

# A Chain Is Only as Strong as Its Weakest Link: A Scoping Review of System Integration Audits in AI

Leah Davis,<sup>1</sup> Dominic Martin,<sup>2</sup> AJung Moon<sup>1</sup>

<sup>1</sup>McGill University, Montréal, QC

<sup>2</sup>Université du Québec à Montréal, Montréal, QC

leah.davis@mail.mcgill.ca, martin.dominic@uqam.ca, ajung.moon@mcgill.ca

## Abstract

As artificial intelligence (AI) systems are integrated into diverse interfaces and applications, model-centric audits are insufficient to address risks arising from their deployment. In safety-critical domains such as aerospace, *system integration* has long been a focal point for technical system audits. Yet, its role in AI auditing remains underexplored. Scanning 4,259 documents, we present a scoping review of AI audits that treat integration as a core tenet of evaluation ( $n = 58$ ). We find that this emerging form of AI auditing extends beyond individual components to evaluate interactions within an AI system, between an AI system and its deployment environments, and across multiple interconnected AI systems. Specifically, these audits aim to identify integration-related risks, coordinate highly distributed evaluation tasks, and support end-to-end assurance. Unlike other AI audits, they also assess whether the intermediaries connecting elements operate in alignment (compatibility), whether sufficient information is available to assess the system as a whole (completeness), and whether responsibilities are clearly assigned (oversight). Despite this progress, system integration audits in AI remain fragmented, with few explicit measures targeting risks that emerge at system interfaces or their interdependencies. Like other AI audits, meeting traditional audit expectations, such as independence, remains limited by upstream access constraints. We call on the AI auditing community to prioritize system integration as a core strategy for capturing failures beyond the reach of component-level evaluation.

## 1 Introduction

In 2018 and 2019, two Boeing 737 MAX 8 aircraft crashed shortly after take-off, resulting in 346 fatalities (Malunga et al. 2022). Two *system integration failures*—defined as breakdowns in the coordinated functioning of interconnected components, subsystems, and the actors responsible for them—were primary contributors to the tragedy (Madni and Sievers 2014b). Incident analyses report that (1) the technical integration of the aircraft’s control software was updated and deployed without adequate review (Nartey 2025; Nicas et al. 2019; Prokop 2024), and (2) the *sociotechnical nature* of the aircraft was undermined, as pilots were not provided with adequate training to manage related failures (Appicharla 2023; Lengyel, Newman, and Yurovich

2023). Incidents like these are rare in highly regulated industries such as aerospace, where safety is understood as a property of a *system*—defined by seminal systems-thinking scholar Donella Meadows (2008) as a set of elements and interconnections with a function or purpose. In these domains, the integration of components and subsystems is subject to dedicated audits, which we collectively refer to as *system integration audits* (Pop, Titu, and Pop 2023). The presence of these incidents in such an established industry foreshadows a stark warning for the artificial intelligence (AI) community, which lacks comparable safeguards.

Integration risks are of particular concern in the context of general-purpose artificial intelligence (GPAI) and foundation models. A small number of upstream providers supply these models as components that downstream system integrators rebrand, modify, and build upon to deploy millions of AI systems (Williams, Schuett, and Anderljung 2025). Yet despite having little control over the upstream technologies on which they depend, integrators often bear legal responsibility for system failures (Ebers and Penagos 2026). These vulnerabilities also affect the safety management of system-level faults. *Belo*, an Argentinian fintech start-up, provides a cautionary tale: without warning, the firm’s access to Anthropic’s Claude had been erroneously revoked, disrupting service for more than 60 of its downstream accounts—with nothing but a Google Form for recourse (Sisodia 2026). The case illustrates a broader problem: the abundance of model-centric evaluation practices is not suited to discover and prevent failures that occur at the site of interdependencies and system interactions.

As a first step in addressing these gaps, we present a scoping review exploring the conceptualization and application of system integration audits in AI. Drawing on a broad set of literature, we address three questions:

- **RQ1** Across which system types and application areas are system integration audits conducted, and which integration sites are targeted by these audits?
- **RQ2** How does the auditing of these sites provide, or fail to provide, evaluation coverage across AI systems?
  - **RQ 2.1** Who is involved in the audit?
  - **RQ 2.2** What qualities do they evaluate?
  - **RQ 2.3** What are the outputs of an audit and what functions do they serve?

- **RQ3** Which factors must be considered in institutionalizing system integration audits in AI?

In this article, we first describe what established system integration audit practices look like in safety-critical industries, then articulate their relevance to integration-related vulnerabilities and regulatory gaps in AI (Section 2). Following the scoping review method (Section 3), we present the results in four parts: the three central sites where these audits are conducted; how audit roles vary across them; the qualities assessed; and the four functions they serve (Section 4). We present our findings in Section 5, emphasizing that system-level audit expectations remain underdeveloped and limited by restricted information and resource access. We conclude by outlining the steps to advance system integration audits in practice (Section 6).

## 2 Background

### 2.1 Existing System Integration Audits

System integration audits have existed in the aerospace industry since the 1960s (Zieke 1963). Designed to identify, manage, and prevent unintended behaviours arising from interactions among components and subsystems, they evaluate numerous integration sites across environmental control systems (cabin pressure), electronic and hydraulic suites (landing gear), and mechanical elements (fuel injection) (Moir and Seabridge 2011). Today, these audits have become well-institutionalized, mandatory protocols for all aircraft through industry standards (e.g., Society of Automotive Engineers (SAE) Aerospace Recommended Practice 4754B) and oversight from accreditation bodies (e.g., the International Civil Aviation Organization, European Union (EU) Aviation Safety Agency, and the Federal Aviation Administration (FAA)) (Pop, Titu, and Pop 2023; SAE 2026). Table 1 lists common integration faults evaluated by these assessments, adapted from Winter’s (2020) aerospace development case study, alongside our interpretation of the risks they pose.

To address these vulnerabilities, aerospace certification processes apply several audit conditions: (1) independent, external auditors; (2) systematic and traceable processes in accordance with a standard or certification framework; and (3) requirements and evidence tested at the same level as the faults themselves (i.e., system-level faults necessitate system-level tests) (Pop, Titu, and Pop 2023; Winter 2020). Together, these conditions promote rigour, enabling credibility, due process, and legitimacy. Despite the aerospace industry being one of the most mature examples, these audits are applied across many kinds of systems. In the medical device industry, supplier audits (e.g., International Organization for Standardization (ISO) 13485) qualify as a form of system integration audit, as they aim to ensure that externally sourced components are satisfactory for use within a device (ISO 2025a; Smogavc Cestar and Harer 2023). The defence industry similarly relies on configuration and subcontractor integration audits (Summers 2013). By contrast, AI systems, which remain vulnerable to comparable integration faults, remain largely unprotected by the same set of audit conditions.

| Integration Faults                                                                                                     | Associated Risks                                                                                                                                                               |
|------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Poor sequencing</b> (e.g., transitions between components and subsystems)                                           | Propagation of software failures, loss of system control, operational shutdowns in critical infrastructure                                                                     |
| <b>Inadequate fail-safe margins</b> (e.g., the addition or removal of components or systems reduces safety margins)    | Increased likelihood of accidents, reduced tolerance for software or sensor errors, greater vulnerability to environmental disturbances                                        |
| <b>Improper indication signalling</b> (e.g., relevant actors and parties are not alerted to system faults)             | Delayed responses to critical system failures, decisions made by auditors or end-users based on incomplete information, escalation of manageable failures into major incidents |
| <b>Unanticipated outputs from unknown inputs</b> (e.g., unforeseen stakeholder responses or emergent system functions) | Unsafe automated responses, unclear attribution of liability, erosion of public trust                                                                                          |

Table 1: Common system integration faults detailed by Winter (2020), followed by our analysis of associated risks.

### 2.2 Integration-Related Vulnerabilities in AI Evaluation

Traditionally, algorithmic systems were developed to operate on in-house machines or as single-vendor stacks, where their quality assurance could be readily provided. Today, software systems involve extensive chains of interdependencies that are increasingly organized around GPAI models (Varoquaux, Luccioni, and Whittaker 2025).

