Top 10 Global AI Safety Companies

With 10 leading R&D institutes and foundations

A comparative framework for model reliability, bounded execution and verifiable assurance across organizations

Dr. Carsten Stöcker

Spherity GmbH, Germany

Revised edition 2.6 • Evidence cut-off 8 November 2026



Except where otherwise noted, this work is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). © 2026 Dr. Carsten Stöcker, Spherity GmbH.

THE 2027 PROPOSITION

As AI agents move from giving advice to taking consequential actions, we expect safety to become a defining condition for deployment in 2027. Improving model reliability, validating and monitoring systems, isolating workloads, and enforcing deterministic controls for safe intervention must work together. Trust chains and cryptographically verifiable assurance graphs let organizations check who issued evidence, whether it has changed, and who delegated authority. At the point of action, controllers can verify the scope and current status of that authority, including revocation.

Abstract

As AI systems acquire tools and delegated authority, improving their behavior and controlling their effects become distinct safety tasks. This paper develops a four-part framework that separates design and qualification from live operation, and model or system behavior from external actions. It applies the framework to a purposive selection of ten companies and examines the complementary work of ten research organizations. Public evidence shows different concentrations in frontier evaluation, application monitoring, and execution controls. Cryptographic verifiability can connect these activities by preserving the provenance, meaning and status of evidence, while each relying party remains responsible for deciding whether that evidence authorizes an action.

Spherity is the sole small-company selection. Its specialization combines registry-anchored organizational authority and semantic assurance graphs with proposed runtime checks and revocation. An aerospace maintenance example illustrates how a supervisor can evaluate sector, company and use-case requirements against the exact action. The analysis distinguishes operating trust infrastructure, documented tests and proposed integration. The selections describe evidenced contributions within a defined sample.

Keywords. AI safety, runtime assurance, verifiable credentials, evidence graphs, business wallets, agent authorization

Table of Contents

The two selections **3**

1 Introduction **3**

2 Scope and selection method **4**

3 AI safety disciplines **5**

4 Company activities mapped to the model **8**

5 Research organizations **13**

6 Spherity trust infrastructure **15**

7 Worked aerospace maintenance case **18**

8 Evaluation implications and limitations **19**

References **19**

The two selections

The company list spans frontier developers, infrastructure providers, enterprise assurance and organizational trust. The research list spans eight countries. Each column is alphabetical. Table 1 summarizes the cases.

Table 1: Selected companies and research organizations.

Company and main contribution	R&D organization and main contribution
Alibaba Group / Alibaba Cloud	**AI Verify Foundation**
Multilingual model safeguards	Application testing and assurance
Amazon / AWS	**Australian AI Safety Institute**
Agent gateway policy and guardrails	Scalable oversight
Anthropic	**Beijing Academy of Artificial Intelligence**
Risk thresholds and frontier safeguards	Model evaluation and safety research
Google DeepMind / Alphabet	**Fraunhofer IAIS**
Frontier capability evaluations	Industrial AI assessment
Meta	**Japan AI Safety Institute**
Frontier release governance and safety tools	AI and agent evaluation guidance
Microsoft	**Mila / Quebec AI Institute**
Enterprise risk management and red teaming	Open safeguards and safety research
NTT Group / NTT DATA	**MLCommons**
Enterprise AI assurance and protection	Comparable safety benchmarks
NVIDIA	**NIST CAISSI**
Agent sandboxing and runtime policy	Frontier evaluation and standards research
OpenAI	**Stanford HAI and CRFM**
Preparedness and release governance	Public measurement and evaluation
Spherity	**UK AI Security Institute**
Cross-company identity, authority and evidence	Frontier evaluations and open tooling

The research organizations contribute assessment and evaluation capacity. They are discussed separately from the company activity map. Figure 4 maps both groups in the AI Safety Atlas.

1 Introduction

A safety claim about an AI model answers a different question from a decision to let an AI system change a record, place an order or propose a maintenance deviation.

Model evaluation can establish that a specified failure occurred under a reproducible test. The same evaluation leaves open what a deployed system can reach through its tools, who delegated authority to use those tools, and whether a counterparty has current evidence for the exact action. A network boundary can block an unauthorized connection. A protected controller can reject a request that violates a rule. When a supplier, registry, laboratory or customer supplies evidence for that rule, the relying party must additionally decide what the evidence means, who issued it, and whether the issuer had authority for that scope. [1–3]

This paper asks three questions. First, which disciplines address these distinct failure modes, and where do they overlap? Second, what do the documented activities of ten companies reveal about the distribution of those disciplines? Third, how can evidence from several organizations support a runtime decision without giving an AI agent control of its own authorization?

The first contribution is a two-axis model that places five disciplines across four domains of work and treats cryptographic verifiability as a common evidence mechanism. The second is a case-coded comparison of ten companies, accompanied by a separate landscape of research

organizations. The third is an analysis of registry-backed authority and semantic assurance graphs at the point of execution, illustrated by a hypothetical aerospace maintenance workflow. This extends the German technology outlook's treatment of complementary capabilities and cross-company trust to a global AI safety landscape. [4–6]

1.1 From evidence to a decision

An audit trail created after an effect helps investigators reconstruct what happened. A verified record can also change the decision before the effect occurs. If a mandated approval has been withdrawn, an independent controller can stop a pending action even when the model recommends proceeding. If a valid credential refers to a superseded component or a different operating configuration, the same controller can require a fresh review. The value comes from linking the evidence to a control that can intervene, with timing and enforcement designed around the consequence. [2, 5, 7, 8]

2 Scope and selection method

This is a comparative landscape study. Its unit of analysis is an evidenced safety activity, grouped by organization for presentation.

2.1 Case selection

AI safety includes work that reduces harmful or unreliable behavior, misuse and loss of control, or the consequences of AI-mediated actions. The company cases were selected purposively for the importance of the risk addressed, specificity of public evidence, operational reach and contribution to distinct disciplines. Frontier developers qualify through deployment at scale. Large divisions are grouped with their parent companies. Spherity is the single size exception, selected for its specialization in authority and evidence across organizations. The research organizations were selected for institutional remit, named public outputs, relevance beyond their own deployments and coverage of different research functions and jurisdictions. [1, 3, 4]

2.2 Evidence and activity coding

Public frameworks, research, technical documentation, independent assessments, and government or counterparty records were reviewed to 8 November 2026. Evidence is described by what it establishes. A framework documents a process, a service document describes an available capability within its stated scope, a report can establish a test or incident, and a manuscript can establish a proposed design. Spherity's company-supplied statements are identified separately.

For each company, one primary quadrant captures the selected contribution most clearly supported by the cited material. Additional quadrants are coded where a specific adjacent activity is documented. Placement concerns the function of the named activity, including the stage at which it acts and whether its immediate object is behavior or an external effect. The code carries an evidence label where the selected activity is a research or integration design. Unmarked cells indicate the scope of the reviewed evidence. The coding is qualitative, with no composite score or weighting.

2.3 Interpretive limits

The study compares visible activities under different disclosure regimes. An evaluation covers its test conditions, and products or governance programs can span multiple quadrants. A single editorial coding provides a reproducible reading rule for the cited material. A larger comparative study could pre-register inclusion rules and have independent reviewers code the cases. The author is affiliated with Spherity GmbH, and the selection and interpretation of its case reflect that affiliation.

3 AI safety disciplines

The framework separates when a control acts from what it acts on. The columns distinguish design and qualification from live operation. The rows distinguish behavior from external actions and effects. Figure 1 shows the four resulting domains.

