The Alan Turing Institute

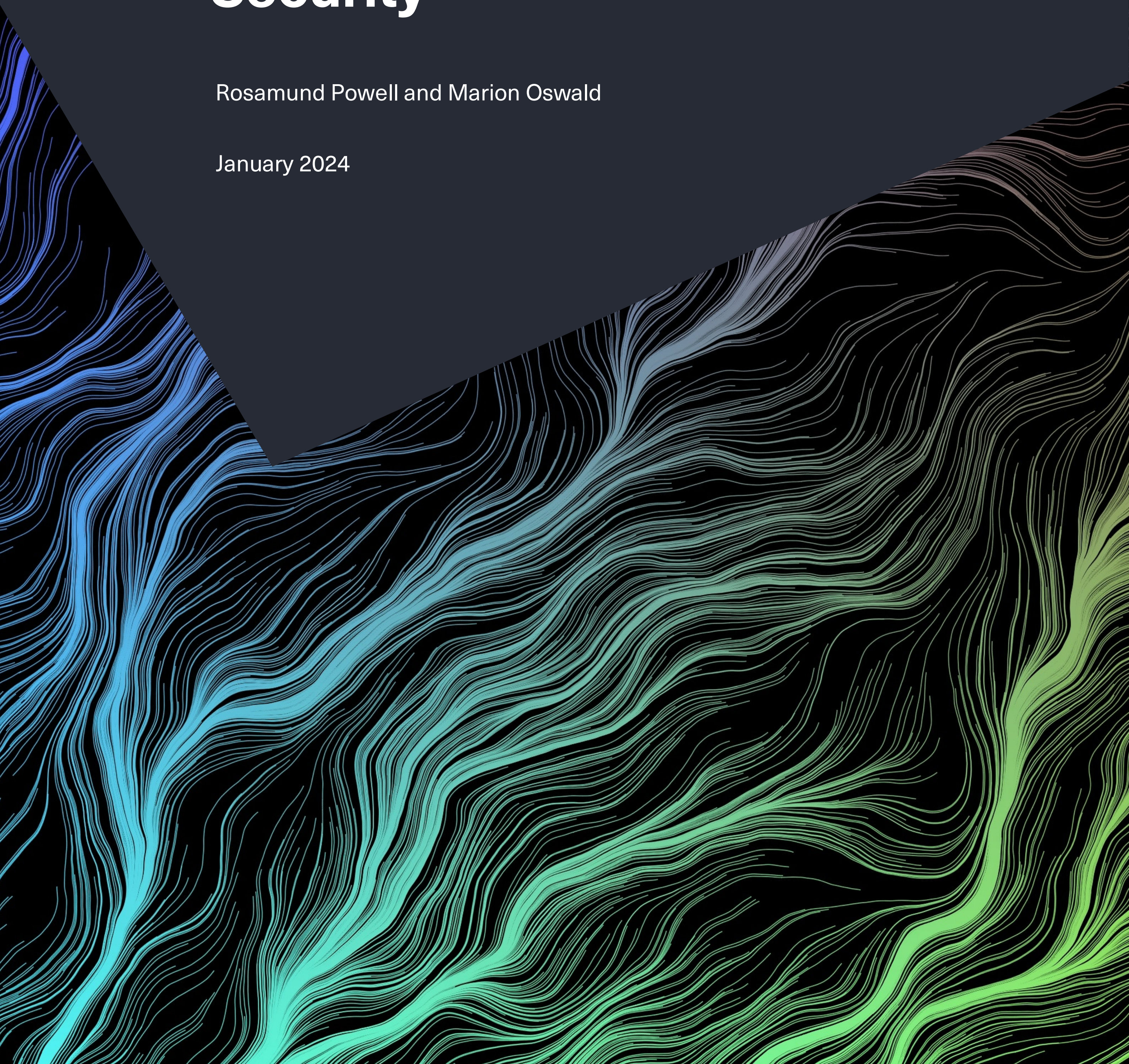
Centre for Emerging Technology and Security

RESEARCH REPORT

Assurance of Third-Party AI Systems for UK National Security

Rosamund Powell and Marion Oswald

January 2024



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 or workshop for this project, without whom the research would not have been possible. The authors are also very grateful to Ian Brown, Elisabeth Mackay and Amy Harland for their valuable feedback on an earlier version of this report. Design for this report was led by Michelle Wronski.

Executive Summary

This CETaS Research Report explores the potential for assurance processes to aid in the responsible adoption of AI across UK national security, with a particular focus on addressing the challenges which arise when industry suppliers contribute to the AI lifecycle.

Involving industry in the design and development of AI capabilities for UK national security brings many benefits, including access to cutting-edge capabilities and the potential to save government time and money. But there are also risks, such as opaque supply chains, insufficient ethical due diligence, and a lack of robust testing for AI security.

While these risks apply to any government use of third-party AI, they are amplified in the national security context where tolerance for error is low, and systems must meet a higher threshold of robustness, security, and compliance. And, while much research has focused on AI assurance in the public sector – including in defence – no concrete proposal has yet been made to account for the specific requirements of UK national security.

We identify three cross-cutting governance challenges which are preventing national security bodies from determining *which* third-party AI systems to deploy. These are:

1. **Disparate access to information and skills,** with government organisations often lagging behind industry in their understanding of the third-party technologies they plan to deploy.
2. **Divergent business models and motivations,** with stronger incentives needed to improve transparency from suppliers on the features of their AI systems.
3. **Distributed responsibility for the introduction of safeguards,** with clearer consensus needed on who should conduct which aspects of the assurance process.

Our report sets out how national security bodies and industry suppliers can tackle these challenges using a tailored framework for AI assurance. Throughout this paper, AI assurance will be defined as:

The portfolio of processes required to evaluate and communicate, iteratively throughout the AI lifecycle, the extent to which a given AI system:

a) Does everything it says it is going to do, and nothing it shouldn’t do.

b) Complies with the values of the deploying organisation.

c) Is appropriate to the specific use case and envisioned deployment context.

Our framework addresses these assurance components through four core pillars:

1. **Robust documentation** protocols in the form of a sector-specific system card template. When complete, this system card constitutes the AI assurance case – the central document of compiled evidence that an AI system meets requirements.
2. **Companion guidance** to clarify what constitutes sufficiently robust evidence to include in the system card. This includes the recommendation for national security bodies to curate a modular portfolio of assurance techniques (e.g. international standards, impact assessments, performance metrics, red teaming protocols) that have been approved for use in the high stakes context of national security.
3. **Investment in skills for evidence review** to enable national security decisionmakers to make thorough assessments of system cards.
4. **Contractual protections** to mandate further transparency from suppliers of third-party AI, where relevant.

While the bulk of this report focuses on detailing this assurance framework, we close by making recommendations for its implementation – both in the near and long term.

In the immediate term, we recommend both industry suppliers and national security bodies trial the framework on specific AI use cases, in place of current model cards, to assess its applicability to a range of AI use cases and identify ways in which the assurance requirements set out here may be adapted to fit the specific risk profile of AI use cases.

In the longer term we recommend national security bodies take the following actions to support implementation of this assurance framework:

Recommendations for implementing AI assurance
Build infrastructure for a sustainable assurance ecosystem, including further investments in platforms to host assurance cases *and* the curation of companion guidance, including a tailored national security portfolio of assurance techniques.
Invest in skills for reviewing assurance cases (technical, ethical, and legal). We recommend government centres of AI expertise (e.g. the AI Safety Institute and Centre for Data Ethics and Innovation) support national security departments in AI assurance.
Connect academic work on assurance to practitioner challenges to increase the availability of practicable assurance techniques that fill persistent gaps e.g. on AI security or data provenance.
Develop exemplar assurance cases across a range of case studies to further specify how recommendations apply in context (such as for LLMs in intelligence analysis or autonomous agents for cyber defence).
Draft bespoke contractual clauses to aid national security customers in ensuring suppliers are transparent about the properties of their AI systems.

Introduction

Designing, developing, and deploying an artificial intelligence (AI) system[1] for UK national security already presents a challenge for policymakers, who must ensure adequate scrutiny of the system at each stage of the AI lifecycle[2] to avoid unintended outcomes in high stakes contexts. This challenge becomes even harder when some of these stages instead occur within third-party organisations, potentially constraining government oversight of aspects of the development process.

Nevertheless, involving third parties, particularly industry, is increasingly essential if national security bodies are to take full advantage of the powerful AI systems now available. Rapidly accelerating capabilities in the private sector, alongside skills shortages within government, mean industry suppliers can in some circumstances be the only option available to national security decisionmakers wishing to deploy the most advanced AI systems.

In this report, we provide guidance on how to assess specific third-party AI systems against suitability criteria for national security deployment. We propose a step-by-step AI assurance framework which guides policymakers and industry suppliers through these decisions.

Prior to the deployment of a third-party AI system, the national security customer must address a three-part challenge. They must:

1. Establish a robust understanding of *what* properties they want the AI system to possess, and *how* these properties might be guaranteed (e.g. through testing, evidence sharing, and contracting).
2. Develop a strategy to maximise evidence available to them (e.g. through industry collaboration).
3. Design clear protocols for assessing evidence, such that risk is minimised and ongoing checks are in place (e.g. through investments in staff expertise and sufficient people resourcing).

The AI assurance framework for UK national security presented here accounts for the three-part challenge described above. This framework centres on a proposal for a tailored system

[1] Since the term was coined in 1955, the parameters of what constitutes 'artificial intelligence' have often been vaguely drawn. For the purposes of this study, our focus is primarily on machine learning technologies, defined throughout as technologies which use patterns in data to make predictions and thus improve performance over time. Please see Appendix 2 'Glossary of key terms' for a full definition.

[2] To include design, development and deployment of AI systems and to incorporate both technical and sociotechnical processes which occur in the AI lifecycle.

card template for UK national security, with the system card serving as the 'assurance case' once populated – the central document containing all relevant evidence that an AI system meets requirements. The report also recommends longer-term actions from UK national security policymakers to help foster responsible AI innovation and thus overcome persistent assurance challenges.

This report is directed at industry suppliers as well as national security bodies. Without industry contributions, the challenges of assuring third-party AI become significantly harder. Suppliers are often in a unique position with control over the project lifecycle, access to commercially sensitive information about the AI system, and significant bargaining power during contractual negotiations. Furthermore, industry bodies also face challenges when it comes to AI assurance. Currently, they lack sufficiently specific guidance on government requirements, leading to uncertainties as to which safeguards should be incorporated into their project lifecycles and communicated to government customers.

Co-creation of the 'AI assurance case' by supplier and customer is presented as the ideal method to facilitate robust assessment. Nevertheless, this framework is also adaptable for circumstances where national security bodies have no relationship with the supplier and consequently must collect and assess all assurance evidence themselves, as is illustrated here through reference to hypothetical case studies.

Research methodology and limitations

This report addresses the following research questions:

- **RQ1:** What benefits and risks come from the deployment of third-party AI systems?
- **RQ2:** What are the trade-offs regarding using third-party systems versus in-house development?
- **RQ3:** What guidance can help UK national security decisionmakers to interpret and interrogate third-party AI systems to ensure due diligence?
- **RQ4:** What should be included in assurance guidance for suppliers developing AI capabilities for use in UK national security?

Data collection for this study was conducted over a four-month period from June – September 2023, including three core research activities:

1. **Literature review** covering academic and policy literature on topics such as responsible development practices, AI supply chains, AI security, AI assurance, and AI procurement.

2. **Semi-structured interviews** with 28 participants from government, industry and academia.
3. **Research workshop** attended by 11 industry representatives with expertise in AI assurance in the national security sector.

The focus of this study was broad, covering the assurance of third-party AI systems across the whole UK national security landscape, with a particular focus on industry partnerships. It is beyond the scope of this report to lay out in depth recommendations for dealing with specific technologies (e.g. biometrics, LLMs, computer vision) or for sectors outside of national security. Further work is needed to investigate the applicability of this framework to real-world use cases and to develop the technical aspects of AI assurance laid out here, in particular regarding AI security. Further work is also needed to examine the extent to which national security bodies should favour third-party or in-house design and development of AI systems in general.

Our recommendations also do not address specific legal frameworks such as the information acquisition and disclosure requirements in the Security Service Act 1989 and Intelligence Services Act 1984, the warrantry and authorisation requirements, data safeguards and notices regime in the Investigatory Powers Act 2016, the Data Protection Act 2018, and the proportionality test in human rights law. However, legality and compliance with warrantry and authorisation conditions will be key aims that assurance can address, thus informing the properties that the AI system must possess. For example, assurance processes can assist in obtaining the dataset and output information needed to determine levels of intrusiveness in relation to a proportionality assessment.[3]

Structure of this report

The remainder of this report is structured as follows. Section 1 outlines the third-party AI landscape, exploring the risks and benefits these technologies bring across the national security sector. Section 2 focuses on the need to address these risks on a case-by-case basis, introducing AI assurance to do this. Section 3 forms the substantive analysis where we present a framework for AI assurance in the national security context. This assurance framework is further specified in section 4 through a discussion of its implementation in the context of hypothetical case studies. Finally, section 5 summarises core recommendations.

[3] Ardi Janjeva, Muffy Calder and Marion Oswald, "Privacy Intrusion and National Security in the Age of AI: Assessing proportionality of automated analysis," *CETaS Research Report* (May 2023).

Appendix 1 presents a compiled system card template for documenting AI system properties while Appendix 2 offers definitions of key terms used throughout this report.

1. Understanding Third-Party AI Systems: Origins, Benefits, and Risks

1.1 Defining third-party AI

This report focuses on procurement and deployment decisions surrounding specific AI systems. Nevertheless, it is first necessary to define third-party AI systems, and understand the benefits and risks they bring to national security in general. Third-party AI systems come in many forms, with third parties contributing at any stage of the AI lifecycle, ranging from data collection and annotation to model training and validation. Increasingly, the modularity of AI systems means there are often multiple actors working together as part of an algorithmic supply chain, each contributing to distinct aspects of the system's functionality.[4]

Throughout this report, third-party AI systems are defined as any AI system where at least one stage of the AI lifecycle (design, development, deployment) occurs partially or wholly outside of the organisation that will deploy the system.

Three factors can be used to roughly map this landscape of third-party AI:

A. **The type and number of third parties involved:** This could include academic institutions, private companies (start-ups, multinational technology companies, defence primes), public sector bodies (including another national security agency), or some combination of these.

B. **The nature of the third-party relationship:** This could include AI systems designed in partnership with companies where formal relationships are established but can also include AI systems made commercially available by multinational tech companies. Even where the relationship with the 'prime contractor' is strong there may be other firms contributing down the supply chain.

C. **The extent of third-party involvement:** Third party suppliers may become involved in any one stage of the AI lifecycle or may have full control over every stage, subsequently impacting how much control the national security body has over each AI lifecycle stage.

