RITHA PUBLISHING

Journal of Emerging Technologies for Society
Research | Innovation | Applications

https://doi.org/10.57017/jets.v1.iss1.02

An Integrated Architecture Combining Blockchain, Artificial Intelligence, and Machine Learning for Decentralised Identity Management in the Internet of Things

Marius Iulian MIHĂILESCU*[1], Ștefania Loredana NIȚĂ**, Valentina MĂRĂSCU*, ***

*Faculty of Engineering and Computer Science, Spiru Haret University, Romania
**Faculty of Information Systems and Cyber Security, Military Technical Academy "Ferdinand I", Romania
***National Institute for Laser, Plasma and Radiation Physics, Romania

Cite: Mihăilescu, M. I., Niță, S. L. & Mărăscu, V. (2026). An Integrated Architecture Combining Blockchain, Artificial Intelligence, and Machine Learning for Decentralised Identity Management in the Internet of Things. *Journal of Emerging Technologies for Society*, Volume I, *1*(1), 27-50. https://doi.org/10.57017/jets.v1.iss1.02

Article info:

Received: 9 March, 2026
Revised: 30 March, 2026
Accepted: 19 May, 2026
Published: 22 May, 2026

Abstract

The identity-management challenge addressed in this study emerges after device enrolment, when an Internet of Things (IoT) device may continue to present valid credentials even though its software state, behavioural patterns, or operating conditions have changed. To address this limitation, the study proposes an integrated architecture that treats cryptographic identity and observed behaviour as complementary but distinct sources of trust evidence. A permissioned ledger anchors decentralised identifiers, credential states, policy versions, model attestations, and decision evidence, while behavioural analysis is performed close to the device and federated learning enables collaborative model development without centralising raw telemetry. At the authorisation stage, credential validity, behavioural risk, and request context are combined into a dynamic trust state, enabling graduated responses as evidence evolves. Low-level observation and temporary restrictions may be automated, whereas suspension is subject to review and permanent credential revocation requires human authorisation.

The study adopts a Design Science Research approach, treating the proposed architecture as an artefact evaluated through requirements-based assessment, threat-model analysis, and feasibility evidence derived from the underlying technological mechanisms. The evaluation identifies both design-level capabilities and unresolved dependencies related to secure key custody on legacy devices, federated coordination at scale, poisoning by credentialled participants, and regulatory conformity. The proposed architecture therefore extends decentralised identity management beyond static credential verification by integrating continuous behavioural assessment, auditable decision-making, and human oversight into a unified IoT trust-management framework.

Keywords: blockchain, artificial intelligence, machine learning, Internet of Things, digital trust, decentralised identity management, cybersecurity.

1. Introduction

The rapid expansion of the Internet of Things (IoT) has increased the complexity of identity management and access control across heterogeneous and distributed environments. Conventional identity-management mechanisms typically establish trust at the point of enrolment by verifying whether a device possesses valid credentials. However, credential validity alone cannot determine whether a device

[1] Corresponding author

continues to operate legitimately throughout its lifecycle. A device may retain valid credentials even after its software state, behavioural patterns, or operating conditions have changed as a result of compromise or unauthorized intervention. This distinction between identity validity and behavioural trustworthiness constitutes a significant challenge for contemporary IoT security.

Blockchain-based and decentralised identity approaches provide mechanisms for establishing verifiable and tamper-resistant identity states while reducing dependence on a single central authority (Allen, 2016; Dorri et al., 2017; Mühle et al., 2018). Machine-learning-based behavioural analysis addresses a different dimension of the problem by identifying deviations from expected device activity. IoT devices frequently exhibit relatively stable operational patterns, allowing anomalous traffic, communication, or resource-access behaviour to provide evidence of potential compromise. Blockchain-based identity management and machine-learning-based behavioural analysis therefore address distinct but complementary dimensions of IoT security.

However, combining identity verification with anomaly detection does not in itself provide a complete trust-management mechanism. A system may verify the identity of a device and simultaneously detect abnormal behaviour while lacking an integrated mechanism through which behavioural evidence can modify the device's effective access rights. In many existing approaches, intervention remains dependent on separate administrative or human processes after an anomaly has been detected (Buczak & Guven, 2016; Hussain et al., 2020). Moreover, distributed-learning environments remain vulnerable to malicious or compromised participants even when their identities are technically valid (Bagdasaryan et al., 2020). The central challenge is therefore not only to establish identity or detect anomalous behaviour, but to connect these forms of evidence to an enforceable and accountable authorisation process.

