International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

Architecting Zero-Trust Enterprises: Identity-First Security, Cloud Governance, and Cyber Resilience at Scale

Sairamesh Konidala[1]
[1]*Visvesvaraya Technological University, Bangalore, India*
Emails: *sairameshkoni@gmail.com*[1]

Abstract

What started as a slogan for perimeter security has now developed into an official principle: according to NIST Special Publication 800-207, zero trust is defined as a set of principles that eliminates the assumption of trust based on network location, and federal, financial-sector and international enterprises are now evaluated using maturity models which require that every request made by a person or a machine be authenticated, authorized and continuously verified, regardless of its origin. Most companies implement zero trust through a series of individual controls—such as an identity provider that enforces multi-factor authentication (MFA) when someone logs on and then treats that identity as trusted for hours after the login, a cloud security posture management (CSPM) tool that checks for misconfigurations every day as part of its scanning process, a network segmentation policy that remains static after it has been put in place, and a security operations centre which only becomes aware that a credential has been compromised after it has already been used to move around the network. No single control is either sufficient or necessary; none of the controls takes into account the risk over the entire journey from identity to resource; and no control is able to explain, as an auditor or regulator would expect, the reason for its decision to grant or deny access. As the name suggests, the AEGIS (Adaptive Enterprise Governance for Identity-first Security) reference architecture brings together the four previously separate zero-trust functions under a single explainability and policy layer and a single governance feedback loop: continuous identity and trust scoring, cloud security posture and entitlement governance, adaptive policy enforcement and micro-segmentation, and cyber-resilience and recovery orchestration. We outline the multi-layered architecture of AEGIS, the data sources and modelling for each of the engines, give an example of how the framework can be used to stop a compromised non-human identity from traversing a hybrid-cloud environment before it reaches a regulated data store, and describe a governance model that offers an end-to-end policy and entitlement remediation lifecycle, receiving a "closed loop" of data that links detection, explanation and remediation. We also set out an assessment plan that is based on similar benchmarks from the zero-trust, identity-security and explainable-AI communities and examine the challenges involved in deploying, regulating and ethically using such a system within a globally regulated financial institution. Just as with other design-framework presentations in this area, AEGIS is not a description of a specific implementation of the design framework, but rather a reference architecture and a practical guide for practitioners, with the recognition that it is meant to serve as a foundation for future empirical validation.

Keywords: *Zero Trust architecture, Identity first security, Identity threat detection and response (ITDR), Cloud security posture management (CSPM), Cloud infrastructure entitlement management (CIEM), Micro-segmentation, Explainable AI (XAI), Cyber resilience engineering, Non-human identity governance, Security governance at scale.*

1. Introduction

For most large businesses the enterprise perimeter—on which traditional security architecture was based—can no longer be regarded as a useful perimeter. Workers' identities now authenticate from home networks and airport lounges, workloads run simultaneously in multiple public clouds and in an on-premises data centre, and an increasing amount of the authenticated and privileged activity in production systems is carried out not by a person but by a service account, an API key, a CI/CD pipeline credential or an autonomous agent on behalf of an organisation. The principles set out in NIST Special Publication 800-207 to meet this change are known as zero trust architecture (ZTA): trust should never be assumed on the basis of network location; every access request must be authenticated and authorised taking into account all available context; and access is granted on

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

a per-session basis with the minimum privileges necessary. [1] The Zero Trust Maturity Model puts these principles into practice by identifying five pillars—identity, devices, networks, applications and workloads, and data—each of which is assessed along a scale from traditional to optimal [2], and the U.S. federal government's zero-trust strategy, initiated by Presidential Executive Order 14028 and detailed in OMB Memorandum M-22-09, has made the implementation of Zero Trust a requirement for federal agencies, not just a best practice [3]. A comparable, though less top-down, pressure exists among financial institutions: the European Union (EU) Digital Operational Resilience Act (DORA) lays down direct regulatory requirements relating to ICT risk management, incident reporting and operational resilience testing for banks and for critical technology suppliers, and similar supervisory expectations are in place in most jurisdictions where a global bank operates [4]. The principle together with the way most companies act on it is too broad. It is usually put into place through a series of individual tools—for instance, a strong-authentication identity provider (IDP) which carries out identity verification at the start of each session and then assumes that the credentials so obtained will remain valid for their entire lifetime, or a cloud posture management (CPM) or cloud infrastructure entitlement management (CIEM) tool that periodically checks for misconfigurations and excessive entitlements rather than doing so continuously, or a network segmentation policy that is seldom updated due to changes in workloads and dependencies, or finally a security operations centre (SOC) that does not detect a compromised credential until a subsequent activity triggers a correlation rule in the security information and event management (SIEM) system, by which time lateral movement has already occurred. All of these are examples of locally defensible controls. Individually or collectively, none of them carries out the function of continuously assessing risk throughout the whole course of an identity's authentication event right up to the resource it ultimately accesses, and none can provide a human-readable, audit-defensible explanation when asked why an access decision was made or why a control failed to be triggered when it should have been. This situation is made worse by the fact that the makeup of the identity population has changed. In the current cloud environment, non-human identities—such as service accounts, API keys, OAuth application grants and workload identities assigned to containers and serverless functions—create significant opportunities for credential abuse and lateral movement; MITRE ATT&CK documents adversary use of valid accounts, including cloud accounts [5]. Since these identities are created in a programmatic way, they are changed only rarely and are considerably overprivileged, and the identity-first approach of zero trust—the idea that identities are now the new perimeter—extends well beyond workforce identity and access management (IAM) to become a governance approach for machine identity that many enterprises still lack. There has been recent research focusing on XAI concerning identity- and risk-security [6][7], and this offers the possibility of continuous, adaptive risk scoring which is based on static rules and can be defended to both the human reviewer and an external examiner. The absence, however, is a comprehensive architecture that integrates identity risk scoring, cloud entitlement governance, network-layer enforcement, and resilience/recovery orchestration into one system with a single explainable alert stream, rather than having four separate products each with its own alert stream. The manuscript makes three contributions. The first is the proposal of AEGIS (Adaptive Enterprise Governance for Identity-first Security), a reference architecture which integrates all four zero-trust functions within a single explainability and policy layer and a single governance feedback loop. The second contribution consists of the identification of the data sources, the methods of feature engineering, the choice of modelling techniques, and the explanation approaches applicable to each engine, all of which are supported by the zero-trust standards literature as well as by other published work in the areas of intrusion detection, cloud attack-graph analysis and cyber resilience engineering. The third contribution is achieved through the use of an illustrative case study in which a compromised non-human entity attempts to move from a low-sensitivity development environment to a regulated production data store, and in which the four engines work together to reconstruct the attack path, predict the next step of the attack, and produce the audit narrative that a governance review would see. The manuscript concludes with some reflections on the deployment architecture, the evaluation approach, and the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

