Concept Paper

Beyond the Questionnaire: A Four-Pillar Reference Model for Continuous Assurance of Public Sector AI Systems

Rajeev Chakraborty *

Posted Date: 2 June 2026

doi: 10.20944/preprints202606.0177.v1

Keywords: AI governance; continuous assurance; public sector AI; lifecycle assurance; post-market monitoring; design science research; responsible AI; UK government; Value Sensitive Design

 doi:10.20944/preprints202606.0177.v1

Concept Paper

Beyond the Questionnaire: A Four-Pillar Reference Model for Continuous Assurance of Public Sector AI Systems

Rajeev Chakraborty

University of Portsmouth, Responsible AI Decision Lab; rajeev@raitracker.com

Abstract

Background. UK central government has built a substantial body of pre-deployment AI assurance practice, anchored in departmental assurance questionnaires, the Artificial Intelligence Playbook for the UK Government (Government Digital Service, 2025), the Algorithmic Transparency Recording Standard (ATRS), and instruments such as Data Protection and Equality Impact Assessments. These instruments establish a strong foundation for assessing whether an AI system is fit to enter live service. The next step, building on this strength, is to extend the same structured rigour across the rest of the system lifecycle. **Aim.** This paper proposes and theoretically grounds a four-pillar reference model for responsible AI (RAI) assurance designed to augment existing UK government instruments across the full lifecycle: Pre-Deployment, Model Activation, Operational Response, and Closed-Loop Learning. The distinctive moves are the treatment of Model Activation as a discrete baseline-setting phase and the specification of Closed-Loop Learning as a cross-government or cross-deployer institutional function, both of which are thinly addressed in existing reference frameworks. **Approach.** The model is developed using a design science research approach. It is derived through structured comparative document analysis of the four reference frameworks UK departments most directly encounter (the NIST AI Risk Management Framework, the OECD AI Principles, the EU AI Act, and the UK AI Playbook) using the AI system lifecycle as the analytical organising frame. It is theoretically anchored in Value Sensitive Design (Friedman, Kahn and Borning, 2006; Friedman and Hendry, 2019; Umbrello and van de Poel, 2021) and refined through expert-informed practitioner engagement at the Department for Science, Innovation and Technology (DSIT) and a second UK central government department in early 2026. **Findings.** The comparative analysis shows that the four reference frameworks converge densely at pre-deployment, address operational response with varying degrees of prescriptiveness, are thin on activation as a distinct lifecycle phase across three of the four frameworks, and consistently underspecify closed-loop learning. The four-pillar model addresses the cells where the leading frameworks are weakest, in a form compatible with each. **Contribution.** The paper contributes a lifecycle reference model for public sector AI assurance, a cell-level mapping of four leading frameworks against that lifecycle, and an institutional fit analysis for UK government adoption. Its distinctive contribution lies in specifying activation as calibration and closed-loop learning as institutional memory. The paper is offered as a basis for cross-government conversation and as the design phase of a research programme whose evaluation phase is described in Section 8.

Keywords: AI governance; continuous assurance; public sector AI; lifecycle assurance; post-market monitoring; design science research; responsible AI; UK government; Value Sensitive Design

 doi:10.20944/preprints202606.0177.v1

1. Introduction

UK central government is increasingly exploring and deploying AI systems across operational, analytical and citizen-facing contexts. Public bodies including the Home Office, HMRC, DWP, the DVSA and NHS organisations have publicly reported, trialled, procured or deployed AI-enabled systems for decision support, document triage, fraud detection, customer interaction and operational analytics. The aggregate footprint is sufficient that the governance of UK government AI is a problem of public administration, not solely of digital policy (Margetts and Dorobantu, 2019; Veale and Brass, 2019; Misuraca and van Noordt, 2020).

The assurance instruments that have accompanied that growth reflect a mature tradition of structured pre-deployment review: departmental AI assurance questionnaires addressing data quality, model documentation, known risks, mitigation strategies, and ethical review; the AI Playbook for Government (Government Digital Service, 2025) with its ten principles and six-stage pathway, mostly concentrated in stages 0 to 3; the Algorithmic Transparency Recording Standard (ATRS); departmental DPIAs and EIAs; and the DSIT Introduction to AI Assurance (Department for Science, Innovation and Technology, 2024) as the central reference for the wider assurance ecosystem. Together these instruments establish a strong foundation for assessing whether a system is fit to enter live service.

The natural next step is to extend the same rigour across the rest of the lifecycle. AI systems are dynamic in service: input distributions shift, model versions change, and usage patterns evolve in ways pre-deployment review cannot fully anticipate (Mökander et al., 2021; Schuett, 2023). The opportunity is to match structured pre-deployment effort with equally structured activity at activation, in live service, and in institutional learning. The risk of leaving this gap unaddressed has been articulated in the UK policy discourse on algorithmic accountability (Ada Lovelace Institute and DataKind UK, 2020; Ada Lovelace Institute, 2023) and in the broader scholarly critique of static, one-shot AI ethics review (Mittelstadt, 2019; Raji et al., 2020; Cobbe, Veale and Singh, 2023).

This paper proposes a four-pillar reference model that addresses the opportunity. The model is developed within doctoral research at the RAID Lab, University of Portsmouth, where continuous lifecycle assurance for public-interest AI is the central object of study. It is derived through structured comparative document analysis of the four reference frameworks UK departments most directly encounter (NIST AI RMF, OECD AI Principles, EU AI Act, and UK AI Playbook), and was refined through expert-informed practitioner engagement at DSIT and a second UK central government department in early 2026. The four pillars (Pre-Deployment, Model Activation, Operational Response, and Closed-Loop Learning) correspond to four phases of the AI system lifecycle, each with distinct assurance activities. The model is intended as an augmentation of existing UK government instruments, not a replacement.

The novelty of the contribution lies not in identifying lifecycle assurance as a general need, the case for which is well established, but in specifying two functions that existing reference frameworks leave thinly addressed: Model Activation as a discrete baseline-setting handover between pre-deployment review and live monitoring, and Closed-Loop Learning as a cross-government or cross-deployer institutional mechanism for converting operational experience into shared knowledge. The cell-level framework mapping at Section 6 shows where existing frameworks are thin on these two functions and provides the analytical basis for treating them as distinct pillars.

The paper is organised around a single overarching research question and three subsidiary questions.

How can existing AI assurance practice be extended into a lifecycle-complete reference model that augments current pre-deployment instruments without duplicating them?

- **SQ1.** What does each of the four leading reference frameworks specify at each phase of the AI system lifecycle, and where do they converge, diverge, or leave gaps? (Section 6.)
- **SQ2.** What set of pillars, theoretically grounded and operationally specified, adequately covers the lifecycle and addresses the identified gaps? (Sections 3 and 5.)

 doi:10.20944/preprints202606.0177.v1

- **SQ3.** How does the resulting model align with existing UK institutional anchors, and what would be required to test that alignment in practice? (Section 7.)

The overarching RQ is answered cumulatively across all three. Section 2 reviews the relevant academic literature. Section 3 sets out the conceptual framework. Section 4 describes the research approach. Section 5 presents the model. Section 6 provides the framework mapping. Section 7 sets out the institutional fit analysis. Section 8 discusses implications, limitations, and conflict-of-interest disclosures. Section 9 concludes.

2. Related Work

The contribution of this paper sits at the intersection of five lines of scholarship: responsible AI principles and their operationalisation; AI auditing and assurance; lifecycle approaches to algorithmic accountability and post-market monitoring; MLOps, model monitoring, drift and incident reporting; and public sector AI adoption and governance. This section locates the four-pillar reference model in each.

2.1. Responsible AI Principles and Their Operationalisation

The last decade has produced a proliferation of high-level AI ethics principles, with comparative reviews identifying substantial convergence on values such as transparency, fairness, accountability, privacy, and human oversight (Jobin, Ienca and Vayena, 2019; Floridi et al., 2018; Fjeld et al., 2020). A now well-established critique argues that principles alone do not translate into the technical, procedural and institutional practices required to make AI systems trustworthy in deployment (Mittelstadt, 2019; Whittlestone et al., 2019; Greene, Hoffmann and Stark, 2019; Morley et al., 2020). Empirical work on practitioners reinforces this operationalisation gap, identifying a shortage of usable methods, checklists, and organisational scaffolds (Holstein et al., 2019; Madaio et al., 2020; Rakova et al., 2021). The technical AI ethics literature has responded with concrete artefacts including model cards (Mitchell et al., 2019), dataset documentation (Gebru et al., 2021), and structured fairness auditing methods (Raji and Buolamwini, 2019; Raji et al., 2020). The reference model proposed here extends that tradition from technical artefact to institutional scaffold: it specifies, at the level of organisational activity, what assurance work needs to happen at each phase of the AI system lifecycle.

2.2. AI Auditing and Assurance

A parallel literature has developed on AI auditing as a discipline. Sandvig et al. (2014) provided an early framing of algorithm auditing techniques; subsequent work has extended audit methodology to ML systems, distinguishing internal, external, third-party and regulatory audits (Brown, Davidovic and Hasan, 2021; Mökander et al., 2021; Costanza-Chock, Raji and Buolamwini, 2022). Ethics-based auditing has been proposed for ongoing assessment of automated decision-making systems (Mökander et al., 2021), and the broader literature increasingly recognises audit as a continuing activity rather than a single event (Raji et al., 2020; Cobbe, Veale and Singh, 2023). In the UK context, the DSIT Introduction to AI Assurance (Department for Science, Innovation and Technology, 2024) is the current operational reference for the wider assurance ecosystem. The literature has been strong on actor mapping and technique cataloguing but less developed on the temporal architecture of assurance: when in the lifecycle each activity belongs, and how findings from one phase inform the next. The four-pillar reference model is positioned in this gap.

