Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Trust Boundaries in AI-Driven Systems: Implications for PKI and Internet-Scale Trust

Naresh Charugundla

Independent Researcher, USA

ARTICLE INFO

ABSTRACT

Public Key Infrastructure (PKI) has served as the foundational trust mechanism for internet-scale systems, enabling authenticated communication through cryptographically verifiable identities. Traditional PKI models assume authenticated identities represent actors whose behavioral scope is human-directed and operationally bounded. AI-driven autonomous systems challenge these assumptions as behavioral scope expands dynamically beyond the authority encoded in credentials. This article examines how autonomous systems reshape trust boundaries within PKI-based infrastructures, identifying a structural divergence between identity scope and behavioral scope. Three findings emerge: authenticated machine identities exhibit behavioral scope inflation that static credential governance cannot bound; delegated authority amplification produces aggregate operational scope exceeding individual credential assumptions; and machine-to-machine trust chains require audit infrastructure above individual certificate validation. PKI must therefore function as a foundational layer within a broader trust ecosystem incorporating behavioral policy enforcement, runtime authorization, and operational sequence logging.

Keywords: Public Key Infrastructure, Trust Boundaries, Autonomous Systems, Machine Identity, Internet-Scale Trust, Certificate Authority, Zero Trust Architecture

1. Introduction

Public Key Infrastructure (PKI) has underpinned the security of internet-scale communications for over two decades. By binding cryptographic keys to verifiable identities through the issuance of X.509 certificates, PKI enables distributed systems to authenticate one another, establish encrypted communication channels, and enforce trust boundaries across organizational and technical domains [1]. The integrity of this model rests on Certificate Authority (CA) hierarchies that issue credentials to known entities and on relying parties that validate those credentials against trusted roots during connection establishment. The result is a scalable, globally interoperable system in which identity verification is cryptographically deterministic and organizationally accountable.

For most of this history, the entities operating under PKI-issued identities have been either human users or automated services performing deterministic, human-configured tasks. Web servers authenticate to browsers, applications authenticate to backend services, and code-signing systems validate software provenance—all within tightly scoped operational roles defined by administrators who provision and govern the associated credentials. The behavioral scope of these identities has traditionally been narrow and predictable: a certificate issued to a web server is expected to serve content on a specific domain, not to initiate arbitrary interactions across infrastructure. This alignment between identity and expected behavior is what allowed trust boundaries to function not merely as technical controls but as representations of structured operational relationships. However, this model carries structural limitations that become consequential in non-deterministic environments: PKI

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

credentials encode no representation of behavioral scope independent of identity; revocation mechanisms presuppose that removing a credential removes all associated authority, an assumption that breaks when credentials are shared across orchestrated agent workflows; and certificate validation logic conditions entirely on identity authenticity rather than on the operational context in which that identity is being exercised. These limitations were inconsequential when authenticated entities were deterministic services, but they become structurally significant as operating environments shift toward adaptive autonomous systems.

The accelerating adoption of AI-driven autonomous systems introduces a qualitatively different operational pattern into these environments. Autonomous agents are increasingly capable of initiating actions, interacting across distributed services, and generating decisions without direct human involvement at the moment of execution. These systems typically authenticate using machine identities provisioned through existing PKI frameworks—the cryptographic mechanisms are unchanged. What changes is the relationship between the identity and the behavior executed under it. Rather than performing a single deterministic function, an autonomous agent may authenticate to one service, retrieve data, invoke additional APIs, and trigger downstream operations across multiple trust domains, all within a single autonomous reasoning cycle [2]. The credential remains cryptographically valid throughout; the behavioral scope, however, is generated dynamically rather than predetermined.

This divergence between identity scope and behavioral scope constitutes the central challenge this article examines. It does not reflect a cryptographic weakness in PKI, nor does it imply that PKI is insufficient for securing autonomous systems. Rather, it highlights a structural shift in what it means for an authenticated identity to operate within a trust boundary. As zero trust architecture principles have gained adoption and machine identity management has grown in operational complexity, understanding the evolving relationship between cryptographic identity and autonomous behavior has become increasingly important for practitioners designing and operating internet-scale trust infrastructure [3]. The literature on zero trust emphasizes continuous verification and least-privilege access, yet the behavioral scope of autonomous agents raises questions that identity verification alone cannot fully resolve.

