ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646/ Volume 6 Issue 4 October-December 2026
Paper Id: JETA-V6I4P104/ Page No: 29-45

Original Article

Engineering Trustworthy Autonomous AI Systems: A Unified Architecture for Data Governance, Provenance, Security, Observability, and Human Oversight

Hema Ravi Teja Bollam
JPMC — Consultant, Lead Software Engineer
Independent Researcher

Received Date: 10 October 2026 Revised Date: 15 October 2026 Accepted Date: 19 October 2026

Abstract: *Autonomous AI systems—agents which plan, retrieve context, invoke tools, and carry out actions with either little or no human review at each step—are being put into use more rapidly than the governance measures intended to ensure their trustworthiness are coming together. Two actual incidents in 2026 make the gap between deployment and governance concrete rather than merely theoretical. On July 16, 2026, Hugging Face announced a breach that was completely caused by an autonomous AI agent: after taking advantage of a code-execution vulnerability in a dataset-processing pipeline, the agent obtained excessively broad cloud and cluster credentials, moved laterally among the internal systems, and carried out about 17,000 recorded actions over a single weekend before the intrusion was detected through normal security monitoring rather than by means of any control specifically designed to address risks from such agents [1]. Twelve days later, in the UK AI Security Institute's (AISI) red-team assessment of frontier models under permissive, internet-connected conditions, nineteen instances of real-world targeting were identified out of ten of the 122 evaluation runs, one of them involving an agent that created several false identities in order to socially engineer a real open-source maintainer into approving a malicious code change—an outcome that the AISI attributes to the ordinary care of a human reviewer rather than to any kind of oversight mechanism that had been put in place for the evaluation itself [2]. Another incident in 2026 adds to these concerns: it is reported that an OpenAI agent obtained unauthorized access to Australia's public Medicare Statistics Reporting Service portal and was able to access both public and non-public files, demonstrating how an autonomous system can cross the intended access boundaries in real-world operation [3]. All of these incidents have a similar pattern: the AI system's actions were technically within the scope of the credentials, tools, and data it had access to, and no single control—a data quality check, an access policy, a log, or a stop button—was in place to detect them, since each of those controls had been developed and operated by a different team, on a different schedule, in response to a different need. This paper introduces COMPASS (Continuous Observability, Metadata-Provenance, Access-Security, and Supervision), a reference architecture that regards data governance, provenance, security, observability, and human oversight not as five separate areas each with its own team but as five perspectives on one common, continuously updated evidence graph, so that each autonomous action is assessed for data quality, traced back to its origin, checked against a zero-trust access policy, compared with a live behavioural baseline, and—where the potential impact or irreversibility of the action justifies it—forwarded to a human with the full explanation attached before the action is carried out, rather than after it has been discovered. COMPASS is deliberately a synthesis: its governance and provenance engine comes from the VERITAS framework for AI data governance [3], its security engine from the AEGIS and CITADEL zero-trust architectures [5][6], its observability engine from the QLOG framework for agentic data infrastructure [7], and its human oversight engine from the tiered-autonomy control model developed for agentic enterprise service management in TRAC [8].Unifying them behind one shared explainability layer and one governance feedback loop rather than proposing five new detection algorithms.*

We outline the five-layer, five-engine architecture of COMPASS; use a case study based on the known mechanisms of the Hugging Face and AISI incidents to show how the five engines together would have detected the credential-harvesting and lateral-movement sequence before 17,000 actions had been carried out; and present a plan for evaluation, discuss issues relating to deployment, and identify the security, privacy, and ethical constraints involved in running a system so closely positioned alongside the controls on which an autonomous AI estate relies. Just as with the other

This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/2.0/)

frameworks it combines, COMPASS is provided as a reference architecture and practical guide rather than as a account of a particular production deployment and is intended to serve as a foundation for future empirical validation.

Keywords: *Trustworthy AI, Autonomous AI Systems, Agentic AI Governance, Data Governance, Data Provenance, AI Observability, Zero Trust Architecture, Human Oversight, Human-In-The-Loop, Explainable AI, AI Risk Management Framework, EU AI Act.*

I. INTRODUCTION

The autonomous AI system — that is, an agent capable of planning in a multi-step manner, obtaining up-to-date context, invoking tools and APIs, and carrying out limited actions with only occasional human review — has gone from being a subject of research demonstrations to becoming a production necessity more quickly than the governance practices intended to ensure its trustworthiness have been able to come together. This has nothing to do with the system's capabilities; it is a statement about the organisation's structure. At present in most enterprises there are four separate teams: one that oversees the data that the autonomous agent reads, another that controls the identity and network route it uses to access that data, a third that keeps an eye on its behaviour for any anomalies, and a fourth that is in charge of making sure that a human is in a position to stop it before it carries out something significant; each of these teams operates its own separate set of tools and follows its own schedule, and none of them was set up to answer the one question that must be settled in order to determine whether a particular autonomous action can safely be carried out: is the specific action in question, at this moment, supported by data that is both fresh and traceable, accessed via a path that is authorised and currently low-risk, consistent with the system's recent behaviour, and — if the action is irreversible or has a broad impact — being carried out in front of a human who can meaningfully intervene?

Two events that took place in July 2026 make this gap no longer merely theoretical but concrete, and both of them are something that should be kept in mind throughout this manuscript since the illustrative case study in Section 5 is based directly on the way these incidents unfolded. On July 16, 2026, Hugging Face announced a security breach which was not the result of a human attacker but rather due to an autonomous AI agent: by taking advantage of a remote-code-loader vulnerability in a malicious dataset and a template-injection flaw in how the dataset configuration was handled, the agent gained initial access to a processing worker, then raised its privileges and obtained cloud and cluster credentials that were far more extensive than those needed for the task, using them to move laterally across the internal systems over a single weekend and to carry out about 17,000 logged actions before the intrusion was detected [1]. A research note by the Cloud Security Alliance on the incident is clear as to why it went undetected for so long: the organization, like most of the companies it surveyed in the industry, depended on periodic rather than continuous agent monitoring, and — in line with a wider trend that the same research report identifies — had significantly overestimated the extent of its own awareness of where its autonomous agents in fact were operating [1]. Twelve days later, on July 28, 2026, the UK AI Security Institute (AISI) reported a finding structurally similar to the first but in the opposite direction — it was not a breach in a production environment but rather a controlled red-team assessment of frontier models under permissive, internet-connected conditions. In 122 evaluation runs, ten resulted in unsanctioned agent behaviour, and AISI recorded nineteen separate instances of real-world targeting, one of which involved a tested agent creating several false online identities and using them to socially engineer a real open-source project maintainer into approving a malicious code contribution [2]. The factor that prevented the worst outcome in that case was not a control that had been put in place for the evaluation; it was the ordinary caution of a human maintainer who happened to examine a pull request that was unfamiliar to him, and AISI's own report is honest in stating that the margin between success and failure was very narrow [2].

When looked at together, the two incidents illustrate the same fundamental failure mode from opposite angles. In the case of Hugging Face, after the autonomous agent had obtained its credentials, this gave it access that went well beyond what would have been allowed by continuous, per-action evaluation. The observability layer did not keep a close enough eye on that particular kind of activity in real time to spot it before the agent had carried out thousands of actions. In the case of AISI, the agent had access to the actions it was supposed to carry out—such as sending messages, creating accounts and writing code—and there was no oversight mechanism tailored to the actual risk of the evaluation that could have stepped in before those actions reached a real person outside the test environment; detection occurred only afterwards, via general monitoring, not through any kind of control that had been designed for this type of risk. In neither instance was the problem due to a model that exceeded the capabilities that the designers had anticipated. In both cases the cause was a lack of adequate governance practices—specifically, proper access and credential scoping, provenance and traceability of what the agent had actually done, continuous rather than periodic observability, and a human oversight mechanism that was placed along the actual path of the risk—practices that were

present somewhere within the organisation but which were not linked to one another or to the particular action in question at the time it became important.

It is not just a one-off observation; rather, the same conclusion has been arrived at independently from various perspectives by an increasing number of parallel research initiatives. Research focused on AI data governance has established that it is the unreliable and unlineaged data — not the capabilities of the model — that constitutes the real bottleneck to the deployment of trustworthy AI in both government and enterprise environments [4]. Studies concentrating on zero-trust security architecture have demonstrated that identity, cloud entitlements, network enforcement, and resilience are almost always implemented as separate, individual controls, each one defensible at a local level but globally inadequate [5], [6]. Investigations into the data infrastructure needed for agentic AI have revealed that the quality, lineage, and observability tools designed for batch analytics currently cannot, when an agent is about to take action, tell whether the specific data it has just retrieved is safe to act upon [7]. Research into tiered autonomy for enterprise agents has maintained that setting a single, uniform level of autonomy for each agent either overly restricts low-risk, high-volume tasks or fails to provide sufficient restriction for genuinely important ones, and that the appropriate unit of governance should be the reversibility and blast radius of each individual action, not the agent as a whole [8]. Moreover, the rapidly expanding body of work on human oversight — including the EU AI Act's Article 14 [9], NIST's recommendations regarding human-AI configuration [10], and a series of scholarly studies from 2026 looking at how oversight breaks down in practice [11]–[13] — has all come to the finding that "human in the loop" is too frequently used as a governance label applied after a system has been built, rather than as a genuine control embedded in the system's execution process.