This challenge is particularly important in IoT environments because connected devices frequently operate under substantial resource and lifecycle constraints. Industrial and infrastructure devices may remain operational long after manufacturer support has ended, while limited processing capacity, energy availability, and network connectivity restrict the security mechanisms that can be deployed locally (Sicari et al., 2015). Centralised security architectures may simplify administration but introduce concentration risks, including dependence on central infrastructure and the aggregation of potentially sensitive device telemetry (Khan & Salah, 2018). Large-scale IoT deployments also increase the consequences of compromised devices, as demonstrated by attacks in which legitimately connected devices have been incorporated into distributed malicious activity (Antonakakis et al., 2017, Kolias et al., 2017).

These limitations create a need for an architecture capable of maintaining trust beyond initial enrolment. This approach is consistent with zero-trust principles, according to which trust should not be implicitly granted on the basis of prior authentication alone but should remain subject to continuing verification and policy evaluation (Rose et al., 2020).

Such an architecture should combine verifiable identity state with continuous behavioural evidence while avoiding unnecessary centralisation of raw telemetry. It should also support graduated responses to changing risk conditions rather than treating trust as a binary and permanent property. This requirement is particularly relevant in healthcare, industrial systems, and critical infrastructure, where an incorrect automated decision may itself disrupt safety-relevant or essential operations. Consequently, automated observation and temporary restrictions must be distinguished from more consequential actions such as suspension or permanent credential revocation.

To address this gap, this study proposes an integrated architecture in which a permissioned ledger records identity state and anchors policy versions, model attestations, and decision evidence; behavioural analysis is performed at the edge; federated learning supports collaborative model development without centralising raw telemetry; and a policy layer combines credential validity, behavioural risk, and request

context before access is granted. The response mechanism is deliberately graduated: observation and temporary restriction may be automated, suspension initiates a review process, and permanent revocation remains subject to human authorisation.

The contribution of the study is architectural rather than experimental. It does not claim to introduce new blockchain, machine-learning, or federated-learning algorithms. Instead, it addresses the dependency between these technological components by specifying how decentralised identity, behavioural assessment, access control, and governance can be integrated into a single trust-management architecture. The design is derived from the identity-management, distributed-ledger, machine-learning, cybersecurity, and governance literature and is evaluated against explicit functional and non-functional requirements, four adversary classes, and feasibility evidence from the underlying component technologies.

Accordingly, the study addresses the following research question:

RQ: How can decentralised identity management and behavioural intelligence be integrated into an auditable and adaptive IoT trust architecture while preserving privacy, human oversight, and accountability?

Addressing this question, the study contributes to the emerging-technology literature in two related ways. First, it proposes a bidirectional relationship between identity and behavioural intelligence: ledger-verified identity establishes accountable participation in distributed learning, while behavioural evidence can subsequently influence the permissions associated with an otherwise valid credential. Second, it incorporates governance directly into the architecture by requiring consequential decisions to remain explainable, contestable, and subject to human oversight. The proposed framework therefore positions digital trust not as a one-time outcome of credential verification, but as a continuously evaluated relationship between identity, behaviour, context, and accountable decision-making.

The remainder of the paper is structured as follows. Section 2 reviews the relevant literature and identifies the capability dependencies and research gap underlying the proposed architecture. Section 3 presents the Design Science Research methodology and derives the functional, non-functional, and governance requirements. Section 4 describes the integrated architecture and its principal components and interaction flows. Section 5 evaluates the proposed artefact against the identified requirements, threat conditions, and feasibility considerations. The final section discusses the implications, limitations, and conclusions of the study.

2. Literature Review

The literature is examined as a source of design constraints and technological dependencies rather than as a catalogue of individual technologies. IoT identity-management research establishes the requirements associated with heterogeneous hardware, large-scale deployment, and extended credential lifecycles; decentralised identity changes how authoritative identity states are maintained and verified; behavioural analytics introduces post-enrolment evidence regarding device activity; and the regulatory literature establishes constraints on privacy, automated decision-making, accountability, and human oversight.

These research streams address complementary dimensions of digital trust but also depend on capabilities that they do not independently provide. Distributed ledgers can establish verifiable identity and credential states but cannot determine whether the entity holding a valid credential continues to behave legitimately after enrolment. Behavioural analytics can identify anomalous activity but requires authenticated participants, reliable telemetry processing, and an enforceable mechanism through which detected risk can affect access rights. The literature is therefore reviewed below with particular attention to these capability dependencies and to the conditions required for their integration.