2.3. Lifecycle, Post-Market Monitoring and Continuous Accountability

A third strand addresses lifecycle and post-market obligations on AI systems. Schuett (2023) analyses the risk-management obligations in Article 9 of the AI Act, and Schuett (2024) extends the argument to a three-lines-of-defence organisational design. Mökander, Sheth, Watson and Floridi (2023) examine how the Act's auditing and conformity assessment provisions translate into operational practice. Cobbe, Veale and Singh (2023) argue that accountability in algorithmic systems

 doi:10.20944/preprints202606.0177.v1

requires attention to the full supply chain and the full life of the system. The post-market monitoring concept itself is borrowed from medical-device and product-safety law, where lifecycle obligations on manufacturers are well established (Black and Murray, 2019; Kop, 2021). ISO/IEC 42001 on AI management systems (International Organization for Standardization, 2023a) and ISO/IEC 23894 on AI risk management (International Organization for Standardization, 2023b) explicitly treat AI assurance as a continuing organisational activity through a Plan-Do-Check-Act cycle. The lifecycle framing is therefore not novel to this paper; what is offered here is its structured application to the institutional machinery of UK government AI assurance, and the identification, at cell level, of where leading reference frameworks are weakest.

2.4. MLOps, Model Monitoring, Drift and Incident Reporting

A fourth strand, originating in the ML systems and software engineering communities, addresses the operational realities of running ML systems in production. Sculley et al. (2015) gave the early account of 'hidden technical debt' that accumulates in production ML systems beyond what routine software engineering addresses. Paleyes, Urma and Lawrence (2022) surveyed deployment case studies and catalogued the challenges practitioners encounter at each stage of the ML deployment workflow. Kreuzberger, Kühl and Hirschl (2023) consolidated MLOps as a definition, architecture, and set of practices; Breck et al. (2017) proposed a rubric for production readiness specifically. On distribution shift and concept drift, Gama et al. (2014) provide the foundational survey, and Klaise et al. (2020) translate monitoring concerns into operational architectures for models in production. The AI Incident Database (McGregor, 2021) is among the more developed cross-organisational mechanisms for cataloguing AI failures and a natural reference point for the Closed-Loop Learning function described at Pillar 4. The four-pillar reference model does not duplicate this technical literature; it provides the institutional scaffold within which MLOps techniques can be located, governed, and reported on for public sector accountability. Pillars 2 and 3 in particular formalise, at institutional level, activities the MLOps community has already developed at engineering level: baseline-setting, monitoring, drift detection, and operational incident handling.

2.5. Public Sector AI: Adoption, Accountability and Administrative Law

A fifth strand, originating in public administration and administrative law, addresses the specific challenges of AI in government. Veale and Brass (2019) argue that public sector ML sits within a tradition of public management that imposes distinct legitimacy and accountability requirements not reducible to private sector AI ethics. Margetts and Dorobantu (2019) call for an AI-aware public administration. Bullock (2019) and Vogl et al. (2020) examine the discretion and bureaucratic implications of algorithmic decision support in government. Misuraca and van Noordt (2020) catalogue AI in European public services. Engin and Treleaven (2019) describe algorithmic government as a research agenda. Yeung (2018) and Cobbe (2019) frame the legal and constitutional dimensions. UK-specific work has examined departmental experience with automated decision-making (Cobbe and Singh, 2021) and the role of procurement as a governance lever (Madan and Ashok, 2023); the Ada Lovelace Institute and DataKind UK (2020) and Ada Lovelace Institute (2023) have shaped the UK policy debate specifically. The reference model draws on all five strands and is positioned to advance the public sector AI literature by providing a lifecycle-complete institutional scaffold against which UK departments can locate their existing instruments and identify extension opportunities.

3. Conceptual Framework

3.1. Lifecycle Thinking for AI Assurance

The intellectual move of this paper is to treat an AI system as something that has a lifecycle with distinct phases, each with distinct assurance requirements, rather than as an artefact that is certified once at deployment. The lifecycle framing is not new; it underpins ISO/IEC 42001 on AI management

 doi:10.20944/preprints202606.0177.v1

systems (International Organization for Standardization, 2023a), the NIST AI Risk Management Framework (NIST, 2023), and the EU AI Act's post-market monitoring provisions (European Union, 2024). What is new is the structured application of lifecycle thinking to the specific institutional machinery of UK government AI assurance, and the foregrounding of two phases that existing frameworks leave thinly specified.

The paper distinguishes four phases.

- **Pre-Deployment.** The period before a system enters live service. Assurance activity here is anticipatory: what might go wrong, and what is being done to prevent it.
- **Model Activation.** The transition into live service. Assurance activity here is baseline-setting: what does 'normal' look like for this system in production conditions, such that future divergence can be detected. Without activation, monitoring becomes observation without calibration.
- **Operational Response.** The period in live service. Assurance activity here is detective and reactive: noticing divergence from the activation baseline, escalating when required, and intervening.
- **Closed-Loop Learning.** The mechanism by which operational findings feed back into design, procurement, and governance. Assurance activity here is institutional: making sure lessons are captured, shared, and acted on across departments.

Each phase has a different temporal character. Pre-Deployment is discrete and time-bounded. Model Activation is a handover event. Operational Response is continuous and open-ended. Closed-Loop Learning is periodic and reflective.

3.2. Theoretical Anchor: Value Sensitive Design

The reference model is theoretically anchored in Value Sensitive Design (VSD), which treats human values as principled design inputs alongside functional and performance requirements (Friedman, 1996; Friedman, Kahn and Borning, 2006; Friedman and Hendry, 2019). VSD's tripartite methodology (conceptual, technical, empirical) maps directly onto the four pillars. Pillar 1 corresponds to VSD's conceptual investigation: explicit articulation of values, risks and impacts. Pillar 2 corresponds to the moment conceptual specification meets technical implementation, at which baselines are set. Pillar 3 corresponds to empirical observation of the system in its sociotechnical context. Pillar 4 corresponds to the iterative loop VSD requires between empirical observation and revised conceptual and technical specification. Umbrello and van de Poel (2021) extend VSD specifically to AI and argue that its iterative, multi-method commitment is well suited to the operationalisation gap in AI ethics.

The model is also informed by sociotechnical systems theory, which insists that any account of a technology in deployment must include the people, procedures and institutions through which it is operated (Selbst et al., 2019; Sambasivan et al., 2021). The pillars are stakeholder-specified throughout: each pillar identifies who owns the activity, not just what the activity is.

3.3. An Illustrative Example: AI-Assisted Document Triage

To make the four pillars concrete, consider an AI-assisted document triage system used by a generic public sector department to route incoming case correspondence to appropriate handlers. At Pillar 1, the department conducts pre-deployment review: completes the departmental AI assurance questionnaire, a DPIA, and an EIA; documents the model in line with current departmental practice; commissions adversarial testing for known failure modes (misrouting of safeguarding-relevant correspondence, demographic disparities in routing speed); and submits an ATRS record. At Pillar 2, on go-live, the department records the input distribution observed in the first weeks of live operation, captures baseline metrics (routing accuracy, time-to-handler, demographic parity in routing), sets operating bands and alert thresholds, documents expected edge cases, and signs off the handover from project team to operational service team. At Pillar 3, the live service is continuously

 doi:10.20944/preprints202606.0177.v1

monitored against the activation baseline: drift detection on input vocabulary and case type distribution; output-side monitoring of routing decisions; alerts when fairness proxies excursion beyond thresholds; SRO-level escalation paths defined. At Pillar 4, when an incident occurs (for example, systematic misrouting of a previously rare correspondence category following a policy change), structured post-incident analysis is conducted, root cause is documented, risk taxonomy and procurement specification are updated as appropriate, and the lesson is shared with comparable departments through a cross-government channel. The same system passes through all four pillars; the pillars are concurrent rather than sequential once the system is in service.

3.4. What the Model Is Not

To avoid confusion, the four-pillar model is not a replacement for existing departmental instruments (questionnaires, the AI Playbook, ethics reviews, and ATRS continue to do useful work; the model is a scaffolding within which these can be located), not a single tool (each pillar involves different stakeholders, evidence, and cadences, and adoption involves assembling a combination of existing and new instruments), and not a maturity model (all four pillars must operate concurrently for any system in service, not in sequence).

4. Research Approach and Methodology

4.1. Design Science Research

This paper adopts a design science research (DSR) approach, concerned with the construction and evaluation of artefacts intended to address identified problems in their domain (Simon, 1996; Hevner et al., 2004; Peffers et al., 2007; Cross, 2001). The artefact may be a method, a model, a system, or an institutional design (March and Smith, 1995). DSR has been applied to public administration and policy design (Barzelay, 2007; Lewis, 2021) and to AI governance specifically (Mökander et al., 2021). The four-pillar reference model is a DSR artefact of the 'model' class: an abstraction of the activities and stakeholders required to make continuous AI assurance operationally tractable. The paper presents the construction of the artefact and its derivation against a reference framework set. Following the DSR convention of distinguishing artefact construction from artefact evaluation (Hevner et al., 2004; Sonnenberg and vom Brocke, 2012), outcome evaluation in live deployment is the proper subject of subsequent empirical work, designed in outline at Section 8.

4.2. Stage One: Comparative Document Analysis

The analytical foundation of the model is a structured comparative document analysis of the four AI assurance frameworks UK central government departments most directly encounter: the NIST AI Risk Management Framework (NIST, 2023), the OECD AI Principles (OECD, 2024), the EU AI Act (European Union, 2024), and the UK AI Playbook for Government (Government Digital Service, 2025). The selection rationale is at Section 6.1. ISO/IEC 42001 and ISO/IEC 23894 are referenced in Sections 2.3 and 6.1 but are not included in the comparative matrix, for the reasons set out at Section 6.1.