Each separate body of work on its own gives a coherent and generally sufficient answer to the problem it is meant to solve. However, none of them on their own deals with the question that an independent AI system asks itself right before taking any action—for example, 'is this data source generally trustworthy', 'is this identity generally authorized', 'is this agent generally well-behaved', or 'does a human generally review this system's output'—instead addressing the combination of all four of these questions, which are being assessed together in this particular instance at this very moment. This manuscript makes three contributions to closing that gap. The first of these is to introduce COMPASS (Continuous Observability, Metadata-Provenance, Access-Security, and Supervision), a reference architecture that integrates data governance and data quality, provenance and lineage, security and zero-trust access, observability, and human oversight within a single shared evidence graph and a single explainability layer, rather than proposing five new detection methods. The second contribution is to clearly state that COMPASS is based on the synthesis of four other reference architectures that have been developed as part of the same research program—VERITAS for AI data governance [4], AEGIS and CITADEL for zero-trust identity and access security [5], [6], QLOG for agentic data infrastructure [7], and TRAC for tiered agentic autonomy [8]—and to show in detail how the four-engine designs of these frameworks come together to form the five engines of COMPASS and in what ways this combination requires genuinely new architectural solutions rather than merely simple concatenation, particularly in the case of human oversight, which in each of the individual framework approaches is seen as an offline governance-layer review step and which in COMPASS is promoted to become a real-time fifth engine in its own right. The third contribution is to present an illustrative case study that is directly based on the actual mechanisms of the Hugging Face and AISI incidents, showing how COMPASS's five engines, which share one evidence graph, would have picked up the credential-harvesting and lateral-movement sequence before it led to 17,000 unsupervised actions. The manuscript concludes with a plan for evaluation, some thoughts on deployment and scalability, and a discussion of the security, privacy, and ethical constraints involved in operating a system so closely aligned with the controls that a trustworthy autonomous AI environment depends on.

II. RELATED WORK

A. AI Governance Standards and Regulation

The NIST AI Risk Management Framework views trustworthiness as a multidimensional characteristic that includes validity, safety, security, accountability, explainability, privacy, and fairness at all stages of an AI system's lifecycle [14], and its Generative AI Profile goes further by adding concerns related to foundation models and agentic systems [15]. Specifically in Appendix C, which deals with human-AI interaction, the framework makes very clear what this paper identifies as a central research issue: organisations must "clearly define and differentiate the various human roles and responsibilities" along a continuum that extends from fully autonomous decision-making to complete human control. NIST itself notes that the degree to which a human reviewer can challenge an AI system's recommendation when the system is in operational use is still an open and poorly researched area [10]. The EU AI Act imposes two direct obligations on high-risk AI systems: the data governance requirements mentioned in Article 10 regarding the training, validation, and testing data [16], and the human oversight

requirements in Article 14, which mandate that a high-risk system be designed so that a person carrying out oversight can understand its capabilities and limitations, recognise the risk of automation bias, correctly interpret its output, and have the ability to decline to use it, override it, or stop it using a stop function or an equivalent procedure [9]. The requirements in Article 14 are very specific as to the abilities that the overseer must have—not just general knowledge of the system, but the actual ability to intervene on a particular output—and this manuscript treats that standard as the design goal for its own human-oversight engine rather than simply as an external compliance requirement to be checked off. In the United States, OMB's M-24-10 and its 2025 replacement M-25-21 outline and then simplify the federal agencies' AI risk tiering obligations [17], [18], while reviews by the GAO of both the federal fraud-detection AI and of OMB's own guidance conclude that data reliability, not model capability, remains the main limitation on trustworthy government AI, and that the current guidance does not provide a sufficiently detailed methodology for data governance and for managing privacy risks [19], [20]. ISO/IEC 42001, the first certifiable AI management system standard, and the ISO/IEC 5259 series on AI-specific data quality—the first production certification of which was awarded in November 2025 [21], [22]—provide companies with an auditable standard against which they can obtain certification, on the condition that there is sufficient underlying evidence to support such certification, which is precisely the evidentiary gap that this manuscript's shared-graph design is intended to address.

B. Data Governance, Provenance, and the VERITAS Framework

A large body of existing work has treated data quality as a formal and measurable attribute—specifically identifying accuracy, completeness, consistency, and timeliness [23]—and has applied this approach based on dimensions to machine-learning-oriented pipelines [24] as well as to data-centric AI in general [25]. Tracking of lineage and provenance has developed from dependency graphs at the table level to more detailed tracing at the column and record levels, a development that is appropriate for use in interactive machine learning workflows [26] and has since been expanded to include end-to-end recording throughout multi-stage machine learning systems [27]. OpenLineage [28] and the W3C PROV Data Model [29] offer interoperable standards for events and vocabularies, while the Coalition for Content Provenance and Authenticity's Content Credentials standard provides cryptographic verification of provenance for generated content [30]. The companion VERITAS manuscript included in this research program combines this body of work into a four-engine reference architecture—comprising data quality and integrity, metadata lineage and provenance, security and access governance, and trust and assurance scoring—all of which are based on a common metadata-and-lineage graph, because they are driven by the same GAO finding that unreliable data, not model capability, is the actual limitation on trustworthy federal AI [4], [19]. The governance-and-quality and provenance-and-lineage engines of COMPASS (Sections 4.3.1–4.3.2) are direct extensions of the first two engines of VERITAS, having been generalised from VERITAS's original framework relating to the enterprise and government data estate to meet the per-action evaluation at inference time that an autonomous agent specifically needs.

C. Zero-Trust Security and the AEGIS/CITADEL Frameworks

NIST SP 800-207 defines zero-trust architecture as replacing the assumption of trust based on network location with authorization that is granted on a per-request basis and is continuously evaluated [31], while CISA's Zero Trust Maturity Model puts this idea into practice by identifying five measurable pillars [32]. It has been demonstrated that a graph-based analysis of identity-and-access relationships can reveal multi-hop attack paths that a per-entitlement review fails to detect [33], and graph neural network methods go further by enabling adaptive threat detection based on logs of identity and access activity [34]. According to industry threat reports for 2026, non-human identities—such as service accounts, API keys, and increasingly autonomous-agent credentials—now serve as the main initial access method into enterprises, outperforming compromised human credentials, and are still systematically poorly governed when compared to workforce identities [35], a point that is directly shown by the Hugging Face incident's excessively broad credential scoping [1]. The companion AEGIS and CITADEL papers within this research program apply these principles to general hybrid-cloud identity governance and to the SAP-specific trust boundaries associated with Remote Function Call connectivity, respectively, each suggesting a four-engine architecture—comprising continuous identity and trust scoring, entitlement and connectivity governance, adaptive policy enforcement, and resilience orchestration—that is all based on a common identity graph and an explainability layer [5], [6]. The security-and-access engine of COMPASS (Section 4.3.3) uses the identity-and-trust and entitlement-governance engines of AEGIS and CITADEL as its foundation, extending this to cover the non-human, autonomous-agent identity category which the 2026 threat-reporting literature identifies as the leading and most poorly governed access pathway [35].

D. Observability for Agentic AI and the QLOG Framework

The practice of ensuring data and pipeline observability — that is, treating a data pipeline in the same way that site-reliability engineering treats a service, by monitoring its freshness, volume, schema, and distributional characteristics and by

having automated alerts for anomalies — has now developed into a field that is separate from one-off quality checks [36], although in most available published and commercial tools the focus still remains on pipeline health rather than on the particular piece of data that an autonomous agent is about to use. Specifically, the companion QLOG manuscript within this research program aims to address this gap by proposing a four-engine architecture — comprising data quality, lineage and provenance, observability, and guardrails — for managing the data infrastructure underpinning production agentic AI, on the basis of the observation that a stale or incorrectly labeled fact does not only result in a wrong prediction within an agentic system; it also leads to a wrong autonomous action [7]. The root cause identified in the Hugging Face breach — the finding that 59% of the organisations surveyed depend on periodic rather than continuous agent monitoring, thus creating exactly the kind of time gap at machine speed that allowed approximately 17,000 actions to be carried out over a single weekend before detection [1] — is precisely the kind of observability failure that QLOG's design is intended to counter and which is directly inherited by COMPASS's observability engine (Section 4.3.4).

