# Building The UK's AI Assurance Infrastructure

*An Independent Verification Organisation (IVO) Ecosystem as the UK's Strategic Lever as a Middle Power in AI Governance*

*Co-Authored by: Sarah Fu, Taro Qureshi & James White*

## Executive Summary

Middle powers face a structural problem in AI governance. The United States and China dominate frontier compute and model development, and the European Union is converting its market scale into regulatory export through the EU AI Act. States like the UK that cannot compete on either dimension risk becoming rule-takers, absorbing governance costs they did not design and forfeiting the returns that flow to those who set the rules.

This paper argues that the UK's strongest available response lies at the deployment layer. This is the institutional architecture that determines how AI systems are verified as trustworthy once they are put to real-world use across healthcare, finance, public services, and other sectors. The central proposition is that an ecosystem of Independent Verification Organisations (IVOs), accredited third-party bodies that assess how AI systems have actually been integrated into operational contexts, represents the UK's most cost-effective marginal contribution to AI governance. It is cost-effective because the institutional foundations are largely in place. The National Physical Laboratory provides measurement science, the British Standards Institution provides standards governance, the United Kingdom Accreditation Service provides accreditation, and the AI Standards Hub provides coordination. Building the ecosystem is primarily a matter of connecting bodies that already exist and reforming their mandates, rather than constructing institutions from scratch.

The case rests on three claims. The assurance ecosystem unlocks the value of investments the UK has already committed elsewhere, because compute, adoption funding, and data infrastructure generate full returns only when deployers, insurers, and procurement officers trust AI systems enough to adopt them at scale, and the absence of a credible quality signal appears to suppress that adoption where no downstream party requires it, a claim the paper treats as conditional rather than settled and flags for further research. Domestic capacity must precede international influence, because coalition-building is slow and a working ecosystem lets the UK arrive at coordination tables with a product rather than a position. And the window is closing, because the EU's conformity assessment machinery and US-based private standards efforts are already generating their own ecosystems, while first-mover advantage at the standards layer is durable as compliance infrastructure builds around whichever benchmark establishes itself first.

The proposed architecture is a pipeline of five connected functions: measurement, standardisation, accreditation, verification, and feedback. Four of them map onto institutions the UK already hosts. The fifth, the function that aggregates deployment failures across sectors and

routes them back into measurement and standards revision, does not yet exist anywhere in the UK system, and no country currently operates a functioning mandatory cross-sectoral AI incident reporting regime. This pipeline supplies the professional infrastructure of assurance. It operates alongside a parallel accountability infrastructure of procurement, due diligence, insurance, and liability, through which assurance findings acquire commercial and regulatory force. Infrastructure alone does not generate influence, rather, it does so only when activated by demand, which is why the paper treats incentive design as inseparable from architecture.

One architectural question is worth surfacing here rather than leaving in the body. The pipeline produces credible assessments, but the artefact through which those assessments are communicated to procurement teams, regulators, insurers and the public is not yet specified, and the choice will materially shape how the architecture transmits into commercial and regulatory decisions. The paper suggests three potential options. A static certification mark is the most legible to procurement and the cheapest to operate, but it ages quickly in a domain where systems are updated continuously. A continuously-updated disclosure preserves currency but requires standing infrastructure to maintain and consume. A structured filing analogous to sustainability reporting sits between the two and aligns with existing investor due diligence practice, but takes longer to embed. The choice is not made in this paper, but it is the kind of decision that, once taken, ostensibly locks in for a long time, which is why it is flagged in advance rather than as an afterthought.

This is a plausibility demonstration rather than an empirical study. It sets out to show that a deployment-layer strategy is coherent, grounded in the UK's real political and economic constraints, and worth pursuing, not to unequivocally prove demand through primary research or to resolve every design question it raises. The assumptions on which the argument rests, as well as areas for further research are set out in the Annex. The proposal’s strategic claim is narrow and firm: the UK is unusually well placed to build this infrastructure, and the cost of beginning is low relative to the cost of letting others define the benchmark first.

**Recommendations**

Five core actions follow from the analysis. The first three require no new primary legislation and can begin immediately.

1. **Issue a DSIT directive formalising the AI Standards Hub’s coordination role.** This is the lowest-cost intervention available to make the pipeline functional and should be the first, requiring neither new legislation nor significant capital expenditure.
2. **Develop a BSI Flex document on AI deployment standards.** None currently exists. A Flex provides the interim operating mode while the slower ISO/IEC cycle runs, directed by NPL’s emerging sector-level metrics, and can later be elevated into a PAS, British Standard, or international new work item proposal.
3. **Secure DRCF agreement on a common AI assurance reference standard.** A joint statement of AI deployment expectations referencing a single accreditation standard prevents sectoral regulators fragmenting into divergent regimes, while each retains full enforcement discretion within its own mandate.

4. **Use the Cyber Security and Resilience Bill as the legislative model for mandatory AI incident reporting.** Its two-stage notification architecture, and the prospect of extending reporting duties through secondary legislation tied to designated sectors, offers a statutory compulsion power without standalone AI-specific primary legislation, subject to developing AI-specific thresholds.
5. **Engage the AI Growth Lab consultation now.** Make the case during the consultation, before the operating model is locked in, that the incident-reporting aggregation function is a candidate purpose for the Lab and that the design constraints drawn from the ISU and COBR should bind whatever institutional home emerges.

# Introduction

Middle powers face a structural problem in AI governance. By middle power we mean a state with the technical capacity, institutional depth and convening credibility to shape outcomes in a given domain, but without the compute, capital, or market scale to set the terms unilaterally. The category binds in AI because the two routes to setting terms, building frontier capability and wielding market gravity, are dominated by actors a middle power cannot match on either axis. The strategic question is therefore not whether to influence the regime but how that may be possible. This means answering which layer of the governance stack offers a lever proportionate to the assets a middle power actually holds.

The US and China dominate frontier compute and model development. The EU is converting market scale into a regulatory export through the EU AI Act. States like the UK that cannot compete on either dimension risk becoming rule-takers, absorbing governance costs they did not design, and losing the economic and strategic returns that flow to those who set the rules. The UK's AI Opportunities Action Plan commits substantial resources to compute, data access, and sectoral adoption. These are necessary supply-side investments, but they compete on terrain where the UK faces structural disadvantages. The question is: where can the UK act largely unilaterally, at low relative cost, in a way that compounds the value of these investments while building durable international influence?

The strategic problem this creates for middle powers — and the range of responses available to them — has been mapped by Chatham House and Anton Leicht among others. This paper argues that the answer for the UK is by building verification and assurance infrastructure at the deployment layer. This is the institutional architecture that determines how AI systems can be evaluated as trustworthy, as they are put to real-world use across healthcare, finance, public services, and other industries. We argue that an Independent Verification Organisation (IVO) ecosystem represents the UK's most cost-effective path to durable governance influence. It is also where the UK holds a strong existing hand: institutional assets that most jurisdictions lack, a deployment lead in critical sectors, and legal and commercial infrastructure with genuine global reach. The institutional foundations to build an IVO ecosystem are already largely in place, and crucially, doing so is the condition under which the UK's existing investments in compute and data generate returns at scale.

The UK's establishment of AISI in 2023 ahead of the Seoul and Paris summits and before most peer governments had acted in the AI safety space, gave it convening authority and technical credibility. That head start is being threatened as the governance space is consolidating: the EU's conformity assessment machinery is already generating its own assurance ecosystem, and US-based private standards efforts are accelerating. The window in which the UK can shape the benchmark is open but closing.

We make three supporting claims:

- **The IVO ecosystem serves a dual purpose**. It not only creates durable infrastructure that encourages safer AI deployment domestically, but also crucially consolidates the UK's influence as rule-setter, rather than rule-taker, on the international stage. Early standard-setting on audit methodology, access protocols, evaluator accreditation, and outcome-metrics creates lock-in effects as compliance infrastructure is built around established benchmarks, furthering the UK's position, as well as meaningfully contributing to global AI safety.
- **This is the UK's most cost-effective marginal strength.** The institutional infrastructure largely already exists: NPL for measurement science, BSI for standards governance, UKAS for accreditation, the AI Standards Hub for coordination, English commercial law and the London insurance market for private enforcement. Creating an IVO ecosystem is primarily a matter of connecting existing bodies and reforming mandates, rather than building institutions from scratch, though the standards those bodies must produce remain demanding technical work. The £11M AI Assurance Innovation Fund and consortium-building commitments in the Trusted Third-Party AI Assurance Roadmap are early steps, but they are framed as market-support measures. We argue they should be elevated to a primary strategic lever.
- **The UK should build domestic assurance capacity first.** Middle-power coalition-building on AI governance is slow, and relies on a climate of collaboration and trust, which cannot be guaranteed in the short-term. Domestic institutional capacity builds credibility in influencing international governance. By demonstrating a flourishing domestic assurance ecosystem, the UK can arrive at international coordination tables with a working product.

# Section 1: The Strategic Case for Building the UK's Independent IVO Ecosystem

## 1.1. The UK's Marginal Strength is at the Deployment Layer

The UK cannot match the US or China at the development layer, and it cannot match the EU at the regulatory layer. The US and China host roughly 74% and 15% of global compute respectively, with the UK below 1%, and a 1GW data centre alone costs an estimated $30B before the model is trained or powered. The UKRI's commitment of ~£1B to the AI Research

Resource (AIRR) compute ecosystem is insufficient to close the capability gap. On the regulatory layer, the EU's Brussels Effect rests on market scale and a binding legislative regime the UK cannot replicate unilaterally. Both are terrain on which the UK is, structurally, a follower.

By contrast, the deployment layer presents an opportunity. It is the layer at which AI systems are integrated into healthcare, financial services, public administration and other operational contexts, and on which the question of whether those systems can be trusted is settled in practice. It is also the layer on which the UK has unusually strong existing assets, none of which require significant new capital expenditure to put to work. Those assets fall into three categories that map onto the architecture this paper proposes: supply-side institutions that produce credible assurance, the operational evidence that makes the standards those institutions issue substantively meaningful, and demand-side mechanisms through which assurance findings translate into commercial and regulatory consequences for deployers. Section 2 sets out the supply-side pipeline. Section 3 sets out the coordination and demand-side levers. This section establishes the underlying assets present.