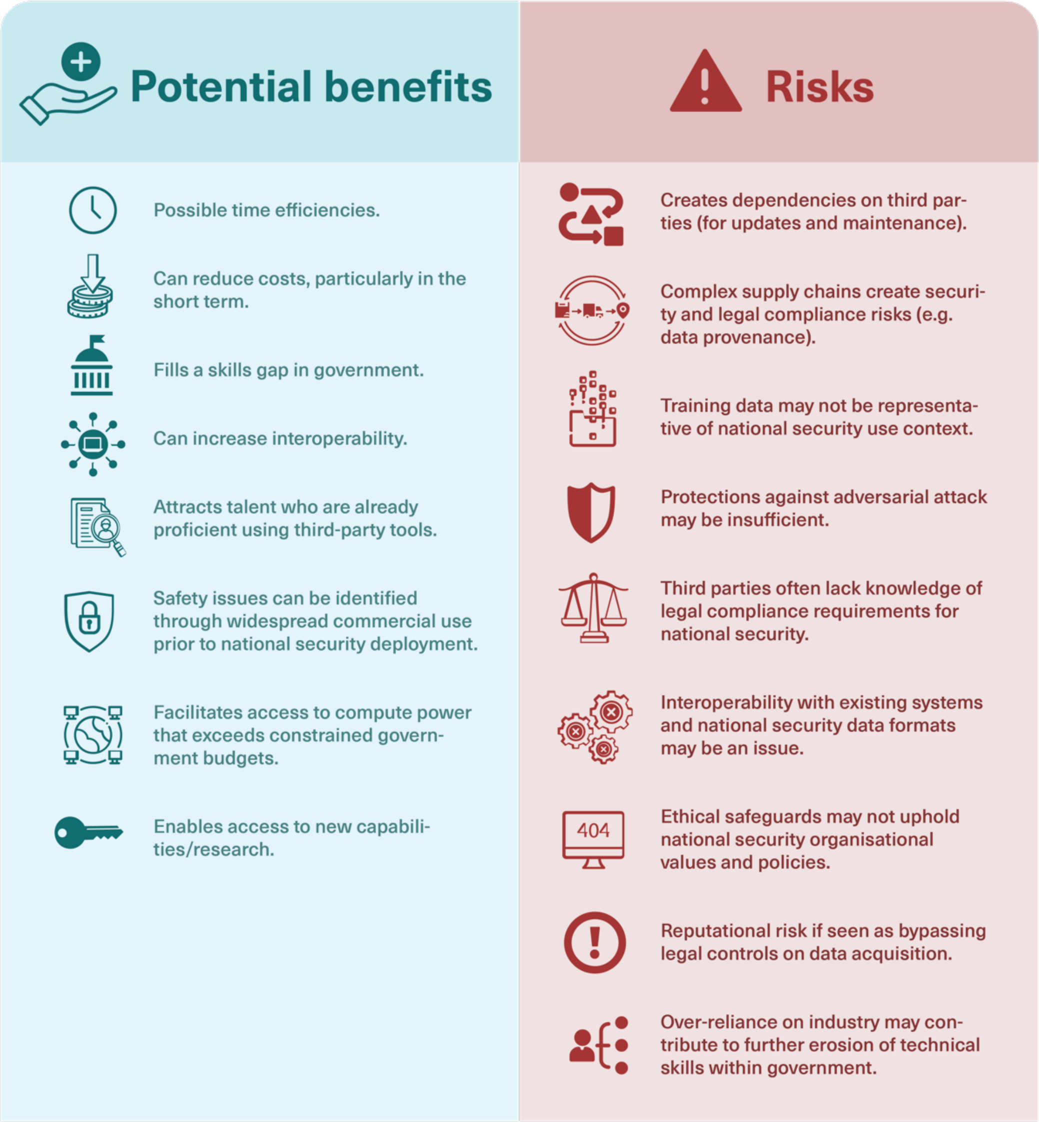
[4] Jennifer Cobbe, Michael Veale and Jatinder Singh, "Understanding accountability in algorithmic supply chains," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2023), 1186-1197.

Section 4 offers guidance on how the assurance framework set out here might be applied across this varied third-party AI landscape, using three hypothetical case studies to structure discussion.

1.2 Cross-cutting risks and benefits

Each AI product raises distinct concerns for national security decision-makers. Nevertheless, several benefits, risks and governance challenges recur across a range of third-party AI systems (Figure 1).[5]

Figure 1: Benefits, risks and governance challenges associated with third-party AI



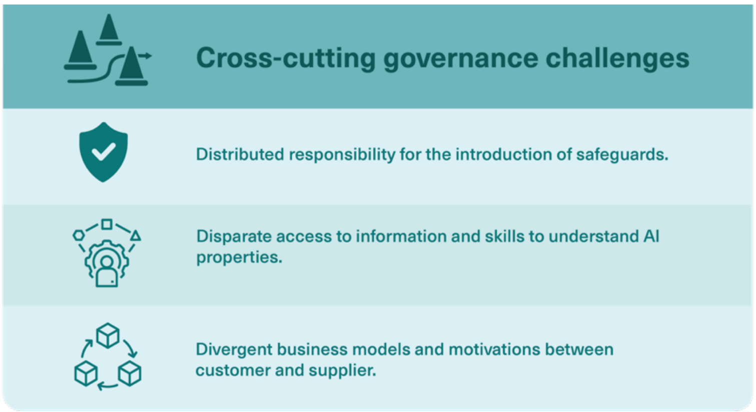
Potential benefits

- Possible time efficiencies.
- Can reduce costs, particularly in the short term.
- Fills a skills gap in government.
- Can increase interoperability.
- Attracts talent who are already proficient using third-party tools.
- Safety issues can be identified through widespread commercial use prior to national security deployment.
- Facilitates access to compute power that exceeds constrained government budgets.
- Enables access to new capabilities/research.

Risks

- Creates dependencies on third parties (for updates and maintenance).
- Complex supply chains create security and legal compliance risks (e.g. data provenance).
- Training data may not be representative of national security use context.
- Protections against adversarial attack may be insufficient.
- Third parties often lack knowledge of legal compliance requirements for national security.
- Interoperability with existing systems and national security data formats may be an issue.
- Ethical safeguards may not uphold national security organisational values and policies.
- Reputational risk if seen as bypassing legal controls on data acquisition.
- Over-reliance on industry may contribute to further erosion of technical skills within government.

[5] Benefits and risks of third-party AI included in figure 1 were identified during research interviews and literature review.



Cross-cutting governance challenges

Distributed responsibility for the introduction of safeguards.

Disparate access to information and skills to understand AI properties.

Divergent business models and motivations between customer and supplier.

1.3 AI supply chain risks

Complex supply chains present one of the most common and challenging risks of using third-party AI in the national security context as 'their complexity makes it difficult to guarantee their security',[6] while further concerns exist around legal compliance and ethical practice all the way down a supply chain.[7] The machine learning lifecycle of systems with complex supply chains is highly sociotechnical,[8] meaning technical, legal, and ethical supply chain risks become intertwined. For example, data provenance has been described as the 'biggest issue' for third-party AI,[9] due to technical, policy and compliance concerns. On the technical side, 'there is a risk of poisoned data, bias, the data misbehaving, risk of attacks,'[10] while on the policy side concerns around copyright and intellectual property are preventing suppliers from being fully transparent.[11] Beyond data provenance, instability in the supply chain with regard to compute and hardware sourcing can present a particular risk for load bearing AI systems, while national security decision-makers also expressed concerns over model provenance, in particular in contexts where they do not have oversight over who else may be using an AI system.[12] As noted by one interviewee, verifying the big data supply chain for compliance and security is not easy, but 'that is the cost of doing things in defence and security'.[13]

[6] Nii Simmonds and Alice Lynch, "Mitigating supply chain threats: building resilience through AI-enabled early warning systems," *CETaS Expert Analysis* (January 2023).

[7] Interview with industry expert, 4 August 2023.

[8] Jennifer Cobbe, Michael Veale and Jatinder Singh, "Understanding accountability in algorithmic supply chains," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2023), 1186-1197.

[9] Interview with government representative, 5 July 2023.

[10] Interview with government representative, 19 July 2023; for more detailed analysis of these technical risks see section 3.2.4 'performance and security'.

[11] Interview with government representative, 5 July 2023.

[12] Interview with government representative, 5 July 2023.

[13] Interview with industry expert, 4 August 2023.

2. Why Assurance?

2.1 What is AI assurance?

The term 'AI assurance' is used in a variety of ways in the UK and internationally, contributing to confusion among experts.[14] During our engagements, consensus emerged on what assurance does and does not involve. Assurance does not involve eliminating risk from AI systems,[15] setting rigid rules for developers,[16] or quantitatively ranking AI systems against one another.[17] Assurance is more nuanced than this and must at times be adaptable to the needs of stakeholders with divergent priorities. For instance, for some interviewees issues of AI security and performance were seen as most central to assurance,[18] while for others the core focus was legal compliance and ethical due diligence.[19]

We define 'AI assurance' as the portfolio of processes required to evaluate and communicate, iteratively throughout the AI lifecycle, the extent to which a given AI system:

a) Does everything it says it is going to do, and nothing it shouldn't do.
b) Complies with the values of the deploying organisation and upholds established ethical principles.
c) Is legally compliant and appropriate to the specific deployment context.

2.2 Challenges to effective AI assurance

Much progress has been made by other parts of the public sector towards successful AI assurance. The UK's Centre for Data Ethics and Innovation (CDEI) has spearheaded this work,[20] while Dstl has outlined how assurance might be implemented in the defence

[14] Across interviews, the scope of assurance varied, both regarding the *processes* involved and the *properties* seen as important to assure for. A government representative cited international variation as a contributing factor, with the US in particular favouring the term 'risk management' over 'assurance'. Interview with government expert, 21 June 2023.
[15] Interview with academic expert, 21 June 2023.
[16] Interview with government representative, 21 June 2023.
[17] Interview with academic expert, 6 July 2023.
[18] Interview with government representative, 5 July 2023.
[19] Interview (2) with law enforcement member of staff, 6 July 2023.
[20] HM Government, *CDEI portfolio of AI assurance techniques* (Centre for Data Ethics and Innovation: 2023), https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques; HM Government, *The roadmap to an effective AI assurance ecosystem* (Centre for Data Ethics and Innovation: 2021), https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem.

context.[21] Despite progress, the below challenges, illustrate why further work is needed for AI assurance to be deployed effectively in the UK national security context:

1. **Existing frameworks have not specifically addressed national security needs:** Existing work on AI assurance (e.g. by CDEI) doesn't adequately address factors such as protection against adversarial attack.[22]
2. **Crowded landscape:** Techniques for trustworthy AI are proliferating. Without structured ways to choose between all the standards, impact assessments and performance metrics on offer, developers and policymakers are left confused and overwhelmed.[23]
3. **Separation of technical vs ethical assessment and a lack of intersecting skills:** Currently, AI assurance tools tend to be *either* technical *or* ethical*.* Ethical and technical assessments need to occur in tandem. This requires a multidisciplinary team.[24]
4. **Accommodating start-ups:** Assurance entails a resourcing requirement, which is likely to favour larger companies.[25] If entry costs are too high, and start-ups are left behind, there is potential for stifled innovation and competition.[26]
5. **Convoluted, theoretical frameworks:** Practitioners expressed frustration at assurance frameworks which fail to specify requirements in terms they understood.[27] One participant claimed that too much investment had been placed in academic work, resulting in assurance frameworks that are 'very confusing for most developers.'[28]
6. **Added bureaucracy:** Procurement is already slow, and assurance could slow it down further. Additional safeguards are needed but should be balanced with efficiency.[29]
7. **Divergent business models hamper communication:** Industry suppliers are often reluctant to communicate transparently, for example due to concerns around trade secrecy and commercial IP. This can limit evidence available to government as part of an AI assurance case.[30]
8. **Complex supply chains are poorly understood:** Existing assurance frameworks struggle to account for disparate information access across complex supply chains.[31] In

[21] HM Government, *Assurance of Artificial Intelligence and Autonomous Systems: A Dstl Biscuit Book* (Dstl: 2021), https://www.gov.uk/government/publications/assurance-of-ai-and-autonomous-systems-a-dstl-biscuit-book.

[22] Interview with government representative, 5 July 2023.

[23] Interview with government representative, 21 June 2023.

[24] Interview with government representative, 21 June 203.

[25] Interview with industry expert, 21 July 2023.

[26] Interview with industry expert, 4 August 2023.

[27] Interview with industry experts, 28 July 2023.

[28] Interview with academic expert, 26 July 2023.

[29] Interview with academic expert, 4 July 2023.

[30] Interview with government representative, 21 June 2023.

[31] Interview with academic expert, 6 July 2023.

addition, increasingly agile development cycles mean assurance must be able to account for 'continuous testing and iterative revision after deployment'.[32]

9. **Risks false sense of security**: The success of AI assurance is ultimately limited by the capability and diligence of the people assessing assurance cases. It can easily become a rubber-stamping exercise, and lead to a false sense of security.[33] In the national security context, this has even turned decision-makers away from the term 'assurance' as it can give false confidence if residual risk is not appropriately communicated.[34]

[32] Jennifer Cobbe, Michael Veale and Jatinder Singh, "Understanding accountability in algorithmic supply chains," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2023), 1186-1197.

[33] Interview with academic expert, 4 July 2023.

[34] Interview with government representative, 22 June 2023.

3. A Model of AI Assurance for UK National Security

Our framework for AI assurance addresses the above challenges. It consists of two core stages, each with two constituent pillars (Figure 2).

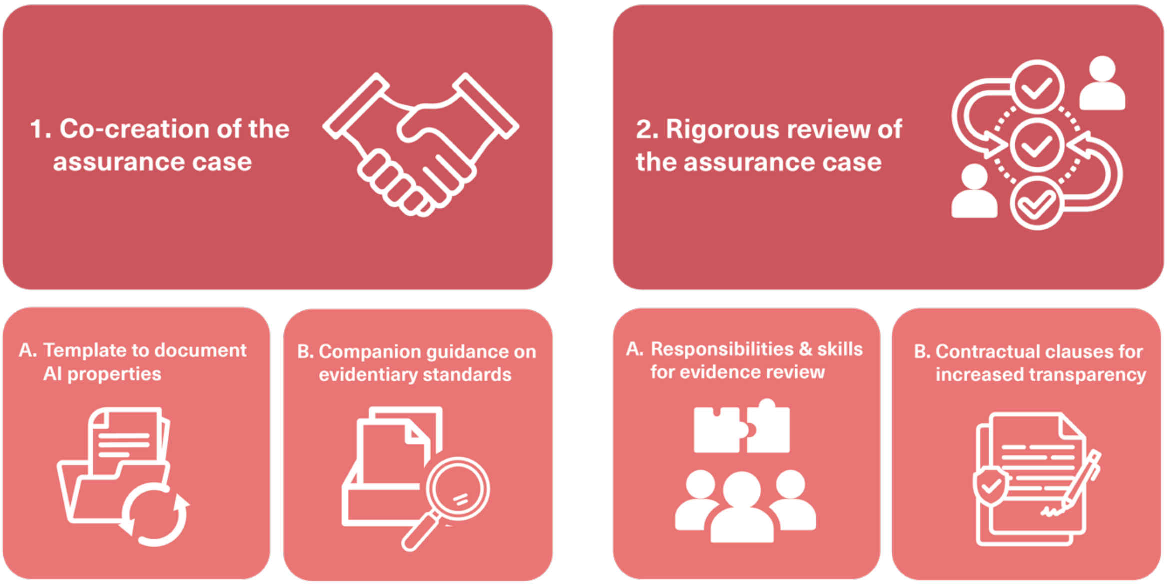
First, the assurance case must be created – ideally through cooperation between suppliers and national security bodies. To support this process, we propose:

a) A template for documenting AI system properties.
b) The creation of companion guidance to support those filling out the template.

Second, the assurance case must be reviewed to assess whether evidence is sufficient. To support this process, we propose:

a) Clarity on responsibilities for evidence review and investment in internal skills.
b) Contractual clauses to mandate transparent sharing of evidence with reviewers.

Figure 2: Two stage model for AI assurance



1. Co-creation of the assurance case
2. Rigorous review of the assurance case
A. Template to document AI properties
B. Companion guidance on evidentiary standards
A. Responsibilities & skills for evidence review
B. Contractual clauses for increased transparency

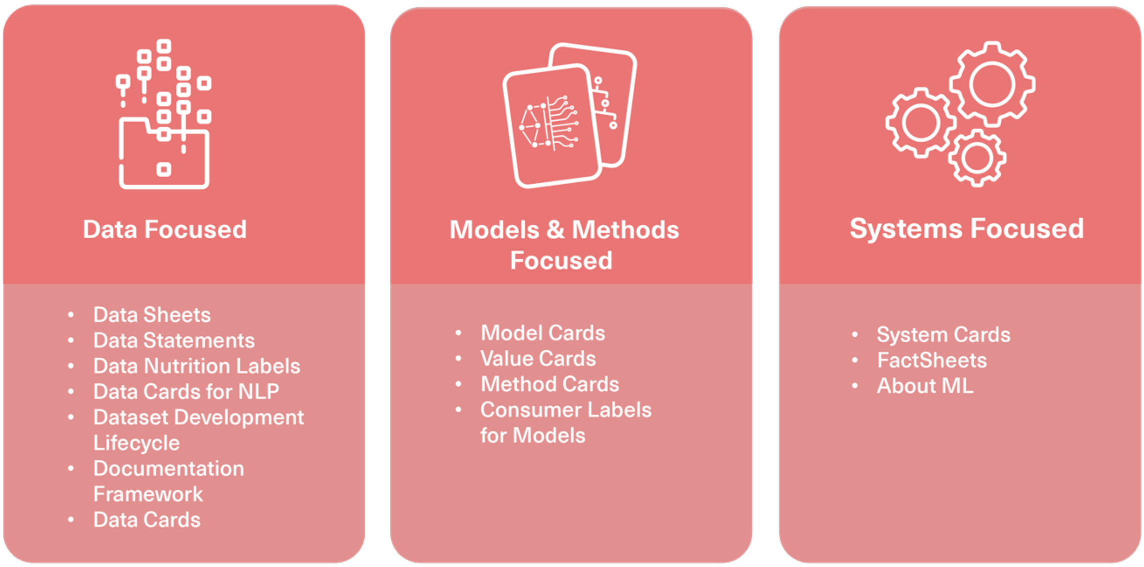
3.1 Co-creation of the assurance case

3.1.1 Existing methods for documenting AI system properties

The assurance case is the central document containing all relevant evidence that an AI system meets requirements, structured into a logical argument supporting some end goal or collection of desired properties.[35] The choice of method for documenting these desired AI system properties is a crucial component of the assurance framework as it sets the standard on *which* properties are included.[36] Selecting a documentation method also presents a particular challenge for third-party systems as evidence is typically generated by multiple actors, necessitating co-creation of the assurance case.