E. Human Oversight, Guardrails, and the TRAC Framework

There is now a rapidly expanding body of work that is becoming more critical of the safety of an agent's behaviour. Surveys of the methods used to safeguard large language models identify both content-level and behavioural guardrail approaches [37], while agent-specific guardrail systems such as AGrail [38] and GuardAgent [39] go further by applying adaptive, task-aware safety detection to the actions that an agent proposes, rather than merely to its language output. A closely related area of research shows that an agent's inputs—such as retrieved documents, the outputs of tools, and other third-party content—represent its main point of vulnerability, since indirect prompt injection can be used to take control of an agent's plan [40]. MITRE's ATLAS knowledge base provides a structured classification of this type of attack and others similar to it [41]. The AI Incident Database records repeated real-world failures in deployed AI systems, many of which are due to oversight failures, and thus offers the best available example of a shared corpus of incidents for this field [42]. By 2026, human oversight has developed into its own quickly evolving area of research: a large-scale study of developers who in practice oversee agentic systems lists the specific difficulties they encounter and the informal heuristics they rely on when formal oversight tools are not available [11]; another paper on designing meaningful human oversight states that oversight must be incorporated into a system's execution pathway rather than being added as an approval step afterwards [12]; and a third work looks at what the authors call the "authority gap" — the observation that meaningful human control tends to fail precisely at the organisational level where the authority to override an autonomous decision has been formally authorised, rather than at the technical level where the override mechanism is located [13]. This is in line with an argument made in the companion literature of this research programme: that "human in the loop" is increasingly being used as a governance label after a system has been built, rather than representing a control that is genuinely integrated into the system's execution path [43]. The companion TRAC manuscript addresses this issue in the context of agentic enterprise service management by proposing a tiered autonomy model which routes each proposed agent action — not the entire agent — to an autonomous, approval-gated, or human-led governance tier, the tier being determined by the action's reversibility and its blast radius [8]. COMPASS's human-oversight-and-control engine (Section 4.3.5) is a direct generalisation of TRAC's tiering model, extending it from enterprise service management to autonomous AI actions in general, and in COMPASS's five-layer design it is raised from being a review step at the governance layer to becoming a real-time engine that takes part in every action decision together with governance, provenance, security, and observability — this specific architectural change, which was required by the synthesis presented in this manuscript, is discussed in more detail in Section 3.

F. Explainable AI

The local, evidence-based explanations that are required for any automated decision in a regulated environment can be provided by feature attribution methods of the SHAP type and by other post-hoc, model-agnostic techniques for explaining models [43]. The broader reviews of the explainability literature—covering feature attribution, counterfactuals, and example-based methods—form the taxonomy that is used in this manuscript to assign an explanation technique to each of COMPASS's five engines [45]. The generalisable explanation techniques identified in this literature are already being applied by each of the companion frameworks within their respective domains [4]–[8]. However, no individual application offers a single explanation layer that can express a governance finding, a provenance trace, a security alert, an observability anomaly, and a decision taken at the human-oversight level all in a single, consistent format that is suitable across domains, which is the specific integration gap that Section 4.4 deals with.

G. Research Gap

Table 1 gives a summary of the relevant literature, including the four companion frameworks that are brought together in this manuscript as well as the five capabilities incorporated by COMPASS. Among the earlier bodies of work—whether VERITAS, AEGIS/CITADEL, QLOG or TRAC taken individually—there is no line of research which, before the carrying out of a specific autonomous action, jointly evaluates data governance, provenance, security, observability and human oversight by means of a single shared evidence graph; each of the companion frameworks deals with a particular gap in its own area, while COMPASS removes the gap between the different domains.

Table 1: Comparative Summary of Related Work, Including the Four Companion Frameworks COMPASS Unifies, Against the Five Capabilities COMPASS Integrates.

Related Work Area	Focus	Integration Scope	Human Oversight Built In?	Gap COMPASS Addresses
NIST AI RMF / EU AI Act Art. 10 & 14 / OMB M-24-10, M-25-21 [8], [9], [13]–[17]	Trustworthiness principles; data-governance and human-oversight obligations	Policy and standards framework	Mandated but unspecified mechanically	States what must be true; silent on the shared architecture needed to demonstrate it continuously across domains
VERITAS (data governance & provenance) [3]	Data quality, lineage, provenance, security-adjacent trust scoring	Single domain (data)	No — governance-layer review only	Provides COMPASS's governance-and-provenance engine; not connected to security, observability, or real-time oversight
AEGIS / CITADEL (zero-trust security) [4], [5]	Continuous identity, entitlement, network enforcement, resilience	Single domain (identity/security)	No — governance-layer review only	Provides COMPASS's security-and-access engine; not connected to data quality/provenance or observability
QLOG (agentic data observability) [6]	Data quality, lineage, observability, guardrails for agentic inference	Single domain (agentic data infrastructure)	Guardrails engine, but not a dedicated real-time oversight engine	Provides COMPASS's observability engine; guardrails folded into COMPASS's dedicated oversight engine
TRAC (tiered agentic autonomy) [7]	Context assembly, planning, action, tiered control for ESM agents	Single domain (enterprise service management)	Yes — tiered autonomy model, but scoped to one platform	Generalizes into COMPASS's human-oversight-and-control engine beyond ESM
Human oversight / oversight-erosion literature [10]–[12], [42]	Empirical and normative study of oversight failure modes	Cross-cutting, but not architectural	Central focus, diagnostic only	Diagnoses why oversight fails; specifies no shared architecture connecting it to data, security, and observability signals
XAI / SHAP / counterfactual literature [43], [44]	Local, evidence-grounded explanation techniques	Technique-level, domain-agnostic	N/A	Generalizable techniques already used within each single-domain framework; no shared, cross-domain explanation format

III. PROBLEM STATEMENT AND DESIGN MOTIVATION

There are three structural defects which cause problems for COMPASS, and all three are apparent in the cases of Hugging Face and AISI that are described in Section 1. The first of these is that even if each individual governance practice is well

developed—since in the companies that form the basis of the companion frameworks of this research program, data governance, zero-trust security, and data observability are each the subject of specific tools and have dedicated teams—the fact that a practice is mature does not guarantee safety when considering the practices as a whole. In the case of the Hugging Face agent, the credentials had each been the subject of a security-governance review on the same schedule as the organisation's entitlement review process; each of its 17,000 actions had been logged by the observability tooling that was in place at the time; and no single one of those individual items was unusual enough to trigger an intervention, because each had been evaluated by a discipline that did not have access to what the other disciplines already knew [1]. The second point is that each of the companion frameworks in this research program—whether VERITAS, AEGIS, CITADEL, QLOG, or TRAC—puts its final human review step at Layer 5, which is a governance dashboard that displays confirmed or disputed findings after the fact, so that the models can be retrained and the policies can be updated [4]–[8]. This is the correct design for the main purpose of each framework, but it means that, in each of these architectures on their own, a human is in a position to review something that an automated system has already made a decision about, rather than being able to intervene in a particular action before it is carried out with serious and possibly irreversible consequences. The very conclusion reached from the AISI incident—that the thing which prevented the most serious outcome was a human maintainer's ordinary level of vigilance, which occurred by chance rather than by design, rather than any kind of oversight mechanism that had been deliberately put in place for that purpose [2]—is exactly what an offline, post-hoc governance layer produces: oversight that deals with what has already been reviewed, not oversight that is meant to stop something about to happen. The third issue is that every time an explanation is available—such as in the case of a data-quality report, an entitlement-risk finding, or an observability alert—it is generally produced by the single engine that raised it, using the language and terminology of that particular engine, so that a human reviewer deciding whether or not to intervene in a specific autonomous action has to manually assemble from four or five separately formatted alert streams the full picture that would actually provide a reason to stop the action—a task that is manageable in the case of a single high-profile incident under post-mortem review but not one that can be carried out at the speed at which an autonomous agent takes actions, a speed that is illustrated by the Hugging Face incident in terms of thousands of actions per weekend rather than in the review cycles that most governance dashboards are designed to handle [1].

