The Alan Turing Institute

Centre for Emerging Technology and Security

POLICY BRIEF

AI Assurance for UK National Security

CETaS Policy Brief

Rosamund Powell and Marion Oswald

January 2024

Overview

This policy brief summarises findings from CETaS research on ‘assurance of third-party AI systems for UK national security’ – a report outlining how national security bodies can effectively evaluate AI systems designed and developed (at least partially) by industry suppliers.

Summary of Findings

- Involving industry in the design and development of AI is essential if UK national security bodies want to keep pace with cutting-edge capabilities – due to both skills and budget constraints within government.
- When stages of the AI lifecycle are outsourced, direct oversight of design and development processes may be reduced. As a result, third-party AI systems may not conform to the security, ethical, legal compliance, and performance requirements set for high stakes national security use cases.
- Our tailored AI assurance framework for UK national security facilitates more transparent communication about AI system properties and robust assessment of whether AI systems meet requirements.
- The framework centres on a structured system card template for UK national security. This provides guidance on how AI system properties should be documented – to cover legal, supply chain, performance, security, and ethical considerations.
- To effectively operationalise the system card template and ensure it facilitates robust review of third-party AI systems, we also propose:
    1. Companion guidance on what evidence should be used to fill out the system card.
    2. Bolstered investment in internal skills to review system cards.
    3. Contractual clauses to mandate transparent information sharing from suppliers.
- We recommend this assurance framework is trialled by national security bodies and industry suppliers in the immediate term.
- Following this, we recommend investment in research, infrastructure, and skills to support implementation of the framework on a larger scale.

Understanding third-party AI systems: origins, benefits, and risks

Third-party AI systems are defined as AI systems 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. We focus primarily on industry suppliers as the external contributor.

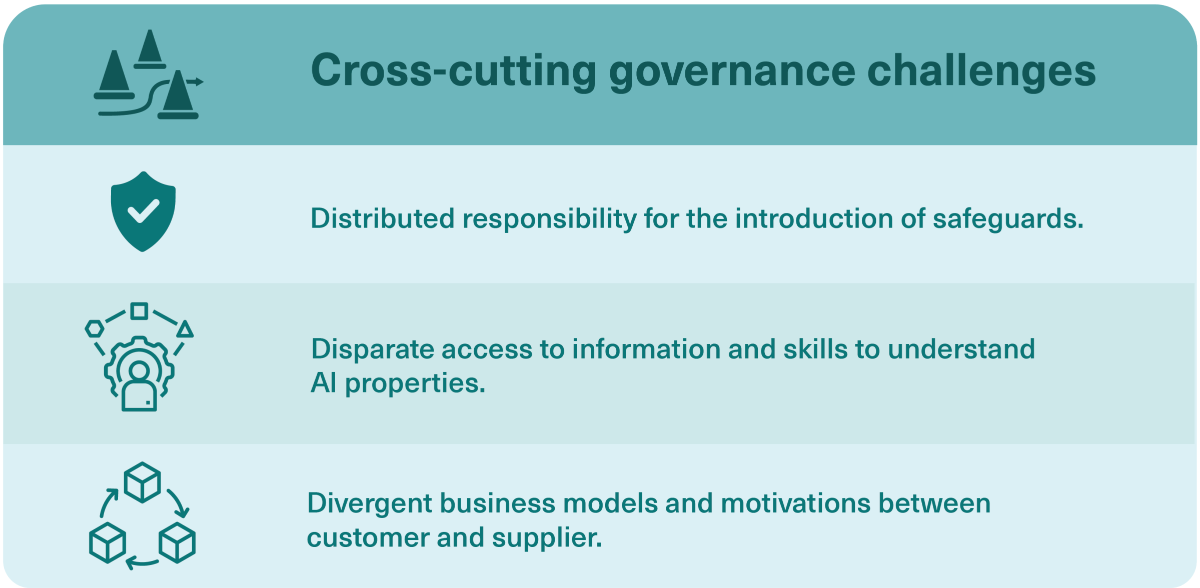
Three factors can be used to map the third-party AI landscape:

A. **The type and number of third parties involved** (e.g. academic institutions, private companies, other government departments, or some combination of these).
B. **The nature of the third-party relationship** (e.g. formal collaborations, open-source AI systems, commercially available products).
C. **The extent of third-party involvement** (e.g. partial contribution to one stage of the AI lifecycle or full control over every stage).

Each AI system raises distinct concerns for national security decision-makers. Nevertheless, several benefits, risks and governance challenges recur across a range of third-party AI systems (as illustrated in *figures 1 and 2).*

The risks emerging from complex supply chains are particularly pervasive because the complexity of AI supply chains makes them difficult to map,[1] and because legal, ethical and security concerns must be evaluated all the way down a supply chain.[2] Our framework for AI assurance therefore prioritises strategies to minimise these supply chain risks.

