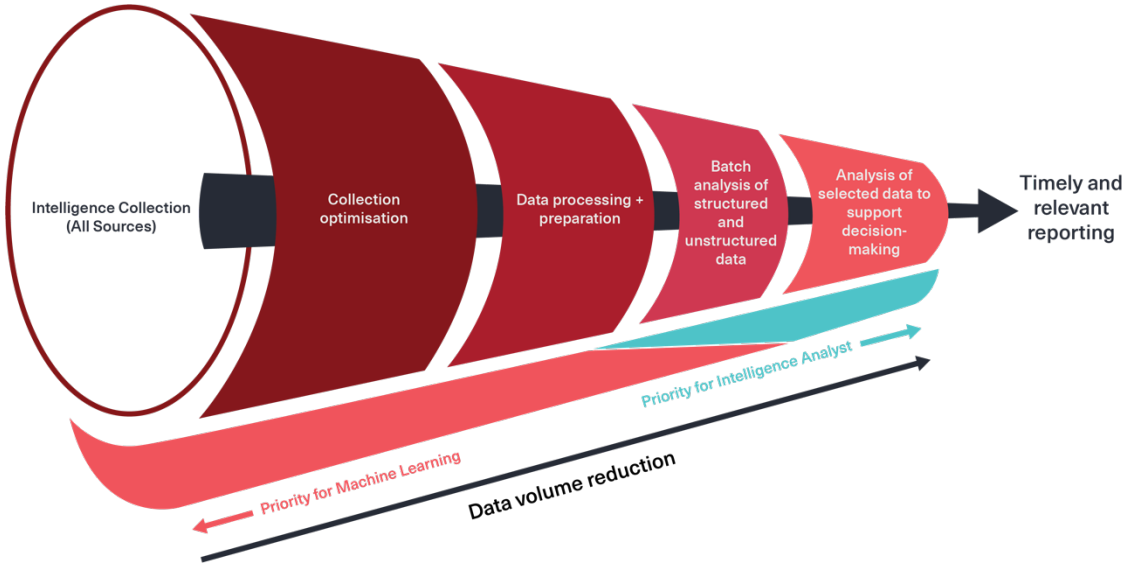
AI tools could potentially offer incremental and transformative benefits to the speed and rigour of all-source assessment. AI can be used to lighten the workload of intelligence analysts (e.g., performing tasks related to data processing), allowing more time for analysts to perform more valuable tasks.[55] Analytical rigour could be improved by using AI to triangulate and validate findings across a larger set of data. Crucially, AI can process large volumes of data (such as bulk data) and identify patterns, trends, and anomalies beyond human capability.

Prior CETaS research on human-machine teaming and intelligence analysis illustrated that AI is likely to be most useful in processing, triaging and prioritising large volumes of data (see Figure 3).[56]

[55] Alexander Babuta, Marion Oswald and Ardi Janjeva, "Artificial Intelligence and UK National Security: Policy Considerations," *RUSI Occasional Papers,* Royal United Services Institute (April 2020).

[56] Anna Knack, Richard Carter and Alexander Babuta, "Human-Machine Teaming in Intelligence Analysis: Requirements for developing trust in machine learning systems," *CETaS Research Reports* (December 2022).

Figure 3: CETaS analysis



A subsequent CETaS article co-authored with GCHQ's Chief Data Scientist identified specific areas where LLMs could be most beneficial to intelligence analysis (see Figure 4).[57]

Figure 4: CETaS analysis

The above examples have the potential to reduce the mounting pressure on human analysts facing an exponentially growing volume of data, and address the risk that key sources are not properly identified and examined due to time constraints.

As the use of AI proliferates and the general population becomes familiar with using AI in their day-to-day lives, SDMs might come to expect AI to be used more extensively in

[57] Adam C and Richard Carter, "Large Language Models and Intelligence Analysis," *CETaS Expert Analysis* (July 2023).

intelligence analysis and assessment.[58] If AI can be demonstrated to provide intelligence insights above and beyond that which could be derived through non-AI methods, then choosing *not* to use available AI tools will contravene the principle of comprehensive coverage. An inability to fully exploit both open and closed-source data may lead to patterns and connections going unnoticed. One research participant emphasised that an inability to use AI tools to access increasing volumes of data could ultimately lead to a risk of "intelligence failure".[59]

Listed below are several potential opportunities for AI usage as suggested by research participants across the assessment community. These opportunities fall within two categories: AI as a support function to automate and increase the efficiency of existing tasks; and AI as a tool to generate additional insights beyond the capability of individual analysts.