The EU AI Act defines a *GPAI model* as one that “displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market, and that can be integrated into a variety of downstream systems or applications” (Art. 3(63), Regulation (EU) 2024/1689). Upstream GPAI model providers supply large-language models (LLMs), text-to-image models, and cyber-physical models to system integrators, who embed them into context-specific deployment environments to create *GPAI systems* (Art. 3(66), Regulation (EU) 2024/1689; Brohan et al. 2024). Rather than developing foundation models in-house, organizations commonly access them through cloud-based model-as-a-service (MaaS) arrangements. By lowering cost and technical barriers, MaaS has made GPAI integration relatively straightforward and inexpensive, drastically expanding the number of downstream integrators (Gan, Wan, and Philip 2023).

As a result, GPAI models are now deployed across a wide range of application domains, from decision-support tools in medical chatbots to software stacks in power grid management (Chow and Li 2025; Xie et al. 2022). Making these capabilities accessible to more organizations is, in principle, a welcome development. However, this widespread adoption also means that, in practice, many more actors face liability exposure for the behaviour of the AI systems they integrate (Williams, Schuett, and Anderljung 2025). Given the diversity of these integrated systems, which gives rise

to novel impacts, regulatory regimes are increasingly concerned with faults stemming from an AI system's overall behaviour, rather than the model itself (Carey 2025).

Yet a significant obstacle remains in providing system-level assurance: the industry's overreliance on component-level evaluations, which restrict the scope of assessment to a single AI pipeline component (Varoquaux, Luccioni, and Whittaker 2025). Rismani et al. (2025) report that over 57.3% of responsible AI measures target the model, while only 3% do so for the full system. Carey (2025) identifies a related regulatory blind spot in the EU AI Act: it continues to emphasize risks stemming from GPAI as *models* rather than *systems*. This distinction has important downstream implications. If an integrator places a GPAI system on the market under its own name or trademark, they become a provider under the EU AI Act. In doing so, they assume significant compliance obligations despite little visibility into a model they did not develop (Art. 25(1)(a), Regulation (EU) 2024/1689). The information, resources, and access needed to identify and prevent harms arising from integration faults (see Table 1) often fail to reach those responsible for managing them.

Without AI governance mandates specifying how responsibilities should travel down the supply chain, downstream actors are left with suboptimal means to evaluate the performance and safety issues they inherit. While model-centric evaluations remain useful for benchmarking components, they provide limited coverage of risks arising from interactions between them and their deployment environments. As Correll et al. (2014) note in their review of evaluation approaches, "...as reductionist techniques attempt to control for more and more components, they become less and less adept at validating particular real-world systems." (p. 2)

### 2.3 Evaluation Gaps Surrounding AI Auditing

Regulators have begun to recognize these system-level dynamics. The EU AI Act recognizes that effective evaluation requires both upstream information sharing practices and downstream quality management audits (Art. 53(1); Annex VII, para 5; Regulation (EU) 2024/1689; Bieger et al. 2016; Rico 2025). Yet the regulation itself offers little guidance on how such audits should be operationalized among these actors: where they should occur within an AI system, who should be involved, and the qualities or functions they should evaluate. Unlike the aerospace industry's highly structured audit regimes, AI auditing has emerged without clear mandates, leading to diverse approaches varying widely in structure, scope, and methodological rigour (Birhane et al. 2024). As a result, the audit conditions traditionally applied, such as independent oversight, are often absent from practices labelled as AI audits. If AI governance frameworks expect audits to deliver holistic assurance, greater clarity is needed on how they should be done.

## 3 Methodology

To provide a landscape overview of the ways in which system integration is operationalized in AI auditing, we conducted a scoping review using the Preferred Reporting Items

for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) (Tricco et al. 2018). The main inclusion and exclusion criteria are as follows:

1. **Reference to audits:** The document must propose, conduct, support, improve, or review an audit or audit artifact. Benchmarks and other evaluations were excluded *unless* they explicitly referenced auditing. We did not prescribe a single definition of AI auditing or distinguish among auditing approaches.
2. **Relevance to AI:** The audit object or target must be an AI or machine learning (ML) system. Documents using AI solely as a tool to evaluate a system were excluded.
3. **Relevance to system integration:** The audit must examine the integration, interaction, or interconnections beyond the AI model at a broader system-level.
4. **Publication type:** Published articles; book chapters; industry white papers; and pre-prints were included. ISO and Institute of Electrical and Electronics Engineers (IEEE) standards were excluded due to inaccessibility or underdevelopment.
5. **Language:** Only full-text documents in English were included.

We did not discriminate against literature based on publication date or geographical location.

### 3.1 Database Search Strategies and Study Selection

To capture a broad perspective of system integration across disciplines, we queried six computing and social science databases: *Scopus*, *IEEE Xplore*, *ProQuest*, *Web of Science*, *EBSCOhost*, and *WorldCat*. Searches took place between mid- and late November 2024. Limiting our search to the abstract, title, and keyword fields, our queries sought at least one of "algorithmic", "artificial intelligence", "ai", "machine learning", "ml", and at least one of "audit", "audits", "auditing", "auditor", "auditors" occurring within a three-word proximity. Direct searches for "system integration auditing" yielded too many unrelated results because the term is not used systematically. Rather, we analyzed a broad set of literature claiming an auditing approach, then narrowed the corpus. The initial queries resulted in 4,182 documents (Scopus: 1,086, IEEE Xplore: 1,018, ProQuest: 744, Web of Science: 673, EBSCOhost: 615, WorldCat: 46). An additional 77 documents were manually added from the authors' existing collection of AI auditing literature, totalling 4,259 documents for screening (see Figure 1). Details of the query results are available as supplementary material online (Appendix A (Davis, Martin, and Moon 2026)).

The screening was performed using the Covidence platform. First, duplicates (1,443) and retracted (25) documents were automatically removed. An extensive screening phase then took place between December 2024 and February 2025. The lead reviewer, along with two other reviewers, screened the titles and abstracts of all documents against the eligibility criteria (Section 3.1). Two "yes" votes from reviewers advanced documents to full-text review. Documents with a "maybe" or a mismatch in votes were discussed by all three

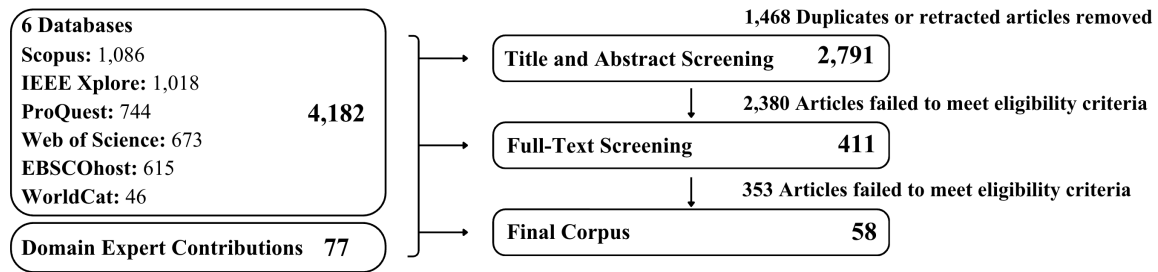


Figure 1: The Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews flow chart of document inclusion.

reviewers to reach a final decision, with an inter-coder agreement of 82% for the first 50 documents. The reviewers met every 100 documents to manage conflicts.

The final corpus comprised 58 documents for qualitative coding, imported into *ATLAS.ti* for analysis. See Appendix B for the full list of included documents (Davis, Martin, and Moon 2026). One-third of the corpus was randomly sampled and independently coded by all reviewers to form an initial coding tree. Reviewers met every 3 to 5 documents to reorganize, edit, merge, and remove codes and broader code groups. Several code groups were mandated for the review, including definitions relevant to “audit” and “system integration”; the system type and application area; the AI system elements implicated; the roles involved; the qualities assessed; and the audit outputs. The lead reviewer then completed a second round of coding across all documents in the corpus.