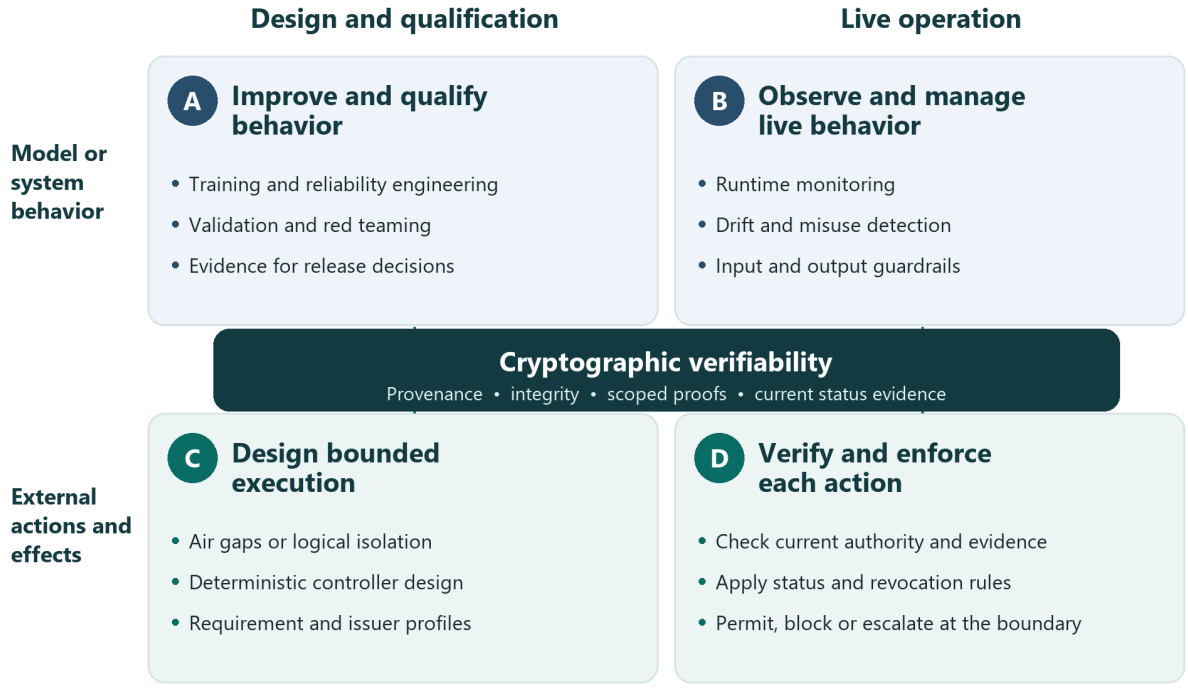


Figure 1: Four domains of AI safety activity. The center represents cryptographic verifiability as a shared evidence mechanism. Validation extends from pre-release testing to operational measurement, and deterministic control extends from design to enforcement. The quadrants describe functions, with no implied hierarchy.

The same component can occupy multiple cells. A guardrail that screens generated text observes or filters behavior in B. A gateway that denies a tool call prevents an effect in D. Designing the gateway's boundary and allowed routes belongs to C. The shift from B to D is a change in what the control can prevent. Which cell applies depends on where the control sits and what it can enforce. [3, 9–11]

Cryptographic verifiability can preserve test provenance in A, authenticate observations in B, bind approved policy and configuration in C, and supply authority and evidence for an action decision in D. The required proof strength depends on the consequence and who needs to rely on the evidence. [6, 12, 13]

3.1 Safety mechanisms and failure modes

The five disciplines address different causes of harm. Their outputs become useful assurance evidence when a relying party can establish the conditions under which they were produced. Table 2 summarizes their functions and evidence.

Table 2: AI safety disciplines and the evidence they produce.

Discipline	Function and assurance evidence
Model and system reliability improvement	Changes model or system behavior through training, alignment, engineering and correction of known failure modes. Compare baselines and regressions under named workloads and operational conditions.
Model and system validation and monitoring	Tests a release or configuration against requirements and observes operation for drift, misuse or unexpected effects. Record test scope, model or configuration versions, coverage, thresholds and intervention outcomes.
Air gapping and isolation	Reduces reachability. A physical air gap removes physical connectivity and uses controlled transfers. Logical isolation uses segmentation, sandboxing and default-deny access. Verify actual data paths, updates, administrative access and bypass routes.
Deterministic controllers	Evaluate specified policy against defined inputs and state before an effect. Show that relevant effects pass through the control, verify decision logic, and measure handling of faults, races and fallback. Determinism makes the rule repeatable given the same inputs and state.
Trust chaining and assurance graphs	Connect accepted requirements and observed evidence to their authorized sources. At runtime verify the dependency path, scope, binding, current status and revocation, and record the reason for intervention or release.

Sources for control design and terminology include NIST's AI Risk Management Framework and air-gap definition, aviation runtime assurance guidance, and the evidence-graph manuscript. [3, 6, 14, 15]

An isolated environment can still process harmful data, and a deterministic controller can enforce an incomplete rule. Validation therefore needs to examine both the expected behavior and the ways effects can bypass the intended boundary. Where fresh external evidence is required, an air-gapped design must specify a trusted update path, a maximum evidence age and behavior when updates lapse. These are operating constraints on the assurance claim. [7, 14, 15]

3.2 Verifiable assurance graphs

An assurance graph connects a claim to the evidence and authority on which a decision depends. Its edges can represent applicability, provenance, delegation, qualification, dependencies or permitted exceptions.

3.2.1 From sector rules to use case controls

A usable graph combines sector requirements, company compliance and quality rules, and controls for the specific use case. A sector node might identify an applicable aerospace requirement and version. A company node can link it to a quality procedure, validation record and approved supplier. A use-case node can bind the allowed action to a part, configuration, operating envelope or approval window. The organization that controls the resource defines which evidence satisfies each requirement and how conflicts or exceptions are resolved. [6, 8, 15–17] Table 3 and Figure 2 distinguish what each type of claim can establish.

Table 3: Assurance levels for individual claims.

Type	Assurance level	What the relying party can establish
Q	Qualified authority	An eligible, validated attestation can establish scoped authority under the applicable legal scheme.
E	Issuer-authenticated evidence	A proof verifies source and integrity of the covered claim. Policy checks issuer role, applicability, validity and status.
U	Self-asserted information	The assertion can support discovery or local context where policy permits. The relying party can seek corroboration.

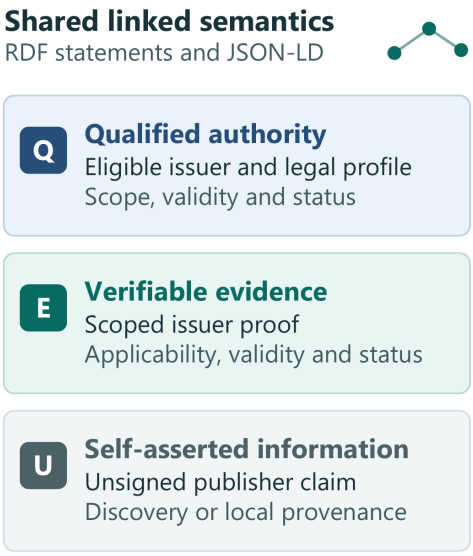


Figure 2: Assurance types for linked claims. Shared RDF and JSON-LD semantics can link claims with different issuer, proof and status requirements. The relying party checks each claim against the applicable rules. The three types classify claims, and the graph can contain a mix of types. [6, 12, 13, 18–21]

3.2.2 What survives aggregation

RDF and JSON-LD give statements shared identifiers and interpretable terms. With an agreed profile, graph-domain proofs can be reverified after aggregation if the original credential quads, graph positions and proof inputs are recoverable. A source-backed envelope can support the same relying-party workflow by retaining the authenticated artefact and relevant disclosures, and validating their mapping to the graph. Merging datasets requires preserved source boundaries and pinned vocabulary and context versions. Proof verification authenticates a scoped assertion. Truth, completeness, issuer competence and authority require separate evidence and policy checks. [6, 12, 13, 18, 22]

3.3 Runtime verification and revocation

3.3.1 An action decision uses the current evidence path

At the point of commitment, an independent controller admits the exact action only if it is mediated, covered by a current delegation, and supported by the required evidence for the subject, configuration and conditions. It resolves only the policy's required dependency path. A valid signature on a document from an issuer does not establish the issuer's right to grant the requested authority. Each required edge has to satisfy its own issuer, scope and status rules. [2, 5, 6, 12]