3.1.1 AI as a support function

Research participants raised a concern that current assessments of all-source intelligence can involve trade-offs based on the scope of a task and the volume of data to consider. Time pressures can also mean the scope of inquiry is inevitably limited to those sources deemed to be most relevant. The use of AI for triaging relevant data would therefore be invaluable for analysts as an efficient tool for casting the net wider to consider a greater number of sources. Alongside triage, accurate summarisation of multiple sources or large amounts of text using LLMs was also identified as a beneficial use of AI.[60] Moreover, AI could strengthen the source validation process by corroborating sources or acting as an alert for abnormalities in source reporting.[61]

Several research participants emphasised that open-source intelligence (OSINT) is not being fully exploited during the intelligence production and assessment process. This is often due to time constraints and the sensitivities that must be considered and navigated when validating and evaluating classified information. AI could support the development of OSINT tradecraft by verifying the content of intelligence reporting against open sources. AI could also be used in the delivery of personalised intelligence to decision-makers. One participant suggested a future system might be capable of recommending or providing

[58] Author interview with government participant, 18 August 2023; Author interview with government participant, 23 August 2023.
[59] Author interview with government participant, 11 August 2023.
[60] Author interview with government participant, 23 August 2023.
[61] Author interview with government participant, 21 August 2023; Author interview with government participant, 23 August 2023.

curated and summarised intelligence relevant to an intelligence consumer's particular interest.[62]

3.1.2 AI to generate insights

Research participants agreed that using AI to draw out trends that might otherwise be missed and are too complex for human analysis would be of value to all-source intelligence analysts. This is particularly the case as comparative, quantitative and trend-related data is valued by SDMs.[63] There is also a demand from decision-makers for future-facing exploratory assessment and predictive models, particularly as private sector capabilities in this area grow.[64] AI could be used to support forward-looking work that is outside the usual scope of analyst work using forecasting methods or predictive analysis.[65] Additionally, AI could be used to estimate the accuracy of past key judgements.[66] AI could also be used to support creative and critical thinking by acting as an alternative form of challenge, or to red-team assessments and offer alternative viewpoints.[67] Similarly, AI could be used to red-team other AI models' outputs to produce competing insights and provide an additional layer of rigour.[68]

In the near future, highly complex 'black box' AI systems will be available which use non-human interpretable modelling techniques and process volumes of data far beyond the capacity of manual analysis. The research has concluded that the use of such complex, non-human interpretable AI systems as the *sole basis* for strategic decision-making would pose significant challenges to the principles of analytical rigour, source validation and transparency in decision-making. Such complex 'black box' systems may be valuable earlier in the intelligence pipeline, but there was an expectation among research participants that such outputs would need to be corroborated by human-interpretable reporting if used to inform high-stakes national security strategic decision-making.[69] This expectation may change over time as familiarity with AI increases.

[62] Author interview with government participant, 18 August 2023.
[63] Author interview with government participant, 23 August 2023.
[64] Author interview with government participant, 18 August 2023.
[65] Author interview with government participant, 23 August 2023.
[66] Author interview with government participant, 18 August 2023.
[67] Author interview with government participant, 18 August 2023.
[68] Author interview with government participant, 18 August 2023.
[69] CETaS workshop, 24 January 2024.

3.2 Assurance

Intelligence assessment outputs are bolstered by rigorous assurance processes for source validation, challenge functions, and quality control processes. At the intelligence analysis level, this particularly involves considerations around the sourcing and derivation of material. This process is based on complex, human-based sense-checking and professional judgement exercised by trained individuals.[70] Intelligence analysts must ensure that assessments are as objective and accurate as possible.[71]

Intelligence is one input to all-source assessment; assessment practitioners must sift through and review all available sources of insight to tackle a defined analytical question. The judgements made in an assessment report must consider any limitations of the evidence base using the standardised lexicon of the Probability Yardstick and the AnCR Framework to convey the degree of uncertainty associated with said judgements.[72]

Assessment products must be accessible to time poor SDMs who trust existing assurance processes and are expected to take analytical confidence ratings in the final product at face value. One research participant suggested that "if a report requires any skill and interpretation on the part of a reader, it has gone wrong."[73]