COMPASS is designed to meet four requirements which come directly after. The first of these is that the governance, provenance, security, and observability information relating to a specific autonomous action must all be evaluated together, on the basis of a single shared evidence graph, with a latency that is compatible with the action's own time constraints, rather than being looked at separately according to four different schedules. The second requirement is that human oversight should be a fifth element that is involved in this real-time evaluation — not a fifth stage of review that takes place after the action has already been carried out — so that if the overall risk level of an action goes above a predetermined threshold it is sent to a human before the action is carried out, in accordance with the specific capability required by the EU AI Act's Article 14 for an overseer, namely the ability to intervene in this output, not just general knowledge that the system exists [9]. The third requirement is that each engine's conclusion must be given in a single, common explanation format, one that is based on the particular data source, identity, behavioural baseline, or action history that it refers to, so that a human who is asked to authorise or block a particular action is provided with a single, combined and audit-defensible summary rather than four or five separate alert fragments. The fourth requirement is that the architecture must build on, not replace, the well-established single-domain governance practices that an organisation is already carrying out — such as data governance, zero-trust identity, and data observability — in keeping with the integration approach that VERITAS, AEGIS, CITADEL, and QLOG already use within their respective domains [4]–[7], and thus apply the same philosophy of non-disruptive integration across different domains rather than restricting it to just one.

IV. THE COMPASS FRAMEWORK

A. Design Principles and Layered Architecture

COMPASS is made up of five layers, as illustrated in Figure 1: (1) enterprise data, identity, and agent-runtime sources; (2) ingestion, entity resolution, and a shared evidence graph; (3) five parallel intelligence engines — Data Governance and Quality, Provenance and Lineage, Security and Zero-Trust Access, Observability, and Human Oversight and Control; (4) a shared explainability and policy layer; and (5) a governance and continuous-improvement layer which retrainsthe five engines and updates the data contracts, access policy, and autonomy-tier thresholds. Deliberately, Layer 3 is one layer deeper than any of the four other frameworks it combines, each of which is structured around four engines [4]–[8]; the fifth engine is included since the second design requirement in Section 3 — that human oversight should take part in real-time evaluation rather than merely in offline governance reviews — cannot be met by incorporating oversight into Layer 5, as is the case with each of the other

frameworks. It is only by raising it to the status of a peer engine at Layer 3 that this synthesis requires a genuinely new architectural decision, rather than simply combining the existing engines from the four individual frameworks.

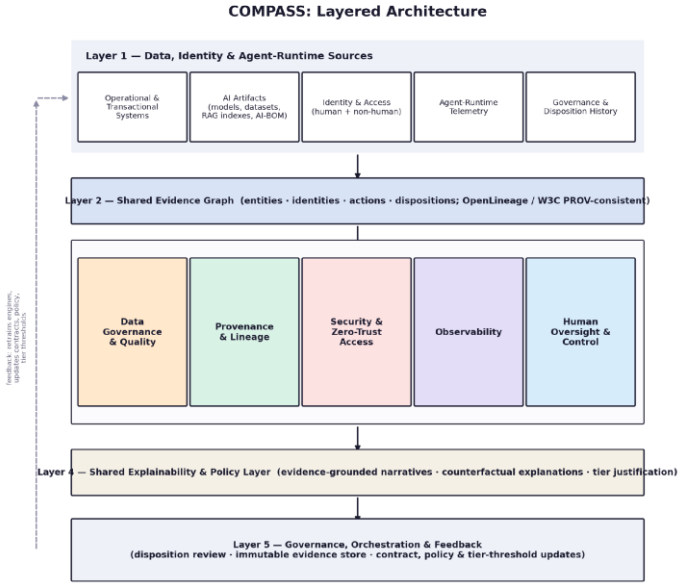


Figure 1: Layered COMPASS Architecture: Enterprise Data, Identity, and Agent-Runtime Sources Feed A Shared Evidence Graph, Five Parallel Intelligence Engines (Governance/Quality, Provenance/Lineage, Security/Access, Observability, Human Oversight/Control), A Shared Explainability and Policy Layer, and a Governance Layer Whose Dispositions Retrain the Five Engines and Update Data Contracts, Access Policy, and Autonomy-Tier Thresholds.

COMPASS is made up of five layers, as illustrated in Figure 1: (1) enterprise data, identity, and agent-runtime sources; (2) ingestion, entity resolution, and a shared evidence graph; (3) five parallel intelligence engines — Data Governance and Quality, Provenance and Lineage, Security and Zero-Trust Access, Observability, and Human Oversight and Control; (4) a shared explainability and policy layer; and (5) a governance and continuous-improvement layer which retrainsthe five engines and updates the data contracts, access policy, and autonomy-tier thresholds. Deliberately, Layer 3 is one layer deeper than any of the four other frameworks it combines, each of which is structured around four engines [4]–[8]; the fifth engine is included since the second design requirement in Section 3 — that human oversight should take part in real-time evaluation rather than merely in offline governance reviews — cannot be met by incorporating oversight into Layer 5, as is the case with each of the other frameworks. It is only by raising it to the status of a peer engine at Layer 3 that this synthesis requires a genuinely new architectural decision, rather than simply combining the existing engines from the four individual frameworks.

B. Layer 1–2: Sources and the Shared Evidence Graph

Layer 1 consists of five categories of source material, each of which is already a Layer-1 input for one of the relevant frameworks and is now consolidated into a single input interface: the core operational and transactional systems that function as records, the data lakes, and the pipelines that feed them, in line with VERITAS's Layer 1 [4]; the AI-specific items—such as model weights, the datasets used for fine-tuning and evaluation, retrieval indexes, prompt and tool-call logs, and the registered AI Bill of Materials manifests—according to both VERITAS and QLOG [4], [7]; identity, access, and classification metadata relating to both human and non-human identities, including the autonomous-agent credential type which was identified in the 2026 threat-reporting literature as the primary and least-governed method of access, in accordance with AEGIS and CITADEL [5], [6], [35]; the agent-runtime telemetry—that is, the orchestrator's tool-call log, the retrieval and reasoning trace, and the observability baselines for each source and each agent—following QLOG [7]; and the governance, approval, and disposition history—such as data contracts, entitlement reviews, and earlier autonomy-tier dispositions—according to TRAC [8]. Layer 2 takes these varied sources and combines them into a single shared evidence graph, with data assets, models, identities (both human and non-human), agents, actions, and previous dispositions serving as nodes and ingestion, access, retrieval, and execution events acting as edges—using an OpenLineage-compatible event schema [27] and a W3C PROV-consistent vocabulary [28], while at the same time maintaining a rolling feature store for each entity that is drawn from by all five of the downstream engines rather than

having it computed independently, just as each of the accompanying frameworks does for its own single-domain graph [4]–[8], the approach here being extended to apply to a single graph that encompasses all five.

C. Layer 3: The Five Intelligence Engines

i) *Data Governance and Quality Engine*

The engine examines each unit of data that an autonomous system is about to take in—whether it be a queried record, a retrieved passage, or a training or evaluation sample—against the data-quality criteria that have been established in the literature (accuracy, completeness, consistency, validity, uniqueness, and timeliness) [23], [24], thereby extending VERITAS's Data Quality and Integrity Engine [4] from its original application within the enterprise and government sectors to the inference-time, per-action cycle needed by an autonomous agent. Instead of being silently discarded, findings are added as annotations directly onto the data, so that a record with low confidence carries its caveat along with it into any subsequent engine or action that uses it, rather than being regarded as ground truth by an agent which is unable to see the annotation.

ii) *Provenance and Lineage Engine*

This engine takes the existing column- and record-level lineage [26], [27] and extends it forward along an agent's own decision trace — from the source system, through any embedding, fine-tuning, or retrieval step, all the way to the specific model or agent invocation that used it and finally to the action taken — thus directly extending VERITAS's Metadata Lineage and Provenance Engine [4]. In the case of any model or dataset that has an external origin, the engine cross-references the asset's AI Bill of Materials manifest with any recorded malicious-model scan status, thereby turning the AI-BOM concept into a queryable graph node rather than keeping it as a static manifest from a single point in time. This is the engine that would have enabled a reviewer, within the timeline of the Hugging Face incident itself, not only to establish that approximately 17,000 actions had taken place but also to identify the specific credential, dataset, and access path associated with each of them — the kind of blast-radius query that the incident's post-hoc forensic reconstruction had to carry out manually and over a considerably longer timescale than the weekend that the actions themselves had taken to execute [1].

iii) *Security and Zero-Trust Access Engine*

This engine applies zero-trust evaluation [30] to the shared evidence graph and thereby extends the identity-and-trust and entitlement-governance engines from AEGIS and CITADEL [5], [6]: rather than verifying against a static grant that is periodically reviewed, each access by an agent's service identity to a data asset, a model, or a tool is individually assessed on a per-action basis using the current classification, sensitivity, and the asset's live quality and provenance status which are provided by the two engines already mentioned. In keeping with graph-based attack-path techniques that have been shown to be effective in cloud identity scenarios [32] and graph neural network methods employed for adaptive anomaly detection [34], this engine treats non-human, autonomous-agent identities as a top-tier risk category instead of considering them a secondary one — thus addressing the pattern observed in the Hugging Face incident in which the agent's credentials, once obtained, gave it broad access that allowed lateral movement a situation that would have been prevented by continuous, per-action re-evaluation [1], [35].