Numerous proposals have been made for how to document AI system properties (as illustrated by Figure 3).[37] We focus on three methods which have shown significant promise, comparing their strengths and weaknesses.[38] Our proposal will build on the strengths of each.

Figure 3: Options for documenting AI properties as illustrated by Hugging Face



Data Focused

- Data Sheets
- Data Statements
- Data Nutrition Labels
- Data Cards for NLP
- Dataset Development Lifecycle
- Documentation Framework
- Data Cards

Models & Methods Focused

- Model Cards
- Value Cards
- Method Cards
- Consumer Labels for Models

Systems Focused

- System Cards
- FactSheets
- About ML

[35] HM Government, *Assurance of Artificial Intelligence and Autonomous Systems: A Dstl Biscuit Book* (Dstl: 2021), https://www.gov.uk/government/publications/assurance-of-ai-and-autonomous-systems-a-dstl-biscuit-book.

[36] Mona Sloane et al., *AI and procurement: a primer* (New York University: Summer 2021), https://archive.nyu.edu/handle/2451/62255.

[37] Hugging Face, "Model Card Guidebook," https://huggingface.co/docs/hub/model-card-guidebook.

[38] In selecting these methods for comparison, we acknowledge the existence of further documentation methods such as data sheets and explainability factsheets.

1. Model cards:

Table 1: Strengths, weaknesses and examples of model cards

Description
Model cards are defined as files 'that provide information about [a model's] purpose and details about its provenance, the data used for training, any known limitations and bias,' or simply as 'files that accompany the models and provide handy information'.[39] They were proposed by Mitchell et al. (2018) to increase transparency through accessible information sharing.[40]
Examples
Hugging Face:[41] Hugging Face model cards are widely adopted across the private sector. Their template is shared publicly and is designed to be filled in with descriptions of the model, its intended uses, limitations, biases and ethical considerations, the training parameters and experimental information, which datasets were used, and evaluation results. These model cards require input from developers, sociotechnical experts, and project organisers.
Bailo:[42] A system introduced by GCHQ to ensure model cards are uploaded to a central repository for easy review. For each model card, two stages of review are required (a. technical assessment, b. policy assessment). The aim is to manage the AI project lifecycle and enable compliance with organisational requirements.
Algorithmic transparency recording standard:[43] While not designed as a model card or to be incorporated into an approvals process, the algorithmic transparency recording standard (developed by CDDO and CDEI) offers many useful insights on the sorts of properties that must be included in a comprehensive overview of a model. It is aimed at

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

[40] Margaret Mitchell et al., "Model cards for model reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220-229.

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

[42] GCHQ/Bailo, "Bailo – managing the lifecycle of machine learning to support scalability, impact, collaboration, compliance and sharing," GitHub, https://github.com/gchq/Bailo.

[43] HM Government, *Algorithmic Transparency Recording Standard Hub* (CDDO and CDEI: January 2023), https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub.

<table>
<tr><td colspan="2">public sector bodies rather than industry, requiring them to input ‘clear information about the algorithmic tools they use, and why they’re using them’. It requires public sector bodies to provide information including a system overview, contact details for the responsible team, details on how the tool will be used, mechanisms for review, details on the datasets used and more. So far, it has been piloted across a range of public sector bodies from healthcare to policing and the cabinet office – with updates made in response to user feedback.</td></tr>
<tr><th>Strengths</th><th>Weaknesses</th></tr>
<tr><td>• Succinct, with the potential to act as ‘boundary objects, a single artefact that is accessible to users who have different backgrounds and goals when interacting with model cards.’[44]
• Ease of completion by developers.
• Existing uptake from developers indicates familiarity with this approach,[45] and Bailo indicates similar familiarity among the national security community.[46]
• Existing versions promote breaking down performance criteria into results for individual demographic, cultural or domain relevant conditions.[47]</td><td>• Frequently tailored for interpretation ‘by individuals with AI or NLP expertise’, offering insufficient context for non-experts.[52]
• Focus on ‘model’ can be oversimplistic given that safeguards are often built into the broader system, for instance covering the front-end graphical user interface as well as the model behind it.[53]
• The integrity of the model card is highly reliant ‘on the integrity of the creator(s)’,[54] and there is not enforcement of transparency from developers associated with this documentation method.
• Typically, there is no distinction between the space for ‘claims’ versus</td></tr>
</table>

[44] Hugging Face, “Model Card Guidebook,” https://huggingface.co/docs/hub/model-card-guidebook.

[45] Hugging Face, “User Studies,” https://huggingface.co/docs/hub/model-cards-user-studies.

[46] GCHQ/Bailo, “Bailo – managing the lifecycle of machine learning to support scalability, impact, collaboration, compliance and sharing,” GitHub, https://github.com/gchq/Bailo.

[47] Margaret Mitchell et al., “Model cards for model reporting,” in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220-229.

[52] Anamaria Crisan et al., “Interactive Model Cards: A Human-Centred Approach to Model Documentation,” in *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2022), 427-439.

[53] Interview with industry expert 04/08/23.

[54] Margaret Mitchell et al., “Model cards for model reporting,” in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220-229.

• Includes room for ethical considerations, most often highlighting the importance of fairness.[48] • High degree of standardisation allows for ease of comparison between cards.[49] • Adaptations can be introduced to allow for context-specific factors to be included on a model card.[50] • Many completed versions are publicly available, offering a starting point for developers.[51]	'evidence', making them subjective.[55] Suppliers tend to pitch their product rather than lay out limitations, meaning clear distinctions are needed.[56] • Often lack interactivity,[57] for example offering means to further question what is written on the model card.[58] • May not capture complex supply chains or 'chaining' of multiple machine learning models in sequence. • Software packages can automate the completion of model cards, 'but it takes the responsibility away from people to consider whether there is more that needs to be communicated.'[59] • Insufficient guidance is given on ethical components, 'it is hard to think about whether you have created a fair system, a sustainable one, an explainable one'.[60]

[48] Margaret Mitchell et al., "Model cards for model reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220-229.

[49] Ibid.

[50] Evidently AI, "A simple way to create ML model cards in Python," Evidently AI Tutorials, 15 June 2023, https://www.evidentlyai.com/blog/ml-model-card-tutorial.

[51] See HM Government, *Collection: Algorithmic Transparency Reports* (CDDO and CDEI: 2023), https://www.gov.uk/government/collections/algorithmic-transparency-reports.

[55] Interview with government representative, 5 July 2023.

[56] Christopher Burr and Rosamund Powell, *Trustworthy Assurance of Digital Mental Healthcare* (Alan Turing Institute: 2022), https://zenodo.org/records/7107200.

[57] Interview with government representative, 19 July 2023.

[58] Anamaria Crisan et al., "Interactive Model Cards: A Human-Centred Approach to Model Documentation," in *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2022), 427-439.

[59] Interview with government representative, 8 June 2023.

[60] Interview with government representative, 8 June 2023.

2. System cards:

Table 2: Strengths, weaknesses and examples of system cards

Description	
Some have argued in favour of a shift from model cards to system cards.[61] Currently there is little consensus on what should comprise a system card, with the only common feature being that instead of documenting a single model, system cards aim to document the features of all the models and other components which make up the final AI system.[62] Meta, for instance, propose the system card approach is crucial to document how components interact in complex AI systems,[63] applying their system card methodology to their Instagram Feed technology.[64]	
Example	
GPT-4 System card:[65] OpenAI's system card for GPT-4 aims to detail the testing and safeguards put in place to address safety challenges. It spans model and system-level interventions, discussing adversarial testing, red teaming, and expert consultations. It is a long, free-form document (as compared to model cards, e.g. Hugging Face).	
Strengths	**Weaknesses**
• Uptake from big tech shows a willingness to adopt this approach.[66]	• Examples so far (see: ChatGPT) offer insufficient structure. This makes it easy to be convinced by the evidence

[61] Meta, "System cards, a new resource for understanding how AI systems work," Meta Blog, 23 February 2022, https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/; Interview with industry expert, 4 August 2023.

[62] Meta, "System cards, a new resource for understanding how AI systems work," Meta Blog, 23 February 2022, https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/; Interview with industry expert 4 August 2023; OpenAI, "GPT-4 System Card," 23 March 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[63] Chaves Procope et al., "System-level transparency of machine learning," Meta Research, 22 February 2022, https://ai.meta.com/research/publications/system-level-transparency-of-machine-learning/.

[64] Meta, "What is the Instagram Feed?," Meta Tools, 23 February 2022, https://ai.meta.com/tools/system-cards/instagram-feed-ranking/.

[65] OpenAI, "GPT-4 System Card," 23 March 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[66] Meta, "System cards, a new resource for understanding how AI systems work," Meta Blog, 23 February 2022, https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/; Interview with industry expert, 4 August 2023; OpenAI, "GPT-4 System Card," 23 March 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

• Accounts for features beyond the model, to include safeguards which can be introduced at different points in the AI lifecycle and/or as part of the final interface.[67] This is particularly useful in a context where 'software development now often involves, to various degrees, integrating pre-built modular components provided as services and controlled by others into a complete product: not simply a system, but a system-of-systems.'[68]	that is there, but difficult to see what is missing, especially for non-technical audiences.[69] • As with model cards, there can be insufficient distinction between evidence and claims.[70] • Current examples focus on technical rather than sociotechnical assessments (e.g. on properties such as fairness and explainability), and little attention is paid to the importance of legal compliance.[71]

[67] Interview with industry expert, 4 August 2023.

[68] Jennifer Cobbe, Michael Veale and Jatinder Singh, "Understanding accountability in algorithmic supply chains," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2023) 1186-1197.

[69] Interview with industry expert, 4 August 2023.

[70] OpenAI, "GPT-4 System Card," 23 March 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[71] Ibid.

3. Argument-based assurance:

Table 3: Strengths, weaknesses and examples of argument-based assurance

<table>
<tr><th colspan="2">Description</th></tr>
<tr><td colspan="2">A process of using 'structured argumentation to provide assurance to another party (or parties) that a particular claim (or set of related claims) about a property of a system is warranted given the available evidence.'[72] This has been widely deployed in safety-critical domains to assure features of complex engineering systems, and since expanded to software and AI.[73] Most recently, it has been expanded to address AI ethics.[74]</td></tr>
<tr><th colspan="2">Example</th></tr>
<tr><td colspan="2"><b>Argument pattern for explainability:</b>[75] The conceptual gap between ethical AI principles, for example 'fairness' or 'explainability', and concrete evidence is large. Argument-based assurance uses structured flowcharts to break these broad goals down into sub-goals which are then each supported by multiple pieces of evidence. For a single goal, such as explainability, a highly complex flowchart is needed to fully justify and communicate how each piece of evidence comes together to support the stated goal. Due to the complexity of these assurance cases, it would be infeasible to expect developers to start from scratch for each new AI system they wish to assure. It has instead been suggested that argument patterns for common goals like explainability should be developed to offer a starting point for developers, who then simply need to adapt them for the specific AI use case they have in mind.</td></tr>
<tr><th>Strengths</th><th>Weaknesses</th></tr>
</table>

[72] Christopher Burr and David Leslie, "Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies," *AI and Ethics* 3 (2023): 73-98.

[73] John McDermid, Yan Jia and Ibrahim Habli, "Towards a Framework for Safety Assurance of Autonomous Systems," in *Proceedings of the Workshop on Artificial Intelligence Safety 2019* (CEUR: 2019), 1-7.

[74] Christopher Burr and David Leslie, "Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies," *AI and Ethics* 3 (2023): 73-98; Christopher Burr and Rosamund Powell, *Trustworthy Assurance of Digital Mental Healthcare* (Alan Turing Institute: 2022), https://zenodo.org/records/7107200; Zoe Porter, Ibrahim Habli and John McDermid, "A principle-based ethical assurance argument for AI and Autonomous systems," *Arxiv* (March 2022).

[75] Christopher Burr and Rosamund Powell, *Trustworthy Assurance of Digital Mental Healthcare* (Alan Turing Institute: 2022), https://zenodo.org/records/7107200.

• The structure of arguments places more pressure on suppliers to back up claims with evidence.[76] • Offers options to involve impacted groups in determining what should be included as a top-level goal for an AI assurance argument.[77] • Provides clarity on how evidence relates to claims about system properties, in particular in more complex scenarios where multiple pieces of evidence back up a single claim.[78] • Argument patterns can be repurposed for multiple AI systems, allowing exemplars to be adapted for future AI systems.[79]	• Can be complex and onerous to complete, and challenging to understand for those who need to interpret assurance cases.[80] Limited uptake by suppliers will be a key limitation for this approach, as will the lack of sufficient internal skills and resources to review complex assurance cases. • Suggested concepts (e.g. fairness, trustworthiness) can be uncertain and difficult to evidence due to both subjectivity and complexity.[81] • Would need to be adapted to reflect different priorities and principles which are relevant in a national security context (e.g. the specific definition of proportionality used in this context).[82] • The structure of each assurance case is bespoke, to the point where comparing AI systems becomes challenging.

Each of these three documentation methods offers key insights into best practice for UK national security, in particular highlighting the need to:

1. Balance accessible, concise documentation with detail to aid interpretation by non-experts.
2. Ensure documentation structure is consistent and facilitates easy comparison and identification of gaps by reviewers.

[76] Christopher Burr and Rosamund Powell, *Trustworthy Assurance of Digital Mental Healthcare* (Alan Turing Institute: 2022), https://zenodo.org/records/7107200.

[77] Christopher Burr and Rosamund Powell, *Trustworthy Assurance of Digital Mental Healthcare* (Alan Turing Institute: 2022), https://zenodo.org/records/7107200.

[78] Christopher Burr and David Leslie, 'Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies,' *AI and Ethics* 3 (2023): 73-98.

[79] Interview with academic expert, 8 June 2023.

[80] Interview with academic expert, 8 June 2023.

[81] Zoe Porter, Ibrahim Habli and John McDermid, "A principle-based ethical assurance argument for AI and Autonomous systems," *Arxiv* (March 2022).

[82] Interview with academic expert, 8 June 2023.

3. Accommodate flexibility to adapt AI assurance to emerging technologies and specific use contexts, while establishing consensus on properties which must always be documented.
4. Implement a framework which builds on industry practices so that uptake across the sector is maximised, while also putting sufficient pressure on industry to increase transparency.
5. Clarify how assurance builds on related processes (e.g. legal compliance and procurement).

We take forward the strengths of each documentation method in our proposal for a tailored system card template for the UK national security context, with our system card incorporating features from each of the above examples (see *Section 3.2*).

3.1.2 Companion guidance on evidentiary standards

Documentation methods such as model and system cards would not go far towards solving the challenges associated with third-party AI if they were simply filled in with descriptive claims. This is arguably the most significant limitation of model cards in their current form. As noted by one interviewee, ‘it is possible for people to just write down an opinion on the model card. It is very subjective still’.[83] This is especially problematic when subjective input is communicated across multiple organisations where underlying assumptions differ.[84]

Consequently, in proposing a tailored system card template for national security, we must be more stringent about the sorts of evidence that may be used to support claims set out within it, providing companion guidance to users of the system card on how to fill it out. Table 4 below, reproduced from CDEI’s work on AI assurance,[85] summarises some of the key assurance techniques which may be used to generate 'evidentiary artefacts' to support claims made within an AI system card.