Increased use of AI in analytical processes may a) require an evolution in the way assessments are communicated, and b) demand some degree of basic technical interpretation skills from SDMs. During the Covid-19 pandemic, the Scientific Advisory Group for Emergencies (SAGE) was activated to provide scientific advice to decision-makers, often based on epidemiological modelling.[74] The Covid-19 Inquiry heard that former Prime Minister Boris Johnson struggled to understand key terms, statistics, and data visualisation.[75] According to Patrick Vallance, Chief Scientific Advisor to the UK Government during the pandemic, scientific advisors in Europe had also complained of a lack of scientific

[70] Author interview with government participant, 11 August 2023; Author interview with government participant, 18 August 2023.
[71] Author interview with government participant, 11 August 2023.
[72] "Professional Development Framework for all-source intelligence assessment," PHIA, Cabinet Office, 2019, https://www.gov.uk/government/publications/intelligence-analysis-professional-development-framework/professional-development-framework-for-all-source-intelligence-assessment-html.
[73] Author interview with government participant, 11 August 2023.
[74] "About SAGE and COVID-19," Government Office for Science, 2022, https://www.gov.uk/government/publications/about-sage-and-covid-19/about-sage-and-covid-19.
[75] "'Bamboozled' Boris Johnson struggled to understand COVID-19 stats, UK inquiry hears," Politico, 2023, https://www.politico.eu/article/bamboozled-boris-johnson-struggled-to-understand-covid-19-stats-uk-inquiry-hears/.

understanding among European leaders.[76] This emphasises the need for some degree of technical upskilling to enable senior decision-makers to make effective load-bearing judgements on the basis of statistical or mathematically-derived information. The communication of uncertainty must therefore adapt to incorporate new sources of information and data inputs in a simple, standardised manner.

During the TTX, the research team tested which factors increased SDMs' confidence in AI-enriched intelligence insights.[77] AI-enriched intelligence insights were subject to much greater scrutiny than is typical for other sources of intelligence. A minority of TTX participants requested additional technical detail on elements such as the historic use of the system, the technical evaluation of the models, and how the models were trained. Participants were universally uncomfortable with the inherent uncertainty of *non-interpretable* models and outputs, and requested further interpretable verification and corroboration. Participants also suggested that additional context from open source, closed source and secret intelligence would be valuable to corroborate or provide collateral for any AI-enriched intelligence insights.

TTX participants generally had greater confidence in the ability of AI to identify events and occurrences than the ability to determine causality. AI-enriched intelligence was therefore viewed as useful for triggering investigation and determining the direction for further information gathering, but alone did not meet the threshold for taking high-stakes action. Ultimately, SDMs were unwilling to treat AI as in the same way as other, established sources of insight and sought more assurance than is usual to feel comfortable in making decisions based on AI-enriched intelligence.

As AI continues to become more widely used in day-to-day life, the level of assurance sought by SDMs may naturally reduce.

3.3 When to communicate AI-enriched intelligence

The necessity of explicitly communicating the use of AI to SDMs will vary based on context, and the degree to which AI-enriched intelligence influenced the judgements and conclusions in the final assessment product. In certain cases (e.g. source corroboration),

[76] "'Bamboozled' Boris Johnson struggled to understand COVID-19 stats, UK inquiry hears," Politico, 2023, https://www.politico.eu/article/bamboozled-boris-johnson-struggled-to-understand-covid-19-stats-uk-inquiry-hears/.
[77] CETaS workshop, 24 January 2024.

the role of AI may be so peripheral that explicitly communicating its use could complicate the reporting process and overload SDMs with unnecessary information. Providing detailed information about certain tools could inadvertently lead to decision-makers giving said tools more weight and discarding other sources of information.[78] In other cases, AI-supported intelligence insights could be a crucial factor in reaching the conclusions and judgements presented in an assessment product – and any inaccuracies in the AI output may render the assessment invalid. Wider societal perceptions of AI are also an important consideration. If scepticism of AI exists across the broader policy community, more detail on model limitations and assurance processes may be required to overcome general anxieties regarding AI.

Formal guidance for the assessment community is needed, to determine the threshold for explicitly communicating the use of AI-enriched intelligence to SDMs. Research participants generally agreed that in most cases final products issued to decision-makers may not need detail beyond stating that AI has been used in the process. This is because the analyst producing the assessment product remains responsible for evaluating relevant technical metrics (e.g. accuracy and error rates) in the underlying AI methods, and taking any limitations and uncertainty into account when producing their conclusions and judgements.