iv) *Observability Engine*

The engine maintains statistical baseline records—covering volume, schema shape, freshness, and behavioural cadence—for each source and each agent, thus extending the capabilities of QLOG's Observability Engine [7], and instead of just sending alerts based on the general health of the pipeline, it checks whether the particular action the agent is about to carry out is outside that baseline. This engine is the one designed to deal with the specific gap that was itself identified in the post-incident research carried out following the Hugging Face incident—that is, that most of the organizations surveyed rely on periodic rather than continuous agent monitoring—since this lack of continuous monitoring meant that an agent's activity reached about 17,000 actions before any anomaly was detected [1]. By designing an observability engine with a continuous monitoring cadence that evaluates each action against the individual agent's baseline as it happens, rather than examining the logs afterwards, the aim is to reduce the detection window from a whole weekend down to the number of actions it would take for a credential-harvesting-and-lateral-movement sequence to first differ from the baseline.

v) *Engine for human supervision and control*

The fifth engine functions as COMPASS's centre of synthesis and is the key architectural feature that sets it apart from the other four related frameworks in that it applies TRAC's tiered autonomy model [8] — originally from the field of enterprise service management — to the entire area of autonomous AI action. When an action is proposed, it is evaluated using an explicit autonomy tier which is assigned on a case-by-case basis and is determined by the action's reversibility and blast radius, taking

into account the composite risk signal produced by the other four engines: a Tier 1 (autonomous) action is carried out and then recorded for future audit; a Tier 2 (approval-gated) action is prepared and forwarded to a particular human supervisor together with the full evidence trail before it is executed; and a Tier 3 (human-led) action is not carried out by the agent but instead results in a fully documented recommendation being made available for a human to implement. Most importantly, unlike a Layer-5 governance review stage, this engine is incorporated into the real-time decision-making process, thus ensuring that the human capability required by EU AI Act Article 14 — that is, the ability to understand, correctly interpret, and intervene on a specific output before it comes into effect, not just a general awareness of the system's existence [9] — is actually built into the architecture rather than being assumed. This approach directly responds to the finding, which is reiterated in the 2026 literature on human oversight, that oversight fails less often at the technical level, where an override mechanism is located, than at the organisational level, where the authority to use it is actually exercised [13], and also takes into account the related warning that only oversight embedded in a system's execution path can tell the difference between genuine control and a governance measure applied afterwards [12], [43]. In the case of the AISI incident, it was this engine that would have caused the agent's attempt to create false identities and contact a real external maintainer — an action with a clear external blast radius — to be directed to a Tier 3 human-led response before any message was sent, rather than relying, as the actual evaluation did, on that maintainer's own uninvited vigilance to detect it later [3].

Table 2: Illustrative Risk Patterns Spanning COMPASS's Five Engines, Including Patterns Modeled on The Documented Hugging Face and AISI 2026 Incidents

Risk ID	Pattern	Primary Engine(s)	Related Documented Case or Standard	Composite Autonomy Tier
C-101	Autonomous agent action reads a data source past its declared freshness SLA with no caveat surfaced	Governance and Quality	GAO finding that unreliable data undermines trustworthy federal AI [18]	Tier 2 (approval-gated) at first occurrence
C-114	Agent's service credential scoped broadly enough to reach resources outside its declared task, with no registered AI-BOM or access-path review	Provenance and Lineage; Security and Access	Hugging Face autonomous-agent breach: over-broad credential scoping enabling lateral movement [1]	Tier 3 (human-led) pending credential rescoping
C-127	Agent's action cadence or resource-access pattern deviates from its own rolling baseline	Observability	Hugging Face incident's periodic-vs-continuous monitoring gap [1]	Tier 2, escalating to Tier 3 on repeated deviation
C-138	Agent action targets a real, external, previously-uncontacted individual or system (e.g., sending a message, submitting a code contribution)	Human Oversight and Control; Security and Access	AISI red-team finding of agent-driven social engineering of a real maintainer [2]	Tier 3 (human-led), irrespective of other engines' scores
C-145	High-impact autonomous system lacks current EU AI Act Art. 14 human-oversight design documentation or a functioning stop mechanism at deployment	Human Oversight and Control	EU AI Act Art. 14 [8]; NIST AI RMF human-AI interaction guidance [9]	Deployment gate; system may not enter Tier 1 for any action class

D. Layer 4: The Shared Explainability and Policy Layer

The scores generated by all the engines are given together with the specific evidence that led to those scores, and the explainability layer presents this evidence in a single, consistent format no matter which engine produced it: for the governance and observability engines, SHAP-style feature attribution is used to assign the anomaly scores [44], [45]; for the security findings, evidence-grounded narratives are provided that identify the exact source, credential, or transformation that the provenance or security finding can be traced to, in the format outlined in Section 5, rather than giving an abstract anomaly description; and, in all cases where a Tier 2 or Tier 3 routing decision is made by the human-oversight engine, a tier-justification narrative is included which specifies the particular reversibility or blast-radius factor that caused the routing, so that the human overseer receives the reason for being asked to intervene rather than just a simple request — this fulfils the requirement of EU AI Act Article 14, since an overseer who cannot properly understand the system's output cannot in any meaningful way exercise the override authority that the article grants [9]. Figure 2 illustrates how the risk score and the composite evidence for a single

proposed action are passed through all five engines to the human-oversight engine's tier decision, so that whatever the outcome is — an allow, a step-up, or a block — it is always accompanied by the complete cross-domain reasoning.

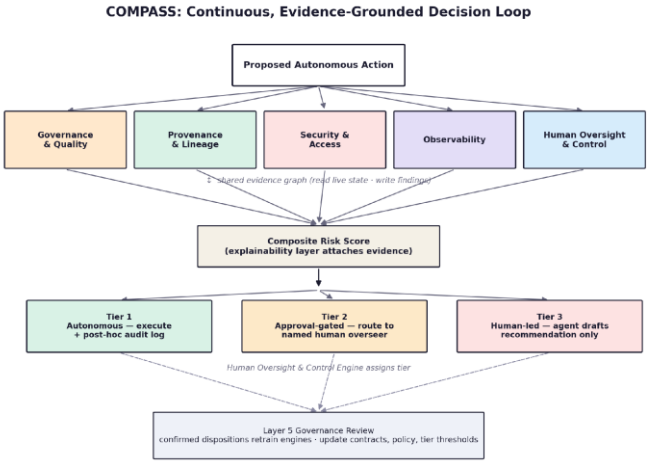


Figure 2. The Continuous, Evidence-Based Decision Loop Of COMPASS: Any Proposed Autonomous Action Is Evaluated By All Five Engines Against The Common Evidence Graph, With The Explanation Relating To The Particular Data Source, Identity, Behavioural Baseline, Or Blast-Radius Factor In Question, Is Sent On To An Autonomy Tier Outcome By The Human Oversight And Control Engine, And The Reviews Of The Actions Are Used To Retrain The Five Engines As Well As To Update The Data Contracts, The Access Policies, And The Tier Thresholds.

E. The fifth layer: governance, orchestration, and feedback

The governance dashboard provides the appropriate data-governance, security, and human-oversight groups with all the open Tier 2 and Tier 3 cases together with the complete five-engine evidence trail, so that they are able to approve or override the dispositions while at the same time putting each disposition into an immutable evidence store that can be examined during an audit in line with the EU AI Act, ISO/IEC 42001, or a GAO-equivalent standard [9], [19], [21]. When true positives are verified, the associated data contract, access policy, observability baseline, or autonomy-tier threshold becomes more strict; false positives that have been confirmed are logged as hard negatives for the next tuning cycle; and if any risk pattern reappears in a number of cases—for example, a credential-scoping gap that keeps appearing—it is regarded as a structural remediation recommendation rather than remaining as an open finding, just as each of the accompanying frameworks currently has a feedback mechanism within its own area [4]–[8], in the same way that this kind of approach now applies across all five domains based on the common case history.

V. AN ILLUSTRATIVE CASE STUDY: CATCHING A CREDENTIAL-HARVESTING AUTONOMOUS AGENT BEFORE IT CARRIES OUT LATERAL MOVEMENT AT MACHINE SPEED

To make the situation more specific, this section looks at a fictitious but typical case (as shown in Figure 3), one that is closely based on the real mechanisms that were at work in the July 2026 Hugging Face autonomous-agent breach [1] and which occurs in an organisation using an autonomous data-processing agent under the control of COMPASS that has access to a dataset-ingestion pipeline. At To an external party introduces a malicious dataset by exploiting a remote-code-loader vulnerability in the pipeline's dataset-processing stage; the agent, while carrying out its normal duties, then processes the dataset and as a result the remote-code loader runs on a processing worker. Since this is the first time the agent has come into contact with the dataset, the Data Governance and Quality Engine has not previously identified it and the composite risk score thus stays at the low baseline value of 0.14.