The procedure followed established conventions for qualitative document analysis (Bowen, 2009; Cardno, 2018), adapted to the requirements of DSR where the coding output is itself a designed artefact. Each framework was retrieved from its primary publisher (NIST, OECD, the Official Journal of the European Union, and gov.uk) in the version current at May 2026. The unit of analysis was the framework text; sections, articles, principles, or functions were the granular units at which coding decisions were made. The four-phase AI system lifecycle (Section 3.1) was applied as a deductive coding frame: each granular unit was assessed for relevance to each of the four phases (Pre-Deployment, Model Activation, Operational Response, Closed-Loop Learning), with multi-phase coding allowed where appropriate. For each cell of the resulting four-by-four matrix, a coverage code was applied using the four-category schema set out in Table 1. The schema was developed iteratively:

 doi:10.20944/preprints202606.0177.v1

an initial three-category schema (specified, implicit, absent) proved insufficient to discriminate between partial and prescriptive specification, so the four-category schema was adopted.

Table 1. Coverage coding schema for the framework mapping.

Code	Definition
Dense	The framework specifies activity at this phase explicitly and prescriptively, with identifiable articles, principles, or functions; operational form is articulated; stakeholders or responsibilities are indicated.
Partial	The framework addresses the phase but with limited prescriptiveness; activity is named but operational form is not specified, or only a subset of the relevant activity is addressed.
Implicit	The framework does not address the phase as a distinct activity but coverage can be inferred from broader provisions or lifecycle framing.
Absent	The framework does not address the phase, even implicitly; no provision is reasonably available at the lifecycle phase concerned.

Cell-level content was synthesised into the comparative matrix at Table 3 (Section 6.2), with each cell summarising substantive content, citing the applicable article, principle or function, and appending the coverage code. Four structural observations were derived and are set out at Section 6.3. The complete coding audit trail, with framework source extracts and rationale per cell, is provided at Appendix A.

Analytical Safeguards

The coding was performed by the author as a single coder, which is a limitation acknowledged at Section 4.5. Three safeguards were applied to reduce confirmation bias and to make the interpretive basis of the mapping transparent and reproducible. First, all framework claims were anchored to specific article, principle, function or stage references, so that any reader can verify each cell against the cited primary source. Second, multi-phase coding was permitted where a provision plausibly addressed more than one lifecycle phase, avoiding forced allocation to a single cell. Third, the distinction between Dense, Partial, Implicit and Absent was applied conservatively: a cell was coded Dense only where the framework specified both the activity and some operational form; Partial where activity was named but operational form was not specified; Implicit where coverage could be inferred but was not addressed as a distinct activity; and Absent only where no provision was reasonably available at the phase concerned. These safeguards do not remove the limitation of single-coder analysis but they make the analytical basis of each cell open to inspection.

4.3. Stage Two: Doctoral Research Development

The model was developed within ongoing doctoral research at the RAID Lab, University of Portsmouth, in which continuous lifecycle assurance for public-interest AI is the central object of study. The doctoral work grounds the model theoretically in Value Sensitive Design (Friedman, 1996; Friedman and Hendry, 2019; Umbrello and van de Poel, 2021) and in the public administration literature reviewed at Section 2, and articulates the operational specification of activity at each pillar.

The model presented in this paper is a version of the design developed in the doctoral research. The doctoral work covers additional material not in scope for this paper, including a detailed theoretical derivation of the pillar structure, comparative analysis against international frameworks

 doi:10.20944/preprints202606.0177.v1

beyond the four examined here, and extended specification of metrics under each pillar. The presentation here is restricted to the components necessary for the reference model to be understood and considered as an institutional design.

4.4. Stage Three: Expert-Informed Artefact Refinement

The reference model was discussed with senior practitioners responsible for AI assurance at the Department for Science, Innovation and Technology (DSIT) and a second UK central government department during a series of structured engagements in early 2026. The engagements took the form of approximately six bilateral and small-group discussions of between sixty and ninety minutes each, conducted between January and April 2026. Participants held senior practitioner-level roles in AI assurance, AI policy, and AI delivery in their respective organisations; precise role titles and the number of individual participants are withheld to protect attribution.

Each discussion was organised around a working version of the four-pillar model and a draft of the four-by-four framework mapping. Topics covered included: the coherence of the pillar structure with current departmental practice; the operational definition of the boundary between Pillar 1 (Pre-Deployment) and Pillar 2 (Model Activation); the realism of the Pillar 3 (Operational Response) specification given current departmental tooling and team structures; the institutional location of Pillar 4 (Closed-Loop Learning) responsibilities; and the appropriateness of the DSIT Responsible AI Unit, departmental service assurance functions, and the Crown Commercial Service as institutional anchors for adoption.

Insights from each engagement were captured through contemporaneous notes by the author, supplemented by written follow-up where a specific refinement had been discussed. Notes were used to revise the working version of the model and the draft mapping; no attributable quotation, no organisational confidential information, and no system-specific detail is reported in this paper. Where a refinement is reported (notably the operational specification of Pillars 2 and 3, and the four areas for augmentation set out at Section 5.1), it is the refinement itself that is reported, not its attribution.

These discussions are best understood as expert-informed artefact refinement consistent with the DSR tradition of expert review of artefacts under construction (Peffers et al., 2007; Sonnenberg and vom Brocke, 2012). They helped test whether the pillar structure was intelligible to senior practitioners familiar with public sector AI assurance, and they shaped the operational specification of the pillars. They do not constitute interview-based qualitative research with formal sampling, transcription, and thematic analysis, and they do not demonstrate institutional acceptance, implementation feasibility, or outcome improvement. They should not be read as endorsement by either department named or any of its officials.

4.5. Methodological Limitations

The methodology has three acknowledged limitations.

First, DSR produces designs; it does not, in its construction phase, produce outcome evidence. This paper does not report whether the reference model improves AI assurance outcomes in operational practice. A separate pilot and evaluation is required for that. The conditions for such an evaluation are set out at Section 8.4, with a proposed design and candidate metrics at Section 8.5.

Second, the coding was performed by a single coder. The analytical safeguards described at Section 4.2 mitigate but do not eliminate this limitation. A subsequent coding pass by an independent researcher, with documented inter-coder agreement on a sample, would strengthen the comparative analysis; the audit trail at Appendix A is structured to support such a pass.

Third, the practitioner engagement drew on DSIT and a second department. A broader cross-government consultation would strengthen external validity and is part of the institutional fit work set out at Section 7.

5. The Four-Pillar Reference Model

 doi:10.20944/preprints202606.0177.v1

This section describes each pillar in turn, setting out what assurance activity belongs to it, what evidence it produces, who its stakeholders are, what existing UK instruments already sit there, and the consequences of its omission.

5.1. Four Areas for Augmentation

The expert-informed refinement reported at Section 4.4 identified four areas where AI-specific augmentation to existing practice would add most value: continuity of assurance activity through live service; feedback between production and assurance through an explicit mechanism for operational findings to inform future cycles; consistent coverage across transactional services, shared platforms, and B2B systems; and AI-specific capability in runtime monitoring alongside conventional IT service management. Each is addressed directly by one of the four pillars.

5.2. Pillar 1: Pre-Deployment

Purpose. Establish that an AI system is fit to enter live service, with known risks identified and mitigation strategies in place.

Activity and evidence. Structured review of training data quality and provenance, model documentation (Mitchell et al., 2019; Gebru et al., 2021), risk identification against a recognised taxonomy, mitigation strategy design and review, ethics and equalities impact assessment, DPIA, adversarial testing and red-teaming (Brundage et al., 2020), and, where required by ATRS, the transparency record. The evidence is a set of structured documents: completed questionnaires, impact assessments, red-team reports, and the transparency record.

Stakeholders and existing UK instruments. Project teams, ethics reviewers, data protection officers, information assurance functions, and the SRO who authorises go-live. Existing instruments include departmental AI assurance questionnaires, the AI Playbook principles and stages 0 to 3 of its implementation pathway, DPIAs and EIAs, ATRS submissions, and departmental ethics review boards.

Value added. Pillar 1 provides the structured anticipation that allows go-live with confidence and produces the documented specification on which every subsequent pillar depends. Without it, the system enters service without a clear specification of what it is expected to do.

5.3. Pillar 2: Model Activation

Purpose. Establish the operational baseline against which future divergence will be measured.

Activity and evidence. Recording the pre-live input distribution; capturing baseline performance metrics on live-like data; setting operating bands and alerting thresholds; documenting expected behaviour and edge cases; agreeing the cadence and ownership of continuous monitoring; and completing the handover from project team to operational service team. The evidence is a set of operational baselines and runbooks: recorded input distributions, initial metric values, threshold specifications, edge case inventories, and the signed handover.

Stakeholders and existing UK instruments. Project teams handing over; operational service teams receiving; data scientists confirming baselines; the assurance function signing off. Existing instruments include service transition processes (ITIL-aligned in most departments), operational runbooks, and departmental pre-production sign-off.

Value added. Pillar 2 is the bridge between anticipation and observation. It establishes the calibrated reference against which live behaviour can be compared. Without it, operational monitoring has nothing to compare against and divergence cannot be detected systematically. This is one of the two pillars where the comparative framework analysis at Section 6 finds the most consistent thinness, and it is one of the two pillars where the contribution of this paper is most distinctive.

5.4. Pillar 3: Operational Response

 doi:10.20944/preprints202606.0177.v1

Purpose. Detect and respond to divergence from baseline in service.

Activity and evidence. Continuous monitoring of input distributions and output behaviour; drift detection against the activation baselines; monitoring of fairness proxies and other AI-specific signals; alerting on threshold breaches; escalation to the SRO; and intervention (retraining, restriction of use, or temporary withdrawal). The evidence is continuous: monitoring telemetry, alert logs, incident records, intervention decisions, and the operational service review record.