regulatory, privacy and ethical boundaries within which this system needs to be protected, since it is so closely linked to the controls of a global financial institution.

2. Related Work

2.1. Zero Trust Architecture standards and maturity models

NIST SP 800-207 remains the standard definition of zero trust architecture and describes a logical architecture consisting of policy decision points and policy enforcement points that evaluate each access request using dynamic rather than static trust in network locations [1]. CISA's Zero Trust Maturity Model adopts this logical approach and divides it into five pillars that cover a spectrum of maturity going from traditional through initial to advanced and then to optimal, and it has become the industry standard for comparing and measuring the degree of zero trust maturity in both federal agency and private-sector security programs [2]. The existing literature on zero-trust deployments in different fields shows a limited number of zero-trust implementations in these various areas, the reason being that they are generally in the initial to advanced stages, specifically because there is no comprehensive view of identity and network as a surface for continuously assessing and controlling risks [8][9].

2.2. Identity-first security and non-human identity governance

More and more projects are considering identity to be the true zero-trust control plane, not the network. Identity threat detection and response (ITDR) addresses risks that point-in-time authentication alone may fail to surface; NIST zero-trust guidance instead calls for dynamic evaluation of access context and continued monitoring [10]. An examination of continuous and behavioural authentication demonstrates how biometric and behavioural signal streams can be used to establish a trust score during the entire duration of a session, not just at the one point in time of the initial authentication, and it has now become technically feasible to use this continuous scoring in real-world applications [11]. Non-human identities (such as API keys, service accounts, OAuth grants, and workload identities) also require access governance: MITRE ATT&CK documents misuse of valid and cloud accounts, without establishing that non-human identities exceed human credentials as an initial-access vector [5]. Although the literature supports the idea that identity is the most important element in zero-trust maturity models, it fails to connect identity risk to the cloud entitlement and network-enforcement layers on which a compromised identity ultimately depends to cause damage.

2.3. Governance of cloud security posture and entitlements

Cloud security posture management (CSPM) and cloud infrastructure entitlement management (CIEM) address cloud configuration and access risk. For example, AWS IAM Access Analyzer identifies externally accessible, internally accessible, and unused permissions and can evaluate complex access paths [12][13]. Analysis based on graphs of cloud identity and access management (IAM) relationships has shown that 'low-privilege' identities can be part of chains which lead to 'high-value' identities and resources several hops removed, for instance via a sequence of role-assumption and trust-policy relationships, and that traversing the IAM relationship graph is able to reveal these paths far more effectively than examining table-based IAM entitlements individually. More recently, machine learning has been applied to cloud IAM activity logs by using graph neural networks in order to detect adaptive threats that overcome rule-based anomaly detection [14]. Although this body of literature is good at spotting entitlement risks and attack paths in the cloud control plane, like the identity-security literature mentioned above, it is generally confined to a single pillar and not connected to identity-session risk or network-enforcement state.

2.4. Explainable AI

Explainable AI has been found to be both useful and effective in the area of intrusion detection, and more recently also in connection with decision points associated with Zero Trust. Many studies focus on the increasing integration of AI into zero-trust technologies; one of the most frequently cited requirements for trusting these technologies is explainability [6]. Techniques based on feature attribution, such as those involving deep learning and the use of explainable AI for zero-trust access control in unmanned aerial systems, demonstrate that feature attribution can be applied directly to the zero-trust policy decision point, not merely as an afterthought [7] in the case of narrow, single-domain deployments. The present manuscript applies the taxonomy from

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

general surveys of explainability techniques (namely feature-attribution, counterfactual, and example-based methods) to assign an explanation method to each of the AEGIS engines [15][16]. As shown in Table 1 Comparative summary of related work with regard to the capabilities integrated by AEGIS.