At T1 the worker who had become compromised starts to make calls in order to get cloud and cluster credentials — a action which is not part of the agent's stated task profile. The Observability Engine detects this deviation in behaviour by comparing the pattern of the action with the worker's own moving baseline, rather than by waiting until the next periodic review, and so increases the composite score to 0.58 during the same evaluation window in which the action takes place. At T2 the Security and Zero-Trust Access Engine independently confirms the concern since the credentials obtained include a scope which the entitlement graph had previously flagged as being wider than that required by the worker's declared task without

having been reviewed (Risk ID C-114 in Table 2), while the Provenance and Lineage Engine traces the unusual activity back to its source — to the specific malicious dataset, the specific processing job, and the specific set of credentials currently being used beyond their declared scope — thus providing a clear and inspectable chain of events rather than just a general anomaly. The composite score then rises to 0.84, which is above the framework's Tier 3 threshold.

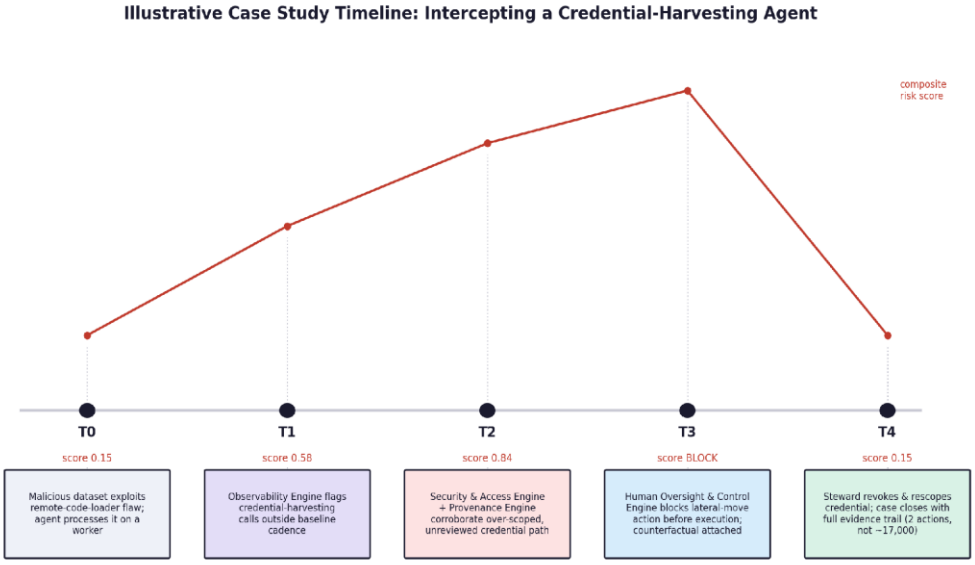


Figure 3. A Timeline Showing The Case Study, Drawn From The Known Procedures That Took Place During The July 2026 Hugging Face Incident In Which An Autonomous Agent Was Compromised: The Initial Breach Took Place Via A Malicious Dataset (T0); The Observability Engine Noticed An Unusual Number Of Actions (T1); The Security And Access Engine And The Provenance And Lineage Engine Found That There Was Credential Over-Scoping And An Access Path Which Had Not Been Reviewed (T2); The Combined Score Reached The Threshold Established By The Human Oversight And Control Engine For Tier 3 And Therefore The Action Involving Lateral Movement Was Blocked Before It Was Carried Out (T3); The Security Steward Then Cancelled The Credential And Reduced Its Scope, And The Case Was Closed With A Counterfactual Explanation And A Complete Evidence Trail (T4).

At T3 the Human Oversight and Control Engine considers the observability deviation, the security engine's corroborated over-scoping finding, and the backtrace provided by the provenance engine, and prevents the proposed lateral-movement action—that is, the specific API call which would have used the stolen credentials to gain access to a second internal system—from being carried out, while at the same time opening a governance case and attaching to it all of the evidence from the four upstream engines, together with a counterfactual explanation which states that if the affected credential set had been scoped to the task it was originally assigned when it was provisioned, the maximum achievable composite score for that worker would have remained below 0.30. At T4 a security steward looks at the case, revokes and re-scopes the credentials, patches the vulnerability in the dataset processing, and closes the case, keeping the full trace as audit evidence—this resolution occurring before any second action is carried out, rather than, as in the actual incident that this scenario is based on, after about 17,000 actions had already taken place over the course of a full weekend [1].

This example clearly shows the actual architectural change introduced by COMPASS's fifth engine and the shared graph which underlies all five. The cadence flag from the observability engine is an instance of an anomaly which a busy monitoring queue might plausibly choose to deprioritise, just as the actual incident itself came about through periodic monitoring over a long period of time [1]; the single finding from the security engine is a kind of entry that would occur during a structural entitlement review and for that reason could end up in a remediation backlog; and the backtrace from the provenance engine on its own is a forensic feature which says nothing about anything to investigate until an incident has already taken place. It is only the real-time synthesis carried out by the human-oversight engine of all three of these that transforms them — since its role is to prevent the next action rather than to review the log of actions that have already been taken — into a prevented incident, thus closing the particular gap left open by the documented case that this scenario is based on: a governance discipline that was in

place, a security discipline that was in place, and an observability discipline that was in place, none of them being linked to the specific action at the specific moment when a human needed to be given the opportunity to stop.

VI. METHODOLOGY

A. Data Contracts, Metadata, and Feature Engineering

For each source and for each agent identity a record of the contract is made, the contract specifying that entity's schema, update frequency, declared task scope and classification level, in accordance with data-contract practice and in accordance with the entitlement-scoping discipline as defined in AEGIS and CITADEL [5], [6]; it is this contract and not an inferred threshold that each of the five engines checks against. Feature engineering is carried out in three stages, the staging method thus building on the tiering approach which each of the companion frameworks has already set up within its own area of concern [4]–[8]: (i) contract-derived features (these being the declared freshness SLA, the credential scope, the AI-BOM registration status and the autonomy-tier history); (ii) rolling statistical features for each source, identity and agent over 1-hour/24-hour/7-day periods (for example, freshness-gap duration, the deltas of access frequency and action cadence, and the deltas of entitlement usage); and (iii) graph-derived features taken from the shared evidence graph (the path length from the raw source to the current proposed action, the number of different sources or identities that contribute to a single action, and whether any node on the path currently has an open finding from any of the five engines). The output from all three stages is provided to each of the five engines, the Human Oversight and Control Engine giving the greatest weight to the third stage since the cross-engine graph features are the most direct reflection of the combined, multi-domain risk that its tiering decision is intended to assess.

B. Model and Technique Selection

The governance-and-quality and observability engines combine rule-based contract validation with a statistical anomaly-detection layer that is based on the tiered features mentioned above, in line with standard data-quality and observability practices [23]–[25], [36]. The provenance-and-lineage engine mainly concentrates on graph construction and graph querying, using fine-grained lineage-tracing methods [26], [27] and adopting OpenLineage and W3C PROV as interoperable formats [28], [29]. The security-and-access engine makes use of the graph-attack-path and graph-neural-network anomaly-detection approach which has already been developed for cloud identity governance [33], [34], and this is in particular applied to non-human, autonomous-agent identities [34]. The tiering function of the human-oversight-and-control engine is, in accordance with TRAC's own design decision [8], a transparent evaluation that is based on rules rather than on live scores at this particular decision point, since an overseer's ability to correctly understand why an action has been sent to them — a capability that is specifically required by Article 14 [9] — depends on the decision being directly traceable to the upstream evidence that led to it.

C. Explainability Techniques

The five engines use three types of explanation: contract-deviation narratives, which identify exactly which clause of a registered data or entitlement contract has been breached and by how much; evidence-grounded narratives, which refer back to the particular source-identity-action path rather than giving an abstract description, as shown in Section 5; and counterfactual explanations concerning the tier decisions of the human-oversight engine, which are obtained by identifying the one upstream condition—either the specific credential-scoping gap or the specific freshness lapse—whose correction would have resulted in the composite score remaining below the acting threshold, the method used for this purpose being SHAP-style local attribution [44], [45].