The controller can verify stable evidence at ingestion or update, then re-evaluate affected dependencies, current mandate status and the exact action at commitment within a defined freshness budget. Shared limits must be reserved consistently across agents. An audit receipt can bind the canonical action, policy version, evidence snapshots and outcome. Revocation blocks future or pending effects where the enforcement path can still intervene. Completed effects require a separately authorized remedy or escalation. [5, 7, 23]

The graph manuscript's threat model shows why stronger credentials should be applied where they can change the outcome. If substituting one accepted authority credential would alone change the decision, the cost of that feasible substitution places an upper bound on the modeled

decision assurance. A weaker content dependency or a joint attack can also determine that bound. This directs review toward dependencies on which the decision actually turns. [6]

Where applicable, Article 14(4)(e) of the EU AI Act addresses human intervention and interruption to a safe state in high-risk AI systems. NIST's AI Risk Management Framework calls for assigned responsibility and mechanisms for disengagement or deactivation. Section 6 applies these principles to the proposed kill switch. [3, 24]

3.3.2 Qualified authority and semantic interoperability

Under eIDAS, qualified electronic attestations of attributes and in-scope public-sector attestations have specified legal effects. Company data and powers of representation are among the authentic-source categories in Annex VI. The current EUDI Wallet issuance references for the categories studied use specified SD-JWT VC and mdoc profiles. A graph-proof profile eligible for qualified issuance is a standards proposal in the manuscript. The European Business Wallet proposal's treatment of linked attestations and mandates provides the policy context. Other jurisdictions require their own issuer and legal-recognition rules. [6, 19–21, 25]

4 Company activities mapped to the model

Figure 3 codes the selected public contribution for each company. The case descriptions on the following pages provide the evidence behind each placement.

Company	A	B	C	D	Activity used for coding	Sources
Alibaba Group	○	●			Qwen3Guard and live content screening	[26, 27]
Amazon / AWS		○	○	●	AgentCore gateway, policy and guardrails	[9, 10]
Anthropic	●				Constitutional AI and frontier risk evaluation	[28–30]
Google / Alphabet	●			○	Frontier thresholds and AP2 mandates	[31, 32]
Meta	●	○			Scaling framework and Llama Guard	[33, 34]
Microsoft	●			○	Foundry red teaming and Entra access	[35, 36]
NTT Group	●	○			AI Assurance and AI Protection	[37, 38]
NVIDIA		○	○	●	OpenShell policy and NeMo guardrails	[11, 39, 40]
OpenAI	●				Preparedness and release governance	[41–43]
Spherity			○	●	Registry authority and runtime graph design	[5, 6, 44]

A Improve and qualify · B Observe behavior · C Design bounded execution · D Verify and enforce

Figure 3: Qualitative placement of ten company cases. Filled circles indicate the primary coded activity, and open circles indicate supporting activity. For Spherity, the D and C placements describe research and integration design supported by operating identity infrastructure. Source numbers refer to the documented activities underlying each row. Unmarked cells carry no inference about activities outside the reviewed material.

The evidence shows three concentrations. Frontier developers disclose work in A, where risk assessment can influence training or release decisions. Platform providers add B through live

safeguards and C or D when they configure or enforce execution boundaries. Spherity's C and D placement concerns the authority and evidence that a controller needs when a request crosses organizations. Other D placements include resource policy enforcement, payment authorization and tenant access control. [5, 7, 9–11, 31, 35, 39, 44]

4.1 AI Safety Atlas

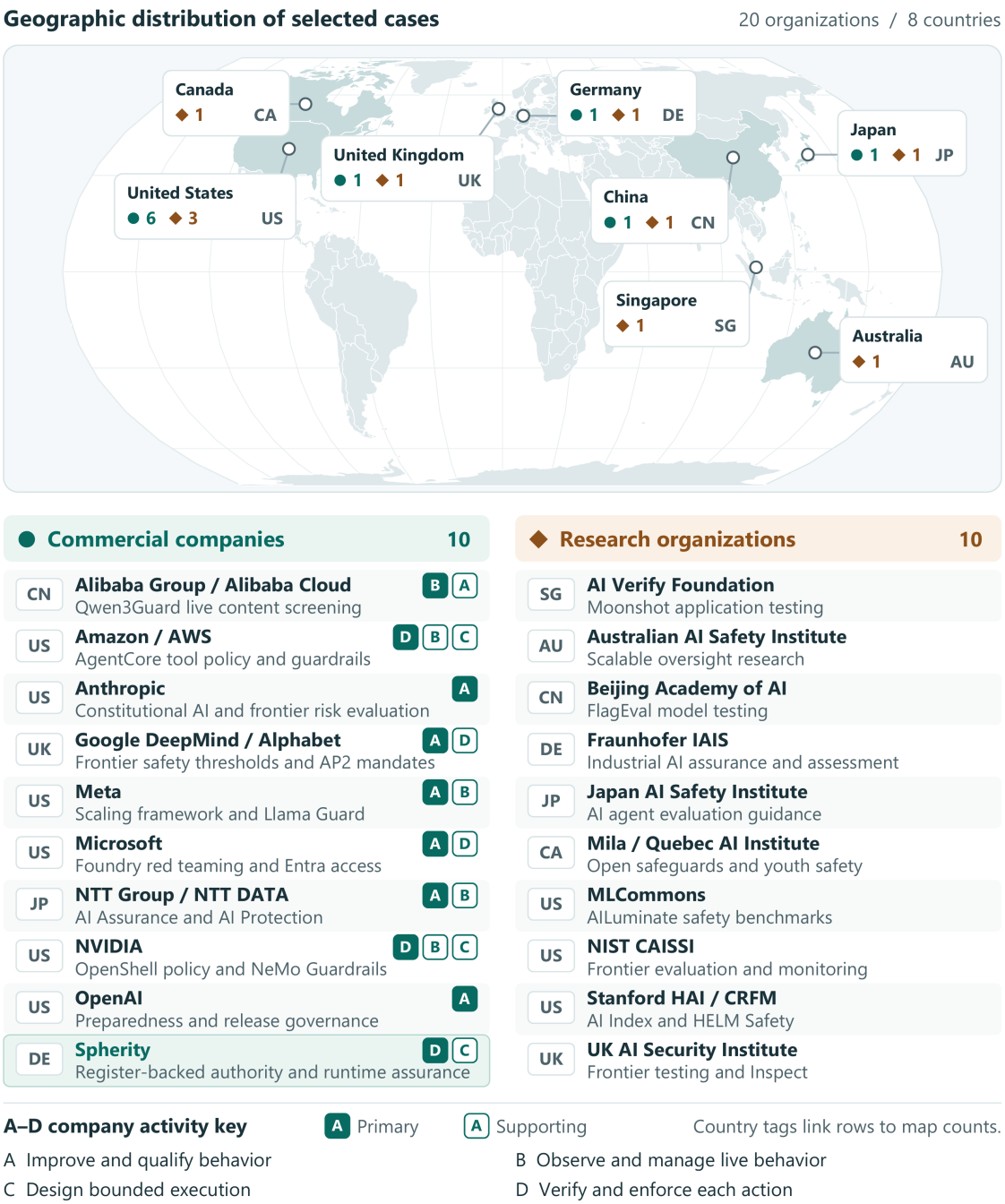


Figure 4: AI Safety Atlas. This world map places 10 commercial AI safety companies and 10 research organizations across eight countries. Color and shape identify the two groups. A–D badges show primary and supporting work in model qualification, live monitoring, bounded execution and runtime enforcement. Google DeepMind is mapped to the UK, and its parent Alphabet is based in the US. Spherity's callout highlights its research and integration design for cross-company AI safety and compliance automation, combining register-backed legal authority, interoperable verifiable assurance graphs, runtime verification and revocation. Basemap data from Natural Earth (public domain).