Stakeholders and existing UK instruments. Operational service teams; AI-specialist engineers or a central AI assurance function providing specialist cover; the SRO for escalations; model owners for interventions. General IT service management provides strong coverage of conventional operational signals; a growing number of departments are adding AI-specific observability capability alongside this foundation.

Value added. Pillar 3 provides continuous visibility that matches the continuity of service itself, complementing general service management with AI-specific signals (drift, output anomalies, fairness proxies). Without it, operational evidence is not systematically collected, which limits the material available to Pillar 4 and to the next system's Pillar 1 cycle.

5.5. Pillar 4: Closed-Loop Learning

Purpose. Feed operational findings back into assurance, design, procurement and governance, across departments as well as within them.

Activity and evidence. Structured post-incident analysis; root cause analysis linking operational events to pre-deployment assumptions; update of risk taxonomies where new failure classes emerge; update of procurement specifications where supplier-level causes are identified; model retraining decisions informed by operational evidence; publication of lessons across departments; and update to the ATRS record where appropriate. The evidence is a set of learning artefacts: incident reviews, updated risk registers, refreshed procurement specifications, retraining decisions, and cross-departmental learning publications.

Stakeholders and existing UK instruments. Model owners, procurement leads, risk functions, departmental assurance functions, and the DSIT Responsible AI Unit as a cross-government publication channel. Internal incident review processes operate in most departments and provide a foundation that can be extended with AI-specific learning; cross-departmental publications and CCS engagement on AI contract updates are emerging areas.

Value added. Pillar 4 converts operational experience into institutional learning. It is the mechanism by which what is observed in one department's production system informs the next department's pre-deployment review, the next procurement specification, and the next iteration of the reference model. The cross-government or cross-deployer dimension is the distinguishing element: existing frameworks address organisational learning, but none specifies a mechanism by which lessons travel from one public sector deployer to another. Without that mechanism, learning stays local rather than becoming shared knowledge across the system.

5.6. The Pillars as a Structure, Summarised

Table 2 presents the pillars side by side with their purpose, evidence, primary stakeholders, and the distinct value each adds to the lifecycle.

Table 2. The four pillars in summary.

Pillar	Purpose	Primary Evidence	Primary Stakeholders	Value Added
1. Pre-Deployment	Fitness for service	Questionnaires, impact	Project teams, ethics, DPO, SRO	Structured anticipation;

 doi:10.20944/preprints202606.0177.v1

Pillar	Purpose	Primary Evidence	Primary Stakeholders	Value Added
		assessments, red-team reports		documented specification for live service
2. Model Activation	Baseline for live service	Baselines, thresholds, runbooks	Project and service teams, data scientists	Calibrated reference for recognising normal behaviour in service
3. Operational Response	Detection and response in service	Telemetry, alerts, incidents	Service teams, AI specialists, SRO	Continuous AI-specific visibility alongside general service management
4. Closed-Loop Learning	Cross-government institutional learning	Reviews, updated registers, cross-departmental publications	Model owners, risk, procurement, DSIT RAI Unit	Conversion of operational experience into cross-government knowledge

6. Mapping the Four Pillars Against Four Reference Frameworks

This section places the four-pillar reference model alongside the four frameworks that UK government policy most directly references: the NIST AI Risk Management Framework, the OECD AI Principles, the EU AI Act, and the UK AI Playbook for Government. The purpose is to show what each framework specifies at each phase of the AI system lifecycle, and where the lifecycle structure of those frameworks is dense, thin, or absent.

6.1. Why These Four Frameworks

The four frameworks are chosen because they are the reference points that UK central government departments encounter most directly in current practice. Together they cover the technical-practitioner reference (NIST), the international-policy reference (OECD), the regulatory reference (EU AI Act), and the UK government's own operational guidance (the Playbook). A test of a lifecycle reference model proposed for UK government adoption is most informatively conducted against this set.

Other significant frameworks (notably ISO/IEC 42001 on AI management systems, ISO/IEC 23894 on AI risk management, and the Singapore Model AI Governance Framework for Generative AI) are referenced where relevant but are not included in the comparative matrix. ISO/IEC 42001 and ISO/IEC 23894 are general-purpose AI management standards rather than direct departmental references; their Plan-Do-Check-Act lifecycle structure is consistent with the four-pillar model, but they do not function as the primary operational reference for UK central government departments.

 doi:10.20944/preprints202606.0177.v1

A test of a UK-government-facing reference model is most informatively conducted against the frameworks departments actually encounter in current practice. A subsequent paper could extend the mapping to ISO and Singapore frameworks where useful for international comparability.

- **The NIST AI Risk Management Framework (NIST, 2023)** is the most widely cited cross-sectoral framework for trustworthy AI in technical and assurance discourse. It defines seven characteristics of trustworthy AI (valid and reliable; safe; secure and resilient; accountable and transparent; explainable and interpretable; privacy-enhanced; fair with harmful biases managed) and four functions (Govern, Map, Measure, Manage) that organise the actions required to address those characteristics. UK departmental practice does not formally adopt NIST AI RMF, but several of its constructs appear in departmental governance materials.
- **The OECD AI Principles (OECD, 2024)** are the first intergovernmental standard for trustworthy AI and are now adhered to by 47 jurisdictions, including the European Union. They set out five values-based principles: inclusive growth, sustainable development and well-being; respect for the rule of law, human rights and democratic values; transparency and explainability; robustness, security and safety; and accountability. The 2024 update relocated the provisions on traceability and systematic, ongoing-lifecycle risk management to the accountability principle, which strengthens the principles' lifecycle relevance for the present argument.
- **The EU AI Act (European Union, 2024)** is, of the four, the only one with regulatory force. Although the United Kingdom is not bound by the Act, UK suppliers placing AI systems on the EU market are, and the Act's structure (notably the Chapter III obligations on high-risk systems and the Chapter IX provisions on post-market monitoring and market surveillance) is becoming an important reference for what comprehensive lifecycle assurance looks like in the regulated case. Article 72 is among the more explicit and prescriptive provisions across the four frameworks for continuous post-deployment activity (Schuett, 2023; Mökander et al., 2023), although a definitive comparative ranking depends on how 'requirement' is operationalised across frameworks of different legal force.
- **The UK AI Playbook for Government (Government Digital Service, 2025)** is the central operational guidance for AI use within UK central government. It sets out ten principles covering knowledge of AI capabilities and limitations, lawful and ethical use, secure deployment, meaningful human control, lifecycle management, fitness of tool to task, openness, commercial engagement, skills, and assurance, and a six-stage implementation pathway from problem definition through to decommissioning. Of the four frameworks, the Playbook is the one departments are most directly accountable to in practice.

6.2. The Mapping

Table 3 sets out, for each pillar, the coverage code assigned to each of the four frameworks at that phase of the lifecycle, with the substantive coding rationale provided immediately below the table and the full coding audit trail at Appendix A. The mapping concerns what each framework requires or recommends at each phase. It does not claim equivalence between frameworks of different types.

Table 3. The four pillars mapped against four reference frameworks, with coverage codes per Table 1. Coding rationale per cell follows below; full coding audit trail at Appendix A.

 doi:10.20944/preprints202606.0177.v1

Pillar	NIST AI RMF	OECD AI Principles	EU AI Act	UK AI Playbook
1. Pre-Deployment	Dense	Partial	Dense	Dense
2. Model Activation	Implicit	Implicit	Partial	Implicit
3. Operational Response	Partial	Partial	Dense	Dense
4. Closed-Loop Learning	Implicit	Implicit	Partial	Partial

Coding Rationale per Cell

Pillar 1 (Pre-Deployment). NIST AI RMF: Map function (context, risks, benefits, impacts), supported by Govern (cross-cutting policy and accountability) and Measure (initial metric specification). Densely articulated. OECD: Principles 1.4 (robustness, security, safety) and 1.5 (accountability, traceability) require risk assessment, documentation and impact analysis before deployment, but with limited operational specification. EU AI Act: Chapter III, Section 2, Articles 8 to 15 (risk management system, data governance, technical documentation, transparency to deployers, human oversight, accuracy and robustness); Article 27 (Fundamental Rights Impact Assessment for high-risk public-sector deployments); Article 43 (conformity assessment). UK Playbook: Principles 1, 2, 3, 5 and 6 (knowing AI's limitations, lawful use, secure use, lifecycle management, fitness of tool); stages 0 to 3 of the implementation pathway, from problem definition through development and testing; ATRS submission at deployment.

Pillar 2 (Model Activation). NIST AI RMF: the Measure function applies at activation but the framework does not separate baseline-setting from continuous measurement. OECD: within the call for risk management at each phase of the lifecycle on an ongoing basis (principle 1.5); no distinct activation specification. EU AI Act: Article 72(3) requires the post-market monitoring plan to form part of the technical documentation specified in Annex IV, which arguably requires baseline specification to occur at activation rather than later; Article 19 (automatically generated logs) and Article 26 (deployer obligations) supply supporting structure. UK Playbook: Stage 4 (deployment with monitoring and incident plans) covers activation but does not separate baseline-setting from continuous monitoring; Principle 4 (meaningful human control at the right stage).

Pillar 3 (Operational Response). NIST AI RMF: Manage function (treat risks, respond to incidents, allocate resources); Measure function applied continuously; the framework directs that the Map function be re-applied as context, capabilities, risks and impacts evolve over time. OECD: Principle 1.5 calls for AI actors to apply a systematic risk management approach to each phase of the AI system lifecycle on an ongoing basis, but does not specify operational form. EU AI Act: Article 72 post-market monitoring by providers, which is among the more prescriptive provisions across the four frameworks for continuous post-deployment activity; Article 73 reporting of serious incidents; Chapter IX, Section 4 market surveillance and corrective action. UK Playbook: continuous monitoring guidance (dashboards, drift detection, logging for reproducibility, periodic re-validation, triggered re-assessment); stage 4 of the implementation pathway; Principles 4 and 5.