Table 1 Comparative summary of related work with regard to the capabilities integrated by AEGIS

Related Work Area	Focus	Technique	Explainability	Gap AEGIS Addresses
NIST SP 800-207 / CISA ZTMM [1],[2]	Zero-trust principles and maturity pillars	Policy-based access evaluation	N/A (standards)	Defines what to evaluate, not how to score or explain continuously
ZTA systematic reviews [8],[9]	Cross-domain ZTA implementation status	Survey / maturity assessment	Low	Confirms pillars remain siloed in practice; no integration model offered
ITDR / continuous authentication [10],[11]	Post-authentication identity risk	Behavioral biometrics, session analytics	Limited	Identity-only; not linked to cloud entitlement or network enforcement
CSPM / CIEM & cloud attack-graph research [12],[13],[14]	Cloud misconfiguration and entitlement risk	Graph traversal, GNN anomaly detection	Low–Moderate	Cloud-control-plane only; not linked to identity-session risk
XAI-in-ZT / XAI-IDS literature [6],[7],[15],[16]	Explainable ML for access control and intrusion detection	SHAP / LIME / counterfactual	High	Narrow single-domain applications; no cross-pillar shared explanation layer
NIST SP 800-160 Vol. 2 / DORA [20],[4]	Resilience engineering and regulatory obligations	Engineering techniques, compliance regime	N/A (standards)	Prescribes resilience goals; no automated linkage to identity/entitlement risk signals

2.5. Attack graphs, analytics for insider threats, and the reconstruction of attack chains

By using attack-path and attack-chain modelling, zero-trust architectures can analyse identity and entitlement risk in the form of a graph of findings rather than as a simple list of findings. There are instances in which a graph-theoretic analysis of the identity relationships of a cloud tenant has been capable of identifying critical multi-hop attack paths that a single permission review cannot detect [12]. Moreover, research into user and entity behavior analytics (UEBA) for the purpose of detecting insider threats, which makes use of sequential models and graph features, has demonstrated that a combined sequential and graph-based method greatly improves insider-threat detection when compared to using either method on its own [17]. In Section 4.3.4, the technique vocabulary employed in this manuscript is

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 593
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

modified to correspond to the identity-and-cloud-specific attack stages that a zero-trust environment has to reason about [18], these stages being derived from the broader MITRE knowledge base and the surveys used to operationalise it within the context of detection engineering.

2.6. The context of cyber resilience engineering and regulation

Even though cyber resilience is not defined in NIST Special Publication 800-160 Volume 1 (SP 800-160, rev. 3), Volume 2 makes it clear that cyber resilience is not the same as cyber security, but instead represents a separate approach, being defined as "the ability to anticipate, withstand, recover from, and adapt to adverse conditions", and includes engineering techniques such as segmentation, diversity, non-persistence, and adaptive response, none of which are part of cyber security. It is now going beyond being just a best practice and has become a regulatory requirement: the EU's Digital Operational Resilience Act (DORA) mandates that financial institutions comply with standards relating to identity, network and workload controls that treat these as part of an overall resilience controls posture, not merely as compliance checkboxes [4]; similarly, the U.S. federal zero-trust strategy in OMB M-22-09 also stresses the way in which identity, network and workload controls should be viewed as constituting an overall resilience posture rather than as a compliance checklist [19]. However, the resilience-engineering or regulatory literature does not explain how the automated detection-and-response system should connect the actions it takes back to the entitlement and identity-risk signals that a zero-trust identity and cloud-governance layer has already calculated.

2.7. Research Gap

Table 1 below provides a summary of five capabilities that are relevant to designing a zero trust architecture at an enterprise level. It is not the case that any one line of work covers all of the following capabilities: (a) continuous, session-long risk scoring for both human and non-human identities, (b) governance of cloud entitlements and attack paths, (c) adaptive, risk-based enforcement of network and workload policies, (d) resilience-engineered detection and recovery orchestration, and (e) a common explanation layer that is audit-defensible and covers all four areas. The AEGIS that is proposed is not about introducing a new detection algorithm into any one of the four constituent areas but is instead intended to address the integration gap.

3. A problem statement together with a design motivation is given

The framework is based on three structural weaknesses. Firstly, although the zero-trust maturity models are correct in identifying identity, network, cloud, and data as the four key areas that need to be dealt with, most implementations deal with these one at a time—so that the risk engine of an identity provider does not know that the identity it has just authenticated is about to access a cloud entitlement chain which has already been marked as high-risk by another CIEM system, and a network segmentation policy also fails to become more strict as the risk level of the identity's session increases during the session. Secondly, the kinds of identities that a zero-trust architecture needs to manage have changed more rapidly than most governance programmes have been able to keep up with: non-human identities can retain standing privileges and introduce risks from compromised credentials, while continuous evaluation and limitation of access remain essential to identity-based security [5][10]. Thirdly, whenever machine learning has been applied to address one of those gaps, it is typically used as an unexplained anomaly score, which may be enough for a SOC analyst to deal with a general alert, but is not adequate for a decision that blocks a production transaction or revokes access for a service account to resources: such a decision must be justifiable to internal audit, to a banking regulator in light of the supervisory requirements under DORA, and finally to a board risk committee [4]. AEGIS has been designed to meet four requirements that arise directly out of those weaknesses: (i) identity risk—whether that is human or non-human—must be assessed continuously throughout the whole session, and not merely at the time of authentication; (ii) the four functional components (identity, cloud entitlement, network/workload enforcement, and resilience/recovery) have to use the same risk context and feature schema (they should not function as separate, independent tools); (iii) every action, whether it involves granting approval for authentication, blocking a transaction, or revoking an entitlement, must be accompanied by a human-readable explanation that is based on the same features employed in making the decision; and (iv)

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 593
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

the system must be closed-loop, with both accepted and rejected decisions being fed into the model retraining algorithm as well as into the entitlement and segmentation policies, so that governance helps to improve the access model rather than just responding to it after the fact.

4. The AEGIS Framework

4.1. The design principle and layered architecture

Figure 1 shows the AEGIS layers as comprising five parts: (1) the sources of identity, cloud, network and workload data and events; (2) the ingestion process, entity resolution and a common risk feature fabric; (3) four parallel intelligence engines; (4) a shared explainability and policy layer; and (5) a governance and human-in-the-loop review layer whose decisions are fed back to layers 1 to 3. Just as in other high-stakes automated-decision scenarios involving similar layered governance models, this separation enables each engine to make its own modelling decisions without influencing the others, although it does require all four engines to publish their risk scores and features into a single shared schema so that the explainability layer and the governance dashboards can access them consistently.