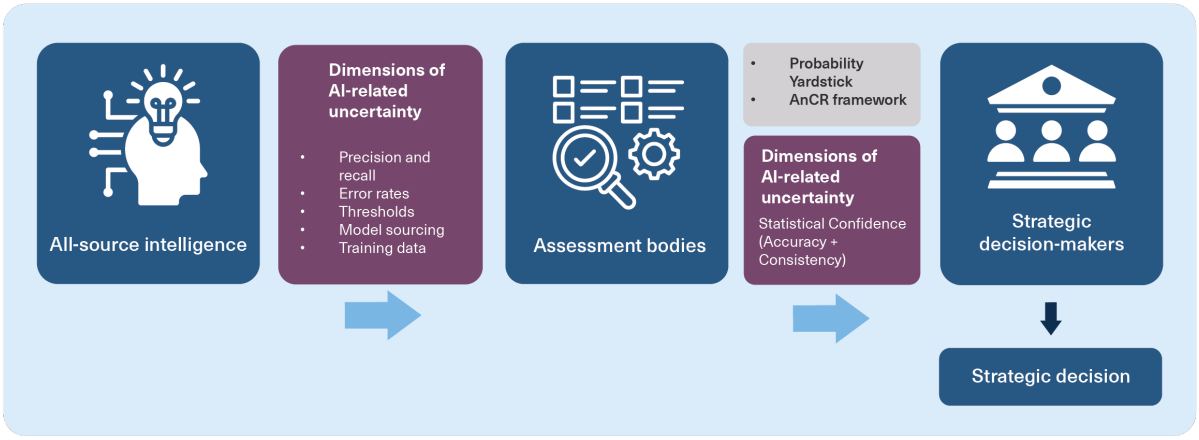
Figure 5 outlines a core concept identified during the research: **dimensionality reduction**. This concept relates specifically to communicating uncertainty in AI-enriched intelligence to SDMs. At each stage in the intelligence cycle, the number of dimensions of technical complexity reduces – as metrics for communicating technical limitations and sourcing information are simplified and eventually combined into one single dimension for communicating overall uncertainty to SDMs. This aligns with current practices, as all-source assessment communicates multiple dimensions of uncertainty in the sourcing and content of intelligence (using the Probability Yardstick and AnCR frameworks).

Intelligence analysts should expect to have access to several metrics specifying technical uncertainty (e.g. error rates, or precision and recall at different classification thresholds) and sourcing information (e.g. origin of training data, model sourcing and provenance) about the AI model. These metrics should be simplified into two dimensions for communicating uncertainty: the model's accuracy and the model's historic consistency. Analysts should then use the accuracy and consistency dimensions to generate one final 'statistical confidence' rating that conveys the level of overall uncertainty relating to the AI model's

[78] Author interview with government participant, 21 August 2023.

outputs. This rating would in turn contribute to the selection of Probability Yardstick terms and AnCR statement content.

Figure 5: CETaS analysis



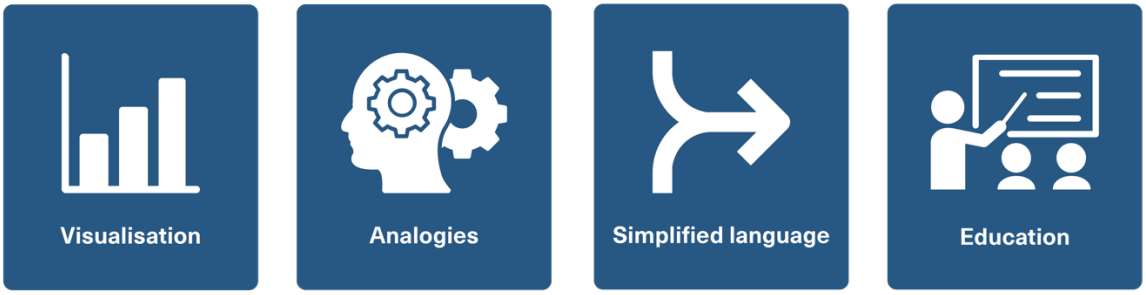
4. How to Communicate AI-enriched Intelligence to Strategic Decision-Makers

This section outlines recommendations and identified best practice for communicating uncertainty in AI-enriched intelligence to SDMs. These recommendations include practices for increasing the accessibility of technical detail for both the assessments and SDM communities, as well as functions for education, governance, and oversight.