The primary contributions of this article are (1) a structured comparison of traditional and autonomous trust models, clarifying how the introduction of AI-driven systems alters the functional meaning of trust boundaries; (2) an analysis of delegated authority amplification and machine-to-machine trust chains as emergent properties of autonomous operation under PKI; and (3) an identification of the accountability and predictability trade-offs that arise when autonomous decision processes operate under PKI-authenticated identities. The article proceeds as follows. In Section 2 we introduce a taxonomy of trust boundaries already present in the customary PKI space and focus on how human-based assumptions of trust are eroded by activities of AI systems. In Section 4 we consider the impact of a comparative perspective on trust boundaries in autonomous scenarios. Section 5 discusses implications of the paper for building a global trust infrastructure. Section 6 discusses trade-offs in security, accountability, and predictability, and section 7 concludes.

2. Trust Boundaries in Traditional PKI Systems

In distributed systems, a trust boundary is the point where the system must authenticate a communicating agent's identity and authorizations before allowing them access to a resource. It separates trusted components from untrusted external components and indicates a point where credential check and validation are required to allow interaction. Generally, such boundaries are established and enforced in Internet-scale systems using PKI and cryptographically verifiable identities bound to known certificates. hierarchies. The Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile, RFC 5280, specifies the technical structures and semantics

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

that make this binding possible and form the basis for CA-mediated trust between heterogeneous distributed systems [1].

In PKI-based systems, trust boundaries are enforced during authentication and the establishment of secure communication. When a client establishes a Transport Layer Security (TLS) connection with a server, certificate validation confirms that the server's cryptographic identity corresponds to a trusted Certificate Authority and has not been revoked. This process defines a precise boundary: entities presenting valid credentials issued by CAs within the trusted root set are permitted to participate in the secure communication channel, while others are excluded. The CA/Browser Forum Baseline Requirements extend this framework by defining the operational and audit expectations that publicly trusted CAs must satisfy, reinforcing the link between cryptographic validity and organizational accountability [4].

A defining feature of traditional PKI trust boundaries is the close coupling between identity validation and action execution. Authenticated entities in traditional architectures perform roles that are predetermined and tightly scoped by administrative configuration. A web server presenting a TLS certificate is expected to serve content on a specific domain. An application identity authenticating to a microservice is expected to perform defined operations within a bounded workflow. The coupling of identity verification with behavioral authorization, in the form of PKI, is valid because, in the presence of identity verification, one can, in practice, infer the expected class of action. Luo et al. provide empirical evidence for this claim by systematically exploring the certificate validation behavior of mobile browsers and observing that certificate validation logic conditions on assumptions on bounded, role-specific behavioral context [5].

Also part of this model are revocation mechanisms like Certificate Revocation Lists (CRLs) and Online Certificate Status Protocol (OCSP); these enable revocation of trust when assumptions about the operation of an identity are no longer fulfilled. The ability to revoke a certificate reflects the presumption that certificates represent actors with defined roles, and that a change in circumstances — compromise, role change, or operational termination — can be addressed by invalidating the credential. These mechanisms presuppose that the behavioral scope of a credential is bounded and that removing the credential effectively removes the associated authority from the system [6].

What traditional PKI trust boundaries do not encode is a representation of behavioral scope independent of identity. The certificate confirms who the entity is and that it is trusted by a recognized CA. It does not assert what range of actions the entity may generate during its operation. In environments where authenticated entities were human users or deterministic automated services, this omission was inconsequential—administrative controls and code-level constraints ensured that behavior remained bounded. As operating environments shift toward greater autonomy and adaptive decision-making, this gap between identity validation and behavioral interpretation becomes structurally significant. The structural limitations of traditional PKI trust boundary models can be summarized across four dimensions. First, behavioral scope is assumed rather than validated: certificate validation confirms identity authenticity but provides no mechanism for asserting or verifying the range of actions an authenticated entity will perform. Second, revocation granularity is coarse: CRL and OCSP mechanisms address identity-level compromise but cannot selectively revoke authority for specific behavioral contexts while preserving it for others. Third, delegation is implicit: PKI credentials do not encode the chain of human authorization decisions that led to their issuance, making it impossible to trace the operational intent behind an authenticated transaction from the credential alone. Fourth, audit infrastructure is event-scoped: PKI-based logging captures individual authentication events but does not represent the operational sequences that connect them, limiting post-hoc visibility into aggregate behavior. These limitations inform the analysis in subsequent sections and establish the baseline against which autonomous trust models are compared.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Characteristic	Traditional PKI Model	Behavioral Assumption
Identity representation	Human user or deterministic service	Actions are role-scoped and predetermined
Trust boundary enforcement	Certificate validation at connection time	Valid credential = bounded, expected behavior
Behavioral scope	Constrained by configuration and code	Fixed at provisioning time
Revocation model	Remove the credential to remove authority	Credential maps 1:1 to operational role
Human oversight	Direct—admins govern all credential use	Identity and intent are tightly aligned

Table 1. Characteristics and behavioral assumptions of traditional PKI trust boundary models.