**Supply-side institutional infrastructure.** The UK already hosts the technical and standard-setting bodies a credible assurance regime requires. The National Physical Laboratory (NPL) holds the measurement science capacity to convert high-level safety goals into auditable engineering criteria, and its newly established Centre for AI Measurement is backed by £10.5m through DSIT's AI Assurance Innovation Fund. The British Standards Institution (BSI) led the development of ISO/IEC 42001, the world's first international AI management system standard, and in November 2025 UKAS became the first accreditation body certified to issue ISO/IEC 42001 certificates. AISI provides frontier evaluation methodology, including the Inspect AI Framework which functions as an internationally recognised reference for how AI benchmarks are run and reported. Its establishment followed a deliberate strategic choice to convene the world's first major AI Safety Summit at Bletchley Park in 2023, moving ahead of peer governments to claim an institutional role before the governance landscape consolidated — and that positioning has since translated into concrete access: pre-deployment testing agreements with Anthropic, OpenAI and Google that no other middle power has secured. The AI Standards Hub provides a coordination forum that could be empowered into an operational role. Most jurisdictions hold one or two of these institutional pieces, whereas the UK nearly has the full set.

Whilst this presents a real advantage for the UK in terms of institutional coverage, possessing the bodies that can write standards is not the same as having the standards themselves, with the technical content forming the greatest challenge. ISO/IEC 42001 illustrates both sides of this. BSI led its development and it became the world’s first international AI management system standard, which demonstrates the UK’s convening capacity. Yet it took years of committee negotiation to produce, and what it produced is an organisational management standard rather than the deployment-specific technical standards this pipeline depends on. Writing those is slower, more contested and more technically demanding than the existence of the institutions implies. The UK’s advantage is therefore best understood as a head start on the institutional preconditions for credible standards, not as a solved standards problem. Section 2 details why this makes the measurement-to-standards linkages the decisive constraint.

**Operational deployment lead.** The UK is a leader in AI deployment in several critical domains: 75% of financial services companies reported adopting AI into their operations in 2024, 68% of NHS organisations were in 'early-stages' of AI system adoption by 2025; and UK public bodies were awarded 1,690 AI contracts worth a total of £4 billion between 2018 to 2026.The scale of adoption generates valuable, sector-specific operational data about how AI behaves once deployed. This is the input that standards need to be credible by reflecting real operational conditions. Data around the deployment experience in different sectors is also a direct input into the planned UK National Data Library, intended to drive the UK's global leadership in data-driven innovation. The NDL's governance layer, with its curated access, clearer permissions, authoritative datasets difficult to assemble elsewhere, is directly complementary to the IVO ecosystem: both are concerned with creating the trusted foundations on which deployment at scale becomes possible. The UK's professional services density across financial services, law and clinical practice provides the domain expertise to make those standards operationally meaningful. In addition, it provides the institutional partners like FCA-regulated firms, actuarial bodies, law societies, whose co-production of standards would lend practitioner legitimacy and a natural dissemination channel. This input quality is what differentiates UK-anchored standards, informed by real operational data on AI performance post deployment.

**Demand-side transmission mechanisms.** English commercial law governs approximately 40% of global commercial contracts. This creates a transmission mechanism for UK-anchored standards that operates through private commercial relationships instead of regulatory mandate. Model contract clauses requiring independent AI assurance as a condition of reasonable care could extend the reach of UK assurance standards into global commercial practice wherever parties choose English law. The London insurance market reinforces this. As the global reference for novel risk pricing, Lloyd's and the London market are well placed to develop the AI risk pricing frameworks that would make IVO engagement a commercial necessity creating demand-side pull. Both operate on longer timescales than supply-side institution-building, but neither requires legislative action to begin functioning. This is discussed further in Section 3.

## 1.2 The Case for Elevating Assurance from Market Support to Strategic Lever

The UK's existing AI strategy treats assurance as one of several competitiveness measures. The £11M AI Assurance Innovation Fund and the consortium-building commitments in the Trusted Third-Party AI Assurance Roadmap are framed as ways to grow a £1.01 billion market projected to reach £6.53 billion by 2035. We argue this framing understates the strategic position assurance occupies, for three reasons.

> **1. The assurance ecosystem unlocks the value of investments the UK has already committed to elsewhere.** Compute, sectoral adoption funding and data infrastructure generate full returns only when deployers, insurers and procurement officers trust AI systems enough to adopt them at scale. The harder question is whether genuine

demand exists or whether observed activity reflects firms adopting AI defensively while assurance lags behind. We cannot resolve this with the evidence available, and establishing it properly would require dedicated market research beyond this paper's scope. What can be said is more qualified. A sizable assurance market already exists, with DSIT identifying 524 firms supplying assurance goods and services in 2024, and DSIT's market analysis attributing uptake among purchasing firms to liability concerns, reliability issues and the absence of credible quality signals. Industry evidence suggests demand is real but uneven. TechUK's account of the market identifies three recurring drivers, namely the need to justify trust in order to drive adoption, compliance with sectoral duties of care and existing regulation, and access to capital through procurement and investor due diligence.[1] The same account is candid about the limits, since some firms keep assurance internal and engagement falls away where the buyer or investor in a procurement or investment conversation does not ask for it. Demand is therefore strongest where a downstream party, whether a regulator, an insurer, an investor or a counterparty, creates a reason to seek it, and weakest where no such party does. This is consistent with the paper's central claim. The binding constraint on demand becoming durable is the absence of a credible quality signal that procurement teams, insurers and regulators can rely on, which is precisely the function a common technical floor provides. Without it the market risks crystallising into compliance theatre, with an assessment by one of those 524 firms not reliably comparable to another's.

**2. Domestic capacity precedes international influence.** Middle-power coalition-building on AI governance is slow and depends on conditions of collaboration that cannot be guaranteed in the short term. A working domestic ecosystem allows the UK to arrive at international coordination tables with a working product, and creates the operational evidence base that international standards work requires. Section 3.2 sets out three concrete transmission channels through which that domestic capacity converts into international influence: the ISO/IEC standards pathway, EU AI Act compatibility-by-design, and English commercial law through model clauses and Lloyd's policy terms. None of these requires the UK to wield superpower leverage or EU market scale to operate.

**3. The standards-setting window of opportunity is closing.** The EU AI Act's conformity assessment requirements are generating their own assurance ecosystem, the US is developing evaluation infrastructure through nonprofits including AVERI, and OpenAI has called for the US's Center for AI Standards and Innovation to lead on frontier auditing standards. First-mover advantage at the standards layer is durable precisely because compliance infrastructure – accreditation pathways, contract templates, regulatory references – is built around whichever standard establishes itself as the credible baseline. Displacing an established standard is costly for all parties, so those who arrive second tend to adopt rather than revise. The UK's deployment lead

[1] Interview with Tess Buckley (Senior Programme Manager, Digital Ethics and AI Safety at TechUK), 6 May 2026; corroborated by TechUK, 'A Maturing AI Assurance Ecosystem' (December 2025).

provides operationally informed data the other contenders lack, but that lead is a depreciating asset.

The implementation path this argument implies runs through existing sectoral regulators instead of new horizontal AI legislation. Sectoral regulators hold the statutory authority to reference UKAS-accredited assurance as a compliance expectation within their existing mandates, and the DRCF provides the coordination forum through which their expectations can be aligned on a common reference standard and prevent fragmentation into divergent regimes. Mandatory incident reporting and statutory safe harbours are the medium-term legislative pieces this architecture requires, but the near-term work is coordination and mandate clarification and does not require any new primary legislation. This implementation path, and the levers that operate within it, are the substance of Section 3.

# Section 2: The IVO Ecosystem - Architecture and Minimum Viable Design

## 2.1 The Pipeline

Translating the UK's deployment-layer advantage into durable governance influence requires building an architecture credible enough to support it. This section sets out what that architecture is.

The UK's deployment-layer assurance infrastructure, as proposed here, is a pipeline of five connected institutional functions: measurement, standardisation, accreditation, verification, and feedback. Together these functions make it possible to verify whether AI systems are being deployed safely and in accordance with applicable standards, and to keep that verification current as deployment conditions change. Four of the five functions map onto institutions the UK already hosts, with NPL hosting measurement, BSI standardisation UKAS accreditation and the emerging IVO market verification. What no jurisdiction yet has is the feedback function: a body that aggregates deployment failures across sectors, synthesises them into discernable patterns, and routes them back into measurement and standards revision. That gap is the subject of Stage 5.

The pipeline's logic is sequential, and the robustness of the pipeline is spread throughout all the components without one node exerting undue influence:

1. Safety goals must be translated into observable metrics before they can be encoded as standards.
2. Standards must exist before verifiers can be accredited against them.
3. Verifiers must be accredited before they can produce credible assessments.
4. Assessments must generate feedback so the system can remain agile and integrate learnings.

Assurance is a nascent market that already exists. DSIT's November 2024 analysis identified 524 firms supplying AI assurance goods and services in the UK, generating an estimated £1.01 billion in gross value added and employing approximately 12,500 people, with the market projected to grow to £6.53 billion by 2035 if quality and adoption barriers are addressed. Without common standards, information-sharing practices and a shared technical floor, an assessment by one firm cannot be relied upon as equivalent to an assessment by another, and procurement teams, regulators, and insurers cannot use assurance findings as a basis for decisions that need to scale. The pipeline's purpose is to make the existing market comparable and credible without compressing the methodological diversity that has emerged in response to legitimate sectoral and use-case variation.

Independent Verification Organisations sit at Stage 4 of this pipeline. They are the accredited third-party bodies that engage directly with deployers to examine how AI systems have been integrated into operational contexts, and produce independent technical assessments against published standards. In practice this means assessing systems at the point where a hospital has integrated a diagnostic tool, where a bank has deployed a credit-risk model, or where a law firm is using AI to review contracts. An IVO examines what safeguards govern the system's operation, what decisions it is authorised to make, what data it draws on, and what mechanisms exist to detect and respond to failures. The output is an independent technical assessment, produced by a body with no commercial interest in the outcome, against standards the IVO did not set and cannot revise.

Four structural features of the architecture are worth establishing here:

**1. The object of assessment is the deployment, not the underlying model.** The choices a deployer makes about integration, configuration, access, and accountability cannot be evaluated by testing the AI model in isolation. An IVO assesses how a specific organisation has embedded a specific AI system into a specific set of processes, and whether those choices meet the applicable standards for that sector and use case. This is what makes the pipeline tractable without requiring development-layer access. A residual dependency on version integrity, the question of whether the model evaluated is the model being run, is an open research question addressed in Annex - A23.

**2. The architecture's independence rests on structural separation.** The bodies that set the technical content of standards are not the bodies that verify against them, and the bodies that verify are not the bodies that accredit verifiers. NPL and BSI set the standards. UKAS accredits IVOs against those standards. IVOs operate above a fixed technical floor with no authority to revise it. This separation is the architecture's primary independence feature In Dean Ball's private governance proposal, private bodies authorised by government provide certifications to frontier AI developers, with the race-to-the-bottom restrained by the government's power to revoke a private body's licence if negligence or misfeasance is found. That mechanism operates as a deterrent after competitive pressure has already shaped behaviour. The UK pipeline's answer is structural rather than deterrent-based: the race to the bottom on standards development cannot begin because standards themselves are not in the market. The separation is also the architecture's primary coordination problem. Five institutions performing five

connected functions cannot align themselves on outputs, schedules, or sectoral priorities through the pipeline structure alone. How that coordination is achieved without compromising the independence the separation creates is addressed in Section 3.