4.1 Balancing accessibility and technical detail

Decision-makers need to understand the limitations of AI-enriched insights without being overwhelmed by too much technical complexity. Previous CETaS research has proposed the following techniques for increasing the accessibility of technical detail related to AI models and outputs:[79]

Figure 6: CETaS analysis



Across the primary research, three additional practices emerged as useful for balancing technical detail and accessibility in the context of intelligence and strategic decision-making:

(i) New guidance for communicating AI-enriched uncertainty;

(ii) A layered approach to communicating technical detail to decision-makers; and

(iii) Timely access to technical expertise.

[79] Anna Knack, Richard Carter and Alexander Babuta, "Human-Machine Teaming in Intelligence Analysis: Requirements for developing trust in machine learning systems," *CETaS Research Reports* (December 2022).

The topic of training and upskilling will be addressed separately in Section 4.2.

4.1.1 Guidance for communicating AI-enriched uncertainty

New guidance is required for communicating uncertainty within AI-enriched intelligence into all-source assessment. This guidance should establish a standard lexicon to clearly and concisely communicate confidence levels in the overall performance of AI models, as well as the inherent uncertainty in model outputs. This guidance will need to be reviewed and updated periodically as the use of AI increases and decision-makers become more familiar and comfortable with its use.

Guidance should also be provided on the threshold at which assessments should communicate the use of AI-enriched intelligence to SDMs. It should make clear that communicating the use of AI-enriched intelligence to SDMs is a context-specific requirement. Every single use case of AI in the intelligence cycle does not necessarily need to be labelled.[80]

Any new guidance must complement and not duplicate existing professional standards. Guidance relating to all-source assessment should be developed and updated by the PHIA. Cross-organisational consistency is key to create a common understanding of AI-related risks, particularly if data is shared between organisations.[81]

4.1.2 Layered approach to communicating technical detail

The main aim of the TTX was to assess the level of technical detail required in intelligence reporting for SDMs to trust AI-enriched intelligence outputs when making high-stakes decisions. The level of participants' technical expertise varied widely. Some demanded a much higher level of technical detail regarding the system, while others were less confident in interpreting technical information. Participants with technical expertise led the conversation, meaning those with less technical knowledge were excluded from parts of the discussion. One participant stated:

"I know so little about AI, I just didn't feel confident enough to make a decision."

Across all levels of technical expertise in the room, participants required a high level of assurance relating to the model's performance and integrity to feel comfortable in making

[80] Author interview with government participant, 11 August 2023.

[81] Author interview with government participant, 11 August 2023.

decisions based on AI-enriched intelligence. This demonstrates the need for a layered approach to communicating AI-enriched intelligence insights. Any assessment in a final intelligence product delivered to SDMs should always remain interpretable to non-technical audiences. However, additional technical information regarding system performance and limitations should be available on request to provide further assurance to those with more technical expertise. This information could take the form of technical annexes to assessment reports. A layered approach would help to ensure all SDMs feel comfortable in interpreting the caveats and confidence ratings associated with AI-enriched intelligence, and the conclusions from any model assurance and testing processes.

4.1.3 Access to technical expertise

Access to technical expertise throughout the intelligence cycle should lend confidence to AI systems and AI-enriched insights.[82] Technical experts who can assess and evaluate a model and its outputs during the intelligence production and analysis processes will be vital to intelligence analysts and SDMs alike. During the TTX, one participant stated they would be unable to make a policy decision based on AI-enriched intelligence reporting "without prior expert discussion and assurance about the model used to deliver the [...] verdict".

The presence of a technical subject matter expert in the room during the TTX was seen as essential to answer SDMs' questions and clarify technical details. Participants also expressed a desire for expert briefings to be provided in advance of decision-making sessions. Acknowledging the many demands on the time of national security SDMs, short, optional briefings on the limitations of models and their outputs should be coordinated immediately ahead of high-stakes decision-making sessions. These briefings should draw on the network of Government Chief Scientific Advisers and Scientific Advisory Councils. The need for briefings should be continuously assessed; as SDMs become more comfortable with consuming AI-enriched intelligence, the level of desired assurance may reduce and briefings may eventually become unnecessary.