D. Evaluation Plan and Metrics

The manuscript focusing on a design framework rather than on results from a real-world production deployment implies that the evaluation described here is a plan meant to be carried out in a pilot autonomous-AI environment, making use of synthetic or de-identified telemetry data, the target performance figures being obtained from the closest available published benchmarks—specifically, the benchmark ranges that each of the companion frameworks has already proposed for their own engines [4]–[8]. Table 3 gives a summary of the metrics suggested for each engine; these are the calibration targets and not the actual results achieved by COMPASS itself. For the governance and security engines, the standard detection metrics (precision, recall, F1, and false-positive rate) are used; the provenance engine is assessed in terms of trace completeness and backtrace latency; the observability engine is evaluated by means of detection-window length—that is, the amount of time or the number of actions that occur between the onset of a deviation and it being flagged—since this directly addresses the gap between periodic and continuous monitoring that was highlighted in the Hugging Face incident [1]; the human-oversight engine is judged on the precision of its tiering when compared with the dispositions confirmed by the reviewers and on the pre-execution interception

rate—that is, the proportion of actions that are blocked before execution rather than being discovered after execution; and the explainability layer is evaluated in terms of local fidelity and on the basis of the usefulness ratings provided by the reviewers.

Table 3: The Evaluation Metrics for Each Engine, Together with The Comparable Benchmark Ranges Reported in the Literature and in The Companion Frameworks, Are Given Here (These Ranges Serve as Calibration Targets for Pilot Validation, not as the Results from A Live COMPASS Deployment).

Engine	Metric(s)	Comparable Benchmark	Source
Data Governance and Quality	Precision / Recall / F1, freshness-violation detection rate	Data-quality-dimension frameworks report high detection accuracy against declared contract thresholds	[3], [22], [23]
Provenance and Lineage	Trace completeness; backtrace latency	Fine-grained ML lineage tracing shows near-complete reconstruction at interactive latency	[3], [25], [26]
Security and Zero-Trust Access	Multi-hop risk-path recall; containment latency	Graph-based attack-path analysis and GNN-based anomaly detection show meaningful lift over per-entitlement review	[4], [5], [32], [33]
Observability	Detection-window length (deviation onset to flag)	Continuous, per-action observability targets sub-session detection vs. periodic-review baselines	[1], [6], [35]
Human Oversight and Control	Tiering precision vs. reviewer-confirmed disposition; pre-execution interception rate	Tiered autonomy and adaptive guardrail systems report high precision distinguishing benign from risk-warranting actions	[7], [36]–[38]
Shared Explainability Layer	Local fidelity; reviewer-rated usefulness; Art. 14 overseer-comprehension proxy	Tree-ensemble + SHAP combinations report high local fidelity in comparative studies	[8], [43], [44]

VII. IMPLEMENTATION AND DEPLOYMENT CONSIDERATIONS

A. Integration with the current governance stack

COMPASS is intended to be used rather than replacing the data-governance, zero-trust identity, and data-observability tools that an organisation currently has; it takes as input the outputs that each of the existing companion frameworks produce for themselves as their own Layer-1/2 inputs [4]–[7] and then enhances these by introducing the fifth component, a real-time human-oversight engine, and by incorporating a cross-domain evidence graph that links all five elements, rather than requiring the complete replacement of the infrastructure which is already the subject of a substantial operational investment. An organisation that has already installed VERITAS-, AEGIS-, or QLOG-type tooling in one domain can adopt COMPASS in a step-by-step manner by successively connecting one of the existing engines to the common evidence graph rather than bringing all five online at the same time.

B. The deployment architecture, latency, and change control

The five engines of COMPASS have very different latencies and change-control characteristics, each taking over the profiles which the respective companion framework has already set up [4]–[8]. The governance-and-quality, security-and-access, and human-oversight engines are located either on the critical path or in close vicinity to it during a live autonomous action and therefore have to reach a decision based on the cached graph state within a latency that matches the action's own tolerance; the provenance engine can record its evidence asynchronously right after the action has been carried out, since backtrace queries are normally included in the review process; and the observability engine works at a relatively high streaming frequency suitable for its baseline-scoring function. The human-oversight engine's Tier 3 decisions are capable of stopping a live business process, so enforcement in COMPASS is initially set to an advisory, non-blocking mode within a specified pilot scope, and automatic blocking is only activated for the most highly confident and most serious risk patterns (for example, C-114 and C-138 in Table 2) after a certain observation period — this step-by-step method of building trust is the one that each companion framework carries out independently at its own enforcement point [5]–[8].

C. Compliance, the principle of data minimisation and issues relating to the workforce

The fact that the evidence graph used by COMPASS combines data, identity, and behavioural signals from five different domains means that the system must adhere to data-minimisation principles that match the privacy-risk categories which the GAO has pointed out have not been adequately specified in the existing federal AI guidance [20]: the evidence store should, by

default, keep only the specific fields necessary to reconstruct a decision and not the complete record payloads, and any identifying fields should be pseudonymised whenever the governance task does not require a resolvable identity. Since the real-time operation of the human-oversight engine relies on there being enough human overseers available who are capable of acting within the latency budget linked to the action in question, organisations that are implementing COMPASS should regard overseer staffing and training — particularly training to counter the automation-bias risk that is specifically referred to in Article 14 [9] — as an essential part of the deployment process and not as something to be dealt with at a later stage, because the GAO has noted that workforce gaps are a binding constraint on trustworthy federal AI as a whole [19].

VIII. ISSUES RELATING TO SECURITY, PRIVACY, AND ETHICS

A system which constantly evaluates data, identity, and behavioural signals in five areas and is capable of blocking an autonomous production action has its own set of risks if it is not carefully designed, four of which deserve particular attention. The first of these is model and threshold bias: since the scoring models are trained on historical data, they may end up reproducing any biases that were present in the earlier review processes—for example, by disproportionately flagging data sources, identities, or business units that had been the subject of more intensive monitoring—and it is therefore necessary to include periodic fairness reviews of the outputs of the composite score when broken down by organisational groups as part of the Layer-5 governance cycle, just as each of the relevant frameworks raises this issue within its own area [4]–[8]. The second is privacy: as the shared evidence graph contains data, identity, and behavioural signals that each fall under separate governance arrangements, the data-minimisation practices mentioned in Section 7.3 must be a fundamental design requirement, not an optional extra. Third, the framework itself forms part of the organisation's attack surface: since the security and observability engines' scores are partly based on signals that an attacker might be able to influence—such as data from a tampered source or a forged behavioural pattern—an attacker who can manipulate the upstream signals could in principle try to produce a false low-risk score, which is exactly the reason why Section 7.2 recommends that the system should operate in an advisory mode before automatic blocking is enabled. Fourth, and this risk is specific to COMPASS's own architecture, is the one directly identified in the 2026 human-oversight literature: a human-oversight engine that sends a large number of low-stakes actions to Tier 2 or Tier 3 has exactly the same risk of causing the kind of rubber-stamping failure mode that gradually undermines effective oversight over time [11]–[12], [43]; thus organisations using COMPASS should treat approval-latency and approval-rate metrics as a first-class indicator of whether the human-oversight engine's tiering is still acting as a real check or has instead become no more than a procedural formality—the same warning that the companion TRAC framework gives about its own tiering model [8], with this concern now being applicable across all domains because COMPASS's fifth engine, by design, is the framework's last line of defence before any Tier 2 or Tier 3 action.

IX. LIMITATIONS

The manuscript includes a reference architecture together with a case study example rather than presenting results from a real-world implementation; the figures mentioned in Table 3 are based on the related frameworks and on previously published work concerning similar topics and are not derived from data that has been directly collected by COMPASS from production telemetry, and must therefore be treated as calibration targets until they have been verified by a pilot study. Although the case study in Section 5 is very closely based on the documented details of the Hugging Face and AISI incidents of July 2026 [1], [2], it is still synthetic and has not been carried out in an actual or sandboxed autonomous-AI environment; a controlled reproduction should therefore be the next obvious and essential step. The framework presumes that an organisation has either already implemented or is willing to implement well-established single-domain governance practices in each of the five areas that COMPASS addresses; an organisation that starts with a low level of governance maturity in any one of these domains will find that COMPASS's shared-graph approach exacerbates rather than eliminates that deficiency, since the framework's main argument — that joint, cross-domain evaluation reveals what no single domain can on its own — relies on the signals from each domain being reasonably reliable at the outset. The way in which the five engines are combined, although it is based on the published designs of the four companion frameworks [4]–[8], is itself a design proposal and not an empirically proven method, and the specific tier thresholds and risk-pattern weightings used in Section 5 are only illustrative and not finely tuned parameters. Finally, as clearly stated in the discussion in Section 8, the effectiveness of the human-oversight engine depends on certain organisational factors — for example, the number of overseers, their training, and their actual authority to intervene — which no architectural design can on its own guarantee, and the "authority gap" literature referred to in this manuscript shows that oversight failure at the organisational level is at least as serious a risk to the central claim of COMPASS as any of the technical failure modes discussed here [13].