4.2. The Layer 1–2 section: the data sources, ingestion, and the shared risk feature fabric

Layer 1 makes use of six types of source material, namely the identity provider and directory, authentication events, session tokens, group and role membership, and most crucially, inventories of non-human identities and API keys; the cloud provider's identity and access management (IAM) control plane for each connected cloud account or subscription, covering the definitions of roles, policies, and trust relationships; network and workload telemetry such as the state of segmentation policies, east-west flow logs, and software-defined-perimeter (SDP) gateway logs; and security information and event management (SIEM) and endpoint detection platforms, which AEGIS regards as a source to enrich rather than as a system to replace. Change-management and infrastructure-as-code repositories are also a significant source since frequent changes are the main cause of entitlement and segmentation drift and not the result of actions carried out in the console. As shown in Figure 1 The AEGIS layered architecture.

The various entities from these diverse sources are combined into a single organizational identity graph — with both human and non-human identities, roles, cloud resources, network segments, and so on, acting as nodes; authentication, assumption, and connectivity events serving as edges — and a single rolling, per-identity feature store is kept (for example, typical session duration and reach of resources, typical entitlement usage, typical network paths, and how quickly privileges change, etc.), this feature store being accessed by all the downstream engines without them having to recalculate it themselves.

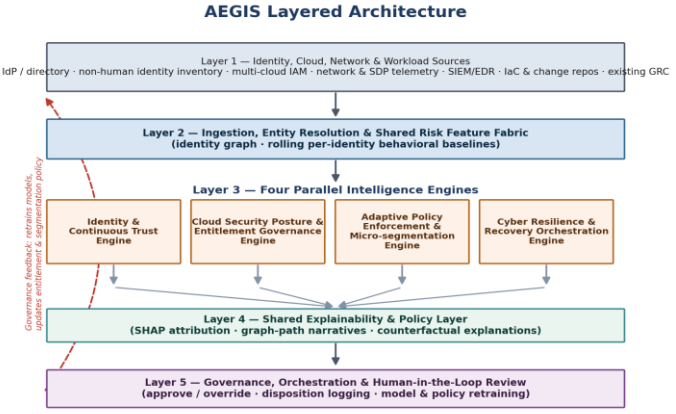


Figure 1 The AEGIS layered architecture

4.3. The Four Intelligence Engines

4.3.1. The Identity and Continuous Trust Engine

This engine is doing more than simply calculating a continuous trust score for each session — whether that session is initiated by a human or not — since it is also applying the continuous and behavioural authentication methods described in the literature [11] by including non-human identity behaviour such as the cadence of API calls, the geography of credential usage, and the rotation and age of the underlying secret among the human behavioural signals. It doesn't just keep a peer group and a self-baseline for each identity; it also detects any deviation as it happens — for instance, a new login location, an API key suddenly accessing a set of scopes which it doesn't usually access, or a service account logging in from a host that is not part of their normal deployment fleet — thus filling the gap in ITDR that is left by static, point-in-time login authentication. As shown in Table 2 An illustrative excerpt of the high-risk identity-cloud entitlement patterns used to initialise the Cloud Security Posture and Entitlement Governance Engine. The engine is not a lower-tier one because machine-identity risk is treated as a first-class output: compromised non-human credentials

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 593
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

can enable unauthorized access, consistent with the valid-account techniques documented by MITRE ATT&CK [5].

Table 2 An illustrative excerpt of the high-risk identity-cloud entitlement patterns used to initialise the Cloud Security Posture and Entitlement Governance Engine

Risk ID	Entitlement / Identity Pattern	Example Cloud Construct	Business Risk	Typical Mitigating Control
R-101	Standing admin-equivalent role held by a service account across non-production and production environments	IAM role / service principal with wildcard action scope	Cross-environment blast radius; prod compromise via a lower-trust dev pivot	Environment-scoped roles; just-in-time elevation
R-118	Long-lived API key with unrestricted object-store write access and no rotation policy	Access key + storage bucket policy	Credential theft leads to silent data exfiltration or tampering	Short-lived tokens; mandatory rotation SLA
R-204	Third-party SaaS OAuth application granted organization-wide read/write scope	OAuth application consent grant	Supply-chain compromise of the SaaS vendor cascades into the tenant	Scope allow-listing; periodic OAuth consent review
R-231	Federated human identity retaining entitlements unused for 90+ days	IAM policy attachment	Privilege creep; dormant excess access widens attack surface	Automated entitlement right-sizing; access recertification
R-309	Cross-account trust relationship without an external ID or session condition keys	IAM role trust policy	Confused-deputy attack; cross-tenant privilege escalation	Mandatory external ID and condition-key enforcement

4.3.2. Engine for cloud security posture and entitlement governance

This engine combines the capabilities of traditional CSPM/CIEM systems for detecting misconfigurations and excess entitlements [12] with continuous graph mining of the Layer-2 identity and resource graph in order to identify multi-hop entitlement paths that a review on a per-entitlement basis would miss: for example, a service account with a moderately sized role in a development account that, through a series of cross-account trust relationships, can eventually be given write access to a production financial ledger account. In Table 2, an example section of the risk matrix for this engine is shown prior to any attack-path evidence being added by the graph mining process. Each known or newly discovered entitlement risk is assigned a dynamic weight that takes into account the severity of the risk (namely, the length of the path to a designated 'crown jewel' resource in the cloud and the breadth of access at the other end of such a path) and the evidence regarding how the entitlement is being used obtained from the identity-and-trust engine, just as has been done in recent machine-learning research in the area of cloud IAM, where cloud IAM activity is scored for adaptive threats rather than simply depending on static rule matching [14].