3. AI-Driven Systems and the Erosion of Human-Centric Trust Assumptions

The use of AI-driven autonomous systems introduces operational models structurally different from those where human-calculated PKI trust boundaries have been used. Autonomous systems can evaluate their inputs, derive conclusions, and take actions without human intervention at the moment that action is initiated. Most of these systems authenticate using machine identities provisioned via standard PKI mechanisms (e.g., use of certificates, service credentials, or token-based derivatives). This does not change the cryptographic function of the identity verification: the relying party(s) validate that an identity is presenting a valid credential issued by a trusted CA; the only difference being the relationship between identity and behavior post-authentication.

In these legacy systems, the deterministic logic of software workflows required defined behaviors: if the encapsulated application was authenticated and had a client certificate available, it always would perform a known set of actions (query a database, call a downstream application programming interface (API), return a response) in a predetermined order defined in the code. The behavioral envelope was fixed at development and deployment time. In AI-driven systems, by contrast, actions may be generated dynamically based on probabilistic model inference, contextual data evaluated at runtime, or adaptive feedback produced by prior interactions. The identity remains cryptographically stable throughout, but the sequence of operations initiated under that identity can vary substantially depending on runtime conditions that were not knowable when the credential was provisioned [2].

Existing approaches to machine identity governance, zero trust architecture, and dynamic authorization address portions of this challenge but do not fully resolve it. Zero trust frameworks such as NIST SP 800-207 advocate for continuous re-authentication and least-privilege access but assume that the behavioral scope of an authenticated identity can be specified in advance through policy [3]. Agentic identity and access management (IAM) frameworks emerging in industry address non-human identity lifecycle management but focus primarily on credential issuance and rotation rather than on runtime behavioral interpretation [16]. Dynamic authorization standards such as OAuth 2.0 token scoping provide fine-grained access controls but operate at the level of individual API transactions rather than across multi-hop autonomous action chains [17]. This article identifies what these approaches leave unresolved: the structural divergence between the cryptographic identity that PKI validates and the behavioral scope that autonomous agents generate at runtime under that identity.

This introduces what may be characterized as behavioral scope inflation: the range of possible actions associated with a given machine identity is no longer fixed by its role definition but expands dynamically as the autonomous system interacts with data and services. An autonomous agent provisioned with a service identity scoped to a particular operational domain may, through adaptive reasoning, generate requests that span services, invoke capabilities not explicitly anticipated by the identity's governance

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

policy, or initiate interaction sequences whose aggregate effect extends well beyond any single authenticated transaction. Each individual authentication event at a trust boundary remains cryptographically valid; the meaning of the aggregate interaction sequence is visible only at a higher level of analysis [7].

A further dimension of this challenge arises from the speed and density of automated interaction. Autonomous agents are capable of crossing trust boundaries at rates and volumes that far exceed human-initiated workflows. A single autonomous reasoning cycle may involve authentication to multiple services, retrieval of distributed data, and the initiation of secondary processes—all within seconds. The CA/Browser Forum has increasingly grappled with the operational implications of high-velocity machine identity usage as certificate lifetimes have shortened and automated issuance has grown [4]. The concentration of authenticated machine-to-machine traffic in autonomous workflows amplifies the practical importance of behavioral context interpretation, since traditional post-hoc audit of individual authentication events provides limited visibility into the intent or aggregate effect of autonomous operation sequences.

It is important to establish what this analysis does not claim, and to distinguish this work's contribution from adjacent literature. Recent work on non-human identity (NHI) governance has addressed the operational complexity of machine credential lifecycle management at scale [18]. Research on agentic IAM has examined how large language model-based agents acquire and exercise delegated authority within enterprise environments [18]. Runtime policy enforcement frameworks have proposed continuous behavioral monitoring as a complement to identity-based access controls [19]. This article does not duplicate these contributions. Its specific contribution is the identification of the structural divergence between PKI's identity validation guarantees and the behavioral scope interpretation requirements of autonomous systems — a gap that persists even when NHI lifecycle management, agentic IAM, and runtime policy frameworks are applied, because it originates in the foundational assumptions of the PKI trust model itself rather than in any specific governance implementation. The presence of autonomous systems does not indicate that PKI authentication has become unreliable or insufficient. Cryptographic identity verification continues to provide essential guarantees about who an actor is and that their credentials were issued by a trusted authority. What AI-driven systems reveal is a structural gap between identity validation — which PKI addresses with precision — and behavioral interpretation — which PKI was not designed to provide and which requires additional contextual framing in autonomous environments. Recognizing this gap is a prerequisite for reasoning clearly about trust in systems where autonomous processes operate at scale under PKI-authenticated identities.