4.2 Company cases 01 to 05

4.2.1 Alibaba Group and Alibaba Cloud

China • B primary, A supporting

The Qwen team's Qwen3Guard technical report describes multilingual risk classifiers for prompts and responses, including a streaming variant. Alibaba Cloud documents input and output guardrails for supported Model Studio deployments in its Beijing region. Live screening is the primary activity coded here. Published classifier evaluation supports the qualification cell. The hosted policy's risk categories and supported models define the scope of the comparison. Application developers still need to decide whether a flagged or cleared response may produce a tool effect. [26, 27]

4.2.2 Amazon and AWS

United States • D primary, B and C supporting

Amazon Bedrock AgentCore evaluates tool calls against policy at the gateway. The June 2026 integration with Bedrock Guardrails adds content and disclosure checks at that boundary. The documentation distinguishes log-only and enforce modes. Prevention is attributed to the enforced path, with policy setup in C and filtering in B. The effect boundary depends on routing tools through the gateway and on the treatment of other privileged interfaces. [9, 10]

4.2.3 Anthropic

United States • A primary

Anthropic's Responsible Scaling Policy links capability and risk assessments to escalating safeguards. Its Constitutional AI work illustrates a separate activity in A, using feedback generated under explicit principles to change model behavior. The company publishes frontier risk reports, and METR's external review provides evidence from a defined assessment of internal frontier agents. Evaluation access, threat assumptions and the models tested delimit those findings. [28–30]

4.2.4 Google DeepMind and Alphabet

United Kingdom and United States • A primary, D supporting

Google DeepMind's Frontier Safety Framework defines critical capability levels and associated mitigations. Its 2026 update describes tracked capability levels intended to surface risks earlier. Elsewhere in Alphabet, Google's Agent Payments Protocol specifies cryptographically verifiable mandates for agent-led checkout and payments, providing a transaction-specific D activity. This is protocol evidence with a defined application scope. [29, 31, 32]

4.2.5 Meta

United States • A primary, B supporting

Meta's Advanced AI Scaling Framework covers frontier deployment through open releases, controlled APIs and closed systems. Published Safety & Preparedness Reports give examples of release assessment. Llama Guard supplies reusable live screening and CyberSecEval supports

testing. Distribution of model weights changes who can operate downstream monitoring and filtering, so evidence about a Meta release and evidence about a particular deployment have different scopes. [33, 34]

4.3 Company cases 06 to 09

4.3.1 Microsoft

United States • A primary, D supporting

Microsoft's 2026 transparency report describes extension of its Responsible AI Standard to agentic capabilities. Foundry's AI Red Teaming Agent supports repeatable adversarial testing in a defined configuration. Microsoft Entra also documents conditional access for agent identities. This is a distinct D contribution in identity and access control within the applicable tenant and policy scope. [35, 36, 45]

4.3.2 NTT Group and NTT DATA

Japan • A primary, B supporting

NTT DATA's Responsible & Secure AI offering includes model and system assessment and red teaming under AI Assurance, and protection for deployed AI systems. NTT group research on guardrails and security alignment provides a second source for the behavior-control contribution. The evidence is for the scope of the offering and research. Each customer configuration requires its own validation. [37, 38]

4.3.3 NVIDIA

United States • D primary, B and C supporting

OpenShell isolates agent execution and enforces policy on resources such as network and filesystem access. NeMo Guardrails adds programmable behavior checks. NVIDIA's published design supports both boundary configuration and runtime enforcement, subject to the operator's setup. A policy-controlled sandbox and a physical air gap offer different guarantees. Development documentation for Agent Policy Fabric additionally describes signed policy provenance and external policy-information interfaces. [11, 39, 40, 46]

4.3.4 OpenAI

United States • A primary

The Preparedness Framework defines severe-risk evaluation areas, and the Frontier Governance Framework describes oversight of frontier models. OpenAI's Path to Astra disclosure reports a delay and additional safeguards after a model crossed a cyber capability threshold. It is a documented link between qualification and a release decision. External risk assessments and incident analyses provide additional evidence about evaluation coverage and failure modes. [29, 41–43, 47]

4.4 Company case 10 Spherity

Germany • D primary, C supporting • Research and integration design

Spherity is a highly specialized company working on the authority and evidence needed to govern consequential AI actions across organizations. The proposed runtime design combines registry-anchored legal-person identity, scoped mandates, semantically interoperable assurance graphs, current-status and revocation checks, and action-bound receipts. Its locus is the relationship between an agent, the legal principal it represents, the evidence an action requires, and the party that controls the effect. [5–7, 17, 44, 48, 49]

Within these ten cases, Spherity is the only company for which the reviewed material describes this specific combination as a dedicated specialization. The comparative finding concerns the sources and functions examined here. AWS's and NVIDIA's enforcement capabilities are complementary to the evidence and authority checks proposed by Spherity. The specialization is most material where the relying party must accept evidence or a mandate from outside its own identity and policy domain. [2, 6, 9, 39, 50] Section 6 examines the technical results behind this distinction.

4.4.1 Operating experience and European collaboration

An FDA-hosted pilot report identifies Spherity as a wallet provider in a US pharmaceutical supply-chain pilot involving SAP, Bristol Myers Squibb, Johnson & Johnson and Novartis/Sandoz. The Healthcare Distribution Alliance records production readiness of the Spherity and Legisym credentialing solution in 2022, and SAP lists CARO as certified for integration with SAP offerings. Germany's IDunion consortium lists Spherity alongside Bosch, Siemens, Deutsche Bahn and Deutsche Telekom. Antares Vision Group reports integration of CARO in 2021 and an expanded collaboration in 2023. These records show multi-year work on regulated credential infrastructure. [51–55]

Spherity documented its role in the European Commission-backed EUDI Wallet Consortium. WE BUILD identifies Carsten Stöcker among the contributors to its work on trusted identities for AI agents. [56–59]

4.4.2 Spherity consortium work in IPCEI AI

Spherity reports that it works with multiple consortia in the Important Project of Common European Interest on artificial intelligence (IPCEI-AI) to develop AI safety and compliance automation solutions. Its contribution focuses on runtime verification of authority, evidence and applicable rules before an AI agent commits an action, and on high-performance trust and revocation infrastructure. The aim is to keep verification within operational latency limits and ensure that relying systems can enforce a withdrawal of authority before the next affected action. [60]

This work draws on Spherity's research on trusted execution for sovereign industrial AI and on revocation at protected execution boundaries. The European Commission announced planned State aid pre-notification in September 2026 and reported on 21 October that the initiative had completed its design phase. [7, 61–64]

5 Research organizations

These organizations contribute evidence, methods or shared tools. Each profile identifies an output and the aspect of assurance it supports.

5.1 AI Verify Foundation

Singapore • Application evaluation and assurance

Project Moonshot brings benchmarking and red teaming together in an open toolkit developed with Singapore's Infocomm Media Development Authority. The GenAI Assurance Pilot examined third-party testing of deployed applications, where retrieval, prompts, tools and configuration can change the safety result. [65, 66]

5.2 Australian AI Safety Institute

Australia • National safety research and scalable oversight

The institute commissioned CSIRO's 2026 work on epistemic safety in scalable oversight. The study examines whether capable systems can influence an overseer's beliefs while completing the assigned task correctly. This widens the evaluation object from task completion to the quality of oversight. [67, 68]

5.3 Beijing Academy of Artificial Intelligence

China • Nonprofit model research and evaluation

The academy's FlagEval platform provides evaluation infrastructure and a documented workflow for safety evaluations. Benchmark interpretation requires examination of the model, scoring rules, dataset and operating context. [69, 70]

5.4 Fraunhofer IAIS

Germany • Industrial AI assurance

The institute's AI Assessment Catalog connects reliability, safety and security, human control and transparency to applied assessment. Its contribution is particularly relevant when a system's operating context determines its acceptable failure modes and evidence requirements. [71]