[83] Interview with government representative, 5 July 2023.
[84] Christopher Burr and Rosamund Powell, *Trustworthy Assurance of Digital Mental Healthcare* (Alan Turing Institute: 2022), https://zenodo.org/records/7107200.
[85] CDEI, “Techniques for assuring AI systems,” https://cdeiuk.github.io/ai-assurance-guide/techniques.

Table 4: Examples of assurance techniques (reproduced from CDEI AI assurance guide)

Assurance technique	Description
Impact assessment	Used to anticipate the effect of a system on environmental, equality, human rights, data protection, or other outcomes.
Risk assessment	Similar to impact assessments but are conducted after a system has been implemented in a retrospective manner.
Bias audit	Assessing the inputs and outputs of algorithmic systems to determine if there is unfair bias in the input data, the outcome of a decision or classification made by the system.
Compliance audit	A review of a company's adherence to internal policies and procedures, or external regulations or legal requirements. Specialised types of compliance audit include system and process audits and regulatory inspection.
Certification	A process where an independent body attests that a product, service, organisation or individual has been tested against, and met, objective standards of quality or performance.
Conformity assessment	Provides assurance that a product, service or system being supplied meets the expectations specified or claimed, prior to it entering the market. Conformity assessment includes activities such as testing, inspection and certification.
Performance testing	Used to assess the performance of a system with respect to predetermined quantitative requirements or benchmarks.
Formal verification	Establishes whether a system satisfies some requirements using the formal methods of mathematics.

However, it is rarely easy to choose exactly which assurance technique should be applied and how they should be combined, especially given that the above list is not exhaustive. Additional techniques such as red teaming and international AI standards may be used as evidence in the system card. And, for each category, there will be many specific examples available (e.g. AI impact assessments focus on overlapping but varied priorities – for instance human rights, fairness, data protection, security, safety, privacy, and sustainability).

Furthermore, national security-specific techniques will also be needed, such as structured frameworks to assess proportionality of AI systems.[86]

Two opposing approaches currently exist on how such techniques for evidence generation should be combined as part of an end-to-end assurance process:

1. **Pipeline of standardised benchmark protocols and evaluations** to embed and assess each of the features you want to document. In this scenario, there would be a single checklist of techniques to implement for every AI system to assess each property within the system card in turn.
2. **Modular 'portfolio of assurance techniques'[87]** selected in a context-specific manner depending on the AI use case. In this scenario, for each property you wish to assure for (e.g. fairness) there will be multiple options for how to evidence it with distinct assurance techniques selected depending on the use case.

Each approach comes with advantages and weaknesses. A pipeline of standardised metrics facilitates easy comparison between the different third-party AI systems on offer. This is particularly useful for technical properties such as performance where quantitative techniques are available to rank AI systems against one another.[88] This approach also allows more resources to be dedicated towards verifying whether a smaller number of evaluation metrics are truly robust, resulting in more confidence in the techniques which do get used.

However, it is harder to standardise tests for qualitative properties such as fairness and explainability. A preoccupation with standardised benchmarks could even lead to an overreliance on technical evaluation as opposed to sociotechnical and qualitative tests, something cited as a problem by numerous interviewees,[89] as these do not produce such clear-cut results in the form of scores or rankings.[90] Furthermore, a pipeline of standardised evaluations offers little flexibility for evidentiary standards to be adjusted to specific use cases. Finally, even for technical properties, standardised benchmarks risk overconfidence based on tools which are not fully interpretable, and which don't sufficiently disaggregate results for distinct tasks.[91] Overall, reliance on a pipeline of standardised tests can lead to

[86] Ardi Janjeva, Muffy Calder and Marion Oswald, "Privacy Intrusion and National Security in the Age of AI: Assessing proportionality of automated analysis," *CETaS Research Report* (May 2023).

[87] HM Government, *CDEI portfolio of AI assurance techniques* (Centre for Data Ethics and Innovation: 2023), https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques.

[88] HuggingFace, "Open LLM Leaderboard," https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

[89] Interview with government expert, 21 June 2023; Interview with academic expert, 8 June 2023; Interview with academic expert, 6 July 2023; Interview with academic expert, 21 June 2023.

[90] Interview with academic expert, 6 July 2023.

[91] Ryan Burnell et al., "Rethink reporting of evaluation results in AI," *SCIENCE* 380, no. 6641 (April 2023): 136-138.

assurance being viewed as a tick-box exercise, without sufficient room for reflection on whether these are the right tests for the specific technology under consideration.

In contrast, a modular portfolio of techniques allows national security bodies to adapt their risk appetite for high-stakes versus low-stakes use cases *and* to accept a broader range of evidence submissions for qualitative system properties where multiple methodologies may be available.[92] A modular repository of AI assessment techniques can also more easily accommodate regular revision as technologies evolve and the techniques to assess them are rendered outdated.[93]

This approach is particularly suited to third-party AI for two reasons. First, suppliers are developing their own assurance techniques. Palantir for example have released 'AI on RAILS'[94], a responsible AI lifecycle framework, while Microsoft have an 'AI Fairness Checklist'.[95] But, each organisation does assurance differently. This variation in industry approaches is further illustrated by "frontier AI" labs' published policies.[96] National security bodies need to be flexible enough to accept evidence submitted from suppliers where it is provably robust, even if the evidence is not submitted in the form the agency would have deemed ideal had they developed the AI system themselves.

Second, for each third-party AI system, the approach to AI assurance will need to be adjusted depending on which stage of the AI lifecycle the national security body is overseeing. For instance, adversarial testing may offer useful evidence of AI security if the national security body only has oversight of the deployment phase. But, if they instead have control over the design phase, they can promote security-by-design, perhaps adopting a security standard such as ISO/IEC AWI 27090 'Cybersecurity – Artificial intelligence'.[97] A repository of ex-ante and ex-poste assurance techniques gives the flexibility to test for AI system properties in different ways, depending on the specificity of the use case and the relationship with the third-party supplier.

[92] Jacqui Ayling and Adriane Chapman, "Putting AI ethics to work: are the tools fit for purpose?," *AI and ethics* 2 (2022): 405-429.

[93] Jacob Mökander et al., "Auditing large language models: A three-layered approach," *Arxiv* (June 2023).

[94] Palantir, "AI on RAILs: A responsible AI lifecycle framework," Palantir Whitepaper, 2023, https://www.palantir.com/assets/xrfr7uokpv1b/4nVc0FDbOrqeVHUZQdIcwZ/21b4e3f13479ecf87c4da4fcc0e8c1a0/RAILS_Whitepaper-FINAL-.pdf.

[95] Michael Madaio et al., "Co-designing checklists to understand organizational challenges and opportunities around fairness in AI," in *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery, 2020), 1-14.

[96] Department for Science, Innovation and Technology, "Emerging processes for frontier AI safety," October 2023, https://assets.publishing.service.gov.uk/media/653aabbd80884d000df71bdc/emerging-processes-frontier-ai-safety.pdf.

[97] ENISA, *Cybersecurity of AI and standardisation* (ENISA: March 2023).

There is a tension between the need to provide more prescriptive and practical guidance, while also showing adaptability to evidence submitted in a variety of forms. We propose a compromise between a narrow pipeline of standardised metrics and existing extensive portfolios of generalist assurance techniques.

Specifically, we propose companion guidance for AI assurance in the national security context. This guidance should include:

A. A narrow, curated repository of assurance techniques which are appropriate for the national security context. This repository should cover the full range of AI system properties laid out in the below system card. It should include specific examples of impact assessments, audit methodologies, performance metrics, red teaming protocols, and more.

B. Comprehensive guidance on choosing between techniques where multiple may be available, to help suppliers and/or national security staff to choose evidence that is most appropriate to their circumstances. This should build on CSET Georgetown's 'matrix for selecting responsible AI frameworks'.[98] For example, which audit methodologies are best suited for generative AI? Which impact assessments are best for high-stakes AI applications? Which performance metrics are appropriate for particular use domains?

C. Exemplar assurance cases for specific AI technologies to ensure recommendations are grounded in the specific and distinct challenges raised by LLMs as opposed to computer vision, or AI for intelligence analysis as opposed to AI for business operations.

This companion guidance should be shared with suppliers alongside the system card template to allow them to evidence their claims more easily. It should also be shared with internal national security staff to aid them in filling out system cards.

It is beyond the scope of this report to produce this companion guidance, and we recommend this for future work. However, we recommend policymakers look to the example set by the CDAO 'Responsible AI Toolkit' released by the US Department of Defence.[99] This interactive toolkit 'guides users through tailorable and modular assessments, tools, and artifacts throughout the AI product lifecycle' and offers guidance to 'current and future DoD industry partners'. It is a living document which will be regularly

[98] Mina Narayanan and Christian Shoeberl, *A matrix for selecting responsible AI frameworks* (CSET Georgetown: June 2023).

[99] US Department of Defence, "CDAO Releases Responsible AI (RAI) Toolkit for Ensuring Alignment With RAI Best Practices," US Department of Defence Press Release, 14 November 2023, https://www.defense.gov/News/Releases/Release/Article/3588743/cdao-releases-responsible-ai-rai-toolkit-for-ensuring-alignment-with-rai-best-p/.

updated.[100] A similar approach is needed in a national security context to support industry partners and direct them towards sufficiently thorough assurance techniques.

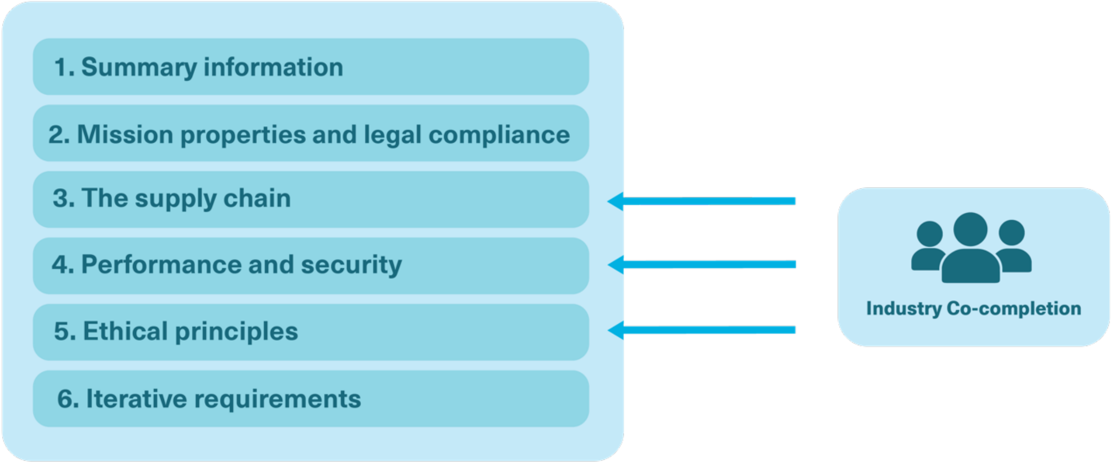
[100] Ibid.

3.2 System card template for UK national security

Drawing on the above analysis, alongside insights from research engagements, we propose a tailored approach to system cards for third-party AI systems in the national security context. The framework presented here should be viewed as a starting point, with regular iterations needed to keep pace with developments in AI.

Six sections form the core of the system card, with industry collaboration advantageous for the completion of three of these. Each section will be covered in depth below as we detail the rationale behind the structure of the system card before presenting instructions for filling it out, directed at both industry suppliers and national security customers. A compiled system card template can be found in *Appendix 1* which illustrates how these sections would be presented to users in practice.

Figure 4: System card structure


1. Summary information
2. Mission properties and legal compliance
3. The supply chain
4. Performance and security
5. Ethical principles
6. Iterative requirements
Industry Co-completion

Before detailing the specifics of this system card template, it is necessary to clarify the status of legal compliance within this framework. Often, 'ethical and legal issues are on the same document (model card)' but this raises tricky questions,[101] as ethics and compliance play distinct roles in assurance. As noted by one expert participant, 'assurance shouldn't just be about setting the ground of what is legally acceptable, but instead encourage people to go beyond this. Just because it is legal, doesn't mean it is moral'.[102] It is crucial for system cards to address both legal and ethical concerns, but their role must be carefully distinguished. Ethical due diligence must build upon the foundations of legal compliance, not the other way around. Internal review teams will need to be meticulous in ensuring legal

[101] Interview with government representative, 19 October 2023.

[102] Interview with academic expert, 8 June 2023.

compliance *and* go beyond this to actively promote ethical best practice. To enable this, legal compliance is separated from ethical due diligence in this system card. It should be borne in mind that the completed system card itself may have more general legal and compliance relevance, for example as information or evidence in a subsequent inquiry into the use of technology or in relation to compliance with authorisations or warrants in respect of data handling or analysis.

3.2.1 Summary information

To enable successful communication about AI systems, including among non-technical users, assurance cases need to be accessible to a range of stakeholders. To enable this, our system card begins with summary information before delving into more in-depth analysis of system properties. As with the algorithmic transparency recording standard, the existence of this summary information should not be taken as an excuse for the remainder of the system card to become inaccessible to non-expert audiences.[103] Nevertheless, providing summary information is important to encourage senior decision-makers who are pressed for time to engage with the assurance process. As noted by one senior decision-maker, 'what you need is headline points and the assurance that people who have deep expertise in a trusted role have had a close look.'[104] This summary information is therefore intended as a supplement rather than alternative to the detailed evidence set out below.

Table 5: System card section one

Part 1: Summary Information
Instructions
System details: Please provide AI system name, 1-2 sentence description of the system and its constituent components, version, and implementation so far.[105]
Mission objectives fulfilled and use cases across the organisation: Please summarise the positive contributions made by the system towards the organisation's goals and give an account of how 'load bearing' the AI system may be across the organisation.[106]

[103] HM Government, *Algorithmic Transparency Recording Standard Hub* (CDDO and CDEI: January 2023), https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub.
[104] Interview with government representative, 8 August 2023.
[105] Hugging Face, "Annotated Model Card Template," https://huggingface.co/docs/hub/model-card-annotated.
[106] Interview with government representative, 8 August 2023.

Internal roles and responsibilities: Detail the key internal decisionmakers responsible for filling out and reviewing this system card, including policy, legal, and technical expertise, and clear separation between the roles of filling out the system card with relevant evidence, and assessing the completed system card.
Supply chain summary: Please summarise the information given in part 3, including list of organisations/departments responsible for design, development, and deployment & at least one contact for each organisation/department.
License: If applicable, details of the licensing/procurement arrangement are to be provided here.
Summary and key take-aways: Please summarise key take-aways from the following sections (mission properties & legal compliance, performance & security, ethics). A red/amber/green scale may be used to highlight sections of concern.
Iterative review summary: Provide dates for any anticipated updates to the AI system *and* for next review and update of this system card.

3.2.2 Mission properties and legal compliance

This system card section covers internal foundational checks that must be conducted by the deploying organisation (i.e. the national security body), without industry contributions. Evidence set out here should focus a) on the positive contribution the AI system can bring to the organisation and b) the legal status of its application in its stated use context.

Table 6: System card section two

Part 2: Mission properties and legal compliance
Instructions
Context and scope of use[107] *A) Delineate clear parameters for AI system use:* Set out who in the organisation will be using this AI system, how often, and for what purpose. If the AI system in question is being repurposed by the national security body

[107] HM Government, *Algorithmic Transparency Recording Standard Hub* (CDDO and CDEI: January 2023), available at: https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub.

from the purpose for which it was designed, this should be flagged here. This section should also set out any prohibited uses that have been identified as risky.

B) Account for how the AI system will impact existing organisational processes and existing workers:

Set out the extent of integration of this system, both with existing human decision-making processes,[108] and with existing technology systems. Where relevant, this may include reference to an assessment of the impact the AI system will have on employees' working conditions, for example through a 'Good Work Algorithmic Impact Assessment'.[109]

C) Non-algorithmic options considered:[110]

Please detail why the AI system in question is preferential to the non-algorithmic options available, including a comparison to the current method for completing this task if relevant.

Legal basis

The legal basis, requirements and powers for the development and use of the AI system alongside other legal compliance requirements that the assurance process will help to support should be set out.