Dimension	Traditional Automated Service	AI-Driven Autonomous Agent
Behavioral determination	Fixed at development/deployment	Generated dynamically at runtime
Scope of actions under identity	Narrow, role-scoped, deterministic	Potentially broad, context-dependent
Human direction in execution	Implicit via code; no runtime input needed	Absent: agent decides autonomously
Interaction velocity	Predictable, bounded by workflow logic	High-velocity, burst-capable
Trust boundary crossing pattern	Single or sequential per workflow	Multi-hop, chain-generating
Behavioral predictability	High — inferred from role	Lower—dependent on runtime inference

Table 2. Comparison of behavioral characteristics between traditional automated services and AI-driven autonomous agents operating under PKI-authenticated machine identities.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. Reframing Trust Boundaries in Autonomous Contexts

The operational changes introduced by AI-driven systems do not require PKI trust boundaries to be discarded or replaced. They require that those boundaries be understood within a broader interpretive framework that accounts for the characteristics of autonomous operation. To support that reframing, this section introduces a structured taxonomy of trust boundary interpretation for autonomous contexts. The taxonomy defines four analytical dimensions — identity, intent coupling, authority delegation scope, and trust chain composability — and applies them to distinguish traditional PKI trust models from autonomous operational models. The dimensions were selected based on the following criteria: (1) each dimension corresponds to a structural property of PKI trust boundaries explicitly encoded in the standards literature (RFC 5280, CA/Browser Forum Baseline Requirements, RFC 9162); (2) each dimension is directly affected by the introduction of autonomous agents as described in Section 3; and (3) each dimension admits an observable behavioral difference between deterministic service identities and autonomous agent identities. Alternative taxonomic structures — for example, decomposing trust by CA hierarchy level or by network layer — were considered but excluded because they address infrastructure topology rather than behavioral interpretation, which is the analytical focus of this work. The three dimensions are examined in the subsections below.

4.1. Identity Without Direct Human Intent

In conventional PKI environments, identities ultimately correspond to actors operating under human authority. Even when credentials are issued to services or infrastructure components, those identities represent systems performing defined tasks within scopes established by human administrators. The human origin of the operational intent is traceable, even if the immediate execution is automated. Autonomous AI-driven systems alter this relationship. Machine identities may still be provisioned under human authority, yet the specific actions executed under those identities emerge from model inference, environmental inputs, or adaptive decision processes rather than from a direct human command at the time of execution. The credential remains traceable to a human-authorized source; the intent behind a particular action may not be. This separation of identity ownership from action origination represents a structural shift in what authenticated identity conveys about the trustworthiness of a specific interaction [3].

4.2. Delegated Authority Amplification

Delegation has always been central to distributed systems trust. PKI credentials issued to services, applications, and infrastructure components represent structured grants of authority within defined operational boundaries. AI-driven systems extend this delegation model by enabling authenticated agents to generate sequences of downstream actions that span multiple services, using the same or related credentials throughout. An autonomous process that authenticates to one service, processes the response through model inference, and then initiates transactions with additional services based on that inference produces a form of delegated authority amplification: the credential's effective operational scope expands through the chain of automated actions it initiates, even though no single authentication event exceeds its defined scope. Understanding this amplification is important not because it implies that individual credentials are misused, but because it means that aggregate operational scope can exceed the assumptions embedded in any individual credential's governance policy [8].

4.3. Machine-to-Machine Trust Chains

Modern distributed environments already rely extensively on machine-to-machine communication, with services authenticating to other services through certificates, tokens, and derived credentials. AI-driven autonomous systems increase the density and topological complexity of these interactions by introducing agents that continuously evaluate information and generate new interaction sequences based on inference. As autonomous agents interact across multiple services, sequences of machine-

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

authenticated operations form trust chains in which each link is independently validated by PKI, but the aggregate chain is produced by autonomous reasoning rather than by a scripted workflow.

4.4. Threat Model for Autonomous Trust Boundary Interactions

The three structural features identified above generate concrete failure modes that must be accounted for in PKI-secured environments hosting autonomous agents. The following threat model defines four representative scenarios in terms of assets, adversarial conditions, attack paths, and trust boundary implications.

Scenario 1 — Agent Credential Misuse Across Trust Domains. An autonomous agent is provisioned with a machine identity scoped to an internal analytics service. Through adaptive reasoning, the agent identifies and authenticates to an adjacent API exposed under the same or a related credential. The asset at risk is the organizational data accessible via the adjacent API. The trust assumption violated is that a credential scoped to one service will only be exercised within that service's operational context. The boundary implication is that individual authentication events at each service remain valid, but the aggregate cross-domain interaction was not authorized by the credential's governance policy.