Figure 1: Cross-cutting governance challenges for third-party AI



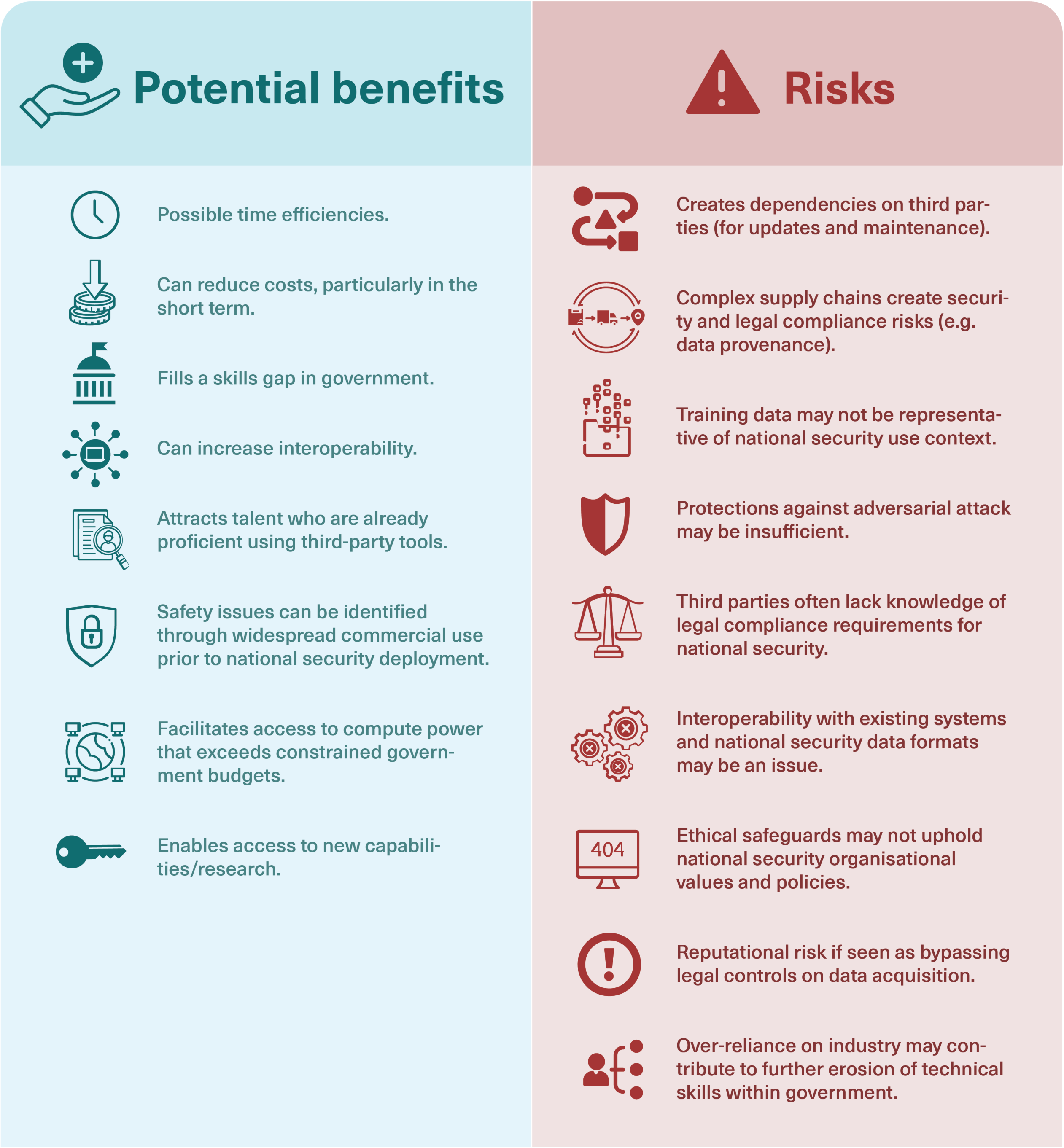
Cross-cutting governance challenges

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.

Figure 2: Benefits and risks of 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.

The promise of AI assurance

'AI assurance' is 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 the supplier 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.

Much progress has been made by other parts of the public sector towards successful AI assurance (particularly by the UK's Centre for Data Ethics and Innovation).[3] Despite this progress, the below challenges illustrate why further work is needed to meet national security needs:

1. **Existing frameworks fail to address national security priorities:** For example, failing to account for existing national security practices and/or failing to address heightened risks from AI in national security contexts.
2. **Crowded landscape:** Techniques for trustworthy AI are proliferating, leaving policymakers and suppliers confused and overwhelmed.
3. **Separation of technical vs ethical assessment and a lack of intersecting skills:** Currently, AI assurance methodologies tend to be *either* technical *or* ethical*.* Ethical and technical assessments need to occur in tandem. This requires multidisciplinary teams.
4. **Accommodating start-ups:** If entry costs are too high, start-ups with limited resources are left behind, and there is potential for stifled innovation and competition.
5. **Convoluted frameworks:** Practitioners expressed frustration at academic assurance frameworks which fail to specify requirements in terms they understood. Additional safeguards are needed but must be balanced with the need for efficient procurement.
6. **Divergent business models hamper communication:** Industry suppliers can be reluctant to communicate transparently about commercial IP and performance metrics.
7. **Complex supply chains are poorly understood:** Existing assurance frameworks struggle to account for disparate information access across complex supply chains.[4]
8. **Risks false sense of security:** The success of AI assurance is limited by the capability and diligence of people assessing assurance cases. It can easily become a rubber-stamping exercise if scrutiny from procuring organisations is not sufficiently robust.