This section may include but is not limited to:

➔ The overarching statutory or legal functions for which the AI system is being developed.

➔ Any limitations, restrictions, or constraints on the exercise of data acquisition and/or analysis for the purposes of national security or other purposes, including those within the Investigatory Powers Act and associated warrants and authorisations.

➔ Consideration of the human rights principle of necessity and proportionality[111] in relation to the development and use of the AI system.

➔ Any requirement for the tool's output to be used evidentially or in legal proceedings.

Licensing/model acquisition

Provide details of the model/software licensing agreement (or other procurement structure such as bespoke development), including details of contractual transparency requirements and other protections. Links to contracts should be included here for

[108] HM Government, *Algorithmic Transparency Recording Standard Hub* (CDDO and CDEI: January 2023), https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub.

[109] Institute for the Future of Work, "Good Work Algorithmic Impact Assessment," IFOW Guidance (March 2023), https://www.ifow.org/publications/good-work-algorithmic-impact-assessment-an-approach-for-worker-involvement.

[110] HM Government, *Algorithmic Transparency Recording Standard Hub* (CDDO and CDEI: January 2023), https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub.

[111] Ardi Janjeva, Muffy Calder and Marion Oswald, "Privacy Intrusion and National Security in the Age of AI: Assessing proportionality of automated analysis," *CETaS Research Reports* (May 2023).

further detail, and details of the invitation to tender (ITT) process should be set out if this took place (including possible assurances which were requested in the ITT process).

3.2.3 The supply chain

System cards can help address the lack of visibility over the supply chain by increasing transparency. However, completing this section will present a challenge for government and suppliers alike. Co-completion of the system card by government and supplier is crucial to getting a full picture, ensuring system cards are not just a 'one-and-done affair' but a place where 'collaborators are working together to collectively address the problem'.[112]

Three system card sub-sections are proposed to account for AI supply chains. First, the system card template must be filled out with a mapping of the supply chain, focused on organisations and individuals who contribute in some way to the final AI system. It is essential to identify relevant contributors to the lifecycle as far as is possible, both to enable communication across multiple organisations about the AI system and to enable future responses to algorithmic harms. Ideally, at least one vetted and cleared individual will contribute to the system card for sensitive use cases to facilitate fully open discussions.[113]

Next, users of the system card must address questions of 'provenance'. This section offers space for suppliers to link to further details on data provenance and also includes considerations of model provenance and system provenance, in addition to sourcing of compute and hardware.[114] As set out by Dstl, a layered approach ensures adequate attention is paid to granular components of a system (data, hardware, compute) in addition to the final products (models, systems, even systems-of-systems).[115] Where industry suppliers are unwilling or unable to supply this information, contractual protections may be used to enforce transparency (as accounted for in section 3.3). If not, it will be up to national security customers to fill this section in as far as possible before deciding whether they can accept residual risk.

Finally, the system card gives room for additional evidence to be submitted in the form of a supply chain risk assessment. This can help provide a fuller picture where evidence gathered on provenance is deemed insufficient. And, it can account for broader supply

[112] Ian Brown, "Expert Explainer: allocating accountability in AI supply chains," Ada Lovelace Institute Paper (June 2023), https://www.adalovelaceinstitute.org/resource/ai-supply-chains/.

[113] Interview with government representative, 19 July 2023.

[114] Interview with industry expert, 21 July 2023.

[115] HM Government, *Assurance of Artificial Intelligence and Autonomous Systems: A Dstl Biscuit Book* (Dstl: 2021), https://www.gov.uk/government/publications/assurance-of-ai-and-autonomous-systems-a-dstl-biscuit-book.

chain concerns, for example issues of ethical due diligence, legal compliance, and secure practices down the supply chain.

Table 7: System card section three

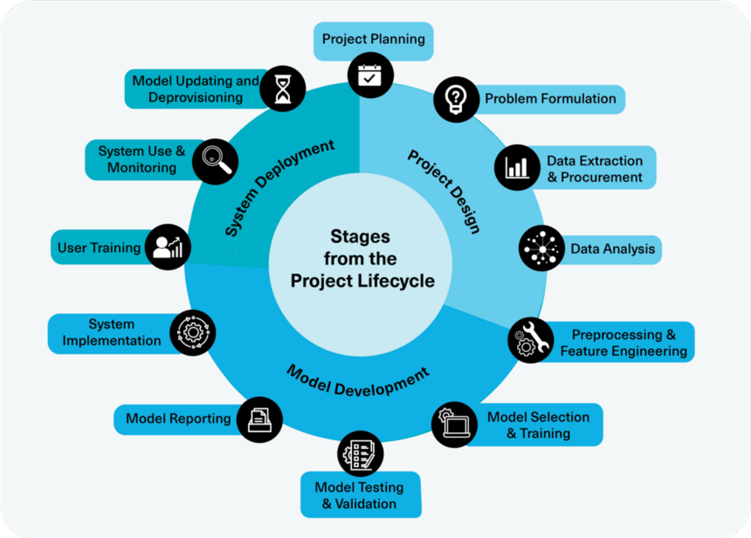
Part 3: The Supply Chain

Instructions

Supply chain mapping & industry contributors

Please identify whether the following stages of the AI lifecycle[116] were government-led or industry-led. Please also attribute each stage to a specific organisation, or for organisations over 100 people, to a specific department.

Additionally, please nominate a point of contact at each relevant organisation, or at each department at larger organisations. Their role should be described, both regarding the project lifecycle itself and the co-completion of this system card. Any vetted and cleared contributors from industry should be identified as potential collaborators on this system card.


Project Planning
Problem Formulation
Data Extraction & Procurement
Data Analysis
Preprocessing & Feature Engineering
Model Selection & Training
Model Testing & Validation
Model Reporting
System Implementation
User Training
System Use & Monitoring
Model Updating and Deprovisioning
Project Design
Model Development
System Deployment
Stages from the Project Lifecycle

Source: Model of the AI lifecycle, reproduced from Burr & Leslie, 2023.

Provenance

Provenance here is defined as the 'chronology of the ownership, custody or location of a historical object',[117] and should be accounted for with regard to:

[116] This model of the AI lifecycle was developed by The Alan Turing Institute and accounts for the highly sociotechnical nature of AI design, development and deployment. See Christopher Burr and David Leslie, "Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies," *AI and Ethics* 3 (2023): 73-98.

[117] Kiran Karkera, "Why is provenance important for AI," Kiran Karkera Medium, 10 July 2020, https://kaal-daari.medium.com/an-example-of-art-provenance-records-for-the-curious-d3a5e4a1dd77.

A) *Data:* What training data was used? Where was it sourced? Please link to full datasets if possible and provide details of any updates to datasets through the AI lifecycle. Please link to audits of relevant datasets where available (for instance through the 'data provenance initiative').[118]

B) *Hardware:* Please detail the hardware feeding into this system, including details of how it was sourced.

C) *Compute:* Please detail the source of compute for this system and how ongoing compute requirements will be met.

D) *Model:* Please provide details of each of the models which feed into this system, including any prior iterations of these models.

E) *System:* Please account for how the above components were combined to create the final system, including details of any further components not accounted for above.

Supply chain risk assessment

Various forms of evidence may be submitted here, to include:

- ➔ Reports from government site visits to assess suppliers.[119]
- ➔ Evidence of compliance with established frameworks for supply chain security e.g. MITRE's system of Trust Framework or the Australian Government's Critical Technology Supply Chain principles.[120]
- ➔ Completed questionnaires from suppliers which detail how their data collection process was a) legally compliant and b) ethical.[121]
- ➔ Assessments of whether suppliers' other customers may raise security concerns.[122]

3.2.4 Performance and security

Performance has been central to model cards since their inception and is a cross-sector priority for AI.[123] As a result, it is one of the most well accounted for features in existing model cards. In contrast, deep consideration of AI security is often lacking. As noted by one government expert, 'I would hope in the real world most companies are good at the other

[118] Edd Gent, "Public AI Training Datasets Are Rife With Licensing Errors," *IEEE Spectrum*, 8 November 2023, https://spectrum.ieee.org/data-ai.

[119] Interview with government representative (2), 19 July 2023.

[120] Australian Government, *Critical Technology Supply Chain Principles* (Government of Australia: 2021); MITRE, "System of Trust Framework," https://sot.mitre.org/framework/system_of_trust.html.

[121] Interview with industry expert, 4 August 2023.

[122] Interview with industry expert, 21 July 2023.

[123] Margaret Mitchell et al., "Model cards for model reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220-229.

testing – what they might be missing is this security stuff.'[124] As a result, we stick closely to Hugging Face proposals for documenting AI performance, adapting them for a national security context. We then make initial recommendations for including AI security on a system card, proposing further work is needed to explore best practice for documenting AI security features.

Performance metrics should be highly tailored to the use context.[125] The risk is not poor performance in general, but rather 'that products are the jack of all trades, but not necessarily the master of what you need.'[126] The system card should not simply detail results from performance evaluations, but the rationale for choosing a particular metric for a particular context.[127] In most cases, the national security body will need to do their own performance tests to supplement supplier assessments which cannot fully replicate the final use context. Performance should be disaggregated across a variety of factors, with careful consideration given to the 'foreseeable salient factors for which model performance may vary.'[128] Users of the system card should justify the way in which performance has been disaggregated.[129] Performance metrics should be taken as just one part of a much larger picture. A 'precision, accuracy, recall, or F1 score delivered without context can give the *appearance* that performance has been robustly assessed, but without explanation of its results in the wider system this can be illusory.'[130]

Existing model cards 'only tell you so much' and they don't 'tell you how you defended the data against poisoning or other more specific things.'[131] In line with this, participants wanted system cards to include more detail on AI security, but recognised this would require longer term research – 'I honestly think the tool that we need is way, way more research in this area'.[132] For instance, AI standards were cited as offering useful evidence that suppliers have done due diligence on AI security. However, with many standards left in draft, suppliers are left in a tricky position where ground rules are not fully established and evidence becomes less clear cut.[133] Below, we account for how features of AI security may

[124] Interview with government representative, 5 July 2023.
[125] Margaret Mitchell et al., "Model cards for model reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220-229.
[126] Interview with law enforcement lawyer, 4 July 2023.
[127] Margaret Mitchell et al., "Model cards for model reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2019), 220-229.
[128] Ibid.
[129] Ibid.
[130] CETaS research workshop, 25 September 2023.
[131] Interview with government representative, 5 July 2023.
[132] Interview with government representative, 5 July 2023.
[133] CETaS research workshop, 25 September 2023.

be evidenced in the short term, with further research needed to offer more robust evidence in this area.

Table 8: System card section four

Part 4: Performance and Security
Instructions
Performance Please provide results from context-specific performance metrics and detail the rationale for selecting these metrics. This section should include details on precision and recall at different classification thresholds, the classification thresholds that have been used, robustness to out-of-sample inputs, live incident rates, and, where relevant, an account of error likelihood.[134] For each result given, the rationale for selecting the specific metric should be given alongside the rationale for disaggregating results in the way that has been chosen (e.g. according to gender, ethnicity, or other relevant considerations).
Security Please detail all available evidence that AI security has been considered throughout the project lifecycle. Evidence presented here may include: ➔ Compliance with international standards on AI security, for example 'ISO/IEC 42001' alongside other relevant ISO and IEEE standards.[135] ➔ Evidence of compliance with NCSC principles on security of AI or guidelines for secure AI system development.[136] ➔ Reports from red teaming exercises and adversarial testing.[137] ➔ Details of data hosting /management plans.[138] ➔ Description of implementation of AI security protocols laid out by MITRE ATLAS or OWASP.[139]

[134] CETaS research workshop, 25 September 2023.

[135] CETaS research workshop, 25 September 2023.

[136] Interview with government representative (2), 19 July, 2023; NCSC, "Principles for the security of machine learning," NCSC Guidance, 31 August 2022, https://www.ncsc.gov.uk/collection/machine-learning; NCSC, "Guidelines for secure AI system. development," November 2023, https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf.

[137] CETaS research workshop, 25 September 2023.

[138] CETaS research workshop, 25 September 2023.

[139] MITRE, "MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems)," https://atlas.mitre.org; OWASP, "AI Security and Privacy Guide," https://owasp.org/www-project-ai-security-and-privacy-guide/.

➔ Where possible, please provide details of residual security risks to facilitate ongoing monitoring.

3.2.5 Ethical considerations

Compared to other properties documented through model and system cards, ethical considerations can be highly abstract, making it even harder for those responsible for AI system design and development, particularly from a technical background, to understand and implement them. The companion guidance discussed in *Section 3.1.* will therefore be particularly important to support users of this system card section. In particular, such guidance will enable future iterations of this system card template to point those filling it out in the direction of national security vetted AI assurance techniques, rather than the more generalist methodologies listed by the OECD and CDEI.

However, in the case of AI ethics there is one further gap that must be filled in by national security bodies. Specifically, they must agree on common definitions of the core principles they wish to prioritise. This has been done in defence,[140] as well as in the context of broader government AI policy, as set out in 'a pro-innovation approach to regulation'.[141] GCHQ have already defined what they consider to be the major challenges to ethical AI – listing fairness, transparency and accountability, empowerment, and privacy. They have also gone further in defining these challenges, often in practical terms. For example, 'fairness' has been defined with reference to three key obstacles which must be overcome for AI systems to be considered fair – namely data fairness, design fairness, outcome fairness.[142] Already, this can help to guide users of this system card template towards which ethical considerations they should address.

However, the national security community is yet to commit to a final set of principles. We recommend they must do so in order to support this assurance process. In doing so, we propose they prioritise principles that are grounded in the real-world impacts of AI systems, that they translate principles such that they correspond directly to the needs of teams who are responsible for assurance, and that they complement and bolster principles which have already been defined in the legal context (e.g. necessity and proportionality). Ultimately,

[140] HM Government, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence* (Ministry of Defence: June 2022), https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence#annex-a-ethical-principles-for-ai-in-defence.

[141] HM Government, A *pro-innovation approach to AI regulation* (DSIT and Office for AI: August 2023), available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

[142] GCHQ, "Pioneering a New National Security: The Ethics of AI," 2021, https://www.gchq.gov.uk/artificial-intelligence/index.html.

such a list of principles and their precise definitions should be included in the below system card section in place of GCHQ's stated ethical AI challenges.

Table 9: System card section five

Part 5: Ethical considerations
Instructions
Please detail how the below set of ethical challenges have been addressed by the project team throughout the AI lifecycle: ➔ Fairness ➔ Transparency and accountability ➔ Empowerment ➔ Privacy In doing so, you should consider drawing on the techniques for responsible AI set out in CDEI's portfolio of assurance techniques and the OECD's tools for trustworthy AI, both of which include reference to a range of assurance techniques from external audits to technical fairness assessments, AI standards, and impact assessments.[143] *Please note that it will often be relevant to include multiple pieces of evidence to evidence a single ethical principle, and to make clear how your evidence supports the stated end goal.*

3.2.6 Iterative requirements

The final stage of the system card sets out plans for ongoing monitoring and future assessment. Model and system cards in their current form have been critiqued for being too static.[144] Simply setting a timeline for future review is insufficient. This system card should be easily updated to account for changes to governance processes, real-world impacts, and system updates from suppliers. For example, one participant noted that the system card should assess 'how people are using the model and the downstream impacts of these interactions.'[145] Others repeatedly emphasised the need to create a process which can eventually accommodate AI systems that are constantly learning and updating.[146] If the

[143] HM Government, *CDEI portfolio of AI assurance techniques* (Centre for Data Ethics and Innovation: 2023), https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques; OECD, "Catalogue of Tools and Metrics for Responsible AI," https://oecd.ai/en/catalogue/tools.

[144] Interview with government representative, 19 July 2023.

[145] Interview (2) with law enforcement member of staff, 6 July 2023.

[146] CETaS research workshop, 25 September 2023.

national security body wishes to re-use a previously supplied algorithm in a new context, or combine it with other AI systems, they will need at the very least to revisit this system card, updating it with new evidence. They may even need to start a new system card if a large quantity of existing evidence has been rendered outdated or irrelevant.