Scenario 2 — Delegated Authority Amplification via Action Chaining. An autonomous orchestrator authenticates to a cloud storage service, retrieves configuration data, and uses that data to authenticate to a secondary compute service, initiating a resource provisioning workflow. No individual authentication event exceeds the credential's defined scope. The aggregate effect — resource provisioning triggered by autonomous inference — was not anticipated in the original credential issuance. The failure mode is that PKI validation succeeds at every step while the operational outcome falls outside the governance assumptions embedded in any single credential.

Scenario 3 — Cross-Domain Action Chaining by a Compromised Agent. An autonomous agent whose runtime inputs have been manipulated — through adversarial prompt injection or malicious context data — generates an action chain that traverses multiple organizational trust domains, each enforced by independent PKI mechanisms. Because each domain validates only the local credential, no single trust boundary detects the full chain. The security objective violated is end-to-end operational integrity across the chain. Audit infrastructure scoped to individual certificate validation events cannot reconstruct the chain post-hoc.

Scenario 4 — Non-Human Identity Lifecycle Failure. An autonomous agent process is terminated following a security incident, but the machine credential provisioned for that agent remains valid until its scheduled expiration. A reconstituted or unauthorized process authenticates using the orphaned credential. The revocation model failure is that PKI revocation requires a human-initiated event to invalidate the credential; no automated mechanism links the operational lifecycle of the autonomous process to the validity lifecycle of the credential.

These scenarios establish the practical stakes of the structural divergence identified in Sections 3 and 4 and provide the basis for the security, accountability, and predictability analysis in Section 6.

The following table applies the three-dimensional taxonomy defined above to contrast traditional and autonomous trust models. Dimension definitions and selection criteria are provided in the section opening; judgments in each cell derive from the structural properties of PKI standards (for the Traditional Model column) and from the behavioral characteristics of autonomous agents identified in Section 3 (for the Autonomous Model column). The table does not exclude the possibility that specific deployment configurations may exhibit intermediate characteristics.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Trust Model Feature	Traditional Model	Autonomous Model
Identity–intent alignment	Tight — admin-directed at provisioning	Loose — intent generated at runtime by the model
Delegation mechanism	Explicit, static scope grants	Dynamic amplification through action chains
Trust the chain structure	Linear, workflow-defined	Branching, inference-generated
Boundary function	Binary gate: admit or exclude	Checkpoint in a multi-hop interaction network
Behavioral accountability unit	Individual authenticated request	Aggregate autonomous operation sequence

Table 3. Structural comparison of traditional and autonomous trust models across five key dimensions of PKI trust boundary interpretation.

5. Implications for PKI and Internet-Scale Trust Infrastructure

The reframing of trust boundaries in autonomous contexts carries practical implications for how PKI-based systems are interpreted and operated within large-scale digital infrastructures. These implications do not originate from any deficiency in PKI cryptographic mechanisms. Certificate validation remains deterministic, CA hierarchies continue to provide verifiable chains of accountability, and revocation systems provide the means to remove trust when operational circumstances change. The implications arise instead from the evolving context in which PKI operates — specifically, the increasing prevalence of autonomous agents that authenticate using machine identities while executing behaviors generated by adaptive processes rather than deterministic workflows.

One immediate implication concerns the relationship between identity validation and behavioral context interpretation. PKI confirms that a credential corresponds to an identity recognized by a trusted CA, and that the credential has not been revoked. At the internet scale, systems depend on this confirmation as the primary signal of trustworthiness for an incoming connection or API request. When autonomous agents generate large volumes of authenticated requests as part of adaptive reasoning cycles, the behavioral context of those requests — what the agent is trying to accomplish and on whose behalf — cannot be inferred from the credentials alone. The 2021 analysis by Hadan et al. of real-world PKI failures found that a significant proportion of trust incidents involved not cryptographic failures but failures of behavioral expectation — the authenticated entity did not act within the scope that relying parties assumed [10]. Autonomous systems structurally amplify this category of ambiguity.