**3. External assurance complements rather than replaces internal capacity.** The architecture does not assume that all assurance is or should be external. It assumes that *credible* external assurance must exist as an option, and that the existence of that option strengthens the quality and comparability of internal practice as well. Many UK firms operate internal assurance teams and will continue to do so. The IVO ecosystem provides a credible external floor against which internal assurance can be benchmarked, and through which deployers can demonstrate independent verification when sectoral regulation, procurement, contractual demand, or insurance pricing requires it. A parallel tension runs one level up, on the state's own side. Government is not only a buyer of assurance but also a maker of it, through bodies such as GDS, the Incubator for AI, OPSS and the MHRA, and how it balances building its own capacity against seeding an external market is a demand-side question addressed in Section 3.3.

**4. Verified assessments must be communicable to be useful.** DSIT defines AI assurance as the process of measuring, evaluating and communicating risk mitigation. An IVO assessment produces information that has to travel from the assurance firm to procurement teams, regulators, insurers, deployers, and in some cases the public. The pipeline as set out below produces evaluation and measurement criteria to assess risk. It does not yet specify the artefact through which those findings are communicated, whether that takes the form of a static certification mark, a continuously-updated disclosure, a structured filing analogous to sustainability reporting, or some other instrument. The form the communication artefact takes will materially affect how the pipeline transmits credibility into commercial and regulatory decisions, and is treated as a research priority in Annex - A27.

This pipeline forms the professional infrastructure of the assurance ecosystem: the institutions, standards, and verification practices that determine how an AI deployment can be credibly assessed. It operates alongside a parallel accountability infrastructure of procurement gatekeeping, investor due diligence, insurance pricing, and liability allocation, through which assurance findings translate into commercial and regulatory consequences for deployers. Section 3 addresses how the architecture transmits into that broader infrastructure. The pipeline's job is to produce assurance findings credible enough to do that transmission.

The ecosystem does not replace sectoral regulators or AISI. The FCA, MHRA, Ofcom, and ICO retain their statutory authorities. The pipeline provides the technically credible floor that regulators can mandate, reference, or incentivise organisations under their jurisdiction. AISI's frontier evaluation work feeds into the standards production in Stage 1 but operates on a different timescale and with a different remit.

AISI is the one institution that could perform several pipeline functions at once, with the depth to measure, to inform standards and to evaluate, and that is precisely why it should operate none of them. Its standing comes from being a government evaluator focused on frontier risk, and

that standing is the asset the separation protects. An evaluator with no stake in the verification market produces standards inputs that carry more weight through their outsider placement. Conscripting AISI into accreditation or verification would not extend that asset across the pipeline, but rather it would spend it down, as credibility would not survive being both the assessor and the assessed. AISI therefore sits alongside the pipeline, with its frontier-evaluation methodology feeding the standards the pipeline produces without it accrediting, verifying or operating any stage. The input relationship into Stage 1 is therefore intentional and contained. Yet, the residual question is whether it stays an input, since methodology that the pipeline adopts could make AISI a *de facto* standard-setter intentionally, which is the proximity worry flagged for further work. A second open question is how AISI develops methodology a market it sits outside comes to rely on without thereby gaining undue influence over it.

The figure below shows how the different stages of the pipeline interact, and the relevant established entities governing activity in each stage. The fifth feedback stage remains ungoverned, and is key to returning incident data to inform measurement criteria in stage 1.

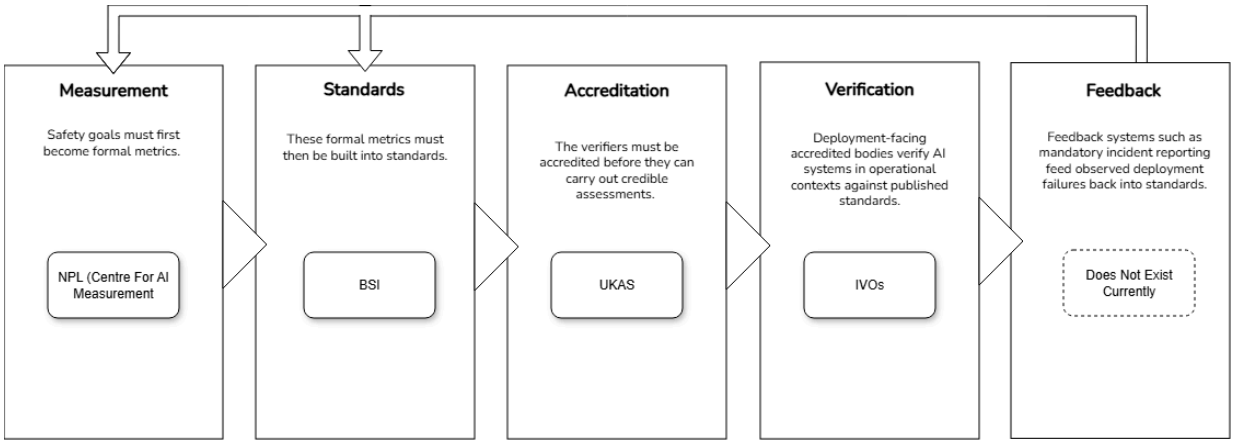


*Figure 1. IVO Pipeline*

The remainder of this section sets out each stage in turn: what each institution's role should be, where intervention is needed to make it operational, and, in the case of Stage 5, what institutional design questions remain open.

## *Stage 1 - Measurement: National Physical Laboratory*

Before a standard can be written or an IVO can assess against it, high-level safety goals have to be translated into observable, auditable engineering criteria: specific, testable metrics that operationalise what safe deployment means in a given sector and use case. This is not itself a standards function, but rather it is the prior technical step that makes standards possible. The Fathom/IASEAI workshop, a full-day session convened by the non-profit Fathom at the International Association for Safe and Ethical AI conference in Paris in February 2026, identified this translation capacity, converting Level 1 outcomes into auditable Level 2 metrics, as a core

requirement of any credible IVO system and one of the least developed components of current practice. The same workshop ran an afternoon tabletop crisis simulation, referenced later in this paper, in which government, IVO and AI-company role groups worked through a scenario of youth self-harm incidents linked to a deployed chatbot.

NPL's newly established Centre for AI Measurement is the natural institutional home for this function. The Centre is backed by £10.5m awarded through DSIT's AI Assurance Innovation Fund, which itself sits within the Trusted Third-Party AI Assurance Roadmap, and represents the most immediate and concrete investment in the pipeline. Whether its budget allocation goes toward deployment-specific, sector-relevant metrics or toward broader foundational research will shape what the rest of the pipeline can do in the near term. If NPL's direction of travel is toward deployment-specific metrics, it represents the clearest near-term opportunity to give the pipeline traction.

One operational question NPL's metric-development work will need to engage with directly is what repeated reassessment means in measurable terms across the range of cadences from periodic to continuous, and how this connects to the consistent backbone of incident reporting introduced at Stage 5. Practitioners across the existing market currently use 'iterative' to cover a range of approaches, from quarterly review cycles to continuous monitoring with threshold-triggered alerts. Without a measurement framework that defines what these mean in specific deployment contexts, downstream stages of the pipeline cannot accredit against them or assess for them. This is treated as an open question.

### *Stage 2 - Standardisation: British Standards Institution (BSI)*

Once NPL has produced auditable metrics, BSI encodes them in formal standards through the ISO/IEC pathway, transforming domestically developed methodology into internationally recognised compliance infrastructure. This is the step that gives IVO assessments their legal and commercial weight outside the UK. BSI has already demonstrated this pathway in an AI context: ISO/IEC 42001, developed under BSI leadership, is the world's first international AI management system standard, and in November 2025 UKAS became the first accreditation body certified to issue ISO/IEC 42001 certificates.

ISO 42001 is fundamentally an organisational management standard rather than a deployment-specific one. The pipeline requires a further generation of standards specifying what safe AI integration looks like in healthcare, financial services, legal practice, and other deployment contexts. BSI's current AI standardisation work operates at the organisational level. Directing it toward deployment-specific standards requires deliberate coordination with NPL's sector-level metric development, so that what NPL produces is designed from the outset to be encodable in BSI standards. This is the upstream coordination problem addressed in Section 3.

Where AI capability advances faster than the formal ISO/IEC cycle, which typically runs across multiple years, BSI Flex provides the pipeline's interim operating mode. BSI Flex standards are designed to develop consensus-based good practice that adapts to fast-changing markets, enabling iterative development to fulfil immediate needs. **No BSI Flex document on AI**

**deployment standards currently exists, and developing one is itself a concrete near-term action the pipeline could take, directed by NPL's emerging sector-level metrics.** The instrument's value extends beyond interim guidance: a Flex can be considered for further development as a PAS or British Standard, or constitute part of the UK input into the development of a European or international standard. This makes it a natural vehicle for submitting new work item proposals to ISO/IEC JTC 1/SC 42, the committee responsible for AI standards, through which BSI is directly positioned to act as the UK's national standards body and ISO member. Compliance with a Flex should not, however, confer immunity from legal obligations, which reinforces the case for the formal elevation pathway rather than treating Flex as a permanent substitute. Whether this model holds under conditions of rapid capability advancement is identified as a research priority in Annex - A31.

### *Stage 3 - Accreditation: United Kingdom Accreditation Service (UKAS)*

UKAS accredits IVO firms against BSI-published standards, providing the structural separation between standard-setting and verification that the pipeline's independence depends on. This is the mechanism that prevents market competition among verifiers from eroding the standards they assess against. Brundage et al. identify this risk as a structural weakness of company-selected, company-paid assurance arrangements, and it surfaced in the Fathom/IASEAI tabletop simulation, where participants raised the 2008 credit-rating-agency dynamic, in which assessors paid by the entities they rate face structural pressure that undermines rigour, as the dominant mental model for IVO capture risk. The Fathom participants concluded that if the IVO is paid by the company it verifies, the same dynamics could emerge. UKAS's certification as the first accreditation body authorised to issue ISO/IEC 42001 certificates demonstrates the function is already operational in an AI context. The task is extending it to cover the deployment-specific standards the pipeline will generate.