Table 10: System card section six

Part 6: Iterative requirements
Instructions
Evidence of internal skills base to effectively use the system AI literacy needs to improve if third-party AI tools are to be effectively assessed and monitored.[147] National security teams should justify that they have plans in place to upskill internal teams to become effective users of new AI systems. This could include descriptions of training to be conducted prior to deployment or of data science and AI policy representation within the team.
Ongoing monitoring provision, protections against accidental misuse & impact mitigation plan: What tests have been put in place to monitor the impacts of the system as it is deployed? Are mechanisms put in place to allow users to report errors? How do these feed into decisions about any updates or potential model retirement? It may be relevant to include a link to an internal plan for impact monitoring and mitigation which sets out in depth protocols for dealing with pre-identified potential adverse impacts.[148] The necessity of this should be determined by national security bodies depending on how high-risk they judge the use of an AI system to be.
Details of timelines: *A) Timeline for system updates:* This system card should account for future updates to AI systems, being updated with each supplier update or retraining cycle. In the future, this system card should be trialled

[147] CETaS research workshop, 25 September 2023.

[148] David Leslie et al., *Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A primer* (Council of Europe: 2021), https://edoc.coe.int/en/artificial-intelligence/10206-artificial-intelligence- human-rights-democracy-and-the-rule-of-law-a-primer.html.

on online learning AI systems to assess the extent to which it can become a living document.[149]

B) Timeline for system card review:

Set a timeline for review of the system card. It may be relevant to review a system even if it has not been updated, for example in response to impact monitoring or to changes in scope of use, or when approaching the end of an authorised data retention period. National security bodies should commit to timelines in advance while also remaining flexible to bring reviews forward when needed.

[149] Interview with government representative, 19 July 2023.

3.3 Assessing evidence

The fitness-for-purpose of assurance processes depends on the existence of tangible evidence that can justify the claims made, and the ability of decision-makers to assess the evidence in the context of their own risk acceptance threshold. The question then shifts to the evaluation of that evidence and who within an organisation – or potentially externally within a regulatory/audit/oversight structure – will assess the validity and robustness of the evidence provided, and thus provide approval for a project to proceed or require mitigations. In other words, the process of deciding what the evidence reveals about the system being assured.

3.3.1 Skills to review system cards

The interconnected and interdependent nature of many models make this assessment role a particular challenge. As Cobbe et al. point out 'Statistical guarantees may not hold when systems are composed together, and it is not straightforward to evaluate a whole system when each individual component may have been evaluated under different threat models (or other criteria).'[150] According to Brown, issues that must be considered not only relate to performance, but to copyright, data protection, product liability/negligence, equalities/bias and human rights (including those of workers involved in developing AI).[151]

Therefore, assessment of evidence put forward to address such broad issues is not likely to be a one-person – or one-discipline – job. Much will depend upon 'transparency mechanisms that enable a flow of critical information', in particular from the supplier(s) to the customer, and on the knowledge, understanding and critical skills of the persons carrying out the evaluation task.

Our interviewees generally agreed that a range of people should be involved in the reasoning and evaluation process, proposing a range of ways in which this could be operationalised in practice:

[150] Jennifer Cobbe, Michael Veale and Jatinder Singh, "Understanding accountability in algorithmic supply chains," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2023) 1186-1197.

[151] Ian Brown, "Expert Explainer: allocating accountability in AI supply chains," Ada Lovelace Institute Paper (June 2023), https://www.adalovelaceinstitute.org/resource/ai-supply-chains/.

- One suggested a multi-disciplinary board which would inform a senior responsible officer – on technical, community issues, legal and wider ethical questions (e.g. source of training data etc.) – in order to understand the risks.[152]
- Another commented that 'I would see it as a sociotechnical thing. It is not just the model itself, checking the accuracy of the model, but also how people are using the model and the downstream impacts of these interactions.'[153]
- It was further suggested that, due to the mission requirements that a model will be designed to achieve, 'people who are close to the use case itself will be best placed to make a lot of these context specific judgments.'[154]

Concern was expressed, however, that the required multidisciplinary in-house expertise was in limited supply, and ethics or multidisciplinary boards may not have the capacity to review all systems.[155] Furthermore, approvers or oversight boards may be reluctant to take ultimate accountability for approving a model, especially in circumstances where no one person may have sufficient understanding or visibility of the whole system.[156] Ultimately, the oversight of third-party AI must become integrated with wider institutional processes for assessing risk, and therefore part of senior management responsibility, as other risks are.

Interviewees had mixed opinions regarding community engagement with assurance. Many were supportive of the idea in principle but uncertain of the feasibility or appropriate process. The danger of 'participation-washing' was mentioned if community engagement was surface-level only or lacking influence.

Within a national security context where specific operational information could not be shared, it was suggested that a scenario or hypothetical context could inspire deliberation of the benefits and harms of technologies, with the outcomes feeding into the assurance process: 'It is easier to involve impacted groups in higher level decisions and steering or policy rather than specifics.'[157]

[152] Interview with law enforcement lawyer, 4 July 2023.

[153] Interview with law enforcement expert, 6 July 2023.

[154] Interview with industry expert, 21 July 2023.

[155] Interview with government representative, 5 July 2023.

[156] Jennifer Cobbe, Michael Veale and Jatinder Singh, "Understanding accountability in algorithmic supply chains," in *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2023) 1186-1197; David Gray Widder and Dawn Nafus, "Dislocated accountabilities in the 'AI supply chain': Modularity and developers' notions of responsibility," *Big Data and Society* (2023).

[157] Interview with academic expert, 6 July 2023.

3.3.2 The process of evidence assessment

Evaluation of evidence was not regarded by our interviewees as a one-off task, but a process of rolling review against set timelines or factors (as acknowledged in GCHQ's BAILO process).[158] This process must cover deployment into a real system[159] and further review if errors or concerns are detected or reported. This view reflects the conclusion in literature that assurance should be seen as an ongoing process to improve practices.[160]

In terms of evidence presentation, observability, combined with live monitoring/audit, and structured, accessible information were regarded as essential. It was suggested that 'old-fashioned' site visits should be included in the assurance process.[161] Having access to cleared/vetted individuals within suppliers was said to be important,[162] as was the ability to obtain independent operational testing results.[163]

However, a key issue for the process of evaluation 'is determining where you set the threshold for risk. At what point do you sign off on capability and determine that this risk is acceptable.'[164] It will be necessary in the overall assurance process to clarify the response to certain evidence presented - such as accuracy levels, performance against specified standards and data provenance - and therefore where the red lines will fall. These levels of tolerance will differ across use cases and depending upon potential harm and urgency. For instance, should a provider failing to reveal the source of their training data be a red flag in defence and national security? [165]

'ALARP' (as low as reasonably practicable) was mentioned by one interviewee, but in the wider context of whether society would regard the decision made as tolerable or acceptable. ALARP is a concept used in safety-critical industries such as aviation as a goal in relation to the management of health and safety risks.[166] Taddeo et al. recommend the ALARP framework as a way to tackling the AI risk 'predictability' problem, with a greater duty

[158] GCHQ/Bailo, "Bailo – managing the lifecycle of machine learning to support scalability, impact, collaboration, compliance and sharing," GitHub, https://github.com/gchq/Bailo.
[159] Interview with academic expert, 4 July 2023.
[160] Jessica Morley et al., "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices," *Sci Eng Ethics* (August 2020).
[161] Interview with government representative, 19 July 2023.
[162] Interview with government representative, 19 July 2023.
[163] Interview with law enforcement lawyer, 4 July 2023.
[164]Interview with industry expert, 21 July 2023.
[165] Interview with government representative, 26 June 2023.
[166] Health and Safety Executive, "Risk management: Expert guidance - ALARP at a glance," https://www.hse.gov.uk/enforce/expert/alarpglance.htm.

of care being required for higher stakes decisions.[167] However, it may be an oversimplistic way of assuring the use of third-party models within the national security context, bearing in mind the categories of information that we recommend are included in the system card approach above, and the importance of the overarching statutory framework, in particular the legal concepts of necessity and proportionality.[168]

Generally, our interviewees did not consider independent oversight bodies as integral to assurance processes, although it was noted that Commissioners and other regulators may require visibility of assurance. It was, nevertheless, regarded as deserving of further consideration, provided the oversight body had the appropriate expertise, particularly on the technical side.[169] Establishing assessor functions internally risks 'marking your own homework favourably'. Checks and balances within the public sector architecture should be established to mitigate this risk, including through industry secondments.[170]

3.3.3 Contractual protections

It was clear from our research interviews that the integration of contractual requirements and protections was essential to the success of any assurance process. As one interviewee put it, 'contract is king.'[171] Contractual warranties might in some circumstances mitigate lack of disclosure due to trade secrecy concerns, although incomplete knowledge may not be sufficient in a national security context.

Contractual clauses may cover specific requirements such as:

- The ability to conduct audits and spot-checks (including by an independent third-party), and to access summary information.[172] Such audits may require the supplier to provide access to their intellectual property via trusted safe harbour or escrow settings.
- Data provenance including evidence trails on sourcing of datasets, particularly in complex supply chains.

[167] Mariarosaria Taddeo et al., "Artificial Intelligence for national security: the predictability problem," *CETaS Research Reports* (September 2022).

[168] Ardi Janjeva, Muffy Calder and Marion Oswald, "Privacy Intrusion and National Security in the Age of AI: Assessing proportionality of automated analysis," *CETaS Research Reports* (May 2023).

[169] Interview with government representative, 22 June 2023.

[170] Interview with industry expert, 21 July 2023.

[171] Interview with law enforcement lawyer, 4 July 2023.

[172] Interview with academic expert, 4 July 2023.

- Requirements to report and remedy faults, and mechanisms to preserve a snapshot of a system when harm occurs. As noted by Brown, 'AI systems can change with new inputs or tweaks to their architecture. This means saving time-stamped versions of systems so that the cause of harms can be examined later, as happens already with self-driving vehicles'.[173]

In national security contexts, consideration should be given to additional or specific transparency and liability requirements in contracts around the use of foundation models,[174] open source, off-the-shelf or generic components and particular training data, or other material that may cause security or misuse concern. This could include restrictions or limitations on use plus mandating extra compliance and transparency responsibilities.

Research participants also raised the need for boilerplate confidentiality and transparency conditions to be enhanced to include transparency around other customers of the supplier which may cause security concerns,[175] and limitations around improvements to the system stemming from the customer's datasets or other information. Contracts and procurement processes also provide the opportunity to set out definitions of key assurance terms, and to specify concrete obligations such as disclosure of bias and accuracy testing results.

For examples of draft standard clauses addressing the above issues, see example clauses drafted by the City of Amsterdam on:

- technical transparency (including technical specifications/source code and data inputs);
- procedural transparency (including the choices and assumptions made) and
- explainability (including, where necessary and appropriate, requirements on the supplier to be able to explain on an individual level why the tool has come to a particular decision and provision of any information required for legal proceedings).[176]

[173] Ian Brown, "Expert Explainer: allocating accountability in AI supply chains," Ada Lovelace Institute Paper (June 2023), https://www.adalovelaceinstitute.org/resource/ai-supply-chains/.

[174] Defined here as 'AI models designed to produce a wide and general variety of outputs,' and 'capable of a range of possible tasks and applications, such as text, image or audio generation'. See: Elliot Jones, "Explainer: what is a foundation model?" Ada Lovelace Institute Paper (July 2023), https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

[175] The National Security and Investment Act 2021 permits scrutiny of corporate acquisitions and mergers that may cause a national security risk. A similar concept could be reflected in contractual terms for changes in the provider that might raise similar risks.

[176] Government of Amsterdam, "Contractual terms for algorithms," https://www.amsterdam.nl/innovation/digitalisation-technology/algorithms-ai/contractual-terms-for-algorithms/.

Standard contractual clauses for AI procurement have also been published by the EU to support trustworthy, fair, and secure AI procurement.[177] These clauses are proposed in the context of the EU AI Act and are available for both high risk and non-high risk AI systems. They include provisions on data and data governance, on technical documentation and record-keeping, on human oversight, on robustness and cybersecurity, and more.[178] Furthermore, it may be likely that Conformity Assessments pursuant to the EU AI Act will be presented by commercial providers as part of the assurance evidence required.

Of course, it would be necessary to consider how the aims and content of such clauses could be incorporated into government procurement contracts under the contract law of England and Wales. However, the robustness of any assurance process will rely to a considerable extent on how it is underpinned by appropriate and relevant contractual requirements.

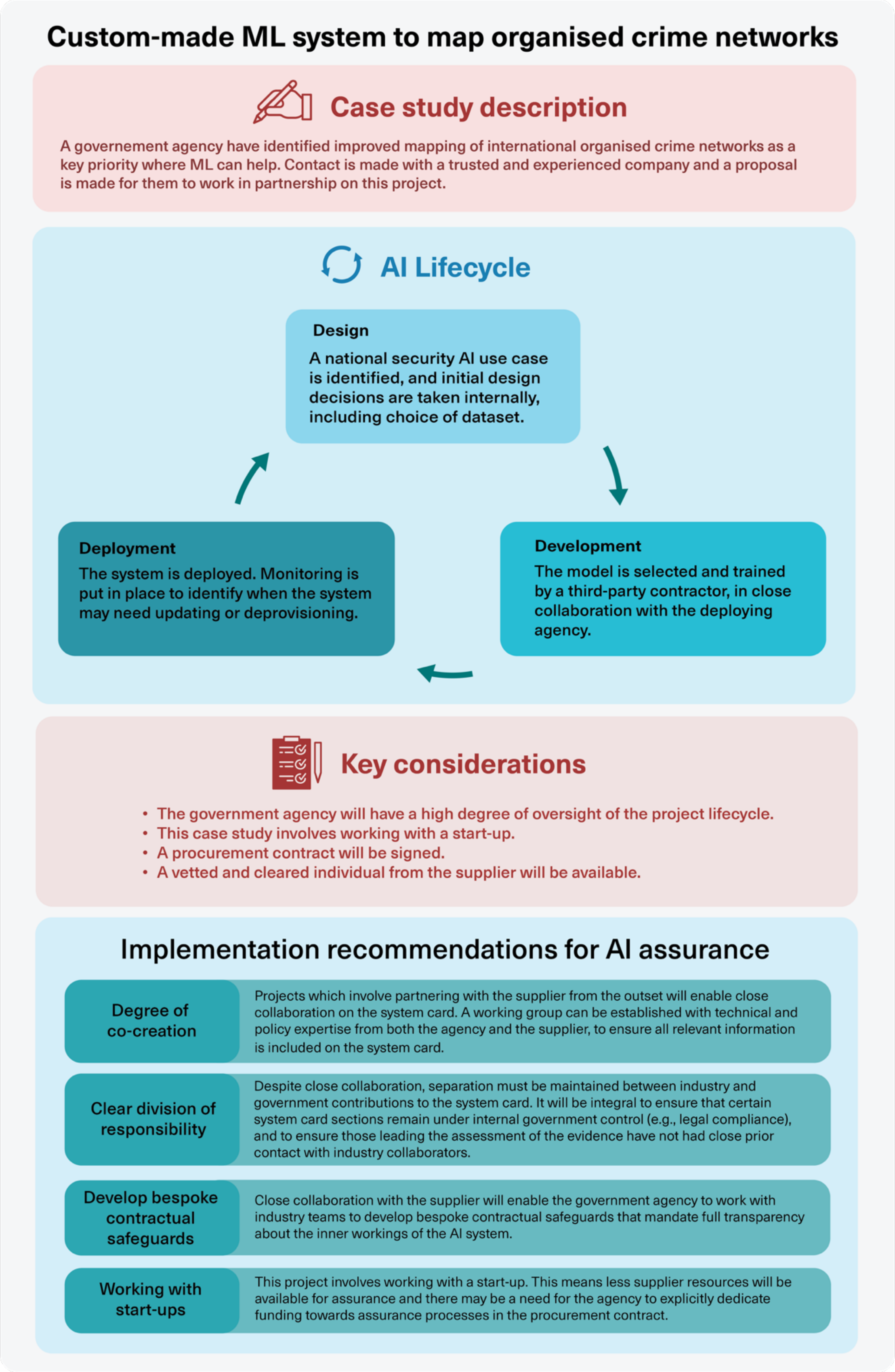
[177] European Commission, "EU model contractual AI clauses to pilot in procurements of AI," European Commission Templates & Guidance, 29 September 2023, https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/eu-model-contractual-ai-clauses-pilot-procurements-ai.

[178] Ibid.