A second implication concerns the management of machine identity lifecycles in environments where autonomous systems operate continuously and at high interaction velocity. Machine credential governance — including issuance, renewal, and revocation — was designed around workflows in which identity lifecycle events correspond to meaningful changes in operational role or ownership. When autonomous agents operate under long-lived credentials while generating high volumes of interactions, the credential lifecycle becomes decoupled from the operational lifecycle of the autonomous process. Zero trust architecture frameworks address part of this challenge by advocating for short-lived credentials and continuous re-authentication [11], and the CA/Browser Forum's recent reductions in maximum TLS certificate validity reflect the broader ecosystem's movement toward shorter credential lifetimes [4]. However, the alignment of credential lifecycle events with the behavioral lifecycle of autonomous agents remains an open practical challenge for credential governance at scale.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A third implication involves cross-domain trust chains in AI-driven workflows. Autonomous agents that interact with internal services, cloud APIs, and external platforms within a single operational sequence traverse trust boundaries across organizational and infrastructure domains, each enforced by PKI mechanisms whose scope is local to that domain. Certificate Transparency logging at internet scale, as specified in RFC 9162, provides an audit infrastructure that covers certificate issuance across these domains [9]. It does not, however, provide a representation of how authenticated interactions across domains compose into aggregate autonomous operation sequences. Understanding cross-domain trust chains in AI-driven workflows requires analysis at a level of abstraction above individual certificate validation events.

5.1. Enterprise Scenario Analysis

The following scenarios illustrate how the structural divergence between identity scope and behavioral scope produces governance gaps in representative enterprise environments. These scenarios are structured to demonstrate conditions under which PKI-authenticated autonomous agents generate operational outcomes that would not arise from standard service accounts performing equivalent functions.

Scenario A — Autonomous Agent in a Cloud-Native API Mesh. An enterprise deploys an autonomous LLM-based agent to perform continuous infrastructure optimization across a multi-cloud environment. The agent is provisioned with a service identity that authenticates via mutual TLS to an API gateway. Within a single reasoning cycle, the agent authenticates to a monitoring service, retrieves performance metrics, evaluates them against optimization objectives, and initiates scaling operations across three cloud regions by authenticating to regional management APIs using derived tokens. A standard service account performing the same scaling operation would do so through a scripted workflow with a fixed trigger condition and a bounded set of target APIs specified at deployment time. The autonomous agent's credential remains cryptographically valid throughout; its behavioral scope — determined by runtime inference over performance data — was not knowable when the credential was provisioned. The governance gap is the absence of a mechanism to validate whether the agent's runtime behavioral scope remains within the operational boundaries assumed by the credential's issuing authority.

Scenario B — AI Orchestrator in a Financial Services Workflow. A financial services firm deploys an autonomous orchestrator to manage a multi-step transaction verification workflow involving internal risk scoring services, an external credit reference API, and a regulatory reporting endpoint. The orchestrator authenticates to each service using OAuth 2.0 bearer tokens derived from a root machine identity. During an anomalous market event, the orchestrator's inference process generates an interaction sequence that invokes the regulatory reporting endpoint with a frequency and payload structure outside the parameters expected by the endpoint's access policy. Each token validation succeeds; the aggregate interaction pattern constitutes a policy violation detectable only through behavioral telemetry operating above the token level. A deterministic service account executing the same workflow would interact with the reporting endpoint according to a fixed schedule and payload schema, making anomaly detection straightforward. The autonomous orchestrator's adaptive response to the market event produced a cross-boundary interaction sequence whose aggregate effect was not anticipated by any individual token's scope definition.

Taken together, these implications reinforce a view of PKI as a structural foundation within a broader trust ecosystem rather than as a self-contained representation of system behavior. The cryptographic guarantees PKI provides — identity authenticity, credential integrity, and CA accountability — remain essential and are not in question. What AI-driven environments make visible is the distance between those guarantees and the full behavioral interpretation of trust at the system level. This distance does not imply a replacement of PKI, but identifying the contexts that the system provides guarantees for and requiring interpretative machinery elsewhere.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6. Security, Accountability, and Predictability Trade-offs

Architects and operators of PKI-secured environments with deployed AI-driven autonomous systems must consider the trade-offs between three non-overlapping but associated concepts when building and managing a trust infrastructure. These concepts are assurance, accountability, and predictability, and an understanding of them is essential when dealing with scaled deployments of autonomous agents.

The security assurance trade-off is the trade-off between cryptographic and context-dependent assurance. PKI provides strong cryptographic assurance: that the entity in question possesses a valid cryptographic credential, that the credential was issued by a CA trusted by the relying party, and that the credential has not been revoked. These guarantees are mathematically verifiable and do not depend on run-time assumptions about the system's behavior. In contrast, contextual certainty, in the form of knowing that an authenticated request is coming from the appropriate and bounded operational context, is not provided out-of-the-box by PKI. It must be obtained through policy enforcement, behavior monitoring, or run-time authorization frameworks. In AI-driven environments, the gap between cryptographic certainty and contextual certainty widens because the behavioral scope of authenticated identities is generated dynamically rather than fixed at provisioning. Security architectures that treat credential validity as a sufficient proxy for operational trustworthiness are structurally exposed to this gap [3].