2.1. Identity Management in IoT Environments: Requirements and Constraints

Identity management in IoT environments is shaped by several recurring constraints: the scale of device populations, substantial heterogeneity in hardware capabilities, long operational lifecycles, and deployment in physically accessible or insufficiently protected environments (Sicari et al., 2015; Khan & Salah, 2018). These characteristics interact rather than operate independently. Credential-management procedures that are feasible for a limited number of capable gateways may become costly at fleet scale, while cryptographic operations that impose negligible overhead on permanently powered devices may be significant for resource-constrained sensors.

The lifecycle of IoT devices introduces additional challenges. Devices may remain operational after manufacturer support or security updates have ended, while physically accessible deployments can expose cryptographic material to extraction or tampering (Sicari et al., 2015; Khan & Salah, 2018). Lightweight cryptographic and identity-verification mechanisms can reduce computational overhead, but they do not necessarily address the structural dependence created when credential issuance, verification, and revocation remain concentrated within a single authority.

Centralised identity-management architectures can therefore introduce both operational and governance dependencies. When authentication and credential-state decisions depend on a single authoritative service, disruption or compromise of that service may affect the availability and integrity of identity-dependent operations (ETSI, 2020). Such concentration can also create broader cybersecurity and operational risks where large populations of connected devices depend on the continued availability of the same infrastructure.

A related concern involves information concentration. Centralised identity infrastructures may aggregate information concerning device ownership, connectivity, deployment, and operational status, thereby creating privacy and informational asymmetries between infrastructure operators and individual participants. The governance implications therefore extend beyond technical availability to questions of data control, transparency, and institutional responsibility.

An additional limitation concerns the persistence of trust after enrolment. Conventional credential-based systems generally determine whether the credential presented by a device remains valid, but credential validity does not necessarily establish the current integrity of the endpoint presenting it. A device whose firmware has been compromised may continue to authenticate successfully as long as its credential remains valid (Ferdous et al., 2019). Similar concerns arise where a device has been altered after manufacture or during its supply-chain lifecycle while retaining legitimate credentials.

This distinction is particularly significant in IoT environments because identity management and intrusion detection have traditionally operated as separate security functions. A credential-management system may establish that a device is authorised, while behavioural monitoring simultaneously indicates that its activity is inconsistent with its expected function. Treating these processes as isolated components prevents behavioural evidence from directly influencing the trust state associated with an authenticated device. Consequently, IoT identity management requires not only reliable initial authentication but also mechanisms capable of reassessing trust throughout the operational lifecycle.

2.2. Blockchain and Distributed Ledger Technologies for Decentralised Identity

Distributed-ledger-based identity architectures reduce dependence on a single repository by providing participating entities with a shared and verifiable record of relevant identity states. However, decentralisation does not eliminate governance or institutional responsibility; rather, it redistributes authority across the participants and rules governing the infrastructure.

Within such architectures, distributed records can support verification of identity registration, credential status, and auditable identity events without requiring every verification request to depend on one central record. This can improve resilience and make unauthorised modification more readily detectable (Dorri et al., 2017; Khan & Salah, 2018). Decentralised Identifiers and Verifiable Credentials provide standardised mechanisms for decentralised identity representation, credential exchange, and selective attribute disclosure (Sporny et al., 2022, 2025; Mühle et al., 2018).

Nevertheless, distributed-ledger architectures involve important trade-offs. Permissionless systems emphasise openness and broad participation, whereas permissioned or consortium-based systems restrict participation in exchange for more predictable governance, throughput, and latency characteristics. For resource-constrained IoT environments, edge-assisted and lightweight ledger implementations have therefore been investigated as mechanisms for reducing transaction and computational overhead.

These developments do not resolve all lifecycle challenges associated with digital identity. Registration on a distributed ledger does not itself address the operational conditions under which an initially legitimate identity may later become compromised. Moreover, immutable identifiers, transaction relationships, and timestamps may reveal interaction patterns even when substantive personal or behavioural data are not directly stored on-chain. Credential revocation, delegation, and recovery also require timely propagation across participants without imposing unacceptable delays on routine identity verification.