4. Implementation Considerations for Hypothetical Case Studies

As emphasised throughout this report, our framework for AI assurance is intended to be applicable across a wide range of third-party AI systems, from those developed in partnership with industry suppliers to those fine-tuned internally from commercially available AI products. However, the framework cannot be implemented across the board without careful consideration of context-specific requirements. Here we introduce three hypothetical case studies to illustrate how our framework can be applied in context.

Figure 5: Assurance case study one



Custom-made ML system to map organised crime networks

Case study description

A governement agency have identified improved mapping of international organised crime networks as a key priority where ML can help. Contact is made with a trusted and experienced company and a proposal is made for them to work in partnership on this project.

AI Lifecycle

Design

A national security AI use case is identified, and initial design decisions are taken internally, including choice of dataset.

Development

The model is selected and trained by a third-party contractor, in close collaboration with the deploying agency.

Deployment

The system is deployed. Monitoring is put in place to identify when the system may need updating or deprovisioning.

Key considerations

- The government agency will have a high degree of oversight of the project lifecycle.
- This case study involves working with a start-up.
- A procurement contract will be signed.
- A vetted and cleared individual from the supplier will be available.

Implementation recommendations for AI assurance

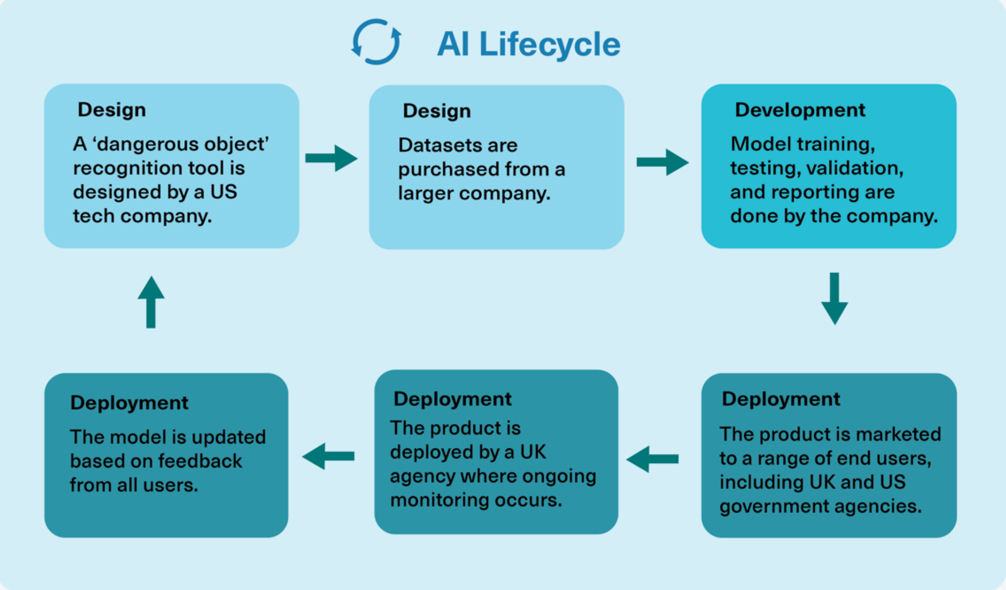
Degree of co-creation	Projects which involve partnering with the supplier from the outset will enable close collaboration on the system card. A working group can be established with technical and policy expertise from both the agency and the supplier, to ensure all relevant information is included on the system card.
Clear division of responsibility	Despite close collaboration, separation must be maintained between industry and government contributions to the system card. It will be integral to ensure that certain system card sections remain under internal government control (e.g., legal compliance), and to ensure those leading the assessment of the evidence have not had close prior contact with industry collaborators.
Develop bespoke contractual safeguards	Close collaboration with the supplier will enable the government agency to work with industry teams to develop bespoke contractual safeguards that mandate full transparency about the inner workings of the AI system.
Working with start-ups	This project involves working with a start-up. This means less supplier resources will be available for assurance and there may be a need for the agency to explicitly dedicate funding towards assurance processes in the procurement contract.

Figure 6: Assurance case study two

Off-the-shelf object recognition tool

Case study description

An AI tool has been developed by a US tech company to identify dangerous objects. The tool was developed for the US context. The company have control over most of the project lifecycle but have limited visibility over the data provenance. The company are interested in licensing the product to UK national security.



AI Lifecycle

Design
A 'dangerous object' recognition tool is designed by a US tech company.

Design
Datasets are purchased from a larger company.

Development
Model training, testing, validation, and reporting are done by the company.

Deployment
The product is marketed to a range of end users, including UK and US government agencies.

Deployment
The product is deployed by a UK agency where ongoing monitoring occurs.

Deployment
The model is updated based on feedback from all users.

Key considerations

- The supplier is not based in the UK and has worked with additional US companies to develop this AI system (e.g., purchasing the data).
- A procurement contract is in place.
- Regular updates to the AI system will be made available to the government agency.

Implementation recommendations for AI assurance

Degree of co-creation	In this case, the supplier is unlikely to begin contributing to the system card right from the design phase of their product. However, industry contributions can still be obtained post-hoc if the national security customer requests this or even mandates it through procurement contracts.
Nature of evidence submissions	As this is an off-the-shelf product, the safeguards introduced by the supplier will reflect their own priorities rather than those of the national security customer. As a result, a greater proportion of evidence in the system card will need to come from post-hoc assessments conducted by the deploying agency, filling in gaps in supplier evidence.
Model updates	Updates to this system will be based on feedback from a range of customers. In this case, careful consideration will need to be given to the nature of these other customers when completing the 'iterative requirements' section of the system card. E.g., how much feedback should the deploying agency share with the supplier in return for the supplier making improvements to suit their needs?
Complex supply chains	In this example, datasets have been purchased from a separate US-based company. It will be essential for the deploying agency to assess the security and legal compliance of this company's practices against their own requirements.

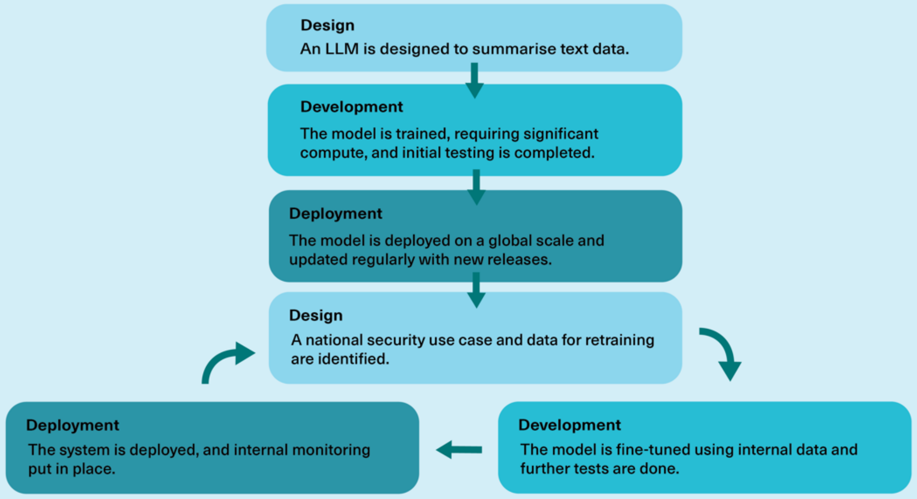
Figure 7: Assurance case study three

Fine-tuned LLM for intelligence analysis

Case study description

An open-source LLM from a mulitinational tech company was released six months ago and has shown promising results across a range of sectors in summarising text. A team in a UK government agency have fine-tuned this model on internal intelligence analysis data.

AI Lifecycle



Design
An LLM is designed to summarise text data.

Development
The model is trained, requiring significant compute, and initial testing is completed.

Deployment
The model is deployed on a global scale and updated regularly with new releases.

Design
A national security use case and data for retraining are identified.

Development
The model is fine-tuned using internal data and further tests are done.

Deployment
The system is deployed, and internal monitoring put in place.

Key considerations

- There is no contact between the supplier and the deploying agency.
- The LLM is commercially available but not fully open source (there is no access to the source code).
- The widespread use of this system in other sectors means it has already undergone significant. safety testing, with reports from this published.

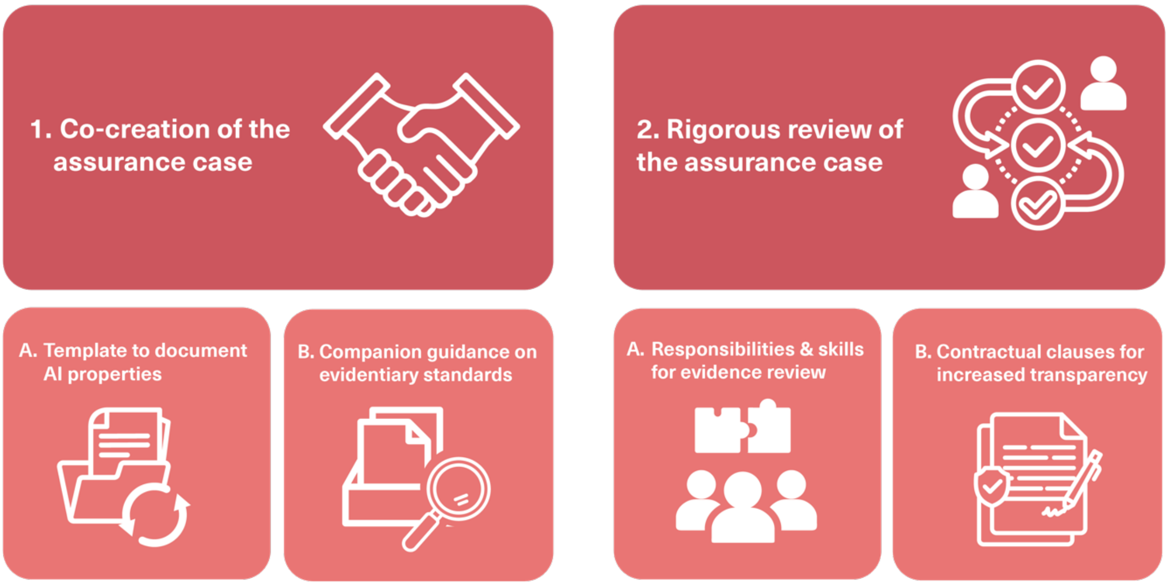
Implementation recommendations for AI assurance

Degree of co-creation	In this case, no input from suppliers can be expected. This does not necessarily mean less assurance information will be publicly available (e.g., from the supplier's own public reports, from other users of the tool publishing outcomes from safety testing). The deploying agency will need to review existing evidence before working to supplement this with their own tests, thus filling out the system card.
Nature of evidence submissions	Evidence included in the system card must always be tailored to the specific AI techniques being used. There are many emergent assessments for LLMs (e.g., audits, red teaming protocols, hallucination assessments). National security teams will need to ensure their repository of assurance techniques keeps pace with developments and includes up-to-date techniques for LLM assessment.
Data provenance	For this third-party system, data provenance will present a particular challenge. Inevitable gaps in knowledge of data provenance must be supplemented by even more rigorous assessments of legal compliance and of performance, security, and potential bias.
Expertise	For complex and emergent AI use cases, especially where supplier expertise is not accessible, it will be crucial for government agencies to seek external expertise elsewhere. In this case, for instance, it may be relevant to seek input to the system card from the UK Government's AI Safety Institute.

5. Recommendations for Implementing AI Assurance

Throughout this report, we emphasise that the first step for national security is to trial this framework on real-world use cases. Nevertheless, we also make recommendations to support effective and sustainable implementation of the assurance framework in the long term. These are summarised below.

Figure 8: Two stage model for AI assurance


1. Co-creation of the assurance case
2. Rigorous review of the assurance case
A. Template to document AI properties
B. Companion guidance on evidentiary standards
A. Responsibilities & skills for evidence review
B. Contractual clauses for increased transparency

1. **Create a template for robust documentation of AI properties:**

   - **Create exemplars to illustrate what complete assurance cases look like in practice:**

The system card template alone presents a challenge for users wishing to apply it to specific AI use cases. Exemplars, in the form of completed system cards, can serve a critical educational role both within government and in communication with suppliers, offering specific detail on what is deemed to be a robust assurance case.

- **Develop a platform to support users of the system card template:**

Already, platforms for the central hosting of model cards (in particular Bailo) are crucial for ensuring information on model cards is widely accessible. However, more can be done to create a platform for real-time system card collaboration, in line with this framework. We recommend updating functionality of these platforms to offer separate digital workspaces for industry suppliers and national security bodies to fill out, edit, and review system cards.

2. **Develop companion guidance to aid users of the system card template:**

   - **Create a national security approved repository of assurance techniques:**

One key piece of infrastructure that should be prioritised by industry suppliers and national security bodies alike is the creation of a searchable repository of techniques for national security AI assurance. Further research is needed to identify appropriate techniques for inclusion in this repository and to map these to the needs of distinct stakeholders within the assurance ecosystem.

   - **Where gaps are identified (e.g. with regard to AI security), invest in academic research to advance assurance:**

National security bodies must make strategic investments to ensure the wider assurance ecosystem serves their needs. For example, interviewees noted that tools to verify and assess AI security were not sufficiently developed.[179] Academic work should be commissioned to fill such gaps. However, current academic work on assurance has been critiqued for being overly theoretical and complex,[180] and priority should be given to academic work which engages directly with practitioners and with impacted communities – to ensure outputs reflect both operational requirements and real-world risks.

3. **Invest in skills for evidence review:**

The challenge of insufficient skills within national security bodies to review assurance cases was raised repeatedly by research participants:

- ‘an underappreciated risk is the challenge of education inside your own organisation to accurately make decisions on adopting third-party models’ [181]
- ‘the single biggest limiting factor on procuring effective systems is ignorance in the D&S sector’[182]
- ‘the key thing is that those in government acquiring these technologies do not have the necessary qualifications.’[183]

This highlights the need to educate not only assurance case reviewers, but also end users of third-party AI systems, building on technical, ethical and legal knowledge within national security bodies. Where relevant, we recommend other government bodies (including the

[179] Interview with government expert, 5 July 2023.

[180] Interview with academic expert, 26 July 2023.

[181] Interview with industry expert, 24 July 2023.

[182] Interview with industry expert, 24 July 2023.

[183] Interview with academic expert, 26 July 2023.

CDEI and the AI Safety Institute) loan experts into national security organisations to further fill this skills gap.

4. **Draft contractual clauses for increased transparency:**

The above framework goes some way to encouraging increased transparency from all parties, but further structural and institutional changes are needed. A promising route towards such transparency may be enabled by standardised contractual clauses which mandate transparency from both parties regarding the properties of third-party AI systems. Legal experts within national security bodies should prioritise work to develop clauses which suit their needs.

6. Conclusion

This report has presented a framework for AI assurance in the context of UK national security, detailing a step-by-step process that enables UK national security bodies to harness the benefits of third-party AI technologies while reducing associated risks. The main recommendation of this report, for both national security decisionmakers and industry suppliers, is to trial this framework in the context of real-world AI applications.

We recommend that national security bodies:

1. Adopt this system card template as a documentation method for third-party AI.
2. Develop clearer guidance for users of the system card on the sort of evidence which can be included in the system card.
3. Establish clear protocols for review of system cards, ensuring that the necessary expertise to review evidence is available.
4. Consider introducing standardised contractual clauses to mandate further transparency from suppliers about the AI systems they develop.

We recommend that industry suppliers:

1. Begin to trial this system card template as a means through which to document the properties of their AI systems.
2. Collaborate directly with national security bodies where possible in filling out system cards to ensure all relevant evidence has been included.
3. Contribute to discussions with national security bodies on how features of AI systems, such as security, performance, and ethics, can best be evidenced within system cards (e.g. through AI standards, audits, impact assessments and red teaming protocols).

To supplement these immediate-term recommendations, we propose further actions in the longer term, including:

Table 11: Recommendations for implementing assurance

Recommendations for implementing AI assurance
Build infrastructure for a sustainable assurance ecosystem, including further investments in platforms to host assurance cases *and* the creation of a tailored national security portfolio of assurance techniques.
Invest in skills for reviewing AI assurance cases, to include technical skills in addition to ethical and legal expertise. We recommend that government centres of AI expertise dedicate time and resource to supporting specific departments in AI assurance, including but not limited to the AI Safety Institute and Centre for Data Ethics and Innovation.
Connect future academic work on assurance to practitioner challenges to increase the availability of practically useful frameworks for AI assurance that fill persistent gaps e.g. on AI security.
Develop exemplar assurance cases across a range of case studies to explore more specific recommendations for real-world AI use cases (such as LLMs in intelligence analysis or autonomous agents for cyber defence).
Draft bespoke contractual clauses which can aid national security customers in ensuring suppliers are transparent about the properties of their AI systems.