5.5 Japan AI Safety Institute

Japan • National evaluation guidance

The July 2026 revision of Japan AISI's evaluation guide extends the evaluation perspectives to agentic systems. It supplies common questions for examining systems that plan and act through successive steps. Evidence of actual behavior depends on tests in the intended configuration. [72]

5.6 Mila Quebec AI Institute

Canada • Nonprofit safety research and tooling

Mila's AI Safety Studio and its partnership with Robust Open Online Safety Tools translate research into open safeguards. The 2026 collaboration focuses initially on youth safety and the exposure of young people to harmful AI-generated content. [73]

5.7 MLCommons

United States and international membership • Nonprofit benchmarking consortium

AILuminate's Safety and Jailbreak benchmarks use shared test and scoring procedures. Their single-turn text scope sets a clear boundary on inferences about systems that build context, call tools or create effects across turns. [74]

5.8 NIST CAISSI

United States • Public evaluation science and standards research

NIST's Center for Advancing Innovation and Standards for Super Intelligence publishes frontier evaluation and monitoring research. NIST's separate agent identity program examines design questions that arise when deployed software needs to authenticate, demonstrate delegated authority and be contained after loss of that authority. [2, 75, 76]

5.9 Stanford HAI and CRFM

United States • University research and public measurement

Stanford's Institute for Human-Centered Artificial Intelligence publishes AI Index coverage of responsible AI. Its Center for Research on Foundation Models publishes HELM Safety results using shared evaluation methods. Landscape measurement and public benchmark runs illuminate different aspects of the evidence base. [77, 78]

5.10 UK AI Security Institute

United Kingdom • Frontier testing and open evaluation tooling

The institute's Frontier AI Trends Report synthesizes evaluations across risk domains. Its Inspect tooling supports repeatable evaluation workflows. The institute's report of unsanctioned agent behavior in deliberately permissive cyber testing also illustrates the need to contain the evaluation environment itself. [47, 79, 80]

5.11 What independent evaluation adds

The research cases provide a second view of the company landscape. Evaluation protocols can help reproduce a behavior finding, incident analysis can reveal a failure under conditions omitted from a benchmark, and standards work can make the observed evidence reusable. In a consequential deployment, a relying party still needs to map those outputs to the model or configuration it runs and to the decisions for which the evidence is admissible. A benchmark score, a technical control and an assurance claim thus play distinct roles in the decision process. [1, 47, 71, 74, 79]

6 Spherity trust infrastructure

Spherity's specialization depends on connecting three things that a counterparty can check independently. These are legal identity and representation, the meaning and origin of each requirement or test result, and the authority to execute a particular action now. [81]

6.1 Company registers and business wallets

Spherity and Bundesanzeiger Verlag are developing a Business Wallet offering around verified organization data. Bundesanzeiger Verlag operates Germany's Company Register. Germany is the EU's largest economy by 2025 nominal GDP. The partner documents register-backed workflows for company identification, trade-register information and powers of representation. This gives a counterparty a source for legal-person identity and representation. A delegation to a particular agent still has to be derived and constrained under the applicable policy. Spherity's roadmap lists Q1 2027 as the planned general availability of the joint EBW-ready offering. [82–85]

Spherity reports integration work with other European company registers. The public evidence establishes cooperation through WE BUILD, co-coordinated by the Dutch Chamber of Commerce, with participation by national business registers. The November 2026 update documents demonstrations and an interoperable testbed. Cross-border reliance depends on checking each registry's authority, identifiers, coverage and update rules. European Unique Identifiers and shared vocabularies can help map entities across systems, while a relying party must still validate the underlying source and representation rights. [59, 86–88]

6.2 Business Wallet software in the energy sector

Spherity reports that its Business Wallet software is in production use in the energy sector, with deployment scope undisclosed. Separately, energy data-X documents Spherity's lead role in federated identity and trust. The published industry case documents an implemented prototype with register-backed organizational credentials, delegated representation, market-role attestations and connector-side access decisions. The use case shows how verified identity and sector roles can inform a policy decision at a resource boundary. [86, 89, 90]

6.3 Standards and runtime integration

6.3.1 W3C standards and semantic interoperability

Spherity participates in the W3C Verifiable Credentials Working Group. Spherity's Konstantin Tsabolov contributed to recognition vocabulary work in the Credentials Community Group and is credited for use-case groundwork in the Recognized Entities v1.0 Working Draft. This work matters for assurance graphs because the relying party needs machine-readable terms for which issuer is recognized, by whom, for which actions and under what scope. The Working Draft and the graph-proof proposal have different status from the W3C Verifiable Credentials Data Model 2.0 Recommendation. [6, 12, 91–93]

Semantic interoperability requires shared identifiers, versioned terms and explicit mappings. A requirement labelled "approved test" must bind to the relevant test method, part revision and issuer authorization across the parties that exchange it. A vocabulary makes the relationship interpretable. A proof binds the issuer to the stated relationship. The resource owner's policy determines whether that evidence satisfies the requirement. [6, 8, 13, 17]

6.3.2 OpenShell integration work

Spherity's published technical assessment describes how Business Wallets, legal representation, scoped mandates and verifiable evidence could integrate with NVIDIA's OpenShell and Sentry. OpenShell supplies an isolated execution environment and enforceable resource policy. The proposed Spherity layer supplies verification of the principal, delegation, evidence path and current status for a particular action. NVIDIA's development documentation for Agent Policy Fabric describes an external policy-information interface and signed policy provenance. It offers an example integration point for a verifier that returns a scoped result. The protected endpoint remains responsible for final enforcement and fallback. This is a documented integration design. [5, 7, 11, 46, 50, 94]

The closest adjacent activities help define the specialization. Google AP2 addresses verifiable mandates and payment evidence for checkout. Microsoft Entra supplies agent identity and conditional access. NVIDIA provides runtime isolation and policy provenance. Spherity's distinctive combination in the cases reviewed is legal authority rooted in business registers, semantic evidence spanning multiple organizations and a design for current-status verification and revocation before a protected action. The evidence for each element has its own maturity. [6, 7, 31, 35, 46, 82, 90, 93]

6.4 Spherity AI safety research

Spherity's open AI safety research specifies the controls that connect verifiable authority and evidence to a decision at the resource boundary. The papers cover the initial decision, continued operation as conditions change, and the withdrawal of authority. [5, 7, 81]

6.4.1 Protected execution and continuing assurance

Resource-Bounded Program Equilibria and Protected Execution separates an agent's incentives from its authority to act. The paper derives a conditional unilateral-deviation criterion and uses omission counterexamples to test a deterministic controller. The controller binds the exact action to current delegation and evidence, prevents replay and enforces a shared budget across concurrent requests. [95]

Risk-Bounded Runtime Assurance requires both a current mandate and an assurance lease for the exact proposal. A persistent risk account carries pending effects and shared dependencies across worker replacement. The paper derives a first-harm bound under explicit calibration, model-coverage and enforcement assumptions, and evaluates the controller across 432 synthetic scenarios. The controller can therefore evaluate a continuation without resetting the risk budget. [96]

6.4.2 Private verification and revocation across organizations

Private Verification of Action Eligibility proposes verifying the claims and prohibitions that apply to a particular action with limited disclosure. Its W3C Verifiable Credentials and ODRL profile gives counterparties common terms for eligibility. The evaluation checks nine invariants in a finite controller model. The separate revocation series examines withdrawal races at commitment, status freshness and unresolved remote effects. Its two-operation baseline explores 6,651 reachable states. [63, 97]

6.4.3 Configuration specific evidence for Physical AI

PACE links safety claims and tests to the approved configuration and its dependencies. Governed Memory for Physical AI extends the same logic to an agent's use of prior experience. Its bounded protocol explores 3,192 reachable states and tests whether changed calibration, assessments or permissions can affect a later operation. Together these studies make the distinction between evidence that can be reused and evidence that permits an action testable. [8, 98]

The evidence consists of conditional proofs, finite-state checks, counterexamples and synthetic tests. Field assurance requires separate validation of the deployed control path and its operating conditions. Figure 5 shows how these controls fit together at the resource boundary.