For the architecture developed in this study, the most significant limitation is therefore conceptual rather than exclusively computational. A blockchain-based identity system may correctly establish that a device is registered, that its credential remains active, and that relevant events have been recorded, while remaining unable to determine whether the endpoint presenting that credential has been compromised after enrolment. In such circumstances, revocation becomes meaningful only when another component provides evidence capable of triggering a reassessment of trust. Behavioural evidence must therefore be connected to credential state rather than operating solely as an independent alerting mechanism.

2.3. Artificial Intelligence and Machine Learning in IoT Cybersecurity

Behavioural machine learning provides time-dependent evidence that static credentials cannot capture, and anomaly-detection approaches have demonstrated the feasibility of identifying abnormal network behaviour generated by compromised IoT devices (Meidan et al., 2018). Because many IoT devices exhibit relatively repetitive operational profiles, changes in communication periodicity, destination selection, payload characteristics, or resource utilisation can provide indicators of anomalous activity (Buczak & Guven, 2016; Hussain et al., 2020).

Existing research includes random forests, support vector machines, neural networks, and transformer-based approaches to IoT intrusion and anomaly detection (Khacha et al., 2024). However, reported predictive performance must be interpreted in relation to the conditions under which models are evaluated. Ageing benchmark datasets, class imbalance, limited adversarial testing, and insufficient assessment under device-level memory, energy, and latency constraints may restrict the transferability of reported results to operational IoT environments. Accordingly, the relevant architectural requirement is not the adoption of a particular model family but the capacity to perform meaningful behavioural assessment within endpoint and deployment constraints.

Federated approaches have also been applied specifically to self-learning IoT anomaly detection, allowing device-type-specific behavioural models to be developed without centralising all observations (Nguyen et al., 2019). Rather than transferring raw telemetry to a central repository, participating devices or edge nodes retain local data and exchange model updates during collaborative training (McMahan et al., 2017). Privacy within collaborative learning can be further strengthened through secure aggregation

mechanisms that allow the server to aggregate model updates without accessing individual client contributions (Bonawitz et al., 2017). This approach is compatible with the decentralisation objectives of the proposed identity architecture because it reduces the need to aggregate sensitive operational data centrally.

Federated learning nevertheless introduces additional technical and governance dependencies. Non-independent and unevenly distributed client data, communication overhead, class imbalance, and model convergence can affect efficiency and performance, while malicious participants may attempt to poison the shared model (Kairouz et al., 2021; Hernandez-Ramos et al., 2025; Belenguer et al., 2025). Moreover, distributed learning requires mechanisms for determining which participants are authorised to contribute and for attributing individual model updates to accountable entities.

Weak participant identity creates particular vulnerabilities. Sybil attacks can allow one adversary to create multiple identities and disproportionately influence collaborative learning, while anonymous or insufficiently attributable updates complicate post-incident auditing (Douceur, 2002; Fung et al., 2020). Identity management therefore becomes a prerequisite for accountable federated learning rather than an unrelated security function.

This dependency provides the principal point of contact between the two research streams. Ledger-backed credentials can restrict participation to recognised edge nodes and support attribution of model-update events, as reflected in existing blockchain-supported federated-learning approaches (Ali et al., 2024). The architecture proposed in this study extends this relationship in both directions: verified identity determines who may participate in distributed learning, while behavioural evidence generated by that learning process feeds back into identity and access control. Consequently, a device may remain cryptographically recognised while receiving reduced authority when its observed behaviour indicates elevated risk.

2.4. Governance and Regulatory Context

The integration of decentralised identity and behavioural intelligence also introduces governance and regulatory requirements. These are particularly significant where immutable audit mechanisms interact with automated decisions that may affect access to safety-relevant or regulated services.

The revised eIDAS framework, Regulation (EU) 2024/1183, establishes requirements relating to interoperable digital identity, verifiable attestations, and selective disclosure (European Parliament and Council of the European Union, 2024a). The Cyber Resilience Act introduces product-security obligations (European Parliament and Council of the European Union, 2024c), complemented at the technical level by established IoT cybersecurity baseline requirements such as ETSI EN 303 645 (ETSI, 2020). The AI Act addresses risk management, transparency, human oversight, and accountability in automated decision-making (European Parliament and Council of the European Union, 2024b). The General Data Protection Regulation (GDPR) introduces additional constraints where identity information, behavioural telemetry, or associated metadata constitute personal data (European Parliament and Council of the European Union, 2016).