UKAS accredits the bodies that perform conformity assessment; it does not itself certify the individuals doing the work within them. Personnel certification is a parallel and currently unresolved question. Several bodies, including the IAPP, the BCS, the Association of AI Ethicists and the International Centre for Ethical AI, have a claim to professionalising the practitioner layer, and the question of which body or combination should certify practitioners is contested in ways the pipeline cannot resolve unilaterally. The architecture's position is that practitioner certification is a function the market needs, but that anointing a single professionalising body prematurely would foreclose legitimate methodological diversity.

A more tractable near-term route preserves the same separation logic. Rather than certifying individuals directly, UKAS could accredit personnel-certification schemes, operated by other bodies under the established ISO/IEC 17024 route, against defined competence standards, so that a designation such as ‘accredited AI red-teamer’ or ‘accredited AI evaluator’ carry independent meaning. This is plausibly faster than firm-level accreditation, not because it escapes the need for standards infrastructure but because the current provider market is immature and dominated by small red-team and boutique consultancies for whom an individual credential is more accessible than full organisational accreditation and for whom no settled

professional pathway yet exists. Two conditions keep this from reproducing the failure modes the pipeline exists to prevent. First, the scheme must certify demonstrated competence against a published standard rather than mere membership of an association, or it risks collapsing into symbolic credentialism rather than competence assurance. Second, accreditation must remain one layer up, with UKAS accrediting the certifying body rather than the practitioners themselves. UKAS is not assumed to be immune from institutional pressure, but its suitability rests on being structurally separated from commercial assurance provision, operating under internationally recognised accreditation norms, and already holding a quasi-public legitimacy role within UK conformity assessment.

### *Stage 4 - Verification: Independent Verification Organisations (IVOs)*

IVOs are the deployment-facing layer. They engage directly with deployers to examine how AI systems have been integrated into operational contexts and produce independent technical assessments against published standards. The UK already has a substantial and growing market of AI assurance providers, ranging from specialist boutique consultancies to large professional services firms with emerging AI assurance practices.

DSIT has acknowledged that existing professional skills in adjacent areas - cybersecurity, data science, software development, internal auditing, and privacy - provide natural pathways into AI assurance for practitioners already operating in the market. Brundage et al. describe the likely mature organisational form as multi-disciplinary: large professional services firms providing sector reach and scale, working through subcontracting arrangements with specialist AI evaluation organisations for technical depth. The UK's density of professional services expertise across financial services, healthcare, and legal practice positions it well for this model. Accredited IVOs would in practice differentiate above the UKAS-accredited floor by sector expertise, assessment methodology, and cost.

Yet, the market currently lacks the structure that would make this meaningful. In the AI Assurance Roadmap, DSIT identified the absence of clarity on skills, competencies, and information access requirements as the primary constraint on market quality. Without settled accreditation criteria, access protocols, and evidentiary standards, UKAS cannot accredit meaningfully and the existing provider base cannot credibly distinguish genuine assurance from compliance tick-boxing. The strategic priority at this stage is therefore feasible design before scaling. It is necessary to establish what a credible IVO assessment requires before the market's growth creates volume without meaningful substance. The unresolved question of what an IVO must demonstrate to satisfy a standard, and what evidence counts as sufficient, is a subject for further research.

The pipeline architecture is sector-agnostic in design but sectoral in operation, with Stage 1 metrics and Stage 2 standards differentiated by sector and Stage 4 verification carried out by IVOs with sector-specific expertise. Cross-sector knowledge sharing across this specialisation is itself a coordination problem that the pipeline's upstream institutions, working with the later proposed coordinator, will need to address and should be treated as a research priority.

The commercial sustainability of IVO services is an open question the architecture deliberately does not resolve. Who pays for verification, how pricing scales with deployment risk, and how the market avoids reproducing the credit-rating-agency capture dynamic in commercial practice rather than only in standard-setting are questions the pipeline raises but does not answer. The structural separation introduced at Stage 3 addresses the standards-side capture risk. The commercial-side risk requires its own analysis and is flagged in the Annex for further study.

### *Stage 5 - Feedback: Mandatory Incident Reporting*

The first four stages of the pipeline map onto existing UK institutions, but Stage 5 does not. Mandatory cross-sectoral AI incident reporting closes the pipeline loop, routing observed deployment failures back into the measurement and standards revision process and keeping standards current between formal BSI cycles. In practice the feedback is unlikely to follow a single path. An incident may reveal that an existing standard was set too loosely, in which case the signal routes to Stage 2, but it may equally reveal that the deployment failed in a way the current metrics could not detect, which is a measurement gap upstream of the standard and routes to Stage 1. Mature incident-investigation regimes in aviation seemingly operate this way, with a single investigation generating findings at the level of measurement, standards and operating practice at once. The Stage 5 aggregation function is therefore what determines, from the pattern across incidents, whether a given failure mode reflects a measurement gap, a standards gap, or both, with the AI Standards Hub coordinating revisions that touch both stages so they remain consistent. Not every reported incident should feed back, since a purely local failure with no systemic lesson would otherwise drive standards churn, so the aggregation function applies a materiality threshold distinct from the reporting threshold below. Without such feedback, the pipeline produces point-in-time assessments that age the moment they are issued, with no systematic mechanism for continuous monitoring. This is the proposal's most open institutional design question and carries a problem the others do not: the feedback loop only works if deployers are required to report, yet no body in the UK currently holds the statutory power to compel them, and a function that nothing requires will not establish itself on its own. The power it needs is the kind that comes from designating a regulated category in law rather than from a new standalone mandate, and Section 3.2 sets out how an existing piece of legislation could supply it. It is also the only stage that requires meaningful new capital expenditure rather than the coordination of existing capacity.

Stage 5 closes the pipeline loop in two ways that operate together. Mandatory incident reporting captures discrete deployment failures and feeds them into measurements and standards revision. It applies consistently across all deployments above a defined threshold, regardless of sector or risk profile, and is the architectural backbone of the feedback function. Sitting alongside it is the practice of repeated reassessment, which can take the form of periodic iterative reassessment, continuous monitoring of telemetry data, or a combination of both depending on the deployment's risk profile and the measurement or standard against which it is assessed. This second layer is where the cadence and intensity of feedback varies. Incident reporting does not. The relationship between iterative reassessment and continuous monitoring,

including how each is triggered, who provides the data each requires, and how their costs are distributed, is treated as a research priority in Annex - A19.

### Function

Stage 5 has to do three things. The Centre for Long-Term Resilience (CLTR) identifies these as gaining real-time visibility of AI safety risks in operational contexts, coordinating rapid responses to major incidents and generating cross-sectoral learnings from their root causes, and identifying early warnings of larger-scale harms for use in risk assessments. These operate on different timescales: rapid response runs on hours and days, early warning runs on months and quarters, real-time visibility runs continuously. No standards revision cycle moves at any of these speeds, which is why the function cannot be folded into Stages 1 and 2.

Sectoral reporting infrastructure already exists in mature form in electricity, nuclear, and aviation, and the MHRA's AI Airlock and the FCA's regulatory sandbox provide deployment-context testing environments whose outputs could in principle feed upward. The problem is therefore not that sectors are unwilling or unable to report. It is that no body in the UK system is funded, mandated, or technically equipped to aggregate those reports across sectors, synthesise them into actionable patterns, and feed the result back into the standards pipeline. To date, **no country has a functioning mandatory cross-sectoral AI incident reporting regime, which means the UK is designing this function from scratch rather than adapting an existing model**.

A parallel feedback loop runs through the insurance market. Insurers evaluating an organisation's assurance practices when pricing coverage create an economic feedback signal on assurance quality that operates independently of regulatory mandate, and Buckley and Szpruch have argued that this insurance-assurance interaction is one of the most underappreciated mechanisms in AI governance. Specialist AI insurance is an emerging market, with companies pricing risk on AI deployments, and academic work by Szpruch and colleagues has sought to formalise the actuarial and structural foundations of this market. Incident reporting and insurance loss data cover overlapping but defined failure modes. Incident reporting captures deployment failures with public-interest significance, while insurance loss data captures claims-related events with commercial significance. How these two information flows interact, and whether the Stage 5 aggregation function should be designed to integrate insurance data alongside regulatory incident reports, is an open architectural question.

## Section 3: Coordination And Incentives - Making The Architecture Work

The pipeline architecture described in Section 2 is a supply-side proposition. It creates the infrastructure for credible AI deployment assurance, but infrastructure alone does not generate governance influence. The pipeline only produces durable influence if it is activated by demand. This requires three things operating in combination:

1. **Institutional coordination to ensure that the pipeline's five stages are actually aligned**: that NPL's measurement science feeds BSI's standards process, that BSI's standards are coherent and sectorally applicable, and that the whole system learns from deployment feedback gained from mandatory incident reporting with the right mandate and institutional home.
2. **Regulatory floors that make assurance non-optional in priority sectors**: so that deployers must genuinely engage with AI deployment assurance and the system does not devolve into tick-box exercises.
3. **Commercial incentives that make choosing to engage with assurance worth the cost:** through insurance pricing, contractual requirements, and liability structures that reward genuine verification over tick-boxing.

Evidence from other assurance regimes consistently reveals that voluntary, market-selected assurance reliably produces lower-quality results and less stickiness than mandated, independently structured assurance. A crisis simulation run by Fathom modelling AI assurance market failure scenarios, reproduced the 2008 credit-rating-agency dynamic in an AI assurance context, with participants repeatedly reaching for the financial-crisis analogy of assessors paid by the entities they rate. It confirmed that structural separation is necessary but not sufficient. Any governance architecture that relies on market self-selection to produce rigorous assurance is designing against the evidence. The absence of credible certification suppresses AI adoption if uncertainty about quality and risk depresses uptake across the board. Building IVO certification capacity directly reduces that bottleneck. The UK's commitments to compute, data access and sectoral adoption programs only generate full returns if the trust infrastructure exists to convert supply-side capability into deployment at scale.

According to an LMA-Barnett Waddingham survey on AI risk management, based on responses representing more than 60% of Lloyd's market stamp capacity, the majority of firms have, or are already developing a structured AI framework. Risk pricing frameworks for AI would likely stimulate demand for the UK's AI assurance ecosystem by making model risk commercially legible. In the same way that Lloyd's market practices translate operational and cyber risks into underwriting decisions, AI-specific pricing would reward firms that can evidence stronger governance, risk tiering, human oversight, and data controls, while penalising opaque or poorly controlled systems. That creates a market signal for third-party verification, because organisations seeking lower premiums or wider cover would need external assessments, monitoring, and documentation to demonstrate that their systems meet a recognised standard.

## 3.1 The Coordination Challenge

The separation of the co-ordination bodies, whilst integral, creates a coordination problem. Five institutions performing five connected functions cannot self-align on outputs, schedules, or sectoral priorities. Without a body mandated to manage the flow between them, NPL develops metrics that BSI cannot yet encode, BSI writes standards that UKAS cannot yet accredit against, and IVOs operate in a market that lacks the consistent technical floor that would make their assessments comparable.