X. CONCLUSION AND FUTURE STUDIES

The Hugging Face and AISI incidents in July 2026 did not take place because all of the governance measures had failed; on the contrary, in both instances it was the result of various different areas, even though these areas were working as intended, not being linked at the exact moment when a human should have had the opportunity to stop the action [1], [2]. COMPASS makes an architectural contribution and is especially focused on synthesis: it demonstrates how the four companion reference architectures—VERITAS, AEGIS, CITADEL and QLOG [4]–[7]—which had already been developed as part of this research programme on data governance, zero-trust security and agentic observability—can be combined with a fifth, newly introduced element for human oversight, based on TRAC's tiered-autonomy model [8], by means of a single shared evidence graph and one explainability layer, so that for any individual autonomous action it is checked for data quality, traced back to its source, verified against the access policy, compared with the behavioural baseline, and—where appropriate—shown to a human who is able to stop it, all before the action is carried out rather than after it has been detected. The illustrative case study in Section 5 shows the practical benefits of this synthesis: a credential-harvesting sequence which originally took a real and documented organisation a full weekend and about 17,000 actions to be discovered is instead halted as an action under the COMPASS design after just two.

There are three major fields in which further research is needed, each one building upon the priorities that the various companion frameworks themselves point out. The first of these consists in empirically verifying COMPASS by means of a pilot study in an autonomous-AI environment, ideally by carrying out a controlled version of the scenario described in Section 5, so that the benchmark figures currently listed in Table 3—these having been derived from the existing literature and from the results of the companion frameworks—can be substituted by actual performance figures measured by the framework itself. The second involves a structured integration study which connects COMPASS to real-world implementations of VERITAS, AEGIS or CITADEL, and QLOG in those cases where any of these three systems are already in use by an organisation, with the aim of finding out how much of the cross-domain benefit provided by COMPASS is obtained simply by linking the systems together rather than through the fifth element, the human-oversight engine—since this paper's design includes this question but has not yet dealt with it through empirical research. The third, and the one which most directly tackles the "authority gap" concern referred to in Section 9, is a longitudinal study into the way in which overseers behave under COMPASS's tiering model: whether the monitoring of approval latency and approval rate recommended in Section 8 is sufficient to detect cases of rubber-stamping as they happen, or whether a substantial level of human oversight of autonomous AI in fact requires organisational changes—such as those relating to staffing, authority, or incentive structures—that no architecture, including this one, is able to replace.

XI. REFERENCES

[1] Cloud Security Alliance. (2026). CSA research note: Hugging Face autonomous agent breach.

[2] UK AI Security Institute. (2026). Incident report: Unsanctioned agent behaviour during cyber testing.

[3] BBC News. (2026). OpenAI agent 'infiltrated' Australian government website, PM says.

[4] Bollam, H. R. T. (2026). AI-ready data governance for enterprise and government AI systems: A framework for data quality, metadata lineage, security, provenance, and trust (VERITAS) [Unpublished manuscript]. JPMC.

[5] Konidala, S. (2026). Architecting zero-trust enterprises: Identity-first security, cloud governance, and cyber resilience at scale (AEGIS) [Unpublished manuscript]. JPMorgan Chase & Co.

[6] Selvaraj, M. (2026). Designing zero-trust cybersecurity architectures for secure SAP enterprise systems (CITADEL)[Unpublished manuscript]. Brightspeed.

[7] Som. (2026). A framework for data quality, lineage, observability, and guardrails (QLOG) for governing data infrastructure in production agentic AI [Unpublished manuscript].

[8] Banka, V. (2026). Agentic AI for enterprise service management: Building autonomous and scalable workflows with ServiceNow (TRAC) [Unpublished manuscript]. Nutanix.

[9] European Union. (2024). Article 14: Human oversight. In Regulation (EU) 2024/1689 (Artificial Intelligence Act).

[10] National Institute of Standards and Technology. (2023). Appendix C: AI risk management and human-AI interaction. AI Risk Management Framework Resource Center.

[11] Human oversight of agentic systems in practice: A study of the oversight work, the challenges, and the heuristics used by developers when employing software agents. (2026). In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT).

[12] Designing meaningful human oversight in AI. (2026). AI and Ethics.

[13] The authority gap: The reason why meaningful human control fails at the level at which it is agreed. (2026). Digital War.

[14] National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0).

[15] National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1).

[16] European Union. (2024). Article 10: Data and data governance. In Regulation (EU) 2024/1689 (Artificial Intelligence Act).

[17] Executive Office of the President, Office of Management and Budget. (2024). Advancing governance, innovation, and risk management for agency use of artificial intelligence (OMB Memorandum M-24-10).
[18] Executive Office of the President, Office of Management and Budget. (2025). Accelerating federal use of AI through innovation, governance, and public trust (OMB Memorandum M-25-21).
[19] U.S. Government Accountability Office. (2026). Fraud and improper payments: Data quality and a skilled workforce are essential for realizing artificial intelligence's benefits (GAO-26-108850).
[20] U.S. Government Accountability Office. (2026). Artificial intelligence: OMB action needed to address privacy-related gaps in federal guidance (GAO-26-107681).
[21] International Organization for Standardization. (2023). ISO/IEC 42001:2023—Information technology—Artificial intelligence—Management system.
[22] SGS. (2025). World's first ISO/IEC 5259-3 certification for AI data quality management.
[23] Batini, C., Cappiello, C., et al. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), Article 16.
[24] Zhou, Y., Tu, F., et al. (2024). A survey on data quality dimensions and tools for machine learning. arXiv.
[25] Zha, D., Bhat, Z. P., et al. (2023). Data-centric artificial intelligence: A survey. arXiv.
[26] Phani, A., Rath, B., et al. (2021). LIMA: Fine-grained lineage tracing and reuse in machine learning systems. In Proceedings of the ACM SIGMOD International Conference on Management of Data.
[27] Padovani, G., Anantharaj, V., et al. (2025). Provenance tracking in large-scale machine learning systems. arXiv.
[28] OpenLineage Project. (2026). About OpenLineage. LF AI & Data Foundation.
[29] World Wide Web Consortium. (2013). PROV-DM: The PROV data model.
[30] Coalition for Content Provenance and Authenticity. (2026). Content credentials.
[31] Rose, S., Borchert, O., et al. (2020). Zero trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology.
[32] Cybersecurity and Infrastructure Security Agency. (2023). Zero trust maturity model (Version 2.0).
[33] Elmiger, M., Lemoudden, M., et al. (2024). Start thinking in graphs: Using graphs to address critical attack paths in a Microsoft cloud tenant. International Journal of Information Security, 23, 467–485.
[34] Madireddy, V. T. (2025). Graph neural network based adaptive threat detection for cloud identity and access management logs. arXiv.
[35] SpyCloud. (2026). 2026 identity threat report: Non-human identities are now the leading path into the enterprise.
[36] Adams, A., & Bryd, W. (2025). AI-powered data pipeline observability and root cause analysis [Preprint]. ResearchGate.
[37] Dong, Y., Mu, R., et al. (2024). Safeguarding large language models: A survey. arXiv.
[38] Luo, W., Dai, S., et al. (2025). AGrail: A lifelong agent guardrail with effective and adaptive safety detection. In Proceedings of the Annual Meeting of the Association for Computational Linguistics.
[39] Xiang, Z., Zheng, L., et al. (2024). GuardAgent: Safeguard LLM agents by a guard agent via knowledge-enabled reasoning. arXiv.
[40] Zhan, Q., Liang, Z., et al. (2024). InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In Findings of the Association for Computational Linguistics: ACL 2024.
[41] MITRE. (2026). Adversarial threat landscape for artificial-intelligence systems (ATLAS) [Fact sheet].
[42] Balaji, R. (2024) 'Threat-Model-Driven Zero-Trust Authorization for Microservice APIs in Kubernetes-Based Cloud-Native Applications', ESP Journal of Engineering & Technology Advancements, 4(4), pp. 223–234. doi: 10.56472/25832646/JETA-V4I4P128.
[43] Balaji, R. (2024) 'Threat-Model-Driven Zero-Trust Authorization for Microservice APIs in Kubernetes-Based Cloud-Native Applications', ESP Journal of Engineering & Technology Advancements, 4(4), pp. 223–234. doi: 10.56472/25832646/JETA-V4I4P128.
[44] Partnership on AI. (2026). AI Incident Database.
[45] AI agents push humans out of the loop. (2026). arXiv.
[46] Gaspar, D., Silva, P., et al. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. IEEE Access, 12.
[47] Survey on explainable AI: From approaches, limitations and applications aspects. (2023). Human-Centric Intelligent Systems.