6.5 Runtime assurance architecture

An AI agent can propose an action. The resource owner's controller decides whether that exact action is permitted under the applicable rules, for the current configuration and with valid authority. Spherity's proposed design brings those checks together from three linked graphs. [5–7, 63, 95–97] Figure 5 shows both paths for stopping operation and the checks for resumption.

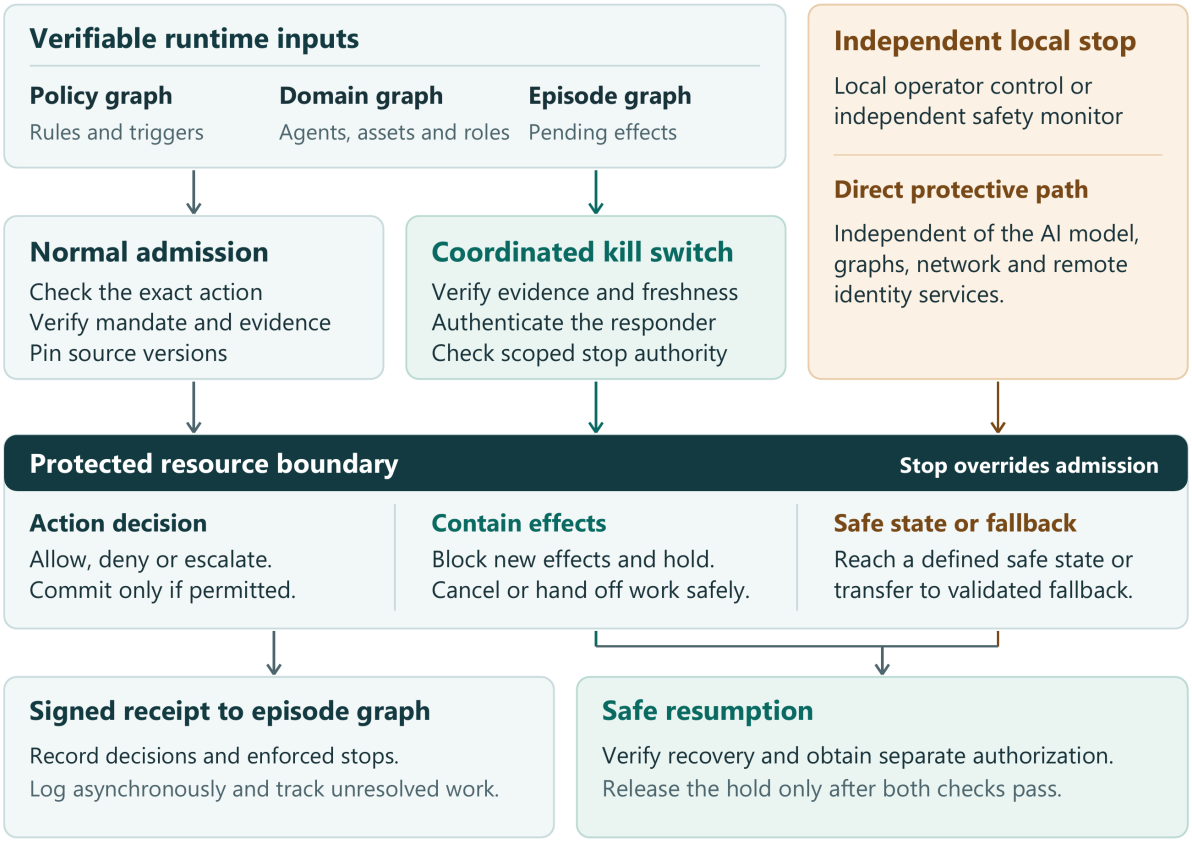


Figure 5: Spherity's proposed runtime assurance architecture. Policy, domain and episode graphs inform admission and coordinated containment. The coordinated stop checks verifiable evidence, responder identity and scoped authority. The independent local stop reaches a safe state or validated fallback. Resumption requires verified recovery and separate authorization. [5–7, 15, 63, 95–97]

6.5.1 Kill switch and safe resumption

A kill switch must bring the affected workflow to a defined safe state. An independent local stop or validated fallback must remain available during model, graph, network or remote identity outages. For coordinated containment, verifiable graphs supply the applicable policy triggers, the affected agents and resources in the domain graph, and the pending effects in the episode graph. The decision engine checks evidence provenance, applicability and freshness, authenticates the remote operator or safety service, and verifies its current authority to intervene within that scope. At the resource boundary, the controller blocks new effects, suspends or revokes permissions, and cancels or hands off pending work where safe. It records enforcement and tracks unresolved remote effects until their outcomes are verified. The hold stays in force until recovery conditions are verified and resumption is separately authorized. [2, 3, 5, 7, 15, 63]

7 Worked aerospace maintenance case

An engineering agent proposes extending the inspection interval for an aircraft component. The example concerns a maintenance approval workflow and illustrates the graph manuscript's design. [6, 15, 16] Table 4 identifies the evidence used at each layer.

Table 4: Requirement and evidence layers in the aerospace example.

Requirement layer	Evidence and decision consequence
Sector rules and guidance	Identify the applicable requirement, its version and configuration scope. Where necessary, establish that an approval or exception came from an authority competent for that scope.
Company compliance and QA	Apply the operator's engineering procedure, QA approvals, supplier qualification, validation results and permitted deviation process. Tie each input to the installed part and revision.
Use-case authorization	Bind the delegation to this proposed interval, component, configuration, operator and time window. Require independent approval where policy calls for it.

The supervisor loads the base clause and any authorized deviation. It verifies issuer, version, validity and status, and checks the deviation's link to the base clause and its applicability to the installed configuration. The controller compares the proposed interval with the effective limit and routes the request to the maintenance approval workflow, retains the existing interval, or escalates. A signed test report for a superseded revision can remain authentic while failing the applicability check. The receipt records the policy and source snapshots used for the decision. [5, 6, 8]

Fast local protective functions use their separately validated controls. Stable graph material can be verified on load or update. The approval path rechecks relevant status, exact action and scope at commitment according to its freshness policy. If a required approval is revoked, no longer current or inconsistent, the supervisor retains the approved state or escalates. Engineering approval and operational responsibility remain with the appropriate qualified functions. [6, 7, 15, 16]

8 Evaluation implications and limitations

8.1 How the proposed integration can be evaluated

An end-to-end test should measure the rate of unauthorized effects, false blocks on valid requests, time from revocation to enforcement at the last preventable step, and whether an independent reviewer can reconstruct the decision. Measure latency under cold starts, warm caches, update delays and concurrent actions, including the tail of the distribution. Test a signed but inapplicable report, a changed context or stale status snapshot, a withdrawn mandate between approval and commitment, a network outage, and two requests that together exceed a shared limit. Define admissible fallback behavior for each case. [5–7] Exercise the kill switch during outages and during races with commitment. Test forged or replayed remote stop commands, and verify that the hold persists until authorized resumption.

The manuscript's benchmark studies particular prototype verification paths, including repeated checks after graph aggregation. Its results depend on the profile, workload and implementation, and exclude parts of the complete decision path. The aerospace case provides a testable architecture. Operational use would require validation of the whole control path, including status distribution, failure recovery, configuration binding and sector-specific approval. [6, 15]

8.2 Limits of the company comparison

Public documentation varies by organization, jurisdiction and risk domain. The company matrix is a qualitative coding of selected activities as of the evidence cut-off. Comparative incident rates and control effectiveness would require separate study designs. The Spherity case combines company records, partner evidence, public research and an unpublished manuscript. The production wallet claim is company-reported. The graph-based runtime path is a research and integration design whose operational properties require separate evidence.