Coordination must occur at two levels: upstream, where the technical consensus on what adequate deployment-layer assurance means must be managed; and downstream, sector regulators must apply those principles consistently across industries, rather than allowing finance, health, transport, and other domains to develop divergent assurance regimes that raise compliance costs and undermine regulatory coherence.

### Upstream Coordination: The AI Standards Hub

The AI Standards Hub is the natural upstream coordinator. Its existing function, maintaining the global AI standards database, facilitating information exchange between NPL, AISI, BSI, and the Turing Institute, is precisely the coordination role the pipeline requires. What it currently lacks is an operational mandate from DSIT that makes its coordination role binding on pipeline participants. **Closing that gap does not require new legislation or significant capital expenditure. It requires a DSIT directive that formalises the Hub's role** as the body responsible for translating NPL's measurement science and AISI's evaluation outputs into BSI-legible standards, and for ensuring those standards remain internationally coherent. **This is the lowest-cost intervention available to make the pipeline functional, and it should be the first.**

### Downstream Coordination: The DRCF

The Digital Regulation Cooperation Forum shows potential as a downstream coordinator, and this is an underused institutional asset. The DRCF currently brings together the FCA, ICO, Ofcom, and CMA with a mandate for regulatory cooperation. Its data sharing provisions were extended under the Digital Markets, Competition and Consumers Act and the Online Safety Act. We propose that the DRCF members publish a joint statement of AI deployment expectations that references the single accreditation standard produced upstream by the Hub and BSI. Each regulator retains full enforcement discretion within its own mandate. What changes is that they align on a common reference standard for AI assurance, avoiding a situation where deployers operating across sectors face contradictory or duplicative compliance requirements. This is consistent with the 2022 government rejection of statutory DRCF powers: the proposal does not ask the DRCF to direct its members, only to coordinate on a shared reference point. The issue of anti-trust risk arising as a result of information sharing remains an issue for further investigation – referenced in Annex A1.

## 3.2 The Institutional Home for Mandatory Incident Reporting

The coordination mechanisms above address the pipeline's internal alignment problem. The legislative mandate for the mandatory incident reporting mechanism in Stage 5 is the pipeline's one genuinely open design question. We identify five institutional options, none of which is a complete fit. Rather than force a premature recommendation, this section does three things: a) draws design constraints from comparable mechanisms that should bind whatever solution emerges; b) assesses the live institutional candidates against those constraints; and c) identifies where active engagement is most productive now.

## Lessons From Comparable Mechanisms

The Investment Security Unit (ISU) and the COBR watchkeeper function present the strongest potential templates from which design constraints can be drawn.

The ISU sits within the Cabinet Office and exercises powers under the National Security and Investment Act 2021 to scrutinise acquisitions of control over qualifying entities or assets in 17 sensitive sectors. The ISU demonstrates the value of combining ministerial authority, all-source cross-departmental inputs, and rapid executive decision-making, rather than placing those powers in a regulatory body. However, ISU only functions because primary legislation specifies a nominated trigger event, a defined timeline, and a bounded consequence. AI incident reporting has none of those features in current law, meaning that the pipeline must work around or close this legislative gap.

The COBR watchkeeper offers a thinner template, but demonstrates that a small standing function can coordinate cross-departmental monitoring without statutory reorganisation. Both suggest that the aggregation function should sit close to executive decision-making not inside a regulator.

The Cyber Security and Resilience Bill, currently progressing through Parliament, is the closest legislative model for how mandatory AI incident reporting could be structured. The bill introduces mandatory incident reporting for operators of essential services and digital service providers through a two-stage structure: a light-touch 24-hour notification to the relevant regulator, with the National Cyber Security Centre informed simultaneously, followed by a full report within 72 hours. An AI equivalent could follow this model, with a sectoral regulator handling primary intake, and the central technical body receiving parallel notification for cross-sector pattern recognition. Two caveats apply. First, whether AI incident reporting can be dovetailed into the Bill's existing 24-hour notification architecture. The risk here is an AI deployment failure regime would need to capture a much wider class of harms than a cyber harm focused model. Second, even if the architecture transfers, the central technical body for AI does not yet exist in the form the NCSC plays for cyber - which is precisely the institutional design question this section is trying to resolve.

The Bill also offers more than a structural template. It points toward the most plausible near-term route for resolving the cold-start problem raised in Stage 5, where the feedback loop requires a compulsory reporting function that no institution currently possesses. The Bill expands incident-reporting duties across operators of essential services and regulated digital infrastructure, creating a statutory reporting architecture tied to sectors designated as critical to national resilience. Secondary legislation made under this architecture could extend AI deployment-reporting obligations to operators within designated sectors, supplying a statutory compulsion power without standalone AI-specific primary legislation. The limit is that this reaches only operators within those sectors, whereas the Stage 5 feedback function ultimately needs cross-sectoral aggregation.

## Potential Candidates

Against those constraints, three institutional candidates have emerged from our consultations. None is currently ready, but understanding the friction points and failure modes should inform where the UK should focus.

A central AI Risk Function within or adjacent to AISI is the most technically credible option, and the organisation already has the cross-government relationships an aggregation function would require. DSIT's earlier Central AI Risk Function (CAIRF) was designed to fulfil much of this role, feeding into the National Risk Register through a COBR lead-government-department architecture. However, reinstating CAIRF now faces two structural problems. The first is mission drift: AISI's voluntary, collaborative posture toward frontier developers does not sit easily with the authority a mandatory reporting function requires, and absorbing it would broaden AISI's remit at a moment when examples like Mythos suggest it needs to keep its narrow focus on frontier evaluation. The second is the political window: when CAIRF was more active, government departments had agreed certain risk domains and the COBR integration was the route into the National Risk Register. Whether those political conditions still hold is unclear. The function may belong somewhere in or near AISI in time, but recommending reinstatement now would be premature without a clear answer to where the statutory authority to compel reporting would come from.

Centre for Emerging Technology and Security (CETaS) at the Alan Turing Institute has topical alignment and the organisational relationships required, but faces a structural timing problem. The Turing Institute is in the middle of a UKRI midterm review that concluded in April 2026 that its strategic alignment and value for money are not yet satisfactory. A new CEO took over in May 2026 with a mandate to deliver a restructuring plan by September 2026. CETaS itself does not hold its own budget line but draws on the Turing's core funding through UKRI/Engineering and Physical Sciences Research Council (EPSRC), a five-year £100m package renewed in 2024. Recommending CETaS as host of a critical national infrastructure function during a period when the host institution is being restructured under public regulatory pressure would build institutional risk on what is already a difficult institutional design problem. The post-September 2026 settlement may make CETaS a stronger candidate, particularly given the Turing's intended pivot toward national security and sovereign capability, but the immediate question of whether the institute can host new operational responsibilities while it is being restructured remains to be seen.

The AI Growth Lab is DSIT's proposed cross-economy regulatory sandbox, currently at the call-for-evidence stage following its October 2025 announcement. Its proposed function is to apply time-limited, supervised modifications to specified sectoral regulations where these are judged to hinder AI adoption. There is a structural connection to the pipeline because a sandbox needs to know what it is measuring, which creates a direct link to NPL's measurement work and to the question of what AI incidents look like under live conditions. It is therefore plausible that the Growth Lab could host an aggregation function, since the data flows it would already need to sustain are similar in shape to the data flows incident reporting would generate. Three

uncertainties remain. Its mandate, governance, and operating model are still being shaped through consultation. Its primary framing is innovation acceleration rather than safety, which may sit awkwardly with the public-interest case for incident reporting. And it is not yet clear whether the Growth Lab will hold statutory authority of its own or operate through borrowed authority from sectoral regulators. The recommendation is therefore to engage the Growth Lab consultation on the explicit assumption that aggregation may be one of its functions, and to make the case during that consultation for the design constraints this section has identified. That is more productive than waiting for the consultation to conclude before engaging, and more honest than nominating the Growth Lab as the answer before its operating model is settled.

Yet, even the question of who should be engaging in the consultation is not immediately clear. The actors with a reason to engage are the same downstream parties the demand-side levers depend on. Insurers and reinsurers would gain the loss data that lets them price AI risk; large deployers in regulated sectors would gain a clearer reporting duty that bounds their liability; assurance providers and professional bodies would gain both a reference point and a customer for their work; and investors and procurers conducting due diligence would gain an authoritative signal they currently lack. Each of these is a party that creates demand for verification at the point where it asks for it, which is the mechanism Section 3.3 relies on. The implication is that this consultation is a venue in which that first-mover demand can surface, but is not a mechanism that manufactures it. If these actors do not yet perceive a return from the regime existing, their engagement will be thin, and that may itself be a reading on the strength of the underlying demand.

### Active Engagement

We deliberately leave the institutional home open. Three things follow from this that are actionable now:

1. **The design constraints drawn from the ISU and COBR be treated as binding constraints on whatever institutional design eventually emerges**
2. **The CSR Bill’s two-stage notification architecture should be used as the legislative model for how mandatory AI incident reporting is structured, subject to developing AI specific thresholds**
3. **The Growth Lab consultation should be engaged actively before the operating model is locked in, on the explicit terms that the incident reporting aggregation function is a candidate purpose for the Lab.**

### Resourcing

Three existing budget lines are worth interrogating before any new spending is proposed. The first is the £10.5m backing NPL's Centre for AI Measurement, awarded through DSIT's AI Assurance Innovation Fund under the Trusted Third-Party AI Assurance Roadmap. The second is AISI's broader research budget and its current capacity to run batteries of simulated failure modes by sector. The third is CETaS's existing AISI partnerships and its draw on the Turing Institute's £100m UKRI/EPSRC core funding.

None of these budgets currently funds the standing aggregation function Stage 5 requires. What does require new investment is a small, technically capable team with cross-departmental data access, statutory authority to receive mandatory disclosures, and a feedback channel into the NPL/BSI standards pipeline. Further work is needed to define that gap precisely and identify where existing resources can be redirected before seeking new capital, but that scoping work should proceed in parallel with the institutional home question.

## 3.3 Domestic Levers

The coordination mechanisms above address the pipeline's supply side architecture. The levers below address demand: the regulatory and commercial conditions that make assurance non-optional and worth doing well. They are organised by what is achievable within existing statutory authority, what requires new statutory provisions, and what depends on ecosystem maturity. The tiers are cumulative: commercial norm-setting and regulatory requirements are mutually reinforcing, and the immediate-tier levers create the conditions under which the statutory levers become easier to operationalise.