## 4 Results

Our final corpus of 58 documents treats system integration as a key consideration of the audit approach. The earliest publication dates to 2019, predating the introduction of GPAI. Publications subsequently increased, with 2024 accounting for the largest share of the corpus (29.3%), suggesting a growing demand for system-level evaluations. The increase in publication volume, however, does not indicate methodological maturity. Possibly reflecting the field’s nascency or a lack of collective awareness of existing system integration audits, nearly two-thirds of documents (65.5%) conceptualize or propose an audit approach rather than apply an existing method. Spanning 28 unique approaches, the literature shows considerable variety across auditing frameworks, protocols, mobile applications, schemas, ontologies, simulations, case studies, catalogues, and scorecards. Most of the corpus consists of peer-reviewed articles (43.1% journal, 37.9% conference), with few industry white papers (5.2%), pre-prints (5.2%), and other document types (book chapters, workshop papers, and a policy report).

While the majority of these approaches target purely algorithmic systems, 17.2% evaluate embodied interfaces, including robotic systems. Despite mounting calls for more comprehensive, end-to-end audits of GPAI and generative AI systems, their explicit evaluation is considered in only 13.8% of documents. Rather than concentrating on foundation model architectures alone, the literature spans ten algo-

rithm types, including recommendation algorithms, image classifiers, and knowledge graph networks, indicating that system integration concerns extend beyond any single technical architecture. In fact, one-tenth of documents state that their approach can be applied to any algorithm.

This evaluation diversity extends into a range of sectoral domains. Of the ten application areas identified, audits of healthcare applications (e.g., clinical algorithms, medical AI platforms, and surgical robots) are most dominant at 34.5%, likely due to their safety-critical nature, risk aversion, and low tolerance for failure—similar to aerospace and defence applications. Hiring follows at 12.1%, with transportation, finance, news and social media, and public service following at 5.2% each, and facial recognition technology (FRT), technology services, and advertising at 3.4% each. 31.0% of the documents discuss application-agnostic audits. Each system type or application area was coded more than once if the audit explicitly targeted multiple within its approach.

As many of these application areas operate under different governance arrangements, we also examined the corpus’s geographic distribution. Where this information was reported, audit data collection and implementation took place in many countries and regions, including China, India, New Zealand, Cyprus, the United States, and Western European countries. This variation highlights that integration concerns appear across regulatory jurisdictions.

### 4.1 Where Audits Take Place: Integration Sites Among AI Systems

Building on this contextual overview, we analyzed the documents in our corpus to identify the following characteristics:

- **System boundaries:** the scope of the audit proposed or conducted;
- **System elements:** the assessment targets encompassed by those boundaries;
- **Intermediaries:** the mechanisms and relationships enabling the flow of information and resources across system boundaries (Madni and Sievers 2014a); and
- **Integration sites:** the recurring points where system elements are integrated and evaluated together.

System boundaries varied across the corpus according to audit scope, but collectively encompassed six recurring system elements: data/input, model, output, user interaction, deployment environment (the place of system operation), and

ecosystem (the broader sociotechnical context). While the first four correspond to components within the AI development pipeline, the deployment environment and ecosystem elements extend beyond it. See Appendix C for detailed descriptions of each element (Davis, Martin, and Moon 2026).

We examined how these system elements were evaluated together within each audit approach, along with the intermediaries supporting their integration. This led us to identify three primary integration sites from the corpus: (1) *inter-component integration*, occurring between two or more pipeline components; (2) *system-environment integration*, occurring between an AI system and an external environment(s); and (3) *multi-system integration*, occurring between two or more AI systems (see Figure 2 for a high-level overview of each site). Below, we introduce each site with an illustrative example and describe its characteristics.

**(1) Inter-component integration:** *E.g., An audit that examines whether API transmission of patient electronic health records to a diagnostic model introduces data formatting errors that affect model outputs.*

Audits at this site operate at the most granular scope, with system boundaries encapsulating two or more development components (e.g., data/input, model, output, user interaction) within a single AI system. Across the corpus (n = 58), system boundaries most frequently encompass the data/input element (87.9% of documents), followed by the model (67.2%), output (65.5%), and user interaction (48.3%). Intermediaries at this site are technical, operating in a linear, sequential fashion with direct input-output relationships (e.g., data-to-model and model-to-output), akin to what system integration audits in aerospace assess. Representative examples of these intermediaries in AI systems include application programming interfaces (APIs), microservices, and graphical user interfaces (see Table 2). These sites align with Rajabalinejad’s (2019) definition of subsystem integration. See (1) in Figure 2.

**(2) System-environment integration:** *E.g., An audit that evaluates how physicians incorporate a hospital’s AI diagnostic support system’s recommendations into patient triage workflows.*

These audits evaluate a single AI system’s integration with an external environment(s). Unlike inter-component integration, which encapsulates development components within an AI system, system boundaries at this site extend beyond the AI development pipeline to user-facing settings. In our final corpus, system boundaries encompass deployment environments in 51.7% of documents and broader sociotechnical ecosystem contexts in 77.6%. Rather than reducing an AI system into discrete components, these audits take a constructivist stance, treating an AI system as a *whole*, “...since a system’s characteristics may be different than an amalgamation of the characteristics of its parts” (Gursoy and Kakadiaris 2022, p. 9). While these audits most commonly target task-specific systems, Figure 2 shows that they can evaluate upstream GPAI systems used directly by end users

| Integrated Components     | Intermediary Examples                                                                                                                   |
|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| Data → Model              | Datasheets, data-sharing agreements, human-annotation platforms, data acquisition systems, schema-enforcing systems, and legacy systems |
| Model → Output            | Model cards, application programming interfaces (APIs), and microservices                                                               |
| Output → User-interaction | Emulators, simulators, wrappers, system prompts, and graphical user interfaces                                                          |

Table 2: Intermediaries in inter-component integration sites.

without downstream modification (e.g., via an LLM API).

Intermediaries at this site are sociotechnical, operating through feedback loops and other coupling effects involving stakeholders and, occasionally, physical spaces for embodied AI (see the circular arrows in Figure 2). They enable the assessment of “the final, social process in which the decision making is embedded” (Hauer et al. 2021, p. 2) by facilitating information flows between an AI system and its stakeholders and/or environment. Example intermediaries include regulatory sandboxes, post-market surveillance analyses, scenario testing, feasibility and ethnographic studies, end-user questionnaires, and impact assessments. Compared with inter-component integration, the intermediaries at these sites place greater emphasis on documentation, stakeholder engagement, and remediation pathways than on technical interfaces alone. As such, these sites closely resemble forms of “human-systems integration” (Madni and Sievers (2014b), p. 48). See (2) in Figure 2.

**(3) Multi-system integration:** *E.g., An audit that assesses how patient-record summaries generated by a commercial LLM (one system) feed into a clinical risk prediction model (another system) used for treatment decisions.*

These audits operate at the broadest scope, evaluating the integration of two or more AI systems. System boundaries encompass both inter-component and system-environment sites, implicating internal development components together with deployment and/or ecosystem elements. Unlike the first two sites, which span a range of AI system types, these sites predominantly involve GPAI, generative AI, or physical AI. Intermediaries at this site reflect the asymmetrical upstream-to-downstream relationships that often characterize AI supply chains. They mediate both technical flows of evaluation information and resources (see Table 2) and broader sociotechnical relationships influenced by power, control, and access. In safety engineering literature, these sites are referred to as system-of-systems integration, in which multiple systems enable capabilities beyond those of any individual system (Madni and Sievers 2014a). See (3) in Figure 2.