Pillar 4 (Closed-Loop Learning). NIST AI RMF: the Manage function gestures at incident response and organisational learning but does not specify cross-organisational coordination. OECD: principle 1.5 (accountability) implies feedback through 'ongoing' risk management, but no cross-organisational learning mechanism is specified. EU AI Act: Article 73 serious-incident reporting to authorities; Article 72(2) requirement to evaluate continuous compliance with Chapter III, Section 2;

 doi:10.20944/preprints202606.0177.v1

the European AI Office (Chapter VII) is a potential cross-deployer learning channel but is not yet operationalised in this role. UK Playbook: Principle 7 (open and collaborative); Stage 5 (decommissioning and knowledge capture); central collation of lessons learned is referenced but the mechanism is not mandated.

6.3. Reading the Mapping

Four observations follow from Table 3.

First, Pillar 1 (Pre-Deployment) is densely populated across the framework set. NIST's Map function is a pre-deployment construct, supported by Govern and Measure; the OECD principles 1.4 and 1.5 call for risk and impact assessment before deployment; the EU AI Act's Chapter III, Section 2 is almost entirely concerned with pre-market obligations on high-risk systems, supplemented by Article 27 on FRIAs; and the Playbook's principles 1, 2, 3, 5 and 6 and stages 0 to 3 sit squarely in pre-deployment. This convergence reflects the maturity of pre-deployment thinking and confirms that this is the phase departments will already recognise.

Second, Pillar 2 (Model Activation) is thinly populated in three of the four frameworks. Neither NIST AI RMF nor the OECD principles speaks explicitly to activation as a distinct phase, treating the transition as implicit within broader lifecycle framing. The Playbook addresses activation through stage 4 but does not separate baseline-setting from continuous monitoring. Only the EU AI Act gives Pillar 2 some explicit textual presence, through the requirement at Article 72(3) that the post-market monitoring plan form part of the technical documentation in Annex IV. The relative thinness of Pillar 2 across the rest of the framework set is the principal motivation for treating activation as a distinct pillar.

Third, Pillar 3 (Operational Response) is the most prescriptively addressed pillar in the EU AI Act (Article 72 and Chapter IX) and is also relatively prescriptive in the UK AI Playbook. NIST's Manage and Measure functions speak to the same activity but in less prescriptive terms; the OECD principles refer to systematic ongoing risk management under principle 1.5 but do not specify operational form. The distinguishing feature is therefore not coverage but prescriptiveness.

Fourth, Pillar 4 (Closed-Loop Learning) is universally underspecified. The EU AI Act addresses incident reporting (Article 73) and continuous compliance evaluation (Article 72(2)) but does not establish a structured cross-organisational learning mechanism. NIST's Manage function gestures at organisational learning without specifying cross-organisational coordination. The OECD principles are silent on the mechanism by which lessons travel. The UK Playbook calls for openness and collaboration under principle 7 but does not mandate a feedback architecture. This consistent gap is the second principal motivation for the four-pillar structure.

The four frameworks therefore converge strongly at pre-deployment, are thin on activation, address operational response with varying prescriptiveness, and consistently underspecify closed-loop learning. The novelty of the four-pillar reference model is concentrated in Pillars 2 and 4: the framework set is mature on Pillars 1 and 3 already, and the contribution of this paper is to give Pillars 2 and 4 the textual and institutional presence they currently lack. This matters because monitoring without an activation baseline risks becoming retrospective observation rather than calibrated assurance. The model is offered as compatible with the frameworks rather than competing with them.

6.4. What the Mapping Is for

The mapping has three intended uses. For departmental adoption, it allows a department to read across the row for each pillar and see which framework's requirements or recommendations apply at that phase: a department subject to the EU AI Act for an EU-market system can locate Article 72 obligations cleanly at Pillar 3, with the linked baseline requirement at Pillar 2. For procurement, the mapping is a specification tool: suppliers can be asked, against each cell of Table 3, which obligations or capabilities they support and which they leave to the buying department, with a supplier claiming post-market monitoring under Article 72 required to evidence the activation baselines at Pillar 2 that

 doi:10.20944/preprints202606.0177.v1

make that monitoring meaningful. For policy and governance, the mapping is a diagnostic: the cells the table leaves thin (Pillar 2 across most of the framework set, Pillar 4 across all of it) are natural targets for cross-government coordination through the DSIT Responsible AI Unit, since these gaps are not unique to any one department or supplier and would benefit from shared infrastructure rather than departmental duplication. The mapping is not a compliance instrument; it is an institutional design tool that locates existing instruments and frameworks in a common lifecycle structure.

7. Institutional Fit Analysis

This section examines how the four-pillar reference model could plausibly align with existing UK institutional anchors, and what would be required to test that alignment in practice. The analysis is offered as institutional fit, not policy prescription; the policy decisions, including whether to take any of the observations below forward, belong to the institutions concerned. Three anchors are considered: the DSIT Responsible AI Unit, departmental service assurance functions, and the Crown Commercial Service.

7.1. DSIT Responsible AI Unit

The DSIT Responsible AI Unit is a plausible location for the cross-government functions of the reference model, subject to its remit and capacity at the time of any adoption. Activities that would fit the Unit's existing role include publishing the reference model as guidance complementary to the AI Playbook rather than duplicating it; hosting a cross-departmental Closed-Loop Learning function with anonymised lessons published at a regular cadence; maintaining a measurability register for each cell of the Table 3 mapping that tracks what is currently feasible under typical provider practice; and convening a standing cross-departmental working group to iterate the reference model as practice matures. Whether any of these would be appropriate is a matter for the Unit; the observation here is structural fit, not endorsement.

7.2. Departmental Service Assurance Functions

Departmental service assurance functions sit in a position that maps reasonably well to the practical ownership of Pillars 2, 3 and 4 in many departments, where Pillar 1 typically sits with project teams and ethics reviewers. Activities that would fit such functions include extending scope to include AI-specific instrumentation at Pillar 2 (baseline capture) and Pillar 3 (continuous monitoring), with either embedded AI expertise or access to a shared central function; integrating AI incident management into standard incident processes, with AI-specific escalation paths to the SRO; operating a structured post-incident review process that feeds findings to a cross-government Closed-Loop Learning function; and coordinating with the DSIT Responsible AI Unit on common tooling where department-level implementation would be inefficient. The fit between AI-specific assurance and general service management would need to be tested department by department, since departmental service assurance maturity and AI maturity vary.

7.3. Crown Commercial Service and Procurement

The Crown Commercial Service and departmental commercial functions could provide a useful lever for advancing the measurability frontier at Pillar 3 (Madan and Ashok, 2023). Procurement standards for foundation model services and AI systems could plausibly include, as conditions of contract for risk-tiered deployments: disclosure of training-data provenance sufficient for jurisdictional Legal and Regulatory Compliance assessment; model versioning and change-log discipline sufficient to detect silent updates; contractual access to raw logs and evaluation telemetry for the department's own audit, subject to data protection safeguards; and standardised post-market monitoring reporting, ideally aligned with the EU AI Act Article 72(3) post-market monitoring plan template once the Commission's implementing act is in force. None of these disclosures is novel; they are inconsistently negotiated at present, and a standardised CCS specification could both raise the

 doi:10.20944/preprints202606.0177.v1

floor and reduce the negotiation cost for individual departments. Whether such standardisation is the right CCS instrument is a question for CCS and the departments it serves.

7.4. What Would Be Required to Test Institutional Fit

Institutional fit, as described above, is structural plausibility, not implementation evidence. Testing the fit in practice would require coordinated pilots across two to three departments using existing institutional anchors, with explicit attention to which pillars current functions already cover, which they could reasonably extend to cover, and which would require either capability development or shared infrastructure. The conditions for such an evaluation are set out in detail at Section 8.4.

8. Discussion

8.1. Contribution to Scholarship

The paper provides a lifecycle-complete institutional reference model for public sector AI assurance that is theoretically grounded in Value Sensitive Design, derived through structured comparative analysis of the four leading reference frameworks, and institutionally located against existing UK government structures. It contributes to three strands of scholarship: the operationalisation-of-principles debate (Mittelstadt, 2019; Whittlestone et al., 2019; Morley et al., 2020), by providing a concrete institutional scaffold against which abstract principles can be instantiated at specific phases with specific stakeholders and evidence artefacts; the AI auditing literature (Raji et al., 2020; Mökander et al., 2021; Cobbe, Veale and Singh, 2023), by extending the temporal architecture of audit from a one-shot pre-deployment activity to a continuing lifecycle activity with explicit baseline-setting and closed-loop learning components; and the public sector AI literature (Veale and Brass, 2019; Margetts and Dorobantu, 2019; Misuraca and van Noordt, 2020), by demonstrating that the institutional machinery of UK government AI assurance already contains anchors that could support lifecycle-complete assurance. The distinctive contribution is the specification of Model Activation as a discrete baseline-setting phase and Closed-Loop Learning as a cross-government or cross-deployer institutional mechanism, both of which are thinly addressed in the existing framework set and neither of which is reducible to a renaming of pre-existing constructs. The model therefore contributes less by adding another assurance principle and more by specifying the institutional work required to make existing principles observable, actionable and revisable in service.

8.2. Platform Independence

The reference model is deliberately platform-agnostic. Each cell of Table 3 identifies an assurance activity, not a tool choice. Departments may implement the activities using a combination of existing tools (service management platforms extended with AI capability), open-source libraries, bespoke internal systems, or commercial RAI observability platforms. The paper makes no claim that any one implementation is superior; what matters is that the activity is in place and its output is usable by the stakeholders the pillar specifies.