The UK government sits on both sides of this market. It produces assurance capacity directly, through GDS and the Incubator for AI building evaluation capability for the public sector's own AI use and through regulators such as OPSS and the MHRA performing assessment as a statutory function, while as a purchaser of AI it is also among the largest potential customers for external verification. These two roles can act in tension, since every function the state brings in-house is one the external market never has to supply, and an undifferentiated preference for internal capacity would narrow the market the pipeline is meant to grow. The tension manifests and resolves differently across bodies. Where the state's evaluation role is statutory, as with the MHRA, it is a legal duty rather than a market choice and cannot be routed outward. Where the state is buying AI for its own operations it has genuine discretion, and that buyer side has its own distinction, between purchasing as a demand signal and conditioning others' access, which the procurement levers below set out.

### 3.3.1. ***Immediate*** *(Existing statutory authority, no new primary legislation required.)*

#### *Regulatory Levers*

**Sectoral regulatory mandates** are the most immediately actionable forcing mechanism using existing statutory authorities. With **financial services**, the FCA has confirmed it will not introduce AI-specific regulations, instead relying on Consumer Duty and the SM&CR accountability framework to govern AI deployment. Regulatory supervision is framed as addressing impacts when they occur. The FCA works closely with industry, hosting the 2025 AI Consortium to gather views from the financial and tech sectors. Preliminary findings from the 2026 Mills Review into AI in retail finance suggest that the FCA may move towards requiring more explicit assurance reports, but without new prescriptive rules. This creates an opportunity for the IVO ecosystem to position itself as the vehicle through which the FCA's evolving expectations are operationalised. The Model Risk Management framework, codified in PS7/24, the FCA's 2024 model risk management policy statement, is already mandatory for AI systems

accessing UK financial services via the Prudential Regulation Authority, the Bank of England and the FCA. It requires that model validation be independent from model development and that performance monitoring be continuous and documented, but it does not specify the technical standard against which that independent validation should be performed. This is precisely where the IVO ecosystem creates value. UKAS-accredited IVO assessment against NPL-developed, sector-specific deployment metrics for financial services would give firms a technically credible way to satisfy PS7/24's independent validation requirement, Consumer Duty obligations around foreseeable harm, and SM&CR accountability expectations simultaneously. The FCA would not need to mandate IVO assessment explicitly: referencing UKAS-accredited assessment as a mechanism through which firms can demonstrate compliance with existing obligations is consistent with its outcomes-focused approach and requires no new regulatory instrument.

**Case Study: Credit-risk Model Deployment In Financial Services**

A mid-tier UK bank integrates an AI credit-scoring model into retail lending decisions. Under PS7/24, the bank must demonstrate that model validation is independent from model development, that performance monitoring is continuous, and that the model does not produce outputs that create foreseeable harm under Consumer Duty. Under SM&CR, a named senior manager holds personal accountability for the model's governance.

The bank engages an accredited IVO. Working against NPL-developed sector-specific metrics geared towards risk outcomes for AI credit decisioning specifically (e.g. fairness across protected characteristics under the Equality Act, model performance stability over time, explainability of individual credit decisions, and distributional shift detection).

The output is an independent assessment the FCA can reference in supervisory conversations under the existing PS7/24, Consumer Duty and SM&CR obligations. No new regulatory instrument required. For the insurer pricing the bank's operational risk, the same assessment provides the informational basis to differentiate premiums. For the FCA, it provides a consistent reference point across the firms it supervises, enabling pattern recognition across the sector without requiring the regulator to prescribe methodology. For the bank's board, the named SM&CR holder now has documented, independently validated evidence of model governance rather than relying on internal assurance alone.

The IVO layer makes existing obligations legible and auditable at scale. Each IVO assessment generates operational evidence that feeds back into standards revision, and each standards revision improves the credibility of the next round of assessments.

With **medical devices**, the MHRA has the clearest existing mandate architecture with its explicit patient safety first mandate providing the political foundation for mandatory assurance. The AI Airlock program is the UK's first regulatory sandbox for AI as a medical device, providing a concrete infrastructure for evidence generation and standards development. The program is

now in its second phase running through 2026, with seven healthcare AI tools selected for structured sandbox testing. Phase 2 outputs are designed to inform the National Commission into the Regulation of AI in Healthcare and international initiatives to harmonise standards. The Airlock is therefore already functioning as the evidence-generation infrastructure the pipeline's upstream stages require: it is producing the operational deployment data in healthcare AI that NPL needs to develop sector-specific metrics.

The International Reliance Framework, introduced in 2026, allows products already approved by the FDA or Health Canada to receive a fast-track UKCA mark, and the MHRA's relationship with the FDA and Health Canada demonstrates how cooperation with allies is still possible. The three core regulatory principles co-developed with the FDA and Health Canada (Good Machine Learning Practices, Predetermined Change Control Plans, and Transparency Guidelines) remain the foundation of the international framework for AI medical devices. In August 2025, that collaboration produced five jointly published guiding principles for PCCPs in machine learning-enabled medical devices, establishing an international framework in which the UK is an equal co-author rather than a downstream adopter.

The ICO's Regulatory Sandbox similarly provides a pre-deployment space for AI and data-intensive products, offering supervised testing of personal-data processing, bespoke data-protection guidance, and, where appropriate, statements of regulatory comfort that help firms build compliance evidence before market launch. This is a complementary mechanism that generates compliance evidence the IVO assessment process can reference.

**The UK government can also act as an anchor customer.** By choosing to buy AI that carries UKAS-accredited assurance for its own use, the government acts as an anchor customer, creating early and reliable demand that gives accredited providers a market to build against before the wider ecosystem matures. The mechanism is pull rather than compulsion: government decides how it buys for itself rather than imposing conditions on anyone else, which is why it meets little resistance and can begin now. Because the state is a large buyer, that demand signal also travels, as suppliers configure to meet it, so anchor purchasing seeds the external market without new statutory authority.

### *Industry Levers*

**Model contract clauses** developed through trade bodies including TechUK, UK Finance, and the Association of British Insurers, could embed accredited IVO engagement as a contractual condition of AI procurement and deployment. The International Swaps and Derivatives Association (ISDA) master agreement and Loan Market Association (LMA) syndicated loan form precedents provide a structural analogy. These are trade body model contracts that became the default commercial expectation across global derivatives and lending markets respectively, and subsequently influenced formal regulation. This should be pursued in parallel with sectoral mandates rather than sequenced after them as commercial norm-setting and the regulatory requirements are mutually reinforcing. Model clause development work can begin immediately

within existing trade body structures without new institutional machinery. The Society for Computers and Law (SCL) AI Group has already published a set of pro-supplier and pro-customer AI contractual clauses. These currently rely on a 'Good Industry Practice' standard across development, training, testing, security, and explainability obligations. Drafting notes explicitly acknowledge the difficulty in effectively benchmarking what that means in AI systems, and call for supplementation with specific requirements and obligations. This gap is where a UKAS-accredited IVO assessment could be held as the standard of care, and the near-term action might be to work with the SLC AI Group to update the clauses. LexisNexis and Thomson Reuters Practical Law precedents widely used by legal practitioners could be another similar vehicle.

This mechanism extends naturally into **sector-specific clause sets and trade body standard terms**. UK Finance, the Association of British Insurers, and equivalent bodies in law, healthcare and the public sector already publish standard contractual terms that function as defaults, providing an entry point for AI assurance requirements. Here is where the viability of downstream incentives is dependent on the pipeline's upstream work - the credibility and therefore uptake of sectoral AI clause sets depend on the sector-specific evaluation metrics that NPL's Centre for AI Measurement is positioned to develop. Clauses that reference UKAS-accredited assessment for algorithmic credit decisioning for example, are only as meaningful as the underlying metrics that a reasonably satisfactory assessment in that context requires. Coordinating the timing of metric development with clause drafting so that clauses reference standards that exist in operationally useful format is therefore a concrete coordination task for the AI Standards Hub. The FCA's existing model of co-produced regulatory expectations with trade bodies provides the institutional template for how that coordination between regulator, trade body, NPL, and IVO ecosystem could work in practice.

### 3.3.2. *Medium-term (New statutory provisions or significant policy development required)*

**Statutory safe harbours** condition tort liability protection on engagement with UKAS-accredited IVOs. The mechanism is structurally analogous to Section 230 of the US Communications Decency Act. The critical design requirement is that the safe harbour references UKAS accreditation and BSI-published standards as the compliance threshold, so that legal protection is contingent on genuine engagement with the standards architecture rather than self-certification. The forthcoming Legal Statement on AI liability will be an important input: the closer English law moves toward treating failure to seek independent AI assurance as a breach of the standard of care in negligence, the more the safe harbour functions as a clarity mechanism for deployers already facing common law liability exposure rather than a primary incentive in its own right. OpenAI's recent industrial policy document independently advocates for this mechanism, which reduces the expected commercial resistance and suggests that frontier developers view the liability clarity it creates as aligned with their interests.

**Mandatory AI incident reporting** requires new statutory authority but has higher political feasibility than previously assumed. OpenAI's recent industrial policy document independently advocates for it in almost identical terms indicating less resistance from frontier model

developers than expected. The institutional design question, including the CSR Bill as the legislative model and the candidate institutional bodies for the aggregation function is addressed in Section 3.2 above. The lever belongs in the medium-term tier because the statutory authority it requires is not yet available, but the design process should begin now through the AI Growth Lab consultation.

**Targeted safeguards** such as the November 2025 amendment to the Crime and Policing Bill introducing proactive testing obligations for AI-generated harms, operate primarily at the upstream model layer rather than the deployment layer. They are noted here as a complementary mechanism, and as evidence that the legislative appetite for AI-specific safety obligations is growing. Where such provisions can be designed to fork from upstream standards, they create a further channel through which the pipeline's technical floor propagates into law.

The London **insurance market** is well-positioned to develop AI-specific risk products that make IVO certification a commercial prerequisite for obtaining cover in high-stakes deployment settings. TechUK notes that assurance can reduce insurance costs, support compliance, and facilitate access to capital. In practice, this means deployers may adopt assurance not only to satisfy regulatory expectations, but also to improve insurability, strengthen procurement prospects, and reduce liability risk. The Lloyd's Market Association is the appropriate body to coordinate model policy terms, following the syndicated lending precedent. This lever has significant international reach: policy conditions travel with Lloyd's paper wherever it is used globally, extending UK assurance standards into commercial practice without requiring regulatory mandate in other jurisdictions. Market development will require a longer maturation timescale to allow the ecosystem to generate actuarially useful data on malfunction rates across deployment contexts.

A compelling **incentive structure** is key, and whilst standardised assurance also gives insurers the informational basis to price AI risk, creating demand-side pull, wider market-sustained ‘carrots’ may be necessary. The tension between the intentional formulation of these through regulatory or state intervention vs permitting them to emerge as natural features of the competitive market is as yet unresolved. This is flagged for further study in Annex A3.