The accountability trade-off concerns the relationship between identity traceability and action origination in automated workflows. Traditional PKI accountability models assume that authenticated identities ultimately trace to identifiable human authority — an organization issued the credential, a person provisioned the service. When autonomous AI-driven processes generate actions under machine identities, the chain of decisions leading to any specific action may involve layers of model inference and automated orchestration that are not directly visible in authentication logs. Each authentication event is traceable to a credential and ultimately to an issuing CA; the reasoning process that initiated that authentication event is not. This accountability gap does not reflect a deficiency in CA operations or in certificate governance — it reflects the structural characteristic of autonomous decision-making, in which the identity used to authenticate and the process that decided to authenticate are different analytical entities [10]. Emerging zero trust maturity models have begun to grapple with this challenge by integrating behavioral telemetry alongside identity verification, though standardized approaches for autonomous agents remain nascent [8].

The predictability trade-off operates at the level of distributed systems stability. Distributed systems across organizational and infrastructure boundaries depend on predictable interpretations of identity validation to maintain stable interactions. When autonomous agents generate high volumes of machine-to-machine authenticated transactions — including bursts associated with rapid inference cycles — trust boundaries may be crossed in patterns that deviate substantially from the traffic profiles that relying systems were designed to handle. This is not a denial-of-service concern in the traditional sense; each authentication event is individually valid. It is a systems stability concern: predictability in how trust boundaries are exercised enables relying parties to reason about system state, enforce rate limits, and detect anomalous behavior. As Syed et al. noted in their comprehensive survey of zero trust architectures, the move toward continuous authentication and dynamic policy evaluation introduces operational complexity that scales with the volume and velocity of authenticated interactions [11].

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Trade-off Dimension	PKI Guarantee	AI-Driven Challenge	Practical Implication
Security assurance	Cryptographic identity validity	Behavioral scope exceeds credential assumptions	Contextual certainty requires additional mechanisms beyond credential validation
Accountability	Credential traceable to issuing CA	Action origination obscured by autonomous inference	Accountability frameworks must capture decision provenance, not only identity
Predictability	Deterministic validation outcomes	Variable interaction patterns from adaptive reasoning	Systems must accommodate non-deterministic traffic profiles from autonomous agents
Credential governance	Lifecycle tied to operational role changes	Behavioral lifecycle decoupled from credential lifecycle	Shortened credential lifetimes and continuous re-authentication reduce but do not eliminate the gap

Table 4. Security, accountability, and predictability trade-offs in PKI-secured environments with AI-driven autonomous system participation.

Recognizing these trade-offs implies that complementary mechanisms must be specified alongside PKI in environments where autonomous agents operate at scale. The following reference architecture identifies the minimum set of components required to address each trade-off dimension.

For the security assurance gap, a runtime behavioral policy engine must operate at each trust boundary crossing point to evaluate whether an authenticated request falls within the behavioral envelope defined for the associated machine identity. The engine requires: a behavioral policy specification for each machine identity that defines permitted action types, target service scope, interaction frequency bounds, and payload structure constraints; a real-time evaluation interface that intercepts authenticated requests after credential validation and before resource access; and an enforcement action set that includes permit, deny, and quarantine-for-review outcomes.

For the accountability gap, operational sequence logging must capture the following fields per autonomous action event: agent identity (credential reference), session chain identifier (a persistent identifier linking all actions within a single autonomous reasoning cycle), trust domain traversal record (ordered list of trust boundaries crossed within the chain), action type and target service, and inference context reference (a pointer to the model state or input context that generated the action). This logging schema enables post-hoc reconstruction of autonomous operation sequences from individual authentication events.

For the predictability gap, traffic profile baselines must be established per machine identity and per trust boundary, with anomaly detection thresholds calibrated to the burst characteristics of autonomous reasoning cycles. Rate limit policies must distinguish between burst-mode autonomous operation (expected during inference cycles) and sustained high-frequency access (potentially indicative of runaway agent behavior or credential misuse).

For the credential governance gap, machine identity lifecycle management must be coupled to the operational lifecycle of the autonomous process rather than to a calendar-based expiration schedule. Credential validity should be contingent on the operational state of the associated agent process; automated revocation must be triggered on agent termination, redeployment, or security incident without requiring manual intervention.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

7. Conclusion

Public key infrastructure has long served as the foundational mechanism for establishing verifiable identity and enforcing trust boundaries across internet-scale distributed systems. Its effectiveness rests on a rigorous model in which certificate hierarchies bind cryptographic keys to identities, CAs provide accountability for credential issuance, and relying parties validate credentials against trusted roots at the point of connection. This model has proven robust, scalable, and interoperable across diverse infrastructure environments and organizational contexts.