[82] Author interview with government participant, 21 August 2023.

4.2 Training, governance, and oversight

4.2.1 Training and guidance

There is a requirement to increase AI literacy across the assessment and SDM communities, as intelligence consumers need to understand how to factor AI-related uncertainty into their decision-making.[83] Workshops, seminars, and training can be effective in improving individuals' understanding of AI and its limitations. Such sessions can also allow users to interact with the technology directly, increasing their comfort and confidence levels.

As AI becomes increasingly used as an additional source of intelligence insights, all-source intelligence analysts will need training on how to interact with models as well as how to interpret, challenge, and evaluate AI-enriched intelligence.[84] Analysts should be given the opportunity to experiment with models in simulation environments to learn and determine where the use of AI might be most useful.[85] A Training Needs Analysis should be conducted to determine the exact requirement for training new and existing analysts across different organisations.

To build confidence and trust in AI-enriched intelligence reporting, SDMs, their staff, and other consumers of intelligence assessments should be offered introductory briefings on the fundamentals of AI and corresponding assurance processes. Where possible, these recommendations should look to build on and enhance existing practices and initiatives.

4.2.2 Governance and oversight

Research participants questioned how the assessment community could build credibility in AI-enriched intelligence reporting, when models may not have a track record of dependable outputs (for example, a certain model may only be suited for deployment in one very specific context). Several TTX participants agreed that a formal assurance scheme to approve models and their outputs would be useful, with one participant stating: "A better understanding of the models, or at least, authoritative statements on the potential strengths and limitations of particular models [are] essential."

[83] Author interview with government participant, 11 August 2023.

[84] CETaS workshop, 10 November 2023.

[85] Author interview with government participant, 11 August 2023.

Two new mechanisms may provide the high level of assurance required for SDMs to make high-stakes decisions based on load-bearing AI-enriched intelligence:

1. **A formal accreditation programme** for AI systems used in intelligence analysis and assessment, to provide a baseline level of assurance that AI systems have met minimum policy requirements of robustness, security, transparency, and a record of inherent bias and mitigation. Despite existing detailed policies for the use of AI within the UKIC, there are no formally agreed minimum technical standards or accreditation processes for AI systems used within the UK Government. This programme will require dedicated resourcing, bringing together understanding of intelligence assessment standards and processes with technical expertise.

2. **Devolved technical assurance functions** within intelligence and assessment bodies across government and intelligence agencies to evaluate and approve the application of an AI system to a specific problem.

5. Conclusion and Recommendations

This study has reinforced existing research that AI is a valuable tool for the intelligence analysis and assessment community. AI could improve productivity and efficiency both as a support function and to generate new insights beyond the capabilities of human analysts. Choosing *not* to make use of available AI tools risks missing key patterns across increasing volumes of data, thereby contravening the guiding principle of comprehensive coverage, and potentially undermining the authority and value of all-source intelligence assessments to SDMs.

However, the use of AI in intelligence analysis and assessment is not without risk. AI could exacerbate existing risks such as bias and uncertainty, and make it more challenging for intelligence analysts to evaluate and communicate the limitations of AI-enriched intelligence. The risks of using AI in intelligence analysis and assessment must be weighed up against a) risks inherent to all intelligence analysis work, and b) the perceived additional benefits of using AI. In addition, there is a critical need for careful design, continuous monitoring, and regular adjustment of AI systems to mitigate the risk of amplifying human biases and errors in intelligence assessment.

Guidance is needed to ensure intelligence analysts can effectively communicate the limitations of AI-enriched intelligence to SDMs in a way that upholds the levels of rigour, transparency, and reliability demanded by intelligence assessment standards. The intelligence analyst producing the assessment product remains ultimately responsible for evaluating relevant technical metrics in the underlying AI model, and taking any limitations and uncertainty into account when producing their conclusions and judgements.

Further upskilling across the assessment and SDM community will help to establish a baseline level of technical understanding of AI models and their limitations. Finally, standardised assurance processes for AI systems are also required to build credibility and trust in assessments informed by AI-enriched intelligence.

It is beyond the scope of this unclassified report to discuss the level of maturity of AI use within the assessment community. However, the research has concluded that the work summarised above should commence now – to ensure the assessment and SDM communities are prepared for any future integration of AI capabilities within the intelligence cycle.