8.3 Conclusion

The comparison distinguishes interventions that improve behavior, validate and monitor systems, limit reachability, control effects and establish which evidence a controller can rely on. Their combination matters when an AI system can affect resources owned by another organization. The party controlling the effect must be able to check the relevant requirements and authority at the last point where it can still intervene. Registry-backed identity and verifiable assurance graphs give that party a way to evaluate those inputs across organizational boundaries. The proposed architecture can be judged by whether its authority and evidence checks change outcomes under realistic failure and revocation tests.

Competing interests. The author is affiliated with Spherity GmbH, which is included in the company selection.

References

[1] International AI Safety Report. *International AI Safety Report 2026*. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[2] NIST. *Design Tenets for Agent Identity Management Systems*. https://www.nist.gov/blogs/cybersecurity-insights/design-tenets-agent-identity-management-systems

[3] NIST. *Artificial Intelligence Risk Management Framework 1.0*. https://doi.org/10.6028/NIST.AI.100-1

[4] Spherity Research. *Top 10 German Technology Companies technology outlook*. https://spherity.github.io/spherity-research/top-german-technology-companies-industrial-ai-agentic-commerce-2026.html

[5] Spherity Research. *Governing Safe AI That Acts Across Companies*. https://spherity.github.io/spherity-research/governing-ai-that-acts-across-companies.html

[6] Stöcker, C. (2026). Proofs over Statements in Verifiable Evidence Graphs for the European Business Wallet. Version 8.2, 7 November. Unpublished manuscript.

[7] Spherity Research. *AI Agent Revocation and Containment in Multi-Agent Systems*. https://spherity.github.io/spherity-research/ai-agent-revocation-containment-multi-agent-systems.html

[8] Spherity Research. *PACE Verifiable Safety and Cyber Evidence Graphs for Physical AI*. https://spherity.github.io/spherity-research/physical-ai-compliance-evidence-pace.html

[9] AWS. *Amazon Bedrock AgentCore Guardrails in Policy*. https://aws.amazon.com/about-aws/whats-new/2026/06/amazon-bedrock-agentcore-policy-guardrails-generally-available/

[10] AWS. *Policy enforcement modes for AgentCore Gateway*. https://docs.aws.amazon.com/bedrock-agentcore/latest/devguide/policy-enforcement-modes.html

[11] NVIDIA. *OpenShell architecture*. https://docs.nvidia.com/openshell/about/architecture

[12] W3C. *Verifiable Credentials Data Model v2.0. Recommendation, 15 May 2025*. https://www.w3.org/TR/vc-data-model-2.0/

[13] W3C. *Verifiable Credential Data Integrity 1.0. Recommendation*. https://www.w3.org/TR/vc-data-integrity/

[14] NIST CSRC. *Air gap glossary entry*. https://csrc.nist.gov/glossary/term/air_gap

[15] NASA. *Runtime Assurance of Aeronautical Products. Preliminary Recommendations*. https://ntrs.nasa.gov/citations/20220015734

[16] EASA. *Artificial Intelligence Concept Paper Issue 2*. https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-issue-2

[17] Spherity Research. *Evidence Graphs for Industrial AI*. https://spherity.github.io/spherity-research/evidence-graphs-industrial-ai-data-plane.html

[18] W3C. *JSON-LD 1.1. Recommendation*. https://www.w3.org/TR/json-ld11/

[19] European Union. *Regulation (EU) 2024/1183. European Digital Identity Framework*. https://eur-lex.europa.eu/eli/reg/2024/1183/oj/eng

[20] European Commission. *Implementing Regulation (EU) 2026/1731*. https://eur-lex.europa.eu/eli/reg_impl/2026/1731/oj/eng

[21] European Commission. *Implementing Regulation (EU) 2026/1735*. https://eur-lex.europa.eu/eli/reg_impl/2026/1735/oj/eng

[22] W3C. *RDF Dataset Canonicalization 1.0. Recommendation*. https://www.w3.org/TR/rdf-canon/

[23] W3C. *Bitstring Status List v1.0. Recommendation.* https://www.w3.org/TR/vc-bitstring-status-list/

[24] European Union. *Artificial Intelligence Act, Article 14 on human oversight.* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02024R1689-20260727

[25] European Commission. *Proposal for a Regulation establishing European Business Wallets. COM(2025) 838 final.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52025PC0838

[26] Qwen Team. *Qwen3Guard Technical Report.* https://arxiv.org/abs/2510.14276

[27] Alibaba Cloud. *Model Studio input and output AI guardrail.* https://www.alibabacloud.com/help/en/model-studio/content-security

[28] Anthropic. *Responsible Scaling Policy and Frontier Risk Reports.* https://www.anthropic.com/responsible-scaling-policy

[29] METR. *Frontier Risk Report on Internal Frontier Agents.* https://metr.org/blog/2026-05-19-frontier-risk-report/

[30] Bai et al. (2022). *Constitutional AI. Harmlessness from AI Feedback.* https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

[31] Agent Payments Protocol. *Agent Authorization Framework.* https://ap2-protocol.org/ap2/agent_authorization/

[32] Google DeepMind. *Strengthening Our Frontier Safety Framework.* https://deepmind.google/blog/strengthening-our-frontier-safety-framework/

[33] Meta. *Advanced AI Scaling Framework and Safety Reports.* https://ai.meta.com/blog/scaling-how-we-build-test-advanced-ai/

[34] Meta. *Responsible Approach to Meta AI and Llama 3.* https://ai.meta.com/blog/meta-llama-3-meta-ai-responsibility/

[35] Microsoft Learn. *Conditional Access for Agents in Microsoft Entra.* https://learn.microsoft.com/en-us/entra/identity/conditional-access/agent-id

[36] Microsoft Learn. *AI Red Teaming Agent in Microsoft Foundry.* https://learn.microsoft.com/en-us/azure/foundry/concepts/ai-red-teaming-agent

[37] NTT DATA. *Responsible & Secure AI service launch.* https://www.nttdata.com/global/ja/news/topics/2026/012000/

[38] NTT R&D. *Security for AI research at AI for Good 2026.* https://www.rd.ntt/e/sil/media/0026.html

[39] NVIDIA. *Open Agent Safety Platform.* https://www.nvidia.com/en-us/solutions/ai/agent-safety/

[40] NVIDIA. *NeMo Guardrails documentation.* https://docs.nvidia.com/nemo-guardrails/

[41] OpenAI. *Updating Our Preparedness Framework.* https://openai.com/index/updating-our-preparedness-framework/

[42] OpenAI. *Frontier Governance Framework.* https://openai.com/index/openai-frontier-governance-framework/

[43] OpenAI. *Path to Astra and frontier safeguards*. https://openai.com/index/path-to-astra/

[44] Spherity Research. *European Business Wallets as the Legal Control Plane for Zero Trust AI Agents*. https://spherity.github.io/spherity-research/ebw-zero-trust-ai-agents.html

[45] Microsoft. *2026 Responsible AI Transparency Report*. https://www.microsoft.com/en-us/corporate-responsibility/topics/responsible-ai/reports/transparency-report/

[46] NVIDIA. *Agent Policy Fabric architecture. Development documentation*. https://docs.nvidia.com/agent-policy-fabric/about/architecture

[47] UK AI Security Institute. *Incident report on unsanctioned agent behavior*. https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

[48] Spherity. *Company profile and founding date*. https://www.spherity.com/about

[49] Spherity Research. *Multi-Trust Architectures for Authorised Agentic Actors*. https://spherity.github.io/spherity-research/beyond-zero-trust-m-trust-authorised-agentic-actors.html

[50] Stöcker, C. (2026). *AI Safety Capability Analysis of the NVIDIA Open Agent Safety Platform*. https://www.linkedin.com/pulse/ai-safety-capability-analysis-nvidia-open-agent-platform-st%C3%B6cker-kfcde

[51] FDA-hosted consortium report. *Credentialized DSCSA Authorized Trading Partner Pilot*. https://www.fda.gov/media/147175/download

[52] Healthcare Distribution Alliance. *2023 Distribution Management Award to Spherity and Legisym*. https://www.hda.org/newsroom/2023/hda-presents-2023-distribution-management-award-to-spherity-and-legisym/