**Worked example across all sites:** To illustrate how an audit can evaluate multiple integration sites in AI, we consider Mökander et al.’s (2025) framework, which coordinates a three-layered audit approach for generative AI sys-

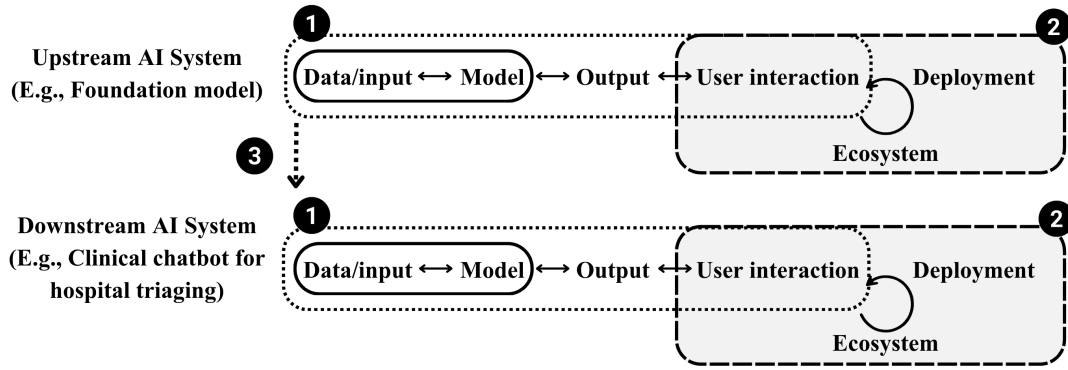


Figure 2: Three system integration sites are identified in this review: (1) inter-component integration (solid lines), (2) system-environment integration (dashed lines and shaded areas), and (3) multi-system integration (dotted lines). The relative sizes of the boxes are not indicative of element proportions.

tems. In the *model-layer*, a foundation model is evaluated by examining how its training data shapes model design and behaviour (inter-component integration). The *application-layer* then evaluates a task-specific application, assessing its performance and contextual suitability at the system-level (system-environment integration). Finally, the *governance-layer* examines the governance, access, and dissemination of the model across organizations and services (multi-system integration). We emphasize that integration sites are complementary, not mutually exclusive. To an extent, these sites provide natural leverage points in the system where interventions (e.g., further testing or remedy) can occur.

## 4.2 Who is Involved in the Audits: Role Types

After mapping the integration sites targeted by the audit approaches, we next examine the roles of the individuals and organizations involved. The corpus reveals at least ten distinct roles: *auditors* (used verbatim) (44.8%), developers (46.6%), domain experts (44.8%), end-users (39.7%), management (10.3%), policymakers (17.2%), external vendors (25.9%), system operators (13.8%), automated AI and robotic agents (5.2%), and a miscellaneous group (6.9%), averaging three per document.

We find that less than half of the documents (44.8%) in the corpus use the term “*auditor*” directly. Among these, the definition of the term varies widely: Benbouzid et al. (2024) and Chadda et al. (2024) describe how an auditor can be internal to an organization or an external third party in their protocol; in contrast, Wilson et al. (2021) introduce the notion of a collaborative auditor, who remains external to the audited organization but partners with it to obtain the information and access needed for the audit to take place. Most often, the term is used without any indication of the auditor’s position, qualifications, or skill set. More commonly, *who participates in an audit seems to be largely predetermined by the level of access and knowledge they hold at each integration site*, rather than by their professional title.

*Developers* of the system being audited, for instance, are involved in 46.6% of audits, particularly when access to the data and model components, and their technical intermedi-

aries (e.g., APIs), is required. This involvement is exemplified by Chadda et al.’s (2024) audit of named entity recognition technology bringing together both data scientists and ML engineers since they “train models and create the housing around them” (p. 23,037). Developers frequently audit inter-component sites.

In 44.8% of documents, *domain experts* are involved when expertise is required to make judgment calls, whether to formulate the initial audit expectations or interpret subjective results. Radiya-Dixit and Neff’s (2023) audit scorecard, for instance, asks how much third-party experts are consulted in FRT implementation to ensure that appropriate oversight mechanisms exist. Surprisingly, in 39.7% of documents, *end-users* even assume lead auditing roles to report user-facing faults in the form of co-audit (Gordon et al. 2023), herd audit (Yang, Zhang, and Zhu 2023), and end-user audit approaches (Lam et al. 2022). Audits led by these actors primarily involve system-environment sites.

In multi-system integration sites, involvement shifts toward *management* (involved in 10.3% of audits) and policy-related roles (17.2%), in which organizational structuring, procurement, and AI governance become central concerns. Consistent with this finding, the above worked example draws on Mökander et al. (2025), whose *governance-layer audit* implicates management in data sourcing and documentation protocols for generative AI systems. Although not exclusive to this integration site, multi-system integration most commonly extends beyond internal actors. 25.9% of documents discuss situations in which *external vendors* (e.g., technology providers, manufacturers, and suppliers) are directly involved in audits, not to lead, but to serve as instrumental contributors to the audit trail. Rather than a single actor or organization possessing end-to-end visibility, different actors and organizations contribute evidence from different parts of the AI supply chain. Ferrell and Anderegg (2020) underscore the importance of external actors as “true safety partner[s]” since “managing suppliers and integrating products from different sources are the most vulnerable points in safety management” (p. 6).

As noted above, role ambiguity may be partly attributable

| Integration Site          | Location                            | System Boundary Encapsulates                             | Form of Intermediaries   | Commonly Associated Roles                       | System Type Targeted                                 |
|---------------------------|-------------------------------------|----------------------------------------------------------|--------------------------|-------------------------------------------------|------------------------------------------------------|
| <b>Inter-component</b>    | Internal development                | Data/input, model, output, and user-interface components | Primarily technical      | Developers                                      | Any system                                           |
| <b>System-environment</b> | External user-facing environments   | Deployment environments and broader ecosystem elements   | Primarily sociotechnical | Domain experts and end-users                    | Any system                                           |
| <b>Multi-system</b>       | Both internal and external settings | All elements                                             | Both                     | Management, policy-makers, and external vendors | General-purpose, generative, and physical AI systems |

Table 3: Summary of integration site and role characteristics in artificial intelligence systems.

to the distributed nature of audit responsibilities. Nearly two-thirds of documents (65.5%) describe *joint auditing processes* in which multiple roles share oversight responsibilities. Audits involving purposeful collaboration often imply that gathering as many perspectives as possible promotes mutual learning, understanding, and communication (Liu et al. 2022; Oala et al. 2020). This is underscored in a study by Mahajan et al. (2020) that audits an AI-enabled radiology system, detailing how performance must be examined by “a radiologist and a data scientist [who] need to sit together for hours at end and read every case where the AI failed...” to examine the implications of erroneous AI behaviour (p. 133). Yet in these cases, who serves as “the auditor”—who leads the process and is ultimately responsible for the audit outcome—often remains implicit. This distribution of roles reflects the distributed nature of AI supply chains: no single actor typically possesses the information, expertise, resources, or access needed to evaluate the entire system. Table 3 summarizes integration site and role characteristics.

### 4.3 What the Audits Assess

Subsequently, we examined the qualities that the audits in our corpus assess, including those specific to integration, the outputs they produce, and the functions those outputs serve.

**Evaluated qualities** We find that our corpus assesses a wide range of qualities (33 in total). On average, an audit approach covers seven distinct qualities, with one document assessing as many as 18. Technical performance (69.0%), transparency (56.9%), accountability (48.3%), robustness (48.3%), lack of bias (44.8%), consistency (39.7%), fairness (31.0%), process improvements (29.3%), trust (27.6%), and non-discrimination (25.9%) are the ten most frequently assessed. This distribution aligns with existing characterizations of AI auditing, which emphasize technical performance or popular responsible AI principles (Koshiyama et al. 2024).

Part of this breadth reflects the tendency of the audits to nest several qualities within broader “umbrella” terms. Helmer et al. (2024), for instance, use *trust* to group six other qualities: *fairness*, *autonomy*, *transparency*, *reliability*, *safety*, and *security*, despite limited discussion about why trust represents these qualities. The tendency to conflate evaluation constructs also extends to the object of evaluation

itself. Several documents infer *system-level* claims exclusively from *model-level* evidence. This pattern is reflected in DeGrave et al.’s (2025) audit framework for a melanoma classification *AI system*, which reports *model-level* performance metrics such as sensitivity. As a consequence, even qualities that become salient through integration remain comparatively underrepresented. Three *integration-specific qualities* do, however, emerge from the corpus.