4.3.3. Engine for adaptive policy enforcement and micro-segmentation

In the sense defined by NIST SP 800-207 [1], this engine constitutes AEGIS's risk score for policy enforcement at the point of decision regarding network and application access on a per-session basis, as opposed to simply having a rule that is set once and changed only infrequently, as explained in [21]. In

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

accordance with the software-defined-perimeter design [22], each resource is initially dark and an identity's network reachability either decreases or increases as the trust score and the entitlement-path risk vary during the session (a decrease meaning it loses reachability and an increase meaning it gains reachability) without having to wait until the next scheduled policy review. It is this engine that turns the outputs of the identity and cloud-governance engines into a real reduction of the blast radius rather than just generating an alert which requires a human to take action before any protective effect is achieved.

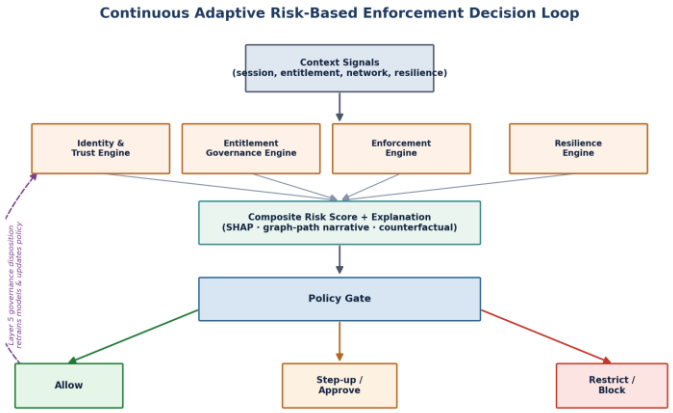


Figure 2 The continuous adaptive risk-based enforcement decision loop involves taking the context signals from all four engines, scoring them, providing explanations for them, routing them to one of three possible policy outcomes, and having the review decisions retrain the models while also updating the entitlement and segmentation policies

4.3.4. Cyber Resilience and Recovery Orchestration Engine (CRROE)

The fourth engine looks at a longer time horizon than session-by-session enforcement because it is able to detect sequences of events across different identities, cloud accounts, and network segments that are typical of the early phases of a lateral-movement or privilege-escalation chain; it can then predict the most probable next step in that chain; and, where allowed by policy, "pre-stage" recovery actions (such as isolating a resource, triggering an immutable checkpoint for a backup before the chain has been fully completed. This engine makes use of a sequence/graph model combined with the fusion of identity- and entitlement-risk signals, a concept that is similar to the graph-based attack-technique prediction [14][12] which is based on an identity-and-cloud-specific stage vocabulary (anomalous authentication, entitlement-path traversal, segmentation-boundary probe, unauthorized data access or exfiltration): Section 5 provides a full illustrative example of how this engine is used.

4.4. The shared explainability and policy layer

The scores from all engines are published together with the feature vector that produced them, and all post-hoc, model-agnostic types of explanations—such as those which attribute features by showing which ones increased or decreased the score and which had the greatest impact, or graph-path explanations that identify a specific path in the graph used to assess the trustworthiness of the identity or a particular trust relationship or number of hops—are applied in the same way, as are the counterfactual explanations, for example the statement that if X had not occurred the score would have been Y, in accordance with the practice of embedding feature attribution directly into a zero-trust policy decision point [7]. As shown in Figure 2 shows the way in which the risk score and its explanation are passed on to the policy enforcement engine, with the score and its explanation always being sent together and never alone, and with the explanation always accompanying the score.

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

A governance dashboard provides open cases to teams in SOC, identity-governance and cloud-security, provides a description of the disposition, allows for approve/override dispositions, and records each disposition to an immutable evidence store that aligns with the audit-trail requirements of a DORA-equivalent regulatory regime [4]. Confirmed true positives reinforce the corresponding model and policy weights, while confirmed false positives weaken the specific pattern and serve as a hard negative to the next training cycle – and any pattern of entitlement, or gap in the segmentation, that repeats across multiple cases is recommended as a role-redesign or network-architecture change, rather than a persistent alert, so as to make the framework increasingly effective at notifying fewer and fewer cases in the future, rather than building up an ever-larger static rule set.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

5. Intercepting a Compromised Non-Human Identity Before It Reaches a Regulated Data Store (IDS): A Case Study of the Problem

To make the framework real, this section will go through a synthetic but representative scenario (Figure 3) within a hybrid cloud estate of a global financial institution. A CI/CD pipeline's deployment API key (a non-human identity with a moderate role that is development scoped) is exfiltrated from a misconfigured build log and used by an external actor at T0. The identity and continuous trust engine detects the key authenticating from an unknown egress area and the number of API calls at a frequency significantly higher than a historical baseline, increasing the composite risk score from the base 0.18 to 0.54, just beyond the advisory threshold . At T1, the attack path is a theoretical one, but becomes an active attack path or traversal when the graph-mined path of three hops from the original development-scoped role to a production data-access role is detected, allowing the attacker to gain access via a cross-account trust policy, which the cloud security posture and entitlement governance engine already identified as an unreviewed structural risk (Risk ID R-309 in Table 2). The shared risk feature fabric enables the adaptive policy enforcement and micro-segmentation engine to read the identity engine's elevated trust score and the entitlement engine's active-path alert from the shared risk and immediately contracts the effective network reachability of the assumed role; the newly assumed production role is no longer able to reach the regulated data store's network segment, although the IAM policy still allows access to the data store, the IAM policy enforcement point evaluates living risk on top of static entitlement. When T2 tries to connect to the data store, the network layer refuses the connection, and the refusal is a high-confidence signal sent back to the shared feature fabric, bringing the composite score to 0.91. The sequence of events observed by the cyber resilience and recovery orchestration engine, T3, at T3 is: anomalous non-human authentication, role assumption across accounts, segmentation-boundary denial, which is expected to be followed by a second attempt to access the system within a few minutes, a pre-stage of an immutable snapshot of the affected production database before any write operation, and a governance case with full evidence chain attached. The case is closed with the corrected entitlement and full trace, recorded as audit evidence, leaving the maximum possible risk score below 0.35 if the R-309 trust-policy gap had been addressed when it was first identified.