This report recommends the following actions to embed and promote best practice when communicating AI-enriched intelligence to strategic decision-makers:

1. The PHIA develop guidance for **communicating uncertainty within AI-enriched intelligence into all-source assessment.** This guidance should outline standardised terminology to be used if articulating AI-related limitations and caveats to decision-makers. Guidance should also be provided on the threshold at which assessments should communicate the use of AI-enriched intelligence to SDMs.

2. **A layered approach should be taken by the assessment community when presenting technical information to strategic decision-makers.** Assessments in a final intelligence product presented to decision-makers should always remain interpretable to non-technical audiences. However, additional information on system performance and limitations should be available on request for those with more technical expertise.

3. The UK Intelligence Assessment Academy should complete a **Training Needs Analysis on behalf of the all-source assessment community** to identify the requirement for training for new and existing analysts. The Academy should work with all-source assessment organisations to develop appropriate training in response to the Analysis.

4. **Training should be offered to national security decision-makers** (and their staff) to build their trust in assessments informed by AI-enriched intelligence**.** Decision-makers should be given basic briefings on the fundamentals of AI and corresponding assurance processes.

5. **Short, optional expert briefings should be offered immediately prior to high-stakes national security decision-making sessions** where AI-enriched intelligence underpins load-bearing decisions. These sessions should brief decision-makers on key technical details and limitations, and ensure they are given advanced opportunity to consider confidence ratings. These briefings should be jointly coordinated by the JIO and National Security Secretariat and should draw from cross-governmental expertise from the network of Chief Scientific Advisers and relevant Scientific Advisory Councils. Guidance on when to offer briefings should be produced, and the need for briefings should be continuously assessed; as decision-makers become more comfortable with consuming AI-enriched intelligence, the level of desired assurance may reduce, and briefings may eventually become unnecessary.

6. **A formal accreditation programme should be developed for AI systems used in intelligence analysis and assessment** to ensure models meet minimum policy requirements of robustness, security, transparency, and a record of inherent bias and

mitigation. Technical assurance for the application of a model to a specific problem should be devolved to relevant organisations, and **each organisation's assurance process should be accredited**. This programme will require dedicated resourcing, bringing together understanding of intelligence assessment standards and processes with technical expertise. PHIA should assist in developing principles and requirements, while technical expertise for accreditation and testing should be drawn from technical authorities in the intelligence community and across government.

About the Authors

Megan Hughes is a Research Associate at the Centre for Emerging Technology and Security (CETaS). Her research explores the impact of AI on intelligence tradecraft and the information environment. Prior to joining the Turing, Megan worked as an Analyst within the Defence and Security research group at RAND Europe. Her research has informed strategy and policy at the UK Home Office, UK Ministry of Defence, the European Commission, and the United Nations Development Programme.

Dr Richard J. Carter is a Senior Visiting Fellow and Strategy Advisor at CETaS. He is a computer scientist and strategic advisor to government and industry on emerging technologies and strategic change, and the CEO and Founder of UK-based human-machine teaming company Tulpa Ltd. Rich has been part of the UK's Defence & Security sector for most of his 25-year career but with 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 Newcastle. Rich is also a Fellow of the Royal Society of Arts, a Fellow of the British Computer Society, and a Chartered IT Professional. He has 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.

Amy Harland is a Senior Visiting Fellow at CETaS, and a senior diplomat with over 20 years' national security and geopolitical experience. She is currently focused on Digital Transformation for the UK Foreign Office (areas of particular interest are AI adoption, and digital upskilling in government). She has worked around the world including in the Middle East, East Africa, and most recently in Israel, where she served as Political Counsellor in the British Embassy in Tel Aviv.

Dr Alexander Babuta is Director of CETaS, and Director of National Security and Policy at The Alan Turing Institute. He previously worked within the UK Government as AI Futures Lead at the Centre for Data Ethics and Innovation, and before this as Research Fellow at the Royal United Services Institute (RUSI). He is Chair of the Essex Police Data Ethics Committee, Honorary Lecturer at University College London (UCL), Research Associate at the National Centre for Gang Research (University of West London), and was previously Associate Fellow at the University of Bristol. He holds a Doctorate in Policing, Crime and Security from the University of West London, an MSc with Distinction in Crime Science and a first class Bachelor's degree in Linguistics, both from UCL.

Centre for Emerging Technology and Security

RESEARCH REPORT