8.3. Conditions for Empirical Evaluation

The reference model has not yet been evaluated against outcome measures in live service. A rigorous evaluation would require a longitudinal study across multiple departments, measuring whether departments adopting the model detect AI-specific incidents earlier, respond to them more effectively, and avoid repetition of similar incidents over time. Following the DSR convention of distinguishing artefact construction from artefact evaluation (Hevner et al., 2004; Sonnenberg and vom Brocke, 2012), such an evaluation could be designed and coordinated across three to five departmental pilots, with anonymisation protocols agreed in advance. Some conditions for

 doi:10.20944/preprints202606.0177.v1

evaluation appear to be in place, but a formal pilot would still require coordination, capability assessment, departmental sponsorship, and agreed anonymisation protocols.

8.4. Future Empirical Evaluation Design

A future empirical evaluation should be designed to test the reference model against outcomes the model itself is intended to improve. The sketch that follows is offered as a starting point.

Design. A controlled longitudinal study across three to five UK central government departments, each operating at least one production AI system, with a baseline measurement period of six to twelve months prior to adoption and a measurement period of at least eighteen months following adoption. Where possible, comparator departments not adopting the model in the same period would provide a contemporaneous comparison cohort. Pilot departments would adopt the four-pillar structure as a working scaffold for existing assurance instruments, with extension activity concentrated on the cells of Table 3 that are thin or absent in current practice (typically Pillars 2 and 4).

Candidate outcome metrics.

- **Time to detection.** The interval between onset of an AI-specific issue (drift breaching a threshold, an output anomaly cluster, a fairness proxy excursion) and the first recorded internal flag. Explicit baseline-setting at Pillar 2 and AI-specific monitoring at Pillar 3 should shorten this interval relative to general service management alone.
- **Time to escalation.** The interval between detection and SRO-level awareness or formal incident declaration. The stakeholder specification at Pillars 2 and 3 should shorten this interval through explicit escalation paths.
- **Number of repeated incidents.** The frequency, within and across departments, with which an incident class observed in one system recurs in a subsequent system. Pillar 4 should reduce this frequency over time.
- **Quality of assurance artefacts.** Completeness, traceability, and downstream usability of documents produced at each pillar, assessed against a structured rubric agreed in advance with participating departments, drawing on existing rubrics for ML production readiness (Breck et al., 2017).
- **Procurement changes triggered by operational evidence.** The number and significance of changes to procurement specifications, contract terms, or supplier requirements traceable to evidence collected under Pillar 3 and processed under Pillar 4. This metric most directly tests whether the closed loop is closing.

Anonymisation protocols would need to be agreed in advance; comparator selection is non-trivial given variation in departmental risk profile and AI maturity, and matched-pairs design is more practical than randomised allocation. Measurement of time-to-detection requires a working definition of issue onset, which itself depends on the activation baselines the model establishes. The design is a sketch, not a protocol; developing it into a formal evaluation protocol is itself future work, and would benefit from input from a cross-departmental working group of the kind described at Section 7.1.

8.5. Limitations

Beyond the methodological limitations at Section 4.5, three artefact-level limitations are worth acknowledging. First, the model is a design proposal: its acceptance by departments and its behaviour under implementation conditions are empirical questions this paper does not answer. Second, the model assumes a functioning DSIT Responsible AI Unit and an engaged CCS; both exist but neither is fully scaled at the time of writing, and the institutional fit analysis is sensitive to the

 doi:10.20944/preprints202606.0177.v1

pace at which those anchors develop. Third, the four-framework mapping at Table 3 reflects frameworks as published at the time of writing; the EU AI Act is being supplemented by Commission implementing acts, including those specifying the post-market monitoring plan template referenced at Pillar 2. The structure of the mapping is more durable than its specific cell content: a department choosing to test the model against a different framework set (for example, ISO/IEC 42001 alongside or in place of NIST AI RMF) can reproduce the same diagnostic exercise.

8.6. Positioning, Supervisory Review and Conflict of Interest

This paper is sole-authored. The author is a doctoral researcher at the RAID Lab, University of Portsmouth, where continuous lifecycle assurance for public-interest AI is the subject of an ongoing doctoral research programme, and is also the founder and CEO of Responsible Systems Ltd, a company building tooling relevant to Pillars 2 and 3 of the reference model. This is a material conflict of interest. The four-pillar reference model is a product of the doctoral research, derived through the comparative framework analysis presented in Section 6 and developed under doctoral supervision. The academic lineage of the model predates the commercial work.

A sole-author paper from someone with a commercial stake has no internal editorial counterweight. Four structural mitigations apply. First, the reference model is platform-agnostic (Section 8.2): it specifies activities, not tool choices, and the paper makes no claim that any particular product, including any built by Responsible Systems Ltd, is required to implement it. Second, the analytical heart of the paper, the framework mapping at Section 6, is built entirely on public-domain reference frameworks, with all factual claims verifiable against publisher sources, and the argument does not depend on any proprietary taxonomy, dataset, or product feature. Third, the model has been developed under doctoral supervision by Professor Mark Xu, Dr Muhammad Awais Shakir Goraya, and Dr Salem Chakhar; Drs Goraya and Chakhar reviewed the manuscript prior to submission and are independent of any commercial activity associated with the author. Fourth, the practitioner engagement at Section 4.4 was expert-informed artefact refinement, not institutional endorsement by any department, and should not be read as such.

What cannot be mitigated is the absence of a formal peer-review process at the point of first publication or independent co-authorship. Readers should weigh the paper's claims with that absence in view. The framework comparison and the analytical observations at Section 6.3 are the parts of the argument best able to stand on independent verification; the institutional fit observations at Section 7 are presented as options rather than recommendations, subject to the policy judgement of the institutions named.

9. Conclusion

UK central government has built a substantial foundation of pre-deployment AI assurance, anchored in departmental questionnaires, the AI Playbook, ATRS, and the departmental review processes that accompany them. The four-pillar reference model proposed here augments that foundation by extending structured assurance activity across the full AI system lifecycle. Pillar 1 (Pre-Deployment) establishes fitness for service and documents the specification against which future behaviour can be compared. Pillar 2 (Model Activation) establishes the calibrated baseline that makes in-service observation meaningful. Pillar 3 (Operational Response) provides continuous AI-specific visibility alongside general service management. Pillar 4 (Closed-Loop Learning) converts operational experience into cross-government knowledge.

The model was developed within doctoral research at the RAID Lab, University of Portsmouth, derived from structured comparative analysis of the four reference frameworks UK departments most directly encounter, and refined through expert-informed engagement with senior practitioners at DSIT and a second UK central government department in early 2026. The four-framework mapping (Table 3) shows where each framework is dense, thin, or silent across the lifecycle. The two structural findings (that Pillar 2 is thinly addressed in three of the four frameworks, and that Pillar 4

 doi:10.20944/preprints202606.0177.v1

is universally underspecified) support treating the lifecycle as a four-pillar structure rather than relying on the implicit lifecycle framing of any single existing framework. The contribution is concentrated in Pillars 2 and 4: the framework set already addresses Pillars 1 and 3 with reasonable density and prescriptiveness, and the value of the model lies in giving Pillars 2 and 4 the textual and institutional presence they currently lack.

Three institutional anchors fit the lifecycle structure reasonably well: the DSIT Responsible AI Unit as a potential cross-government convener, departmental service assurance functions as practical owners of Pillars 2 through 4, and the Crown Commercial Service as a potential lever for standardised procurement terms. The reference model is offered as a proposal and an invitation to cross-departmental conversation. Its evaluation in operational conditions is the proper subject of future work, and the design and metrics for such an evaluation are set out at Section 8.4. The claim made here is that the model extends existing UK government practice in a coherent and theoretically grounded way, with cell-level mappings against the four leading reference frameworks that departments can verify independently.

Acknowledgments: The author gratefully acknowledges the senior practitioners at the Department for Science, Innovation and Technology and at a second UK central government department whose structured discussions in early 2026 informed the refinement of the reference model presented here. The author acknowledges in particular, the input of Dr Nayyab Naqvi (DSIT), in a personal capacity, on the framing of the model as a public administration instrument and on the institutional anchors set out in Section 7. Practitioner contributions are acknowledged at individual level only where the individual has given explicit consent. The author acknowledges the RAID Lab at the University of Portsmouth, where the doctoral research underlying this paper is conducted, and the supervisory team of Professor Mark Xu, Dr Muhammad Awais Shakir Goraya and Dr Salem Chakhar. Drs Goraya and Chakhar reviewed the manuscript prior to submission. The author alone is responsible for the interpretation and argument presented in this paper, including any errors.

Conflicts of Interest: The author is the founder and CEO of Responsible Systems Ltd, which builds tooling relevant to Pillars 2 and 3 of the reference model. The full disclosure, the structural mitigations that bear on the analytical heart of the paper, and what cannot be mitigated, are at Section 8.6.

Data availability: The paper is based on comparative analysis of four publicly available reference frameworks. All sources are listed in the References and are accessible at the URLs and identifiers provided. The full coding audit trail is provided at Appendix A.

Ethics approval: The expert-informed practitioner engagement reported at Section 4.4 was artefact-refinement activity consistent with the design science research tradition, not interview-based qualitative research. It did not collect personal data, attributable quotations, or organisational confidential information, and did not therefore require institutional ethics review.

Appendix A. Coding Audit Trail

This appendix provides the cell-level coding audit trail referenced at Section 4.2. For each of the sixteen cells of the four-by-four matrix (four pillars × four frameworks), the table below records: the framework source reference (article, principle, function or stage), a summary of the relevant framework text, the assigned coverage code, and the rationale for the coding decision. The audit trail is intended to make the analytical basis of each cell open to inspection and to support any subsequent independent coding pass.