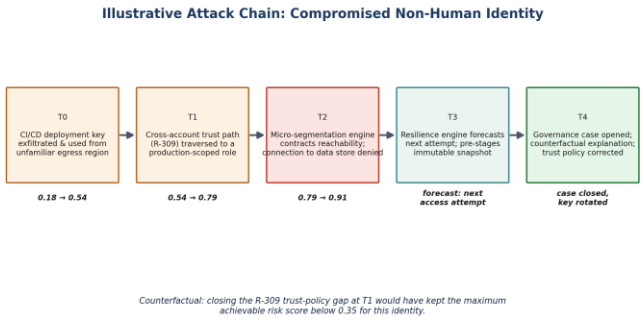


Figure 3 Illustrative attack chain: A compromised CI/CD service-account key traverses a cross-account trust path (stages T0–T2), is contained by adaptive micro-segmentation before reaching the regulated data store (T2), and the resilience engine pre-stages recovery actions while forecasting the blocked next stage (T3–T4)

This scenario highlights the need for the four engines to share context and not work individually. No one, a busy SOC queue in particular, would give it a second thought if an API-call cadence was an unfamiliar pattern on its own; if an unreviewed cross-account trust relationship was a structural finding that could be in a SOC remediation backlog for weeks; or if a single denied connection was network noise. Three signals with low urgency were not enough to be a contained incident, as they couldn't provide a complete explanation that was audit-ready, until combined through the shared risk feature fabric and explanation layer described in Section 4, which is held together by the enforcement engine's live contraction of network reachability, rather than an alert.

6. Methodology

6.1. Data Sources and Feature Engineering

Feature engineering is carried out in three phases: (i) static entity features (identity type, entity classification, role composition, entitlement scope, sensitivity tier); (ii) behavioral time-series features aggregated from the history of each identity, across rolling 1-hour, 24-hour and 7-day windows (API-call cadence, authentication geography, device diversity, rate of entitlement change, session duration relative to baseline); and (iii) relational/graph features derived

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

from the identity-and-resource graph (shortest path length to a designated crown-jewel resource, centrality of a role within the entitlement graph, count of distinct sensitive resources reachable, presence of an open finding anywhere on the current path). As in similar adaptive risk-governance systems in other high-stakes areas, all three tiers feed into each of the four engines and feature subsets are engineered at each tier, using the labeled or weak-labeled objective specific to that engine as a ground truth.

6.2. Model and Technique Selection

The identity and continuous trust engine is a gradient-boosted ensemble model, trained on historical incidents labeled with the identities, and an unsupervised behavioral-anomaly component scored against each identity's own baselines and compared to the baselines of the identity's peers, as per current continuous and behavioral authentication practice [11]. The cloud security posture and entitlement governance engine, as described above, is primarily a graph construction and graph traversal problem, with a graph neural network anomaly component for activity that would not be detected by a static path query [14, 12] as demonstrated in cloud IAM relationship graph (IRG) attack-path analysis. At the heart of the adaptive policy enforcement engine is a low latencies rules-over-live-scores evaluation function (not a learned classifier at the specific decision point, because the critical requirement of the enforcement point is that network or app access changes are immediately traceable up to the rules that caused them that are applied upstream). The orchestration engine for resilience and recovery is a stage-sequence graph/sequence-labeling problem, in the spirit of graph-based attack-technique prediction and attack-technique labeling/aligned detection engineering of MITRE ATT&CK [18] with the stage vocabulary adapted to its context of identity and cloud (Section 4.3.4).

Table 3 Evaluation metrics per engine and comparable benchmark ranges reported in the literature

Engine	Metric(s)	Comparable Benchmark in Literature	Source
Identity and Continuous Trust	Precision / Recall / F1, session-level false-positive rate	Behavioral/continuous authentication studies report high discriminative accuracy on standard user-session benchmarks	[11]
Cloud Security Posture and Entitlement Governance	Attack-path recall; reduction in analyst-confirmed false positives vs. static entitlement review	Graph-based analysis and GNN-based IAM studies motivate comparative evaluation against per-entitlement review; gains require pilot validation	[12],[14]
Adaptive Policy Enforcement and Micro-segmentation	Containment latency (score-threshold crossing to reachability contraction)	Evaluate actual policy propagation time against deployment-specific latency requirements; cited guidance does not establish a universal numerical benchmark	[21],[22]
Cyber Resilience and Recovery Orchestration	Top-k next-stage prediction accuracy; forecast lead time	Attack-technique and sequence/graph studies motivate next-stage prediction metrics; comparative gains need empirical validation	[17],[18]
Shared Explainability Layer	Local fidelity of attributions; reviewer-rated usefulness	Tree-ensemble + SHAP combinations report high local fidelity in comparative IDS and zero-trust studies	[6],[7],[15],[16]

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

6.3. Explainability Techniques

For the identity-trust and entitlement-governance engines, explanations for the scores are provided on a local level in a similar manner as other explainable AI applications to zero-trust and intrusion-detection decisions [6][7][15][16]; For the entitlement and attack-path findings, the graph-path explanations tie the findings to the specific role, trust relationship, or network hop that it traces to, as opposed to an anomaly description; and For the resilience engine's forward-looking outputs, counterfactual explanation is achieved by identifying the single highest-leveraging upstream condition that would have prevented the composite score from crossing the acting threshold, to provide explanation for the findings.