The framework presented here builds on a wealth of existing research on AI assurance. It is intended to be of practical use in the near term while also laying foundations for future expansions that account for the diverse landscape of third-party AI in the national security domain. Continued research into the themes discussed here is essential to establishing and maintaining a robust assurance landscape.

Appendix 1: Compiled System Card Template

Summary information	
Instructions	**Evidence**
System details: Please provide AI system name, 1-2 sentence description of the system and its constituent components, version, and implementation so far.[184]	
Mission objectives fulfilled and use cases across the organisation: Please summarise the positive contributions made by the system towards the organisation's goals and give an account of how 'load bearing' the AI system may be across the organisation.[185]	
Internal roles and responsibilities: Detail the key internal decisionmakers responsible for filling out and reviewing this system card, including policy, legal, and technical expertise, and clear separation between the roles of filling out the system card with relevant evidence, and assessing the completed system card.	
Supply chain summary: Please summarise the information given in part 3, including list of organisations/departments responsible for design, development, and deployment & at least one contact for each organisation/department.	
License: If applicable, details of the licensing/procurement arrangement are to be provided here.	
Summary and key take-aways: Please summarise key take-aways from the following sections (mission properties & legal compliance, performance & security, ethics). A red/amber/green scale may be used to highlight sections of concern.	
Iterative review summary: Provide dates for any anticipated updates to the AI system *and* for next review and update of this system card.	
Mission properties and legal compliance	
Instructions	**Evidence**
Context and scope of use[186] *A) Delineate clear parameters for AI system use:*	

[184] Hugging Face, "Annotated Model Card Template," https://huggingface.co/docs/hub/model-card-annotated.

[185] Interview with government representative, 8 August 2023.

[186] HM Government, *Algorithmic Transparency Recording Standard Hub* (CDDO and CDEI: January 2023), https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub.

Set out who in the organisation will be using this AI system, how often, and for what purpose. If the AI system in question is being repurposed by the national security body from the purpose for which it was designed, this should be flagged here. This section should also set out any prohibited uses that have been identified as risky. B) *Account for how the AI system will impact existing organisational processes and existing workers:* Set out the extent of integration of this system, both with existing human decision-making processes,[187] and with existing technology systems. Where relevant, this may include reference to an assessment of the impact the AI system will have on employees' working conditions, for example through a 'Good Work Algorithmic Impact Assessment'.[188] C) *Non-algorithmic options considered:*[189] Please detail why the AI system in question is preferential to the non-algorithmic options available, including a comparison to the current method for completing this task if relevant.	
Legal basis The legal basis, requirements and powers for the development and use of the AI system alongside other legal compliance requirements that the assurance process will help to support should be set out. This section may include but is not limited to: ➔ The overarching statutory or legal functions for which the AI system is being developed. ➔ Any limitations, restrictions, or constraints on the exercise of data acquisition and/or analysis for the purposes of national security or other purposes, including those within the Investigatory Powers Act and associated warrants and authorisations.	

[187] HM Government, *Algorithmic Transparency Recording Standard Hub* (CDDO and CDEI: January 2023), https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub.

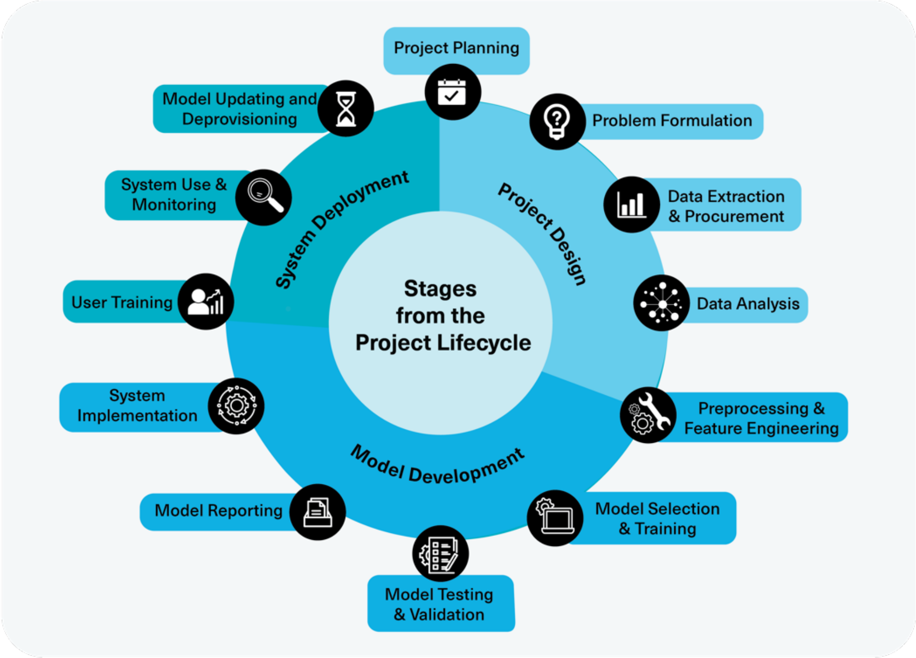
[188] Institute for the Future of Work, "Good Work Algorithmic Impact Assessment," IFOW Guidance (March 2023), https://www.ifow.org/publications/good-work-algorithmic-impact-assessment-an-approach-for-worker-involvement.

[189] HM Government, *Algorithmic Transparency Recording Standard Hub* (CDDO and CDEI: January 2023), https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub.

➔ Consideration of the human rights principle of necessity and proportionality[190] in relation to the development and use of the AI system. ➔ Any requirement for the tool's output to be used evidentially or in legal proceedings.	
Licensing/model acquisition Provide details of the model/software licensing agreement (or other procurement structure such as bespoke development), including details of contractual transparency requirements and other protections. Links to contracts should be included here for further detail, and details of the invitation to tender (ITT) process should be set out if this took place (including possible assurances which were requested in the ITT process).	
The supply chain	
Instructions	**Evidence**
Supply chain mapping & industry contributors Please identify whether the following stages of the AI lifecycle[191] were government-led or industry-led. Please also attribute each stage to a specific organisation, or for organisations over 100 people, to a specific department. Additionally, please nominate a point of contact at each relevant organisation, or at each department at larger organisations. Their role should be described, both with regard to the project lifecycle itself and the co-completion of this system card. Any vetted and cleared contributors from industry should be identified as potential collaborators on this system card.	

[190] Ardi Janjeva, Muffy Calder and Marion Oswald, 'Privacy Intrusion and National Security in the Age of AI: Assessing proportionality of automated analysis,' *CETaS Research Reports* (May 2023).

[191] This model of the AI lifecycle was developed by The Alan Turing Institute and accounts for the highly sociotechnical nature of AI design, development and deployment. See Christopher Burr and David Leslie, 'Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies,' *AI and Ethics (*3, 73-98) 2023.


Stages from the Project Lifecycle
Project Design
Model Development
System Deployment
Project Planning
Problem Formulation
Data Extraction & Procurement
Data Analysis
Preprocessing & Feature Engineering
Model Selection & Training
Model Testing & Validation
Model Reporting
System Implementation
User Training
System Use & Monitoring
Model Updating and Deprovisioning

Source: Model of the AI lifecycle, reproduced from Burr & Leslie, 2023.

Provenance

Provenance here is defined as the 'chronology of the ownership, custody or location of a historical object',[192] and should be accounted for with regard to:

A) *Data:* What training data was used? Where was it sourced? Please link to full datasets if possible and provide details of any updates to datasets through the AI lifecycle. Please link to audits of relevant datasets where available (for instance through the 'data provenance initiative').[193]

B) *Hardware:* Please detail the hardware feeding into this system, including details of how it was sourced.

C) *Compute:* Please detail the source of compute for this system and how ongoing compute requirements will be met.

D) *Model:* Please provide details of each of the models which feed into this system, including any prior iterations of these models.

[192] Kiran Karkera, "Why is provenance important for AI," Kiran Karkera Medium, 10 July 2020, https://kaaldaari.medium.com/an-example-of-art-provenance-records-for-the-curious-d3a5e4a1dd77.

[193] Edd Gent, "Public AI Training Datasets Are Rife With Licensing Errors," *IEEE Spectrum*, 8 November 2023, https://spectrum.ieee.org/data-ai.

E) *System:* Please account for how the above components were combined to create the final system, including details of any further components not accounted for above.	
Supply chain risk assessment Various forms of evidence may be submitted here, to include: ➔ Reports from government site visits to assess suppliers.[194] ➔ Evidence of compliance with established frameworks for supply chain security e.g. MITRE's system of Trust Framework or the Australian Government's Critical Technology Supply Chain principles.[195] ➔ Completed questionnaires from suppliers which detail how their data collection process was a) legally compliant and b) ethical.[196] ➔ Assessments of whether suppliers' other customers may raise security concerns.[197]	
Performance and security	
Instructions	**Evidence**
Performance Please provide results from context-specific performance metrics and detail the rationale for selecting these metrics. This section should include details on precision and recall at different classification thresholds, the classification thresholds that have been used, robustness to out-of-sample inputs, live incident rates, and, where relevant, an account of error likelihood.[198] For each result given, the rationale for selecting the specific metric should be given alongside the rationale for disaggregating results in the way that has been chosen (e.g. according to gender, ethnicity, or other relevant considerations).	

[194] Interview with government representative (2), 19 July 2023.

[195] Australian Government, *Critical Technology Supply Chain Principles* (Government of Australia: 2021); MITRE, "System of Trust Framework," https://sot.mitre.org/framework/system_of_trust.html.

[196] Interview with industry expert, 4 August 2023.

[197] Interview with industry expert, 21 July 2023.

[198] CETaS research workshop, 25 September 2023.

Security Please detail all available evidence that AI security has been considered throughout the project lifecycle. Evidence presented here may include: ➔ Compliance with international standards on AI security, for example 'ISO/IEC 42001' alongside other relevant ISO and IEEE standards.[199] ➔ Evidence of compliance with NCSC principles on security of AI or guidelines for secure AI system development.[200] ➔ Reports from red teaming exercises and adversarial testing.[201] ➔ Details of data hosting /management plans.[202] ➔ Description of implementation of AI security protocols laid out by MITRE ATLAS or OWASP.[203] ➔ Where possible, please provide details of residual security risks to facilitate ongoing monitoring.	
Ethical considerations	
Instructions	Evidence
Please detail how the below set of ethical challenges have been addressed by the project team throughout the AI lifecycle: ➔ Fairness ➔ Transparency and accountability ➔ Empowerment ➔ Privacy In doing so, you should consider drawing on the techniques for responsible AI set out in CDEI's portfolio of assurance techniques and the OECD's tools for trustworthy AI, both of which include reference to	

[199] CETaS research workshop, 25 September 2023.

[200] Interview with government representative (2), 19 July 2023; NCSC, "Principles for the security of machine learning," August 2022, https://www.ncsc.gov.uk/collection/machine-learning; NCSC, "Guidelines for secure AI system development," November 2023, https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development.

[201] CETaS research workshop, 25 September 2023.

[202] CETaS research workshop, 25 September 2023.

[203] MITRE, "MITRE ATLAS (Adversarial Threat Landscape for Artificial Intelligence Systems)," https://atlas.mitre.org; OWASP, "AI Security and Privacy Guide," https://owasp.org/www-project-ai-security-and-privacy-guide/.

a range of assurance techniques from external audits to technical fairness assessments, AI standards, and impact assessments.[204] *Please note that it will often be relevant to include multiple pieces of evidence to evidence a single ethical principle, and to make clear how your evidence supports the stated end goal.*	
Iterative requirements	
Instructions	Evidence
Evidence of internal skills base to effectively use the system AI literacy needs to improve if third-party AI tools are to be effectively assessed and monitored.[205] National security teams should justify that they have plans in place to upskill internal teams to become effective users of new AI systems. This could include descriptions of training to be conducted prior to deployment or of data science and AI policy representation within the team.	
Ongoing monitoring provision, protections against accidental misuse & impact mitigation plan: What tests have been put in place to monitor the impacts of the system as it is deployed? Are mechanisms put in place to allow users to report errors? How do these feed into decisions about any updates or potential model retirement? It may be relevant to include a link to an internal plan for impact monitoring and mitigation which sets out in depth protocols for dealing with pre-identified potential adverse impacts.[206] The necessity of this should be determined by national security bodies depending on how high-risk they judge the use of an AI system to be.	
Details of timelines: A) *Timeline for system updates:*	

[204] HM Government, *CDEI portfolio of AI assurance techniques* (Centre for Data Ethics and Innovation: 2023), https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques; OECD, "Catalogue of Tools and Metrics for Responsible AI," https://oecd.ai/en/catalogue/tools.

[205] CETaS research workshop, 25 September 2023.

[206] David Leslie et al., *Artificial Intelligence, Human Rights, Democracy, and the Rule of Law: A primer* (Council of Europe: 2021), https://edoc.coe.int/en/artificial-intelligence/10206-artificial-intelligence- human-rights-democracy-and-the-rule-of-law-a-primer.html.

This system card should account for future updates to AI systems, being updated with each supplier update or retraining cycle. In the future, this system card should be trialled on online learning AI systems to assess the extent to which it can become a living document.[207] *B) Timeline for system card review:* Set a timeline for review of the system card. It may be relevant to review a system even if it has not been updated, for example in response to impact monitoring or to changes in scope of use, or when approaching the end of an authorised data retention period. National security bodies should commit to timelines in advance while also remaining flexible to bring reviews forward when needed.	

[207] Interview with government representative, 19 July 2023.

Appendix 2: Glossary of Key Terms

Artificial intelligence – For the purposes of this study, AI is defined in line with the OECD as any 'machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'.[208] However, our focus is primarily on machine learning technologies, defined throughout as technologies which use patterns in data to make predictions and thus improve performance over time.

AI lifecycle – The sequence of processes required to take an idea and implement it using AI. To include design, development, and deployment of AI systems and to incorporate both technical and sociotechnical processes which occur during each of these stages.

Third-party AI system – Any AI system where at least one stage of the AI lifecycle occurs partially or wholly outside of the organisation that will deploy the system.

Assurance – The portfolio of processes required to evaluate and communicate, iteratively throughout the AI lifecycle, the extent to which a given AI system does everything it says it is going to do and nothing it shouldn't do, complies with the values of the deploying organisation, is legally compliant, and is appropriate to the specific deployment context.

Assurance case – The central document containing all evidence that an AI system meets requirements, structured into a logical argument and supporting a collection of desired properties. This can take different forms. For the purposes of this report, we recommend a system card is used as the assurance case.

Model card – Files that summarise key information about a model, ranging from key performance results to ethical risks and from training parameters to details on the intended use context.

System card – Files that are closely related to model cards except that instead of documenting a single model, system cards aim to document the features of all of the models and other components which make up the final AI system.

Argument-based assurance - A process of using structured argumentation to communicate all of the evidence that an AI system possesses a particular quality (whether this is a safety-critical feature or an ethical feature).

[208] OECD, "OECD AI Principles Overview," https://oecd.ai/en/ai-principles.

About the Authors

Rosamund Powell is a Research Associate at the Centre for Emerging Technology and Security (CETaS). Prior to joining CETaS, Rosamund worked as an Ethics Research Assistant within The Alan Turing Institute, contributing to the public policy programme research agenda. While at the Turing, she has co-authored research reports exploring topics at the intersection of artificial intelligence, ethics, and policy. She also has experience working in UK Parliament and at UNESCO, both in relation to AI and public policy.

Dr Marion Oswald MBE is a Senior Research Associate at the Alan Turing Institute and Professor in Law at Northumbria University. Marion is a lawyer with over 30 years' experience spanning several contexts: law firms, international technology businesses, central government including national security, academia, and oversight functions. She sits on the Advisory Council for Open Rights Group and chairs the West Midlands Police data ethics committee. She has a particular research interest in the human rights, ethics and use of data analytics within policing and intelligence agencies.

Centre for Emerging Technology and Security

RESEARCH REPORT