**Compatibility (10.3%)** captures whether the intermediaries connecting components and systems operate in alignment. Evaluating compatibility at inter-component and multi-system sites often involves assessing whether technical interfaces remain compatible across the system boundary. Sghaier et al. (2024) illustrate this by evaluating whether the API linking the deep learning module and user interface “conforms to the design principles of the REST architecture (REST API), facilitating user access to available AI models” (p. 2), ensuring standardized communication across a distributed architecture. Technical compatibility can also be reflected in software version synchronization within development practices (i.e., MLOps) and in audits that explicitly account for legacy infrastructure (Helmer et al. 2024; Benbouzid et al. 2024). At system-environment sites, compatibility extends to ‘sociotechnical fit’, evaluating whether the audit remains aligned with the organizational and deployment environments in which the AI system is deployed. Becker and Waihl (2022), for instance, ensure audit activities explicitly support a company’s “digitization and data strategy”, “IT infrastructure goals”, and “cloud data hubs, lakes and warehouses” (p. 288).

**Completeness (8.6%)** refers to whether the information available for evaluation is sufficient. At inter-component sites, this involves assessing the adequacy of information transferred between components for further processing. Knoblauch and Großmann (2023), for instance, draw on the ISO 25012 *Data Quality Model* to evaluate whether the data contains all expected attributes and values it requires, given that the resulting evidence in their audit informs subsequent model quality assessments and, ultimately, system certification (ISO 2025b). At system-environment sites, completeness concerns whether the audit process contains the evidence needed to support its conclusions or any follow-up actions. For example, Hauer et al. (2021) assess whether assurance cases (an auditing approach) contain the assumptions, acceptance criteria, measures, and trade-offs needed to

justify claims that a system will operate as intended within a defined application and environment. Barker et al. (2023) similarly identify completeness as an audit objective, defining it as the need for comprehensive stakeholder feedback and the follow-through of subsequent practitioner updates. Notably, completeness is absent in multi-system sites, likely due to access constraints.

**Oversight (6.9%)** describes the extent to which decision-making authority is established and exercised across integration sites (Lam et al. 2022). At inter-component sites, oversight concerns authority within the development pipeline. Polzer et al. (2022), for instance, incorporate an explainable AI component into their audit approach that translates neural network outputs into visual explanations by highlighting the parts of an input image contributing most to each prediction. This enables evaluators to verify that the AI’s reasoning aligns with predefined decision criteria. At system-environment sites, oversight concerns how authority is retained within operation: to uphold subjective audit judgments (Moreau, Sinatra, and Sekara 2024), govern responsibility over architectural elements (Boltz et al. 2024), sanction deployment decisions (Liu et al. 2022), and determine remediation actions (Sloane, Moss, and Chowdhury 2022). This is illustrated by Moreau et al. (2024) who examine how social workers adjust or maintain their own risk assessment scores relative to those produced by a decision-support algorithm. Oversight also concerns whether audit findings are appropriately communicated to affected stakeholders. Radiya-Dixit and Neff (2023), for instance, develop an audit scorecard assessing how three United Kingdom police forces disclosed their FRT trials. At multi-system sites, oversight extends authority across organizational boundaries. Mökander and Floridi (2023) highlight this in an ethics-based audit of AstraZeneca’s AI development. The audit assessed the extent to which vendor dependence limited auditors’ ability to assess the training data and model architecture. See Appendix D for a summary of all qualities, their proxies, and proportions (Davis, Martin, and Moon 2026).

**Audit outputs and functions** The documents in the corpus produce nine distinct outputs: quantitative metrics (53.4%), descriptive statements (41.4%), binary/discrete outcomes (34.5%), audit summaries/reports (19.0%), question sets (17.2%), comparative analyses (5.2%), audit tasks/requirements (6.9%), standards/certifications (3.4%), and miscellaneous outputs (3.4%). On average, each audit generates three outputs that inform evaluation decisions about the AI system. This raises a further question: what purpose does each output serve? Our analysis reveals four recurring functions (see Table 4).

**(1) Risk exploration** takes stock of vulnerabilities that may arise when an AI component or system is integrated. *Question sets* (in 17.2% of documents) encourage auditees, most often developers and end-users, to consider potential development or deployment risks. Naja et al. (2022), for example, “gathered competency questions relating to accountability of AI systems” to construct a set of eight audit topics and their associated risks (p. 74,385). *Comparative analyses* appear less frequently (5.2%), more specifically examining

| Audit Outputs                                        | Related Function                    |
|------------------------------------------------------|-------------------------------------|
| Question sets and comparative analyses               | Risk exploration—“Guiding”          |
| Quantitative and binary/discrete outputs             | Risk determination—“Concretizing”   |
| Audit tasks/requirements                             | Coordination—“Organizing”           |
| Audit summaries/reports and standards/certifications | Procedural regularity—“Stabilizing” |

Table 4: Mapping audit outputs to functions. Note that the “descriptive statements” output covers all four functions due to its breadth.

potential AI system failures against domain-expert performance, as in Oakden-Rayner et al. (2022), who compare AI classification failure modes with those of radiologists.

These outputs serve an exploratory function by prompting reflexivity: question sets encourage stakeholders to identify potential failures, while comparative analyses reveal discrepancies between AI performance and human judgment. Together, they help evaluators determine which integration sites, elements, and intermediaries warrant closer review (see these relationships in Figure 2). Risk assessments are the most common exploratory instrument, acting as initial scoping and diagnostic mechanisms. Both Benbouzid et al. (2024) and Liu et al. (2022) begin their audits with risk assessments that map which elements (e.g., outputs) and stakeholders should be targeted. Mökander et al. (2025) similarly use risk assessments at multi-system sites, posing open-ended questions to explore upstream inter-component risks and downstream system-environment risks.

**(2) Risk determination** describes the process of characterizing risks based on their priority and severity. Unlike risk exploration, which often involves a wide range of stakeholders, risk determination typically relies on developers or domain experts with the technical access or contextual expertise needed to establish appropriate evaluation thresholds and mitigation strategies (Gerchick et al. 2023).

Two outputs serve a deterministic role: *quantitative outputs* (53.4%), such as performance measures and statistical tests, and *binary/discrete outcomes* (34.5%), represented by pass/fail or categorical ratings. At inter-component sites, audits often employ traditional risk models since failures can be traced through sequential relationships among components (e.g., data-to-model). Adler and Klaes (2022), for instance, implement component fault trees to quantify failure probabilities. At system-environment sites, where performance standards are less agreed upon, domain experts are relied on to interpret findings and make categorical audit judgments. In Gerchick et al. (2023), workers reviewing a child welfare screening tool can either accept or override the system’s risk scores before they are used in practice. Risk determination also appears in multi-system contexts, although visibility constraints often require separate upstream and downstream assessments, as shown by Mökander et al.’s (2025) worked example in Section 4.1. While GPAI providers can evaluate risks associated with their own models, system integrators often lack upstream visibility and as-

sess risks indirectly through their deployments.

(3) **Coordination** refers to the organization, distribution, and alignment of auditing activities across integration sites and their actors. Rather than identifying or confirming risks, this function establishes the operational workflows and communication channels necessary for intermediaries to work: to exchange information, assign responsibility, and manage evidence during the audit. In the corpus, coordination is facilitated by *audit tasks and requirements* (6.9%), which specify the roles and resources required to complete the audit, implicating all three integration sites. Knoblauch and Großmann (2023) coordinate inter-component integration by using audit APIs to standardize the exchange of audit data across MLOps components. At system-environment sites, Krafft et al. (2021) use process flow diagrams to map interactions among stakeholders and AI system elements by analyzing how the output could act as a surveillance monitoring tool—an approach that could be extended to multi-system contexts. Coordination also serves to build consensus on what falls *outside* of the system boundary. As Wilson et al. (2021) note, “just as important as defining what we were auditing is understanding what we were not auditing.” (p. 5)

(4) **Procedural regularity** formalizes traceable and consistent procedures for conducting and documenting audits across integration sites. Whereas coordination aligns audit activities and actors, procedural regularity stabilizes the audit process itself through reproducible chains of reasoning using *audit summaries/reports* (19.0%) and/or *standards/certifications* (3.4%). Knoblauch and Großmann (2023) similarly produce standardized audit reports that regulatory bodies can use to assess compliance for enforcement. A small subset of audits propose certification frameworks, as in Becker and Walzl (2022) and Kazim et al. (2021), who outline how regulators *could* certify an end-to-end system audit.