The proposed design responds to this tension by distinguishing between on-chain integrity evidence and off-chain substantive data. Raw behavioural telemetry and detailed explanatory records are not intended to be permanently stored on the ledger. Instead, integrity commitments can be anchored on-chain to enable later verification of externally stored evidence without permanently exposing the underlying information. A second governance principle concerns the distinction between reversible risk containment and consequential credential actions.

Automated mechanisms may support observation and temporary restrictions, but suspension requires a review process and permanent credential revocation remains subject to explicit human authorisation. This separation limits the authority assigned to automated behavioural assessment and provides a mechanism for contestability where automated risk classification may affect operational access.

The architecture can therefore support traceability by preserving the relevant model version, policy version, and supporting decision evidence. However, such technical traceability should not be interpreted as proof of legal compliance. Regulatory conformity ultimately depends on implementation context, institutional responsibilities, data-processing practices, and the applicable legal framework. The architectural objective is instead to provide the technical conditions necessary for auditability, explainability, human oversight, and accountable decision-making.

Table 1. Comparative assessment of the approaches reviewed against the identified requirements

Requirement	Centralized IdM	Blockchain / SSI	Centralized ML security	Federated ML	Proposed architecture
Elimination of single point of failure	■	●	■	◐	●
Verifiable, tamper-resistant identity	◐	●	■	■	●
Auditability of identity events	◐	●	◐	◐	●
Detection of post-enrolment compromise	■	■	●	●	●
Adaptive, risk-based access control	■	◐	◐	◐	●
Privacy preservation of behavioural data	◐	◐	■	●	●
Feasibility on constrained devices	●	◐	◐	◐	◐
Explainability and human oversight	◐	◐	■	■	●
Regulatory alignment (eIDAS 2.0, CRA, AI Act, GDPR)	◐	◐	■	◐	●

Note: Legend: ● fully addressed; ◐ partially addressed; ■ not addressed

Source: compiled by the authors.

2.5. Synthesis and Research Gap

The literature identifies a research gap that extends beyond the general proposition that blockchain and machine learning should be combined. Decentralised identity reduces dependence on a single issuing or validation authority and provides a shared basis for verifying credential state, but the trust represented by that state remains largely static after enrolment. Behavioural analytics provides runtime evidence that a credential cannot capture, whereas centralised analytics concentrates telemetry and federated analytics depends on reliable identification and accountability of participating clients.

These technologies therefore exhibit complementary dependencies. Decentralised identity can establish who is presenting evidence without determining whether that participant continues to behave legitimately, whereas behavioural analytics can assess how a participant is behaving without independently providing the identity and accountability mechanisms required for distributed learning and enforcement. When identity and behavioural assessment are integrated into the same control loop, the remaining challenge becomes not only technological but also one of governance and accountability. Automated behavioural assessment may identify elevated risk, but reducing a device's authority can affect regulated, industrial, healthcare, or safety-relevant operations. The architecture must therefore distinguish reversible containment from consequential credential actions and provide mechanisms for explanation, review, contestability, and human oversight.

Existing integration approaches frequently use blockchain to authenticate participation in distributed learning, thereby improving participant attribution. However, the reverse relationship remains insufficiently addressed: behavioural evidence generated after enrolment is not consistently connected back to the permissions associated with a device whose credential remains cryptographically valid.

The research gap addressed in this study is precisely this bidirectional relationship between identity and behavioural trust. The proposed architecture integrates ledger-backed identity, behavioural evidence, and enforceable policy at the authorisation stage while maintaining the separate inspectability of each evidence source. This enables behavioural evidence to influence the authority associated with a valid credential without allowing automated risk assessment to become an unrestricted mechanism for permanent identity revocation. Permanent revocation remains subject to human authorisation, while the relevant model, policy, and decision evidence is retained for subsequent review.

Section 3 translates these dependencies into explicit functional, non-functional, and governance requirements and evaluates whether the proposed architecture provides a defensible mechanism for addressing each requirement. Questions that cannot be resolved through architectural analysis alone—including hardware feasibility, scalability, adversarial robustness, and regulatory conformity—remain explicitly identified as matters requiring implementation and empirical validation.