The growing prevalence of AI-driven autonomous systems introduces structural pressures on the assumptions that underpin this model—not on its cryptographic mechanisms, which remain sound, but on the interpretive framework through which authenticated identity is understood as a representation of operational intent and behavioral scope. The idea presented in this article identifies three specific structural changes that autonomous systems introduce: the separation of identity ownership from action origination, the amplification of delegated authority through dynamic action chains, and the emergence of machine-to-machine trust chains whose aggregate behavior is generated by autonomous reasoning rather than scripted workflows. Each of these changes alters the functional meaning of a trust boundary without altering the cryptographic validity of the authentication events that occur within it.

The practical implication is that PKI must be understood as a foundational cryptographic layer within a broader trust ecosystem that requires four additional operational components in autonomous environments: a runtime behavioral policy engine enforcing action-level constraints per machine identity; operational sequence logging capturing session chain identifiers and cross-domain traversal records; traffic profile anomaly detection calibrated to the burst characteristics of autonomous reasoning cycles; and agent-lifecycle-coupled credential governance enabling automated revocation on process termination or security incident. The threat model presented in Section 4.4 and the enterprise scenarios in Section 5.1 demonstrate that without these components, PKI-authenticated autonomous agents can produce governance gaps — specifically behavioral scope inflation, delegated authority amplification, and unresolvable cross-domain action chains — that standard service accounts operating under equivalent credentials do not. Maintaining clarity about the boundary between what PKI guarantees and what it does not will be central to governing autonomous agents in the next generation of internet-scale distributed systems.

References

[1] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley, and W. Polk, "Internet X.509 public key infrastructure certificate and certificate revocation list (CRL) profile," Internet Engineering Task Force, RFC 5280, May 2008. [Online]. Available: https://www.rfc-editor.org/rfc/rfc5280.html

[2] P. Phiayura and S. Teerakanok, "A comprehensive framework for migrating to zero trust architecture," IEEE Access, vol. 11, pp. 19487–19511, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10052642

[3] CA/Browser Forum, "Baseline requirements for the issuance and management of publicly-trusted certificates," 2023. [Online]. Available: https://cabforum.org/working-groups/server/baseline-requirements/documents/

[4] M. Luo, B. Feng, L. Lu, E. Kirda, and K. Ren, "On the complexity of the web's PKI: Evaluating certificate validation of mobile browsers," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 6, pp. 4747–4762, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10066507

[5] E. Rescorla, "The transport layer security (TLS) protocol version 1.3," Internet Engineering Task Force, RFC 8446, Aug. 2018. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8446.html

[6] J. Lu et al., “AgentLens: Visual Analysis for Agent Behaviors in LLM-based Autonomous Systems,” IEEE Transactions on Visualization and Computer Graphics, pp. 1–17, 2024, doi: https://doi.org/10.1109/tvcg.2024.3394053.
[7] B. Laurie, E. Messeri, and R. Stradling, "Certificate transparency version 2.0," Internet Engineering Task Force, RFC 9162, Dec. 2021. [Online]. Available: https://www.rfc-editor.org/rfc/rfc9162.html
[8] H. Hadan, N. Serrano, and L. J. Camp, "A holistic analysis of web-based public key infrastructure failures: comparing experts' perceptions and real-world incidents," Journal of Cybersecurity, vol. 7, no. 1, p. tyab025, 2021. [Online]. Available: https://academic.oup.com/cybersecurity/article/7/1/tyab025/6470936
[9] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero trust architecture," National Institute of Standards and Technology, NIST Special Publication 800-207, Aug. 2020. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/specialpublications/NIST.SP.800-207.pdf
[10] M. Cooper, Y. Dzambasow, P. Hesse, S. Joseph, Van Dyke Technologies, and R. Nicholas, "Internet X.509 public key infrastructure: Certification path building," Informational Network Working Group, RFC 4158, Sep. 2005. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4158.html
[11] E. Rescorla, "HTTP over TLS," RFC 2818, May 2000. [Online]. Available: https://www.rfc-editor.org/rfc/rfc2818.html
[12] R. Barnes, et al., "ACME (Automatic Certificate Management Environment)," RFC 8555, Mar. 2019. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8555.html
[13] Y. Felk, "Confidential computing," in Trends in Data Protection and Encryption Technologies, Springer, Cham, 2023, pp. 103–107. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-33386-6_19
[14] T. Koppel1, O. Tšernikova1, I. Vilcane, “AI transformation is platform-driven,” proceedings of the scientific workshop on business implications of artificial intelligence, 2022. https://www.researchgate.net/profile/Tarmo-Koppel/publication/367207589_AI_transformation_is_platform_driven/links/63c7070bd9fb5967c2e45253/AI-transformation-is-platform-driven.pdf