## 5 Discussion

Motivated by the failures observed in the Boeing 737 MAX 8 incidents, this review began with the premise that neglecting a system integration perspective leaves dangerous technical and sociotechnical gaps. This led us to first identify *where* integration risks (see Table 1) are evaluated in AI systems by characterizing system boundaries, elements, and intermediaries. Across the corpus, this revealed three integration sites within, between, and across AI systems, alongside integration-specific qualities and four audit functions. Collectively, these findings suggest that AI auditing has begun to recognize the importance of system integration. However, this analysis also reveals open challenges for their institutionalization. Building on these findings, we articulate three implications.

### 5.1 Emerging Recognition of System Integration Risks in AI

By characterizing system boundaries, we identified three recurring integration sites and their associated intermediaries (Figure 2), illustrating where integration-related risks in AI

are recognized and evaluated through audit design. Collectively, the integration sites provide distinct opportunities to evaluate different kinds of integration faults and risks.

At *inter-component sites*, *sequencing* faults can be evaluated by examining dependencies among AI pipeline components. Adler and Klaes (2022) evaluate these failure pathways using error-propagation models, including component-fault trees (see Section 4.3). At *system-environment sites*, *indication signalling* faults can be evaluated through sociotechnical intermediaries such as stakeholder engagement, documentation, and remediation actions. These intermediaries enable the assessment of user-facing communication and notification mechanisms, which we explored in Radiya-Dixit and Neff’s (2023) audit of the disclosure mechanisms used in police-enforced FRT trials. Together, these approaches demonstrate that AI audits are beginning to capture and evaluate a range of technical and sociotechnical integration risks analogous to those long addressed in safety-critical industries.

Evaluation was considerably more limited at *multi-system integration sites*, despite the growing importance of auditing GPAI systems. Where these sites were evaluated, risk exploration and risk determination were typically divided between upstream providers (who evaluated technical model risks) and downstream system integrators (who evaluated deployment risks with limited upstream visibility) (see Section 4.3). Consistent with our findings on audit roles, audit evidence was distributed among multiple actors in the AI supply chain, yet few approaches demonstrated end-to-end coordination. This work is nevertheless emerging. Hopkins et al. (2025), for instance, quantitatively show how upstream model provider changes affect downstream task-specific systems by “[modelling] supply chains as directed graphs, where nodes correspond to AI components (and their corresponding actors)” (p. 1266). Such protocols could enable system integration audits to assess upstream and downstream integration risks *simultaneously*, ultimately benefiting from a more unified representation of the supply chain.

However, approaches of this kind remain rare. Unlike aerospace regimes, system integration audits in AI have not yet matured to systematically evaluate integration faults or risks. Few qualities explicitly assess the interfaces, dependencies, and relationships that arise when AI models are integrated into broader sociotechnical systems. The three qualities that do—compatibility, completeness, and oversight—each appear in 10% or less of the corpus. This limited emphasis may reflect the continued dominance of component-level evaluation priorities, which remain highly prevalent even within this corpus. Consequently, institutionalizing system integration audits in AI will require measurement processes and metrics that explicitly target integration faults and their associated risks. In many cases, these metrics may be adapted from existing practices. Lewis and Groth (2022), for example, identify 35 system-level software integration metrics, including *ease of model notification* and *framework coverage for known failure modes*, that could underpin specific audit criteria. Context-specific metrics should also be considered. Time-sensitive healthcare services, for instance, should consider metrics such as *server throughput capacity*

(the total work a system can handle within a given period) to better evaluate *fail-safe margin* faults in clinical AI deployments (Assadi et al. 2022).

## 5.2 Comparison to Traditional Audit Expectations

Given that our analysis centres on audits, which carry higher assurance expectations than AI evaluations, their legitimacy must be examined. In Section 2.1, we identified three audit conditions: independent auditors, systematic adherence to a standard or certification, and system-level evaluation of system-level risks. Given the review criteria, the third condition was readily met; the others, however, were not.

First, most audits in the corpus lack an external, independent auditor (see Section 4.2). In established audit regimes, the role and expectations of external auditors are clearly defined. Under FAA certification, for instance, third-party auditors must complete AS9100 Quality Management System training (International Aerospace Quality Group 2026). By contrast, over half of the documents in the corpus do not specify an explicit “auditor” role; likely reflecting the absence of enforceable standards in the AI auditing profession. More than ten distinct roles are involved in the audits, with more than two-thirds assuming joint evaluation responsibilities. While multi-perspective discourse has benefits, responsibilities were largely determined by the expertise and access available at each integration site (see Table 3). As a result, most actors were neither external nor independent. In many cases, the auditor effectively became the auditee, undermining the audit’s legitimacy (Falco et al. 2021).

The second audit condition reveals similar shortcomings. While individual audits demonstrate procedural regularity on their own, the corpus as a collective does not. We identified more than 28 unique auditing approaches aimed at 33 distinct qualities. Despite many studies proposing requirements or reporting structures, fewer than 5% reference, or even conceptualize, an established standard or certification. As a result, evaluation criteria vary significantly. For an emerging audit class, this goalpost shifting presents a major challenge. Given that over 66% of the audits are conceptualizations or proposals, these findings likely reflect the field’s emerging nature, but may also suggest a gap in translating ongoing standardization efforts in AI auditing research.

As they stand, system integration audits in AI more closely resemble evaluations than formal audits—which are not equivalent. In practice, many audits in the corpus leverage pre-existing AI evaluation types, including those at the component-level, to gather evidence about system behaviour. We encourage the continued development of this audit class to build and extend this rich landscape, serving as a coordinating mechanism capable of addressing gaps that existing metrics and tools cannot capture on their own.

## 5.3 Information and Resource Access as a Constraint

A recurring enabler or blocker in *how* these audits are designed, and *what* they evaluated, was access to information and resources from the auditee. As integration requires dif-

ferent components, systems, and intermediaries to be evaluated *together*, one critical function of these audits is to coordinate activities among multiple parties. Yet only a quarter of audits in the corpus explicitly discuss external vendor involvement, while the remaining 75% use an internal auditing approach. This may explain why developers, managers, and other internal actors currently play significant roles in audits: their involvement is necessary to *overcome* access restrictions to datasets and model parameters and to facilitate remediation processes (see Section 4.2). Even where legal access rights exist, as under the EU’s *General Data Protection Regulation*, weak enforcement makes obtaining such access difficult (Lin 2024).

We emphasize that these access constraints do not hold in other audit domains. Ishibashi et al. (2024), for example, discuss how the harmonized *Medical Device Single Audit Program* enables regulatory authorities and economic agencies to have full access to audited systems. Mökander et al. (2025) similarly elaborate that “this kind of privileged access...is standard practice”, noting that “IT auditors, for example, have full access to material and reports related to operational processes and performance metrics” (p. 13). Intellectual property concerns exist across many industries, but they have not precluded audit access—the AI industry must not be an exception. These concerns will only escalate as GPAI models are integrated into increasingly more diverse interfaces and environments, especially those involving physical hardware. As in other safety-critical domains, downstream actors must have a full systems view of the integration faults, risks, and harms they may face.

## 5.4 Study Limitations

As our queries targeted academic outlets and scientific databases, the screening process may underrepresent industry, grassroots, and community work. Additionally, despite some geographical diversity, the prevalence of European and Western legal frameworks, including the EU AI Act, may limit the representation of Global South contexts. Finally, our review does not include literature published after December 2024, during which integration dynamics for GPAI and physical AI systems have continued to evolve rapidly.

## 6 Conclusion

Over the last few years, system integration audits have emerged in AI, encompassing interactions among pipeline components, deployment environments, and multi-system architectures. Yet these audits reveal significant ambiguity over who bears accountability and how system-level assurance can be obtained. Given the corpus’s emphasis on audit proposals, future work should consider design ethnography to investigate how practitioners adopt, engage with, and resist these technical and sociotechnical structures in practice. As AI systems are embedded in new applications, the auditing community must prioritize dedicated approaches to address integration faults. As the Boeing 737 MAX 8 incidents made clear, system integration is not a peripheral concern; it can be a critical source of failures. Evaluations must be designed to capture the risks that arise when components, systems, and actors are brought together.