### 3.3.3. *Long-term (Contingent on ecosystem maturity)*

**Procurement mandates** face genuine political resistance and government appetite for procurement as a primary policy lever appears to be limited. Here procurement is used not to express government’s own demand but to govern others, conditioning access to public contracts on suppliers meeting accredited assurance requirements. This shapes behaviour across the market rather than within government, which is what makes it powerful and also what makes it resisted.This lever is not viable as the primary mechanism, and its viability increases only once the ecosystem has matured enough that accredited IVOs are readily available and affordable. A 2025 Nuffield Trust review of the NHS AI radiology programme found average deployment

delays of nine to twelve months driven by fragmented governance and layered pre-qualification requirements. Nonetheless, interventions like AI Accelerator tenders, a government procurement initiative designed to fast-track AI adoption across public services by aggregating demand and reducing supplier friction, indicate that the UK government is attempting to improve public procurement speed and efficiency. The November 2025 amendment to the Crime and Policing Bill, which introduces proactive pre- and post-deployment testing obligations for AI systems used in policing contexts, is a concrete example of procurement-driven assurance requirements emerging through sector-specific legislation. This creates a template for how procurement conditions and statutory testing obligations can operate in tandem.

The **National Data Library** offers a related but distinct demand-side mechanism that may not carry the same political resistance as procurement mandates. By structuring and curating fragmented public sector data into governed, accessible form, the NDL creates a competitive advantage for developers who can demonstrate the governance credentials to work with it: better evaluation material, clearer permissions, and authoritative datasets that are difficult to assemble elsewhere. In sectors where data access is the bottleneck over model capability (e.g. health, social care and legal services), NDL access becomes a meaningful commercial incentive to engage with the assurance ecosystem. This makes the NDL a softer but potentially more durable demand-side lever than procurement gatekeeping, and one that can begin operating before procurement mandates become politically viable.

## 3.4 Domestic to International (Exportability)

The strategic problem this paper responds to, that middle powers risk becoming rule-takers as AI governance consolidates around the US, China, and the EU, has been mapped by Chatham House and Leicht. Both identify the range of options available to states in this position; this paper argues that for the UK specifically, the deployment-layer assurance route is more tractable than the alternatives those analyses identify, and that a functioning domestic ecosystem is the precondition for any of the international transmission channels below to operate. These channels convert the domestic IVO ecosystem into a middle-power governance lever. The more that global trust is embedded in UK evaluation methods, auditing standards, accreditation, legal enforceability and post-deployment monitoring, the less the UK has to force recognition through politics or risk.

The four exportability channels proposed below are mutually reinforcing, and should occur sequentially.

**1. ISO/IEC standards developed through stages 1 and 2 of the pipeline feed into compliance infrastructure that is disseminated across jurisdictions whether or not they have bilateral relationships with the UK.** ISO/IEC 42001, to which BSI contributed and against which UKAS is now the first certified accreditation body in the world, already generates commercial pull independently of regulatory mandates. 28% of organisations now require ISO/IEC 42001 compliance from suppliers. This is a significant step forward in creating a shared language for trust, and also the mechanism by which UK methodology propagates without the

UK needing the leverage of a superpower or the market scale of the EU. UKAS already conducts approximately 3,500 to 4,000 assessment days outside the UK annually, including in the US, China, and India, demonstrating that the international transmission capacity exists before the IVO ecosystem is formally established.

**2. Design UK assurance frameworks for compatibility with the EU AI Act as a deliberate strategic choice and an amplifier.** Given the size and regulatory gravity of the EU market, systems that are demonstrably aligned with its risk classification, conformity assessment procedures, and post-market monitoring requirements are more likely to achieve cross-border acceptance with minimal friction. Embedding EU AI Act compatibility directly into UK evaluation methods, auditing standards, accreditation pathways, and legal enforceability mechanisms effectively pre-clears UK-certified systems for one of the world's most stringent regulatory environments making UK certification commercially attractive to developers targeting both markets simultaneously. This is a market-shaping move available to the UK precisely because it is not in the EU: it can design for EU compatibility without being subject to EU political constraints on how that compatibility is structured.

**3. English commercial law is the transmission mechanism that propagates domestically-developed standards into global commercial practice through existing contractual relationships and infrastructure.** Model contract clauses embedding accredited IVO engagement as a condition of responsible AI deployment propagate UK standards into commercial practice wherever parties choose English law as governing law - approximately 40% of commercial contracts globally. This lever is a transmission mechanism with genuinely global reach that operates through commercial relationships already in existence and the least dependent on the domestic ecosystem being fully established before it begins to transmit. The ISDA and LMA precedents confirm that trade-body model clauses can become de facto global governance that subsequently shapes formal regulation.

**4. The AISI international network is the coordination layer that keeps UK methodology feeding all three channels.** NIST and CAISI supply the US standards-and-evaluation node, NPL anchors UK pre-standardisation science, and BSI translates domestic expertise into European and international standards ecosystems through CEN, CENELEC, ETSI, ISO, and IEC interfaces. The result is not linear influence but recursive alignment, in which parallel national systems generate interoperable technical positions that can later be harmonized through bilateral and multilateral standards processes. OpenAI's collaboration with both CAISI and AISI, and industry-wide use of NIST and ISO-based assurance frameworks illustrates this dynamic already happening in practice. The enduring effect of this is difficult to measure and should be the subject of further study. Each national node (UK: NPL/AISI/BSI; US: NIST/CAISI) can independently produce standards-ready technical positions that are still interoperable in principle, since both are feeding into the same international standards-making machinery (ISO/IEC, CEN/CENELEC, etc.). The UK must ensure that the methodology feeding into that machinery originates in UK institutions, and the pipeline described in Section 2 is precisely how it does so.

# Section 4: Conclusion

This paper has argued that the UK's strongest contribution to AI governance is not at the frontier model layer, where it cannot realistically outmatch the United States or China, but at the deployment layer, where it already possesses unusual institutional strengths. By connecting measurement science, standard-setting, accreditation, verification, and incident learning into a coherent assurance pipeline, the UK can turn existing assets into a durable trust infrastructure for AI adoption. An IVO ecosystem would not replace sector regulators or frontier safety work, but would supply the credible, independently structured assurance that makes safe deployment scalable across healthcare, finance, public services, and other high-impact domains. In that sense, the UK's opportunity is not to regulate AI everywhere, but to become the place where AI deployment becomes legible, auditable, and trustworthy. That would strengthen domestic adoption while also giving the UK a genuine route to export influence through standards, legal practice, and commercial norms.

### Time Is Of The Essence

The current moment is especially important because assurance demand is already emerging, but the standards, accreditation pathways, and reporting structures needed to support it are still underdeveloped. If the UK waits until assurance practices are fully consolidated elsewhere — whether through EU conformity systems, US-led private standards, or large vendor-driven norms — it will become harder to shape the benchmark that deployers, insurers, and regulators treat as credible. Early action would allow the UK to lock in its existing institutional advantages, turning NPL, BSI, UKAS, AISI, and sector regulators into a functioning pipeline rather than leaving them as disconnected assets. In short, delay risks allowing the market to define assurance first, while speed would let the UK define it on stronger public-interest terms.

### Agile Standards

BSI Flex is a practical bridge to slower formal ISO/IEC standards. We recommend leveraging this tool, developing standards for sector-specific AI deployment so the system can move quickly whilst still creating a pathway toward formal standards later.

### Co-ordination Bodies

A credible AI assurance ecosystem needs a central co-ordination body because the system will not work if each institution acts in isolation. The missing piece is not another standalone regulator, but a body with the mandate to align technical standards, incident learning, and sectoral expectations across the pipeline. That co-ordination function should ensure that measurement science feeds standard-setting, standards feed accreditation, and verification results feed back into revision and improvement. Without this, the ecosystem risks becoming a set of adjacent initiatives with no shared learning loop or consistent technical floor. With this work we have suggested several natural upstream co-ordinators that may require a DSIT directive to make coordination binding: the AI Standards Hub, the DRCF, CAIRF, the AI Growth

Lab, or AISI themselves. We have left this question open due to the natural flux of Governmental attitudes.

### Mandatory Incident Reporting

Mandatory incident reporting is equally important because it gives the assurance system memory. At present, deployment failures are likely to remain scattered across sectors, trapped inside individual regulatory silos, and lost as one-off events rather than converted into system-wide lessons. A reporting duty would create the evidence base needed to identify recurring failure modes, improve standards, and detect cross-sector risks early. It would also move the UK beyond static compliance towards a living assurance regime, in which real-world deployment experience continuously sharpens the rules. In that sense, mandatory reporting is not just an administrative add-on; it is what makes the entire architecture adaptive, credible, and genuinely safety-improving.

### Exportability Is A Design Goal

Standards, accreditation rules, and verification practices should be designed with international portability in mind from the outset, so the UK can later export its assurance model through standards, contracts, and trade relationships. Building something that other nations can adopt is a form of competitive-collaboration that can increase leverage when the time comes for treaty drafting and trade agreements. We believe that this is preferable to develop in advance, with the upside being greater contributions to AI safe practices.

## Annex (Assumptions, Risks and Further Research)

AI assurance and verification is by nature, a vast, cross-domain research area. With the above, we have focused on the most immediately legible and tractable features, and in the process, found many research streams that are nonetheless integral to the success of the entire system. These are presented here as an a-hierarchical repository of open questions, signposting further work for ourselves and our colleagues in the wider research community.

**A1: Information sharing mandates and antitrust:** The principal unresolved issue is whether DRCF-led coordination on AI assurance could amount to problematic information sharing or standard-setting under UK competition law, even if each regulator retains independent enforcement discretion, especially where discussion might indirectly reveal firm-specific compliance strategies, benchmarking data, or market-sensitive implementation choices.

**A2: Standards pipeline speed:** Can the NPL to BSI cycle be designed to move faster than the ISO/IEC default, and does the incident reporting loop provide sufficient real-time responsiveness?

**A3: Mandate reform feasibility:** What specific changes are required to give the AI Standards Hub an operational rather than coordination mandate, and who champions them within DSIT?

**A4: Optimal configuration of demand-side incentives:** While standardised assurance frameworks improve information symmetry—particularly for insurers, procurers, and downstream deployers—thereby supporting risk-based pricing and liability allocation, these mechanisms alone may be insufficient to generate sustained market demand at scale. Further investigation is required into the relative effectiveness of deliberately constructed incentives (e.g. regulatory preferences, procurement mandates, liability safe harbours, or fiscal incentives) versus those that emerge endogenously through market competition (e.g. reputational differentiation, contractual requirements, or supply chain pressures).