Table A1. Cell-level coding audit trail for the four-by-four framework mapping at Section 6.2 (Table 3).

Pillar	Framework	Source reference	Summary of relevant text	Code	Rationale

 doi:10.20944/preprints202606.0177.v1

1. Pre-Deployment	NIST AI RMF	Map function (1.1 to 5.1); Govern function (1.1 to 6.2); Measure function (1.1 to 4.3)	Map establishes context, identifies and analyses risks and impacts before deployment. Govern provides cross-cutting accountability policy. Measure 1 to 2 covers initial metric specification.	Dense	Activity is explicitly named, operational form is articulated through sub-categories and outcomes, and responsibilities are specified across the four functions.
1. Pre-Deployment	OECD AI Principles	Principles 1.4, 1.5; Accountability principle (post-2024 amendment)	Calls for risk assessment, documentation and impact analysis before deployment.	Partial	Activity is named but operational form is not specified; the principles are values-based rather than procedural.
1. Pre-Deployment	EU AI Act	Chapter III, Section 2, Articles 8 to 15; Article 27 (FRIA); Article 43 (conformity assessment)	Comprehensive pre-market obligations on high-risk systems: risk management, data governance, technical documentation, transparency, human oversight, accuracy and robustness. Fundamental Rights Impact Assessment for public-sector high-risk deployments.	Dense	Activity is explicitly required, operational form is prescribed at article level, and the conformity assessment process specifies responsibilities.
1. Pre-Deployment	UK AI Playbook	Principles 1, 2, 3, 5, 6; Stages 0 to 3 of the implementation pathway; ATRS submission	Knowing AI's limitations, lawful use, secure use, lifecycle management, fitness of tool. Pathway from problem definition through development and testing.	Dense	Activity is named, operational form is articulated through the staged pathway, and ATRS

 doi:10.20944/preprints202606.0177.v1

					provides an evidence anchor.
2. Model Activation	NIST AI RMF	Measure function (1.1 to 4.3)	Measure applies at activation but the framework does not separate baseline-setting from continuous measurement.	Implicit	Coverage can be inferred from the Measure function's application across the lifecycle, but activation is not addressed as a distinct phase with operational form.
2. Model Activation	OECD AI Principles	Principle 1.5 (lifecycle phases, on an ongoing basis)	Risk management at each phase of the AI system lifecycle on an ongoing basis. No distinct activation specification.	Implicit	Activation is implicit within 'each phase of the lifecycle' but not addressed as a distinct activity.
2. Model Activation	EU AI Act	Article 72(3) (post-market monitoring plan as part of Annex IV technical documentation); Article 19 (logs); Article 26 (deployer obligations)	Requires the post-market monitoring plan to be part of technical documentation, which arguably requires baseline specification at activation. Supporting structure via logs and deployer obligations.	Partial	Activity is named through the post-market monitoring plan requirement but operational form for the activation phase specifically is not prescribed; the plan is required but its activation content is not.
2. Model Activation	UK AI Playbook	Stage 4 (deployment with monitoring and incident plans); Principle 4 (meaningful human control)	Stage 4 covers deployment but does not separate baseline-setting from continuous monitoring.	Implicit	Coverage can be inferred from stage 4 but baseline-setting is not addressed as

 doi:10.20944/preprints202606.0177.v1

					a distinct activity with operational form.
3. Operational Response	NIST AI RMF	Manage function (1.1 to 4.4); Measure function applied continuously; Map function re-applied as context evolves	Treat risks, respond to incidents, allocate resources. Continuous measurement. Re-mapping as context, capabilities, risks and impacts evolve.	Partial	Activity is named across functions and operational form is partially articulated, but prescriptiveness is lower than EU Act Article 72.
3. Operational Response	OECD AI Principles	Principle 1.5 (systematic risk management approach on an ongoing basis)	Systematic risk management approach to each phase of the AI system lifecycle on an ongoing basis.	Partial	Activity is named but operational form is not specified; values-based rather than procedural.
3. Operational Response	EU AI Act	Article 72 (post-market monitoring by providers); Article 73 (reporting of serious incidents); Chapter IX, Section 4 (market surveillance and corrective action)	Among the most prescriptive provisions across the framework set for continuous post-deployment activity. Serious incident reporting and market surveillance.	Dense	Activity is explicitly required, operational form is prescribed at article level, responsibilities are specified for providers and surveillance authorities.
3. Operational Response	UK AI Playbook	Continuous monitoring guidance (dashboards, drift detection, logging, periodic re-validation, triggered re-assessment); Stage 4; Principles 4 and 5	Continuous monitoring with explicit guidance on dashboards, drift detection, logging, re-validation.	Dense	Activity is explicitly named and operational form is articulated through stage 4 guidance.

 doi:10.20944/preprints202606.0177.v1

4. Closed-Loop Learning	NIST AI RMF	Manage function (incident response, organisational learning)	Gestures at incident response and organisational learning.	Implicit	Coverage can be inferred from Manage function but cross-organisational coordination is not specified, and a learning loop architecture is not articulated.
4. Closed-Loop Learning	OECD AI Principles	Principle 1.5 (accountability, ongoing)	'Ongoing' risk management implies feedback. No cross-organisational learning mechanism specified.	Implicit	Feedback is implied through the ongoing requirement but the mechanism by which lessons travel is not specified.
4. Closed-Loop Learning	EU AI Act	Article 73 (serious incident reporting); Article 72(2) (evaluate continuous compliance with Chapter III, Section 2); Chapter VII (European AI Office)	Serious incident reporting to authorities; continuous compliance evaluation; Office as potential cross-deployer channel.	Partial	Activity is named through incident reporting and compliance evaluation, but a structured cross-organisational learning mechanism is not established; the AI Office's role in this is not yet operationalised.
4. Closed-Loop Learning	UK AI Playbook	Principle 7 (openness and collaboration); Stage 5 (decommissioning and knowledge capture)	Central collation of lessons learned is referenced; mechanism is not mandated.	Partial	Activity is named through principle 7 and stage 5 but the mechanism is not mandated and operational

 doi:10.20944/preprints202606.0177.v1

					form is not specified.

References

1. Ada Lovelace Institute (2023) Regulating AI in the UK. London: Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/regulating-ai-in-the-uk/ (Accessed: 5 May 2026).
2. Ada Lovelace Institute and DataKind UK (2020) Examining the Black Box: Tools for assessing algorithmic systems. London: Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/ (Accessed: 5 May 2026).
3. Barzelay, M. (2007) 'Learning from second-hand experience: methodology for extrapolation-oriented case research', Governance, 20(3), pp. 521 to 543. doi: 10.1111/j.1468-0491.2007.00369.x.
4. Black, J. and Murray, A.D. (2019) 'Regulating AI and machine learning: setting the regulatory agenda', European Journal of Law and Technology, 10(3).
5. Bowen, G.A. (2009) 'Document analysis as a qualitative research method', Qualitative Research Journal, 9(2), pp. 27 to 40. doi: 10.3316/QRJ0902027.
6. Breck, E., Cai, S., Nielsen, E., Salib, M. and Sculley, D. (2017) 'The ML test score: a rubric for ML production readiness and technical debt reduction', in Proceedings of the 2017 IEEE International Conference on Big Data. Piscataway, NJ: IEEE, pp. 1123 to 1132. doi: 10.1109/BigData.2017.8258038.
7. Brown, S., Davidovic, J. and Hasan, A. (2021) 'The algorithm audit: scoring the algorithms that score us', Big Data and Society, 8(1). doi: 10.1177/2053951720983865.
8. Brundage, M. et al. (2020) Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213.
9. Bullock, J.B. (2019) 'Artificial intelligence, discretion, and bureaucracy', American Review of Public Administration, 49(7), pp. 751 to 761. doi: 10.1177/0275074019856123.
10. Cardno, C. (2018) 'Policy document analysis: a practical educational leadership tool and a qualitative research method', Educational Administration: Theory and Practice, 24(4), pp. 623 to 640.
11. Cobbe, J. (2019) 'Administrative law and the machines of government: judicial review of automated public-sector decision-making', Legal Studies, 39(4), pp. 636 to 655. doi: 10.1017/lst.2019.9.
12. Cobbe, J. and Singh, J. (2021) 'Artificial intelligence as a service: legal responsibilities, liabilities, and policy challenges', Computer Law and Security Review, 42. doi: 10.1016/j.clsr.2021.105573.
13. Cobbe, J., Veale, M. and Singh, J. (2023) 'Understanding accountability in algorithmic supply chains', in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). New York: ACM, pp. 1186 to 1197. doi: 10.1145/3593013.3594073.
14. Costanza-Chock, S., Raji, I.D. and Buolamwini, J. (2022) 'Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem', in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). New York: ACM, pp. 1571 to 1583. doi: 10.1145/3531146.3533213.
15. Cross, N. (2001) 'Designerly ways of knowing: design discipline versus design science', Design Issues, 17(3), pp. 49 to 55. doi: 10.1162/074793601750357196.
16. Department for Science, Innovation and Technology (2024) Introduction to AI Assurance. London: DSIT. Available at: https://www.gov.uk/government/publications/introduction-to-ai-assurance (Accessed: 5 May 2026).
17. Engin, Z. and Treleaven, P. (2019) 'Algorithmic government: automating public services and supporting civil servants in using data science technologies', The Computer Journal, 62(3), pp. 448 to 460. doi: 10.1093/comjnl/bxy082.
18. European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024.