## Acknowledgements

We thank the WeRobot 2026 community for their feedback on an earlier draft of this paper. We acknowledge the financial support of the Observatory on the Societal Impacts of Artificial Intelligence and Digital Technologies (Obvia), the Pierre Elliot Trudeau Foundation, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

- Adler, R.; and Klaes, M. 2022. Assurance Cases as Foundation Stone for Auditing AI-Enabled and Autonomous Systems: Workshop Results and Political Recommendations for Action from the ExamAI Project. In Rauterberg, M.; Nah, F.-H.; Siau, K.; Krömker, H.; Wei, J.; and Salvendy, G., eds., *HCI International 2022–Late Breaking Papers: HCI for Today’s Community and Economy*, 283–300. Springer Nature Switzerland.
- Appicharla, S. 2023. The Boeing 737 MAX 8 crashes. *Safety-Critical Systems*, 2(1).
- Assadi, A.; Laussen, P. C.; Goodwin, A. J.; Goodfellow, S.; Dixon, W.; Greer, R. W.; and Mazwi, M. L. 2022. An integration engineering framework for machine learning in healthcare. *Frontiers in Digital Health*, 4: 932411.
- Barker, M.; Kallina, E.; Ashok, D.; Collins, K.; Casovan, A.; Weller, A.; Talwalkar, A.; Chen, V.; and Bhatt, U. 2023. FeedbackLogs: Recording and Incorporating Stakeholder Feedback into Machine Learning Pipelines. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–15.
- Becker, N.; and Walzl, B. 2022. Auditing and Testing AI—A Holistic Framework. In Duffy, V. G., ed., *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Health, Operations Management, and Design*, 283–292. Springer International Publishing.
- Benbouzid, D.; Plociennik, C.; Lucaj, L.; Maftai, M.; Merget, I.; Burchardt, A.; Hauer, M. P.; Naceri, A.; and Van Der Smagt, P. 2024. Pragmatic Auditing: A Pilot-Driven Approach for Auditing Machine Learning Systems. ArXiv:2405.13191.
- Bieger, J.; Thórisson, K. R.; Steunebrink, B. R.; Thorarensen, T.; and Sigurdardottir, J. S. 2016. Evaluation of General-Purpose Artificial Intelligence: Why, What & How. *Evaluating General-Purpose AI*, 9.
- Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 612–643. IEEE.
- Boltz, N.; Getir Yaman, S.; Inverardi, P.; Lemos, R.; Landuyt, D.; and Zisman, A. 2024. Human empowerment in self-adaptive socio-technical systems. In *Proceedings of the 19th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 200–206.
- Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; and Zitkovich, B. 2024. Rt-2: Vision-language-action models transfer web knowledge to robotic control. ArXiv:2307.15818.
- Carey, S. 2025. Regulating Uncertainty: Governing General-Purpose AI Models and Systemic Risk. *European Journal of Risk Regulation*, 1–17.
- Chadda, A.; McGregor, S.; Hostetler, J.; and Brennen, A. 2024. AI Evaluation Authorities: A Case Study Mapping Model Audits to Persistent Standards. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21): 21.
- Chow, J. C.; and Li, K. 2025. Large language models in medical chatbots: opportunities, challenges, and the need to address AI risks. *Information*, 16(7): 549.
- Correll, M.; Alexander, E.; Albers, D.; Sarikaya, A.; and Gleicher, M. 2014. Navigating reductionism and holism in evaluation. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, 23–26.
- Davis, L.; Martin, D.; and Moon, A. 2026. A Chain Is Only as Strong as Its Weakest Link: A Scoping Review of System Integration Audits in AI. <https://arxiv.org/abs/2608.04921>.
- DeGrave, A. J.; Cai, Z. R.; Janizek, J. D.; Daneshjou, R.; and Lee, S. I. 2025. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nature Biomedical Engineering*, 9(3): 294–306.
- Ebers, M.; and Penagos, E. V. 2026. Upstream, downstream, and in between: navigating the GPAI value chain under EU law. *Information & Communications Technology Law*, 1–46.
- 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 and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union*.
- Falco, G.; Shneiderman, B.; Badger, J.; Carrier, R.; Dahbura, A.; Danks, D.; and Yeong, Z. K. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence*, 3(7): 566–571.
- Ferrell, U. D.; and Anderegg, A. H. 2020. Applicability of UL 4600 to Unmanned Aircraft Systems (UAS) and Urban Air Mobility (UAM). In *In 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*, 1–7. IEEE.
- Gan, W.; Wan, S.; and Philip, S. Y. 2023. Model-as-a-service (MaaS): A survey. In *2023 IEEE International Conference on Big Data*, 4636–4645. IEEE.
- Gerchick, M.; Jegede, T.; Shah, T.; Gutierrez, A.; Beiers, S.; Shemtov, N.; Xu, K.; Samant, A.; and Horowitz, A. 2023. The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. In *2023 ACM Conference on Fairness Accountability and Transparency*, 1292–1310.
- Gordon, A. D.; Negreanu, C.; Cambroner, J.; Chakravarthy, R.; Drosos, I.; Fang, H.; Mitra, B.; Richardson, H.; Sarkar, A.; Simmons, S.; Williams, J.; and Zorn, B. 2023. Co-audit: Tools to help humans double-check AI-generated content. ArXiv:2310.01297.