**A5: English law as transmission mechanism:** On what timescale and under what conditions does the absence of independent AI assurance constitute breach of standard of care in negligence? What is the role of the forthcoming Legal Statement on AI liability?

**A6: Independence architecture:** What independence safeguards (e.g. pooled funding, mandatory rotation, or other mechanisms) are both effective and compatible with market viability? What does the credit rating agency literature tell us about structural solutions?

**A7: Verifiability**: What counts as sufficient evidence in a UK IVO assurance case? This is the gap between pipeline architecture and operational credibility that the AVERI paper also flags as unresolved.

**A8: Deployment IVO Dependencies**: deployment-layer assurance depends on development-layer transparency to be meaningful. If we're making claims about system behaviour in deployment, we need to know whether the model has been updated, whether fine-tuning has shifted its behaviour, and whether the version that has been evaluated is the version being run. That requires access to information that sits at the development layer -- version control, update logs, weight change documentation. This then chances upstream issues of access into issues of upstream dependencies. Does this open up new opportunities for collaboration with developmental IVO ecosystems (e.g. AVERI)

**A9: Information architecture and security conditions**. What specific information flows between which actors, through what gateways, under what security and commercial sensitivity conditions? Antony flagged this as the most technically complex unresolved question. Cryptographic hashing of models, access-controlled incident databases, and the distinction between ex-post misuse reporting and ex-ante capability assessment all require specification. This is a genuine research gap.

**A10: Institutional incentive misalignment**. Why has coordination not happened? NPL and BSI as GovCos with competing commercial interests in the standards space. The AI Standards Hub's operational inertia as a product of specific institutional dynamics, not merely mandate weakness. What would make coordination in each actor's interest? This requires its own analysis rather than a passing acknowledgement.

**A11: What does information sharing between AISI and NPL look like in practice, given commercial sensitivity and securitisation constraints?**

**A12: What is the Growth Lab's actual scope, mandate, and governance? Is it a plausible vehicle for the aggregation function?**

**A13: What is the right whole-of-government architecture for AI risk, and does the COBRA lead-department model translate to AI without modification?**

**A14:The standards pipeline is fast enough to matter**

BSI's ISO/IEC process runs on multi-year cycles; AI capabilities change in months. The UK may build methodology that arrives after the US, EU or other middle powers have set the baseline. However, deployed-system assurance operates on different timescales to frontier evaluation. Sub-frontier standards remain relevant even in a mature AI economy. How do we design for the pacing mismatch between standards development and capability advancement?

**A15:The AI Standards Hub can be operationally empowered**

Currently a coordination forum, not an operational body. Closing this gap requires political will from DSIT. What specific mandate changes are needed, and what is the political pathway?

**A16: English private law creates a self-reinforcing demand signal**

No court has ruled on whether failure to seek AI assurance breaches the standard of care in negligence. UKJT draft statement is suggestive but non-binding. Operates on a timescale of years to decades, However, stronger near-term: model contract clauses (cf. LMA, ISDA), but international reach depends on parties continuing to choose English governing law. How confident can we be in this mechanism, and on what timescale?

**A17: The independence problem can be solved**

Company-selected, company-paid auditors produce lower-quality audits (replicated in the Fathom crisis simulation, reproducing 2008 credit-rating-agency dynamics). Evidence points towards pooled funding or mandated rotation, but both are politically contentious. What independence safeguards are both effective and compatible with market viability?

**A18: National security and the frontier/deployment boundary**

As frontier AI becomes tied to defence, the most capable systems may be governed through closed, national-level arrangements. This proposal focuses on the deployment layer where economy-wide standards remain relevant regardless. How should the architecture account for a shifting frontier/deployment boundary?

**A19: Point-in-time certification vs. continuous monitoring**

AI systems change constantly; audit conclusions may become misleading within days. The UK pipeline is designed for point-in-time assessment. No existing institution owns continuous monitoring. Where does continuous monitoring sit institutionally, and who funds it?

**A20: International adoption follows domestic capacity**

The middle-power payoff depends on international adoption. The domestic case is strong, transmitting to the international stage remains plausible but untested. The UK has no forcing mechanism equivalent to the EU AI Act. However, the UK has successfully exported standards

regimes before (financial regulation, maritime, insurance). But the conditions enabling those exports may not be replicable for AI assurance.Through what specific mechanisms have comparable UK standards regimes achieved international adoption, and are those conditions replicable*?* How could the UK frame this through an economic/trade lens: why would the US or other frontier powers be more likely to recognise these frameworks?

**A21: Access authority is legally grounded and practically exercisable**

Even where a deployer engages willingly, IVO assessment may require information about the underlying model, including version history, fine-tuning records, system configuration, that sits with the developer, not the deployer. The deployer may not hold this information and cannot compel its disclosure. However, the extent to which IVO assessment requires developer-held information depends on the scope of the assurance case. Where assurance is scoped to observable deployment behaviour rather than model internals, the deployer may hold sufficient information. Where deployer-facing access is tractable, what developer-held information is minimally necessary for a valid assurance case, and through what mechanism can it be secured?

**A22:  Evidentiary standards for IVO assurance cases are specifiable**

What must an IVO demonstrate, in a given deployment context, to satisfy the standard? Without knowing what passing the assessment looks like, the IVO concept risks replicating the criticism levelled at Partnership on AI, which is principled architecture with indeterminate outputs. What does a valid IVO assurance case look like at each risk tier, and what is the governance process through which that standard is set, tested, and revised?

**A23: Deployment-layer assurance does not require resolving development-layer dependencies**

If the model evaluated by the IVO is not the model being run, because the developer has updated weights, adjusted fine-tuning, or silently switched the underlying system, the assurance conclusion no longer applies. The deployer may be unaware of such changes and has no contractual guarantee of them being notified. However, version integrity may not require development-layer transparency. Developing verification techniques could address this without requiring access to weights or training data. These are conditions that deployers could negotiate, or that model contract clauses could standardise. UK deployment-layer requirements could function as a demand signal, creating incentives for developers to produce standardised version documentation as a precondition for their models being deployable in accredited contexts. This is a potential point of collaboration with development-layer frameworks such as AVERI, reframing the relationship from dependency to structured complementarity. What minimum version-integrity commitments from developers are necessary for deployment-layer assurance to remain valid over time, and can these be secured through contractual standardisation rather than regulatory compulsion?

**A24: Version integrity**. What minimum version-integrity commitments from developers are necessary for deployment-layer assurance to remain valid over time, and can these be secured through contractual standardisation rather than regulatory compulsion?

**A25: Repeated assessment cadence and cost.** What does iterative reassessment mean operationally, what is the relationship between periodic reassessment and continuous monitoring of telemetry, who pays for each, and how do thresholds for each interact with the incident reporting threshold above which reporting is mandatory?

**A26: The communication artefact.** What form does the assurance communication take? Static certification mark, live disclosure, sustainability-style filing, public register? How does this interact with procurement, investment due diligence, and public accountability? How does the architecture connect to DSIT's third pillar of evaluating, measuring, and communicating risk mitigation?

**A27: Practitioner certification.** Which bodies professionalise individual practitioners, and how does that interact with firm-level UKAS accreditation? How is methodological diversity preserved without one association capturing the certification function?

**A28: Internal versus external assurance complementarity.** Where should each apply? When should external assurance be required, and when is internal capacity sufficient? How do the two interact in regulated sectors?

**A29: Sectoral specialisation and cross-sector learning.** How does the pipeline accommodate genuine sectoral variation in ethical priorities, risk appetite, and methodology while still producing cross-sector learnings? What is the role of the AI Standards Hub in coordinating this?

**A30: Audit-assurance demarcation.** What is the categorical difference between audit and assurance, and what regulatory or professional consequences follow from collapsing them? How is the architecture protected from absorption into expanded audit practice?

**A31: Insurance and incident reporting integration.** How does the Stage 5 aggregation function interact with insurance loss data? Should incident reporting and insurance information flows be unified, kept separate but cross-referenced, or maintained independently? What is the architectural relationship between actuarial pricing of AI risk and the standards-revision feedback loop, and what is the governance framework that allows incident data to inform insurance pricing without compromising the public-interest character of incident reporting?

**A32: Pacing and BSI Flex resilience**. Whether BSI Flex provides sufficient pacing resilience under fast-takeoff scenarios.

**A33: Evidentiary standards for IVO assurance cases**. What does a valid IVO assurance case look like at each risk tier?

**A34: AISI proximity to upstream pipeline functions**. The independence question Section 2.1 deliberately surfaces.

**A35: Institutional incentive misalignment**. Why coordination has not happened across NPL, BSI, and the AI Standards Hub, and what would make it in each actor's interest.

**A36: AI Management Essentials and the question of self-assessment tooling.** DSIT's AI Management Essentials is a tool drawn from NIST RMF, the EU AI Act, and ISO 42001. It is positioned as a sense-check rather than as accredited assurance, but the question of how lightweight self-assessment tools relate to the accredited IVO architecture is genuinely open. Does self-assessment substitute for IVO engagement at lower risk tiers? Does it serve as a pre-engagement readiness check? This is worth a research-priority entry.

**A37: The accountability-infrastructure side of the ecosystem.** The pipeline architecture handles professional infrastructure. The full accountability infrastructure (procurement standards, investor due diligence frameworks, insurance integration, liability allocation) requires its own architectural treatment. This is partly covered in Section 3's lever discussion but the integration question is worth flagging in the annex as an area where Section 3's lever framing and Section 2's pipeline architecture do not yet fully connect.[2]

**A38: The justified-trust definition.** DSIT distinguishes trust, justified trust, and trustworthiness in section 3.1 of "Introduction to AI Assurance." The pipeline architecture is implicitly producing justified trust rather than trust as such. Whether this distinction holds operationally, and what the architecture would need to do differently to produce trustworthiness rather than justified trust, is a definitional research question worth flagging.

**A39: The strength and breadth of latent assurance demand.** Observed assurance spending may reflect a compliance-driven minority rather than broad appetite, and defensive AI adoption could coexist with weak genuine demand for verification. Distinguishing the two would require dedicated market research, including demand-side surveys of deployers rather than only assurance suppliers, and analysis of how demand varies by sector risk appetite. The working assumption is that demand is real but conditional on a downstream party requiring it, which is what the Section 3 levers are designed to supply. Research question: across sectors, what share of assurance engagement is genuinely demand-led versus compliance-forced, and how does that change as regulatory and contractual triggers are introduced?

[2] The two-track framing of the AI Assurance Ecosystem is the paper's own analytical move but partly informed by Tess Buckley, Senior Programme Manager, Digital Ethics and AI Safety at techUK.