6.4. Evaluation Plan and Metrics

As this manuscript contains a design framework instead of results from a production deployment, evaluation is specified here as a plan for implementation of this design in a pilot hybrid-cloud environment (either synthetic or de-identified production telemetry) and target ranges are provided based on the closest comparable published benchmarks. The metrics per engine are summarized in Table 3, with ranges reported in the literature for similar methods supplied as calibration targets, rather than AEGIS measured results. The identity-trust and resilience engines are assessed using standard detection metrics ; the entitlement-governance engine is assessed on the attack-path recall against a labelled or synthetically injected set of attack paths, and on reduction in analyst-confirmed false positives relative to a static-rule-matrix baseline; the enforcement engine is assessed on containment latency ; and the explainability layer is assessed on local fidelity and on reviewer-rated explanation usefulness, collected through structured feedback in the Layer-5 dashboard. As shown in Table 3 Evaluation metrics per engine and comparable benchmark ranges reported in the literature.

7. Considerations for Implementation and Deployment

7.1. Connectivity to the Current Security and Identity Platform

AEGIS is intended to operate with the enterprise's existing IDP (identity provider), CSPM/CIEM (correlation or security policy manager/manager), network segmentation controls, and SIEM/SOAR (security information and event manager/security orchestration and response) platforms: It takes existing enterprise telemetry, policy repositories and findings as Layer-1 inputs and enhances them with the continuous scoring, cross-pillar correlation and explanation capabilities described above, and does not require a rip and replace of existing
security tooling. This is a considered integration posture, where the enterprise's SIEM and IDP already have operational trust and investment that's already in place, which will take years to re-establish if replaced wholesale.

7.2. Deployment Architecture, Latency Budget, and Regulatory Compliance

The four engines of AEGIS have very different latency budgets, that the architecture must explicitly support. The identity-trust and enforcement engines run in the critical path of a live session and must run in milliseconds before a live access decision is made which awaits an authenticated user or workload; the entitlement-governance engine's graph traversal can run on a short cadence (seconds to near real-time) because the structure of an attack-path doesn't change often compared to the access decision; and the resilience engine runs on a multi-event correlation window (seconds to minutes) that matches its forecasting job. This latency separation is kept explicit by deploying a service for the identity-trust and enforcement checks in the identity provider's and network fabric's request path as an identity-trust-and-enforcement policy-decision-point service, and having the service for the entitlement-governance and resilience engines alongside as another service reading from the same event bus. As the framework does process behavioural and HR-related data and related to customers, it is important to implement the principles of data minimisation as outlined by the applicable privacy regimes, and, increasingly, by regimes that prescribe ICT risk-management and audit-evidence requirements for financial institutions, such as DORA [4], including the specific fields necessary to reconstruct a decision, rather than record payloads in their entirety.

7.3. Scalability and Performance

The shared risk feature fabric in Layer 2 is the only place where the Layer-2 engine needs to hold a per-identity rolling state, and the four Layer-3 engines can be scaled and retrained separately from that shared store. This is no different to what is found in similar

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

"layered" risk-governance architectures for other high-stakes domains, where the "low-latency" request-path elements are kept thin, cache-driven, and the "heavy" graph/risk modeling elements are run asynchronously, allowing the system to scale to enterprise identity population of hundreds of thousands without the shared architecture becoming a performance bottle neck.

8. Security and Privacy, and Ethics

If constructed carefully, a system that continually scores identity behaviour and restricts network access to production financial systems poses a genuine threat of its own – and three are worth being stated outright. First of all, model bias – the bias in the historical incident labels can be baked into the model and will be amplified by the risk-scoring model, e.g., the tendency to flag identities of business units or geographies that have been historically over-monitored; the identity-trust engine's outputs should be periodically audited for fairness across business units or geographies, if they exist and it is lawful to do so, as part of the governance cycle in Layer 5, rather than as an afterthought. Second: Privacy by Design: any behavioral or HR related features must be trimmed to what each engine actually requires; must be kept only as long as audit and/or governance require; and access to the feature fabric must be more restricted than to the risk scores and explanations derived from those features. Third, the framework forms a part of the enterprise's attack surface: if an adversary is able to manipulate the signals upstream of the containment engine that feed into the identity or entitlement engines, then, in theory, they could take advantage of the ability to manipulate the containment engine to cause containment as a denial of service attack against the legitimate enterprise business activity, and so the integrity of the inputs into the containment engine should be treated with the same level of scrutiny as the integrity of the inputs into an autonomous system generally. While the framework is designed to be as automated as possible, without a doubt, the final protection will be human, with the explainability layer being present to ensure that human oversight isn't a blind eye to a black box score.