 doi:10.20944/preprints202606.0177.v1

19. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. (2020) Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research Publication No. 2020-1. doi: 10.2139/ssrn.3518482.
20. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. and Vayena, E. (2018) 'AI4People: an ethical framework for a good AI society: opportunities, risks, principles, and recommendations', Minds and Machines, 28(4), pp. 689 to 707. doi: 10.1007/s11023-018-9482-5.
21. Friedman, B. (1996) 'Value-sensitive design', interactions, 3(6), pp. 16 to 23. doi: 10.1145/242485.242493.
22. Friedman, B. and Hendry, D.G. (2019) Value Sensitive Design: Shaping Technology with Moral Imagination. Cambridge, MA: MIT Press.
23. Friedman, B., Kahn, P.H. and Borning, A. (2006) 'Value sensitive design and information systems', in Zhang, P. and Galletta, D. (eds.) Human-Computer Interaction and Management Information Systems: Foundations. Armonk, NY: M.E. Sharpe, pp. 348 to 372.
24. Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014) 'A survey on concept drift adaptation', ACM Computing Surveys, 46(4), Article 44, pp. 1 to 37. doi: 10.1145/2523813.
25. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daume III, H. and Crawford, K. (2021) 'Datasheets for datasets', Communications of the ACM, 64(12), pp. 86 to 92. doi: 10.1145/3458723.
26. Government Digital Service (2025) Artificial Intelligence Playbook for the UK Government. London: Department for Science, Innovation and Technology. Available at: https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government (Accessed: 5 May 2026).
27. Greene, D., Hoffmann, A.L. and Stark, L. (2019) 'Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning', in Proceedings of the 52nd Hawaii International Conference on System Sciences. doi: 10.24251/HICSS.2019.258.
28. Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004) 'Design science in information systems research', MIS Quarterly, 28(1), pp. 75 to 105. doi: 10.2307/25148625.
29. Holstein, K., Wortman Vaughan, J., Daume III, H., Dudik, M. and Wallach, H. (2019) 'Improving fairness in machine learning systems: what do industry practitioners need?', in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York: ACM, pp. 1 to 16. doi: 10.1145/3290605.3300830.
30. International Organization for Standardization (2023a) ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system. Geneva: ISO.
31. International Organization for Standardization (2023b) ISO/IEC 23894:2023 Information technology - Artificial intelligence - Guidance on risk management. Geneva: ISO.
32. Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', Nature Machine Intelligence, 1(9), pp. 389 to 399. doi: 10.1038/s42256-019-0088-2.
33. Klaise, J., Van Looveren, A., Cox, C., Vacanti, G. and Coca, A. (2020) 'Monitoring and explainability of models in production', arXiv:2007.06299.
34. Kop, M. (2021) 'EU Artificial Intelligence Act: the European approach to AI', Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust and IPR Developments, Stanford University, Issue No. 2/2021.
35. Kreuzberger, D., Kuhl, N. and Hirschl, S. (2023) 'Machine Learning Operations (MLOps): overview, definition, and architecture', IEEE Access, 11, pp. 31866 to 31879. doi: 10.1109/ACCESS.2023.3262138.
36. Lewis, J.M. (2021) 'The limits of policy labs: characteristics, opportunities and constraints', Policy Design and Practice, 4(2), pp. 242 to 256. doi: 10.1080/25741292.2020.1859077.
37. Madaio, M.A., Stark, L., Wortman Vaughan, J. and Wallach, H. (2020) 'Co-designing checklists to understand organizational challenges and opportunities around fairness in AI', in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York: ACM, pp. 1 to 14. doi: 10.1145/3313831.3376445.

 doi:10.20944/preprints202606.0177.v1

38. Madan, R. and Ashok, M. (2023) 'AI adoption and diffusion in public administration: a systematic literature review and future research agenda', Government Information Quarterly, 40(1), 101774. doi: 10.1016/j.giq.2022.101774.
39. March, S.T. and Smith, G.F. (1995) 'Design and natural science research on information technology', Decision Support Systems, 15(4), pp. 251 to 266. doi: 10.1016/0167-9236(94)00041-2.
40. Margetts, H. and Dorobantu, C. (2019) 'Rethink government with AI', Nature, 568(7751), pp. 163 to 165. doi: 10.1038/d41586-019-01099-5.
41. McGregor, S. (2021) 'Preventing repeated real world AI failures by cataloging incidents: the AI Incident Database', in Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), pp. 15458 to 15463.
42. Misuraca, G. and van Noordt, C. (2020) AI Watch - Artificial Intelligence in public services: overview of the use and impact of AI in public services in the EU. EUR 30255 EN. Luxembourg: Publications Office of the European Union. doi: 10.2760/039619.
43. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019) 'Model cards for model reporting', in Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). New York: ACM, pp. 220 to 229. doi: 10.1145/3287560.3287596.
44. Mittelstadt, B. (2019) 'Principles alone cannot guarantee ethical AI', Nature Machine Intelligence, 1(11), pp. 501 to 507. doi: 10.1038/s42256-019-0114-4.
45. Mokander, J., Morley, J., Taddeo, M. and Floridi, L. (2021) 'Ethics-based auditing of automated decision-making systems: nature, scope, and limitations', Science and Engineering Ethics, 27(4), 44. doi: 10.1007/s11948-021-00319-4.
46. Mokander, J., Sheth, M., Watson, D.S. and Floridi, L. (2023) 'The switch, the ladder, and the matrix: models for classifying AI systems', Minds and Machines, 33, pp. 221 to 248. doi: 10.1007/s11023-022-09620-y.
47. Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2020) 'From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices', Science and Engineering Ethics, 26(4), pp. 2141 to 2168. doi: 10.1007/s11948-019-00165-5.
48. NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology. doi: 10.6028/NIST.AI.100-1.
49. OECD (2024) Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. Paris: OECD. Originally adopted 22 May 2019, amended 3 May 2024. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (Accessed: 5 May 2026).
50. Paleyes, A., Urma, R.-G. and Lawrence, N.D. (2022) 'Challenges in deploying machine learning: a survey of case studies', ACM Computing Surveys, 55(6), Article 114, pp. 1 to 29. doi: 10.1145/3533378.
51. Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007) 'A design science research methodology for information systems research', Journal of Management Information Systems, 24(3), pp. 45 to 77. doi: 10.2753/MIS0742-1222240302.
52. Raji, I.D. and Buolamwini, J. (2019) 'Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products', in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, pp. 429 to 435. doi: 10.1145/3306618.3314244.
53. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020) 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing', in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). New York: ACM, pp. 33 to 44. doi: 10.1145/3351095.3372873.
54. Rakova, B., Yang, J., Cramer, H. and Chowdhury, R. (2021) 'Where responsible AI meets reality: practitioner perspectives on enablers for shifting organizational practices', Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), pp. 1 to 23. doi: 10.1145/3449081.
55. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M. (2021) 'Everyone wants to do the model work, not the data work: data cascades in high-stakes AI', in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York: ACM, paper 39. doi: 10.1145/3411764.3445518.
56. Sandvig, C., Hamilton, K., Karahalios, K. and Langbort, C. (2014) 'Auditing algorithms: research methods for detecting discrimination on internet platforms', paper presented at Data and Discrimination:

 doi:10.20944/preprints202606.0177.v1

Converting Critical Concerns into Productive Inquiry, 64th Annual Meeting of the International Communication Association, Seattle, WA.

57. Schuett, J. (2023) 'Risk management in the Artificial Intelligence Act', European Journal of Risk Regulation, 15(2), pp. 367 to 385. doi: 10.1017/err.2023.1.
58. Schuett, J. (2024) 'Three lines of defence against risks from AI', AI and Society. doi: 10.1007/s00146-023-01811-0.
59. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F. and Dennison, D. (2015) 'Hidden technical debt in machine learning systems', in Advances in Neural Information Processing Systems 28 (NIPS 2015). Red Hook, NY: Curran Associates, pp. 2503 to 2511.
60. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J. (2019) 'Fairness and abstraction in sociotechnical systems', in Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). New York: ACM, pp. 59 to 68. doi: 10.1145/3287560.3287598.
61. Simon, H.A. (1996) The Sciences of the Artificial. 3rd edn. Cambridge, MA: MIT Press.
62. Sonnenberg, C. and vom Brocke, J. (2012) 'Evaluations in the science of the artificial: reconsidering the build-evaluate pattern in design science research', in Peffers, K., Rothenberger, M. and Kuechler, B. (eds.) Design Science Research in Information Systems. Advances in Theory and Practice. DESRIST 2012. Lecture Notes in Computer Science, vol. 7286. Berlin: Springer, pp. 381 to 397. doi: 10.1007/978-3-642-29863-9_28.
63. Umbrello, S. and van de Poel, I. (2021) 'Mapping value sensitive design onto AI for social good principles', AI and Ethics, 1(3), pp. 283 to 296. doi: 10.1007/s43681-021-00038-3.
64. Veale, M. and Brass, I. (2019) 'Administration by algorithm? Public management meets public sector machine learning', in Yeung, K. and Lodge, M. (eds.) Algorithmic Regulation. Oxford: Oxford University Press, pp. 121 to 149. doi: 10.1093/oso/9780198838494.003.0006.
65. Vogl, T.M., Seidelin, C., Ganesh, B. and Bright, J. (2020) 'Smart technology and the emergence of algorithmic bureaucracy: artificial intelligence in UK local authorities', Public Administration Review, 80(6), pp. 946 to 961. doi: 10.1111/puar.13286.
66. Whittlestone, J., Nyrup, R., Alexandrova, A. and Cave, S. (2019) 'The role and limits of principles in AI ethics: towards a focus on tensions', in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, pp. 195 to 200. doi: 10.1145/3306618.3314289.
67. Yeung, K. (2018) 'Algorithmic regulation: a critical interrogation', Regulation and Governance, 12(4), pp. 505 to 523. doi: 10.1111/rego.12158.