Figure 3: Quotes on the challenges of AI assurance

"Current AI assurance standards are very technical in nature and confusing for most developers."

"I would hope in the real world most companies are good at the other testing - what they might be missing is this adverserial security stuff."

"The OECD has a catalogue of 200 techniques, but it dosen't help me."

AI assurance framework for UK national security

Our model for AI assurance (illustrated in *figure 4)* seeks to address the above challenges while also building on existing national security policies – specifically GCHQ's Bailo process for managing the machine learning lifecycle.[5]

First, the assurance case must be created. This is the central document compiling evidence on whether a particular AI system is suitable for deployment. To support this process, we propose:

a) A system card template for documenting AI system properties which facilitates input from both private sector suppliers and national security customers.

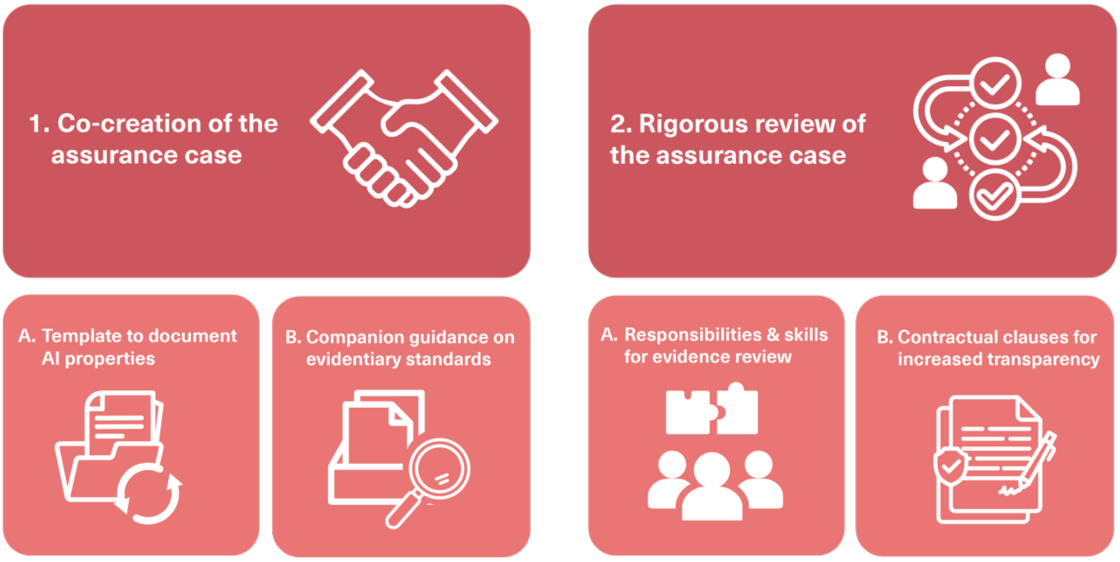
b) Companion guidance to support those filling out the template – directing them towards robust techniques to generate evidence that an AI system meets requirements.

Second, the assurance case must be reviewed to assess whether evidence justifies the deployment of the AI system. To support this process, we propose:

a) Clarity on internal responsibilities for assurance and investment in skills for review.

b) Contractual clauses to mandate transparent sharing of evidence with reviewers.

Figure 4: 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

Implementation recommendations

Immediate term: In the immediate term we recommend industry suppliers and national security bodies trial this AI assurance framework, particularly the system card template, on specific AI use cases.

Medium term: Once the system card template has been trialled, further implementation steps are required to ensure its efficacy on a larger scale:

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, 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 and data provenance).
Develop exemplar assurance cases across a range of case studies to explore more specific recommendations for real-world AI use cases (e.g. 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. These clauses may cover topics such as the ability to conduct audits and spot-checks and data provenance.

For more detail on these recommendations, including the full AI assurance framework for UK national security, please see "Assurance of third-party AI systems for UK national security".

[1] Nii Simmonds and Alice Lynch, "Mitigating supply chain threats: building resilience through AI-enabled early warning systems," *CETaS Expert Analysis* (January 2023).
[2] Ian Brown, "Expert Explainer: allocating accountability in AI supply chains," Ada Lovelace Institute Paper (June 2023), https://www.adalovelaceinstitute.org/resource/ai-supply-chains/.
[3] 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
[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.
[5] GCHQ/Bailo, "Bailo – managing the lifecycle of machine learning to support scalability, impact, collaboration, compliance and sharing," GitHub, https://github.com/gchq/Bailo.