9. Limitations

This manuscript does not report results on a production deployment, but rather a reference architecture and an illustrative case study; the evaluation ranges in Table 3 are based on similar published work on continuous authentication, cloud attack-graph analysis and attack-technique prediction, and should be considered calibration targets until validated in pilot deployments. The design idea for the attack-stage specific vocabulary proposed in Section 4.3.4 would benefit from being validated against a labeled corpus of real zero-trust security incidents at the scale required for supervised learning, which is not to our knowledge, publicly available — this constraint is shared by most enterprise-security research and is why for this class of work, it is common to rely on synthetic or illustrative scenarios. It also assumes that an enterprise's identity provider, cloud accounts, and network fabric can all be instrumented into the shared risk feature fabric, but in reality there will be gaps, such as legacy systems, acquired subsidiaries that were not migrated to the cloud, and third-party managed environments, that the framework can only identify as incomplete visibility. Lastly, there is no straightforward way to objectively assess the quality of the explanation; the fidelity measures suggested in Section 6.4 measure the accuracy of the explanation relative to its model, but a longer-term study of the outcomes of the human reviewers is required to determine whether they make better decisions after the explanation.

Conclusions and future scope of the project

Now, zero trust is not just a new concept, but a well-established architectural principle, with NIST SP 800-207 and CISA's maturity model providing a clear definition of what enterprises are building. What's still very fuzzy and what this manuscript is trying to address, is how a business is going to integrate identity risk, cloud entitlement governance, network enforcement and resilience engineering into a single system that will assess and explain risk on a continuous basis—not four separate point tools on their own timelines. However, AEGIS' contribution to this is architectural, demonstrating the ability for continuous identity and trust scoring, cloud security posture and entitlement governance, adaptive policy enforcement and micro-segmentation, and cyber resilience and recovery orchestration to share one risk feature fabric and one explainability layer, delivering a single, audit-defensible risk narrative rather than four disparate alert streams to one governance team. It's real-world proof of that, from the single "denied

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

connection" that becomes a contained and explained incident, the unreviewed difference in "trust-policies" that becomes a contained and explained difference, and the stolen deployment key that becomes a contained and explained event when read and acted on in real time in the course of a single deployment, not after the fact. There are three priorities for future work. The first is empirical validation of the framework in a pilot hybrid-cloud environment with either injected synthetic attacks or de-identified historical attacks data, which will supplant the benchmark ranges from the literature used in Table 3 with the own measured performance of the framework. The second is creating an identity-and-cloud attack-stage taxonomy that is expanded from the vocabulary introduced in Section 4.3.4, and eventually to become a shared, community-supported taxonomy similar to MITRE ATT&CK, rather than specific to each deployment, so findings and models can be carried from one to another. The third is a longitudinal study of the impact of closing the feedback loop from the confirmed findings back to role and network-architecture redesign on the number of new attack-path findings and alert volume over multiple review cycles, which is what would be most directly tied to the value of the framework's increased complexity compared with today's siloed zero-trust tooling.

References

[1]. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology.

[2]. Cybersecurity and Infrastructure Security Agency. (2023). *Zero trust maturity model* (Version 2.0).

[3]. The White House. (2021). Executive Order 14028 of May 12, 2021: Improving the nation's cybersecurity. *Federal Register, 86*(93), 26633–26647.

[4]. European Insurance and Occupational Pensions Authority. (2026). *Digital Operational Resilience Act (DORA)*.

[5]. MITRE ATT&CK. (2026). *Valid accounts (T1078)*.

[6]. The significance of artificial intelligence in zero trust technologies: A comprehensive review. *(2024).* Journal of Electrical Systems and Information Technology.

[7]. Enhancing UAV security through zero trust architecture: An advanced deep learning and explainable AI analysis. *(2024). arXiv.*

[8]. Zero trust architecture: A systematic literature review. *(2025). arXiv.*

[9]. A systematic literature review on the implementation and challenges of zero trust architecture across domains. *(2025).* Sensors, *25(19), Article 6118.*

[10]. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (NIST Special Publication 800-207, §3). National Institute of Standards and Technology.

[11]. Sensor-based continuous authentication of smartphones' users using behavioral biometrics: A contemporary survey. (2020). IEEE Access.

[12]. Elmiger, M., Lemoudden, M., Pitropakis, N., & Buchanan, W. J. (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.

[13]. Amazon Web Services. (2026). IAM Access Analyzer findings. AWS Identity and Access Management User Guide.

[14]. Graph neural network based adaptive threat detection for cloud identity and access management logs. (2025). arXiv.

[15]. Survey on explainable AI: From approaches, limitations and applications aspects. (2023). Human-Centric Intelligent Systems.

[16]. Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. (2024). IEEE Access.

[17]. Attention-driven fusion of sequential and graph neural models for insider threat detection in UEBA. (2025).

[18]. MITRE ATT&CK: State of the art and way forward. (2024). ACM Computing Surveys.

[19]. Office of Management and Budget. (2022). Moving the U.S. government toward zero trust cybersecurity principles (Memorandum M-22-09).

[20]. National Institute of Standards and Technology. (2021). Developing cyber-resilient systems: A systems security

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5923 - 5936
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0762

engineering approach (NIST Special Publication 800-160, Vol. 2, Rev. 1).
[21]. Cybersecurity and Infrastructure Security Agency. (2025). The journey to zero trust microsegmentation.
[22]. Cloud Security Alliance. (2022). Software-defined perimeter (SDP) specification v2.0.
[23]. National Institute of Standards and Technology. (2026). Digital identity guidelines (NIST Special Publication 800-63-3).
[24]. National Institute of Standards and Technology. (2026). Guide to attribute based access control (ABAC) definition and considerations (NIST Special Publication 800-162).
[25]. Verizon. (2026). 2026 data breach investigations report.
[26]. Gartner. (2026). Continuous adaptive risk and trust assessment (CARTA).
[27]. Cloud Security Alliance. (2026). Zero trust architecture (SP 800-207).