[53] SAP. *Spherity DSCSA Master Data Hub CARO*. https://www.sap.com/germany/products/scm/partners/spherity-gmbh-spherity-dscsa-master-data-hub-caro.html

[54] German Federal Ministry program portal. *IDunion project profile and consortium*. https://www.digitale-technologien.de/Navigation/EN/ProgrammeProjekte/AbgeschlosseneProgrammeProjekte/Sichere_Digitale_Identitaeten/Projekte_Umsetzungsphase/IDunion/IDunion

[55] Antares Vision Group. *Credentialing partnership expansion with Spherity*. https://antaresvisiongroup.com/it/news/post/?blog_id=3&id=2979

[56] European Commission. *Four projects to test the EU Digital Identity Wallet*. https://digital-strategy.ec.europa.eu/en/news/eu-digital-identity-4-projects-launched-test-eudi-wallet

[57] Spherity. *Participation in the EUDI Wallet Consortium pilot*. https://www.spherity.com/post/newsroom-spherity-ewc-eu-digital-identity

[58] European Commission. *EU Digital Identity Wallet large-scale pilots*. https://ec.europa.eu/digital-building-blocks/sites/pages/viewpage.action?navigatingVersions=true&pageId=930449129

[59] WE BUILD Consortium. *Trusted Identities for AI Agents non-paper*. https://www.webuildconsortium.eu/news/non-paper-trusted-identities-for-ai-agents-an-opportunity-for-europe-8396

[60] Spherity GmbH (2026). Company statement on IPCEI-AI consortium collaboration.

[61] European Commission. *Design and pre-notification of the first IPCEI in AI*. https://digital-strategy.ec.europa.eu/en/news/commission-welcomes-design-first-important-project-common-european-interest-ai

[62] Spherity Research. *Sovereign Industrial AI and global cooperation*. https://spherity.github.io/spherity-research/global-cooperation-sovereign-industrial-ai.html

[63] Stöcker, C. (2026). *Runtime Revocation in Multi-Agent Systems*. https://spherity.github.io/spherity-research/Spherity-Reseach-AI-Agent-Runtime-Revocation-in-MAS.pdf

[64] European Commission. *JEF-IPCEI technical meeting minutes, 21 October 2026*. https://competition-policy.ec.europa.eu/document/download/fe416a02-2883-4686-8a9f-54fcd1d19565_en?filename=JEF-IPCEI_minutes_TLM_Oct2026.pdf

[65] AI Verify Foundation. *Project Moonshot*. https://aiverifyfoundation.sg/tools/moonshot/

[66] AI Verify Foundation and IMDA. *GenAI Assurance Pilot main report*. https://assurance.aiverifyfoundation.sg/wp-content/uploads/2025/06/Assurance-Pilot-Main-Report.pdf

[67] Australian Government. *Australia's AI Safety Institute*. https://www.industry.gov.au/science-technology-and-innovation/technology/artificial-intelligence/ai-safety-institute

[68] Australian AI Safety Institute. *Epistemic Safety in Scalable Oversight*. https://www.industry.gov.au/publications/epistemic-safety-scalable-oversight

[69] Beijing Academy of Artificial Intelligence. *Institutional profile*. https://www.baai.ac.cn/en/about-us

[70] BAAI FlagEval. *Safety evaluation workflow*. https://flageval.baai.ac.cn/docs/rules/evaluation-operation-process.html

[71] Fraunhofer IAIS. *AI Assessment Catalog*. https://www.iais.fraunhofer.de/content/dam/iais/publikationen/studien-und-whitepaper/2023/Fraunhofer_IAIS_AI_ASSESSMENT_Catalog_Web.pdf

[72] Japan AI Safety Institute. *Guide to Evaluation Perspectives on AI Safety version 1.20*. https://aisi.go.jp/output/output_information/260707/

[73] Mila. *Open source AI safety partnership with ROOST*. https://mila.quebec/en/news/roost-and-mila-announce-partnership-to-advance-open-source-ai-safety-starting-with-youth

[74] MLCommons. *AILuminate benchmarks*. https://mlcommons.org/ailuminate/

[75] NIST. *Center for Advancing Innovation and Standards for Super Intelligence*. https://www.nist.gov/caissi

[76] NIST. *Challenges of Monitoring Deployed AI Systems*. https://www.nist.gov/publications/challenges-monitoring-deployed-ai-systems-center-ai-standards-and-innovation

[77] Stanford HAI. *2026 AI Index Report chapter on responsible AI*. https://hai.stanford.edu/ai-index/2026-ai-index-report/responsible-ai

[78] Stanford CRFM. *HELM Safety results and methodology*. https://crfm.stanford.edu/helm/safety/v1.5.0/

[79] UK AI Security Institute. *Frontier AI Trends Report*. https://www.aisi.gov.uk/frontier-ai-trends-report

[80] UK AI Security Institute. *Evaluation engineering playbook and Inspect*. https://www.aisi.gov.uk/blog/releasing-aisis-engineering-playbook

[81] Spherity Research. *Open research library*. https://spherity.github.io/spherity-research/

[82] Bundesanzeiger Verlag. *Business Wallet and register-backed business identity*. https://www.bundesanzeiger-verlag.de/evidenzwesen/eubw/

[83] Spherity. *EIDA European Business Wallet*. https://www.spherity.com/eida

[84] Spherity Research. *European Business Wallets Regulation Roadmap and Business Value*. https://spherity.github.io/spherity-research/ebw-roadmap.html

[85] Eurostat. *EU gross domestic product in 2025*. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/wdn-20260806-1

[86] Spherity GmbH (2026). Company statement on Business Wallet deployment and European registry integration work. Scope undisclosed.

[87] WE BUILD Consortium. *Consortium and cross-border business register participation*. https://www.webuildconsortium.eu/

[88] WE BUILD Consortium. *November 2026 General Assembly and interoperability progress*. https://www.webuildconsortium.eu/news/simplifying-cross-border-business-we-build-consortium-meets-in-bucharest-to-take-stock-of-its-first-year-and-plan-the-road-ahead

[89] energy data-X. *Federated service Identity and Trust*. https://www.energydata-x.eu/wp-content/uploads/2025/03/OnePager_06_Identity_Trust_03-2025.pdf

[90] Gößling et al. (2026). *The energy data-X data ecosystem. Energiewirtschaftliche Tagesfragen, 76(7–8), 19–22*. https://spherity.github.io/spherity-research/energy-data-X-European-Business-Wallet-market-role-credentials-BESS-2026.pdf

[91] W3C. *Verifiable Credentials Working Group participants*. https://www.w3.org/groups/wg/vc/participants/

[92] W3C Credentials Community Group. *Verifiable Credentials for Recognition v0.9. Final Community Group Report*. https://www.w3.org/community/reports/credentials/CG-FINAL-vc-recognition-20260320/

[93] W3C. *Recognized Entities v1.0. Working Draft, 2 November 2026*. https://www.w3.org/TR/2026/WD-vc-recognized-entities-1.0-20261102/

[94] NVIDIA. *Agent Policy Fabric identity and delegation. Development documentation*. https://docs.nvidia.com/agent-policy-fabric/how-it-works/identity-and-delegation

[95] Stöcker, C. (2026). *Resource-Bounded Program Equilibria and Protected Execution*. https://spherity.github.io/spherity-research/resource-bounded-program-equilibria-ai-safety.html

[96] Stöcker, C. (2026). *Risk-Bounded Runtime Assurance for Dynamic Multi-Agent Systems*. https://spherity.github.io/spherity-research/risk-bounded-runtime-assurance-multi-agent-systems.html

[97] Stöcker, C. (2026). *Private Verification of Action Eligibility Across Trust Domains*. https://spherity.github.io/spherity-research/private-verification-action-eligibility-trust-domains.html

[98] Stöcker, C. (2026). *Governed Memory for Physical AI*. https://spherity.github.io/spherity-research/governed-robot-memory-physical-ai.html