- Gursoy, F.; and Kakadiaris, I. A. 2022. System Cards for AI-Based Decision-Making for Public Policy. *ArXiv:2203.04754*.
- Hauer, M. P.; Adler, R.; and Zweig, K. 2021. Assuring Fairness of Algorithmic Decision Making. In *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 110–113.
- Helmer, L.; Martens, C.; Wegener, D.; Akila, M.; Becker, D.; and Abbas, S. 2024. Towards Trustworthy AI Engineering—A Case Study on integrating an AI audit catalog into MLOps processes. In *Proceedings of the 2nd International Workshop on Responsible AI Engineering*, 1–7.
- Hopkins, A.; Cen, S. H.; Struckman, I.; Ilyas, A.; Videgaray, L.; and Madry, A. 2025. AI Supply Chains: An Emerging Ecosystem of AI Actors, Products, and Services. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1266–1277.
- International Aerospace Quality Group. 2026. IAQG-Sanctioned Programs - New Auditor Development Process (aligned to 9101:2016).
- Ishibashi, K.; Kondoh, M.; and Kusakabe, T. 2024. Current application of the medical device single audit program (MDSAP) as a global regulatory reliance framework for the inspection of medical devices. *Therapeutic Innovation & Regulatory Science*, 58(6): 1172–1179.
- ISO. 2025a. International Organization for Standardization. *Medical devices — Quality Management Systems — Requirements for Regulatory Purposes (ISO Standard, (13485:2016))*.
- ISO. 2025b. International Organization for Standardization. *Software engineering — Software Product Quality Requirements and Evaluation (SQuARE) — Data quality model (ISO Standard, (250121:2008))*.
- Kazim, E.; Koshiyama, A. S.; Hilliard, A.; and Polle, R. 2021. Systematizing Audit in Algorithmic Recruitment. *Journal of Intelligence*, 9(3).
- Knoblauch, D.; and Großmann, J. 2023. Towards a Risk-Based Continuous Auditing-Based Certification for Machine Learning. *The Review of Socionetwork Strategies*, 17(2): 255–273.
- Koshiyama, A.; Kazim, E.; Treleaven, P.; Rai, P.; Szpruch, L.; Pavey, G.; Ahamat, G.; Leutner, F.; Goebel, R.; Knight, A.; Adams, J.; Hitrova, C.; Barnett, J.; Nachev, P.; Barber, D.; Chamorro-Premuzic, T.; Klemmer, K.; Gregorovic, M.; Khan, S.; Lomas, E.; Hilliard, A.; and Chatterjee, S. 2024. Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. *Royal Society Open Science*, 11(5): 230859.
- Krafft, P. M.; Young, M.; Katell, M.; Lee, J. E.; Narayan, S.; Epstein, M.; Dailey, D.; Herman, B.; Tam, A.; Guetler, V.; Bintz, C.; Raz, D.; Jobe, P. O.; Putz, F.; Robick, B.; and Barghouti, B. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 772–781.
- Lam, M. S.; Gordon, M. L.; Metaxa, D.; Hancock, J. T.; Lantry, J. A.; and Bernstein, M. S. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–34.
- Lengyel, D. M.; Newman, J. S.; and Yurovich, D. 2023. Examining risk management failures: The case of the Boeing 737 Max program. *Enterprise Risk Management*, 8(1): 1–1.
- Lewis, A. D.; and Groth, K. M. 2022. Metrics for evaluating the performance of complex engineering system health monitoring models. *Reliability Engineering & System Safety*, 223: 108473.
- Lin, Y. 2024. More Than an Enforcement Problem: The General Data Protection Regulation, Legal Fragmentation, and Transnational Data Governance. *Columbia Journal of Transnational Law*, 62(1).
- Liu, X.; Glocker, B.; McCradden, M. M.; Ghassemi, M.; Denniston, A. K.; and Oakden-Rayner, L. 2022. The medical algorithmic audit. *The Lancet Digital Health*, 4(5): 384–397.
- Madni, A. M.; and Sievers, M. 2014a. System of systems integration: Key considerations and challenges. *Systems Engineering*, 17(3): 330–347.
- Madni, A. M.; and Sievers, M. 2014b. Systems integration: Key perspectives, experiences, and challenges. *Systems Engineering*, 17(1): 37–51.
- Mahajan, V.; Venugopal, V. K.; Murugavel, M.; and Mahajan, H. 2020. The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms—How We Do It. *Academic Radiology*, 27(1): 132–135.
- Malunga, M.; Ahmed, R.; Raynard, A.; Stoyanova, M.; and Chisala, K. 2022. Key lessons from the Boeing 737 MAX 8 accidents. *Loss Prevention Bulletin*, 287(25): 24–26.
- Meadows, D. 2008. *Thinking in systems: International best-seller*. Chelsea Green Publishing.
- Moir, I.; and Seabridge, A. 2011. *Aircraft Systems: Mechanical, electrical, and avionics subsystems integration*. John Wiley & Sons.
- Mökander, J.; Curl, J.; and Kshirsagar, M. 2025. A Blueprint for Auditing Generative AI. In *Research Handbook on the Law of Artificial Intelligence*, 307–327. Edward Elgar Publishing.
- Mökander, J.; and Floridi, L. 2023. Operationalising AI governance through ethics-based auditing: An industry case study. *AI and Ethics*, 3(2): 451–468.
- Moreau, T.; Sinatra, R.; and Sekara, V. 2024. Failing Our Youngest: On the Biases, Pitfalls, and Risks in a Decision Support Algorithm Used for Child Protection. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 290–300.
- Naja, I.; Markovic, M.; Edwards, P.; Pang, W.; Cottrill, C.; and Williams, R. 2022. Using Knowledge Graphs to Unlock Practical Collection, Integration, and Audit of AI Accountability Information. *IEEE Access*, 10: 74383–74411.
- Nartey, J. 2025. The Boeing 737 MAX Crisis: An Ethical Analysis of Corporate Governance Failures and the Conflict

- Between Profit and Safety. *SSRN Electronic Journal*. Available at SSRN: 5316121.
- Nicas, J.; Kitroeff, N.; Gelles, D.; and Glanz, J. 2019. Boeing Built Deadly Assumptions into 737 Max, Blind to a Late Design Change. *The New York Times*.
- Oakden-Rayner, L.; Gale, W.; Bonham, T. A.; Lungren, M. P.; Carneiro, G.; Bradley, A. P.; and Palmer, L. J. 2022. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: A diagnostic accuracy study. *The Lancet Digital Health*, 4(5): 351–358.
- Oala, L.; Fehr, J.; Gilli, L.; Balachandran, P.; Leite, A. W.; Calderon-Ramirez, S.; Li, D. X.; Nobis, G.; Alvarado, E. A.; Jaramillo-Gutierrez, G.; Matek, C.; Shroff, A.; Kherif, F.; Sanguinetti, B.; and Wiegand, T. 2020. ML4H Auditing. In *From Paper to Practice. Proceedings of the Machine Learning for Health NeurIPS Workshop*, 280–317.
- Polzer, A.; Fleiß, J.; Ebner, T.; Kainz, P.; Koeth, C.; and Thalmann, S. 2022. Validation of AI-based Information Systems for Sensitive Use Cases: Using an XAI Approach in Pharmaceutical Engineering. In *Hawaii International Conference on System Sciences*.
- Pop, G. I.; Titu, A. M.; and Pop, A. B. 2023. Enhancing aerospace industry efficiency and sustainability: Process integration and quality management in the context of Industry 4.0. *Sustainability*, 15(23): 16206.
- Prokop, L. E. 2024. Software error incident categorizations in aerospace. *Journal of Aerospace Information Systems*, 21(10): 775–789.
- Radiya-Dixit, E.; and Neff, G. 2023. A Sociotechnical Audit: Assessing Police Use of Facial Recognition. In *2023 ACM Conference on Fairness Accountability and Transparency*, 1334–1346.
- Rajabalinejad, M. 2019. Safe Integration for System of Systems: The Safety Cube Theory. In *2019 14th Annual Conference System of Systems Engineering (SoSE)*, 323–327.
- Rico, C. V. 2025. General-Purpose AI Models (Chapter V). In *The EU regulation on Artificial Intelligence: A commentary*, 373–402. Wolters Kluwers Italia.
- Rismani, S.; Shelby, R.; Davis, L.; Rostamzadeh, N.; and Moon, A. 2025. Measuring What Matters: Connecting AI Ethics Evaluations to System Attributes, Hazards, and Harms. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2199–2213.
- SAE. 2026. Society of Automotive Engineers. *Aerospace Recommended Practice for Development of Civil Aircraft and Systems*, (ARP4754B: 2023).
- Sghaier, M. O.; Hadzagic, M.; and Shahbazian, E. 2024. Verifiable Human Autonomy Teaming for NORAD C2 Operations. In *2024 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, 137–144.
- Sisodia, S. 2026. Anthropic’s Claude Shut Down Firm Without Explanation, Claims CTO And Issues Warning. *NDTV*.
- Sloane, M.; Moss, E.; and Chowdhury, R. 2022. A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, 3(2).
- Smogavc Cestar, A.; and Harer, J. 2023. Mastering Quality System Audits and Inspections for Medical Devices. In *Medical Devices and In Vitro Diagnostics: Requirements in Europe*, 1–42. Cham: Springer International Publishing.
- Summers, B. L. 2013. *Effective methods for software and systems integration*. CRC Press.
- Tricco, A. C.; Lillie, E.; Zarin, W.; O’Brien, K. K.; Colquhoun, H.; Levac, D.; and Straus, S. E. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of internal medicine*, 169(7): 467–473.
- Varoquaux, G.; Luccioni, S.; and Whittaker, M. 2025. Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 61–75.
- Williams, S.; Schuett, J.; and Anderljung, M. 2025. On regulating downstream AI developers. *European Journal of Risk Regulation*, 1–29.
- Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; and Polli, F. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677.
- Winter, J. 2020. System Integration Testing for Unintended Behaviors in Flight-Critical Aerospace Applications. In *2020 IEEE International Systems Conference (SysCon)*, 1–5. IEEE.
- Xie, L.; Zheng, X.; Sun, Y.; Huang, T.; and Bruton, T. 2022. Massively digitized power grid: Opportunities and challenges of use-inspired AI. *Proceedings of the IEEE*, 111(7): 762–787.
- Yang, Y.-T.; Zhang, T.; and Zhu, Q. 2023. A Game-Theoretic Analysis of Auditing Differentially Private Algorithms with Epistemically Disparate Herd. In Fu, J.; Kroupa, T.; and Hayel, Y., eds., *Decision and Game Theory for Security*, 349–368. Springer Nature Switzerland.
- Zieke, R. P. 1963. The progress function in the aerospace industry—A historical analysis. *IEEE Transactions on Engineering Management*, 2: 74–86.