3. Research Methodology and Design Requirements

3.1 Research Approach

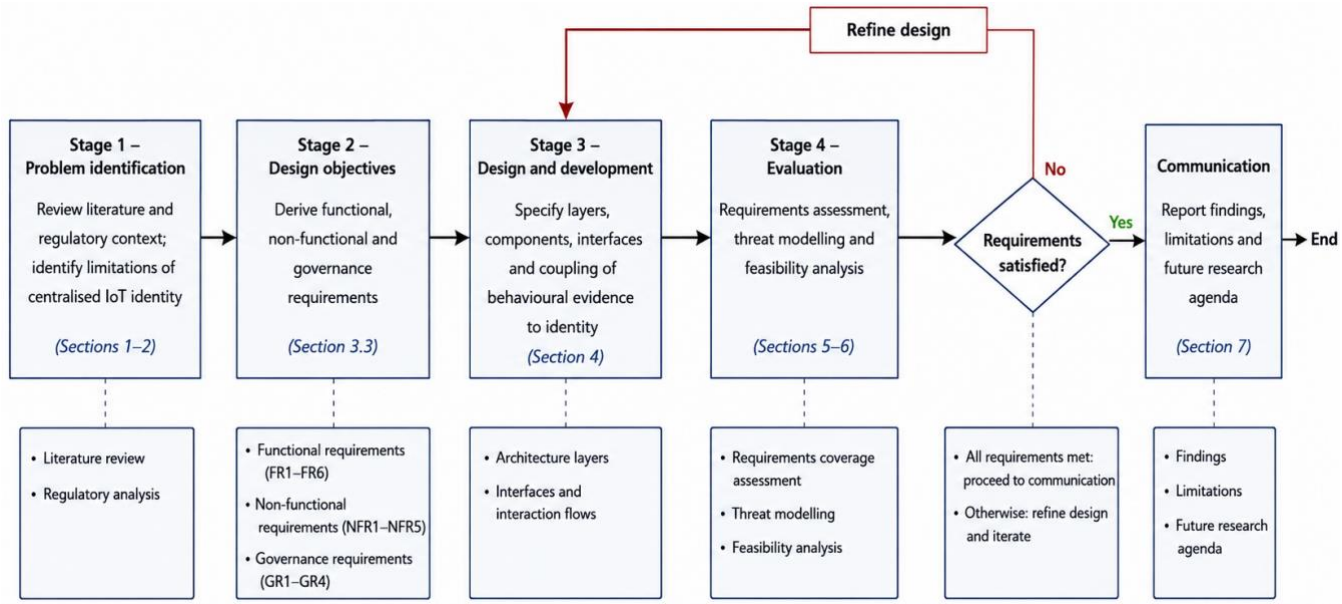
This study adopts a Design Science Research (DSR) approach because its primary contribution is an architectural artefact designed to address a defined class of IoT identity-management and trust problems rather than to test an explanatory theory or report empirical system performance (Hevner et al., 2004; Peffers et al., 2007). The constituent technologies, decentralised identity, distributed ledgers, behavioural machine learning, federated learning, and policy-based access control—already exist and have been investigated independently. The methodological problem addressed here concerns their composition and interaction: where identity, behavioural assessment, and access control should intersect; which functions may be delegated to devices, edge nodes, or the ledger; and which decisions must remain reviewable when the architecture operates in regulated or safety-relevant environments.

DSR is appropriate because it links problem identification, design objectives, artefact construction, and evaluation through explicit criteria. In this study, evaluation is analytical rather than experimental. It examines whether the proposed architecture provides a coherent mechanism for each identified requirement, whether its internal assumptions remain consistent across layers, how it responds to defined adversary classes, and whether the technological assumptions underlying the design are plausible based on existing evidence. The study therefore does not claim end-to-end performance measurements that have not yet been generated through implementation.

The research process follows four connected stages. First, recurring limitations in IoT identity management, distributed-ledger security, behavioural intrusion detection, federated learning, and the European regulatory context are identified from the literature. Second, these limitations are translated into functional, non-functional, and governance requirements. Third, those requirements guide the design of the architectural layers, responsibilities, interfaces, and interaction flows presented in Section 4. Finally, the resulting artefact is evaluated against the same requirements, four adversary classes, and feasibility evidence derived from the underlying component technologies.

This sequence ensures that the evaluation remains directly linked to the problem definition and design objectives rather than introducing assessment criteria after the architecture has been developed. Figure 1 summarises this process, including the iterative possibility that an unmet or partially addressed requirement may require design revision.

Figure 1. Research Methodology Workflow Based on the Design Science Research (DSR) Cycle



Source: Generated/sketched by authors using PlantUML

3.2 Knowledge Base Construction

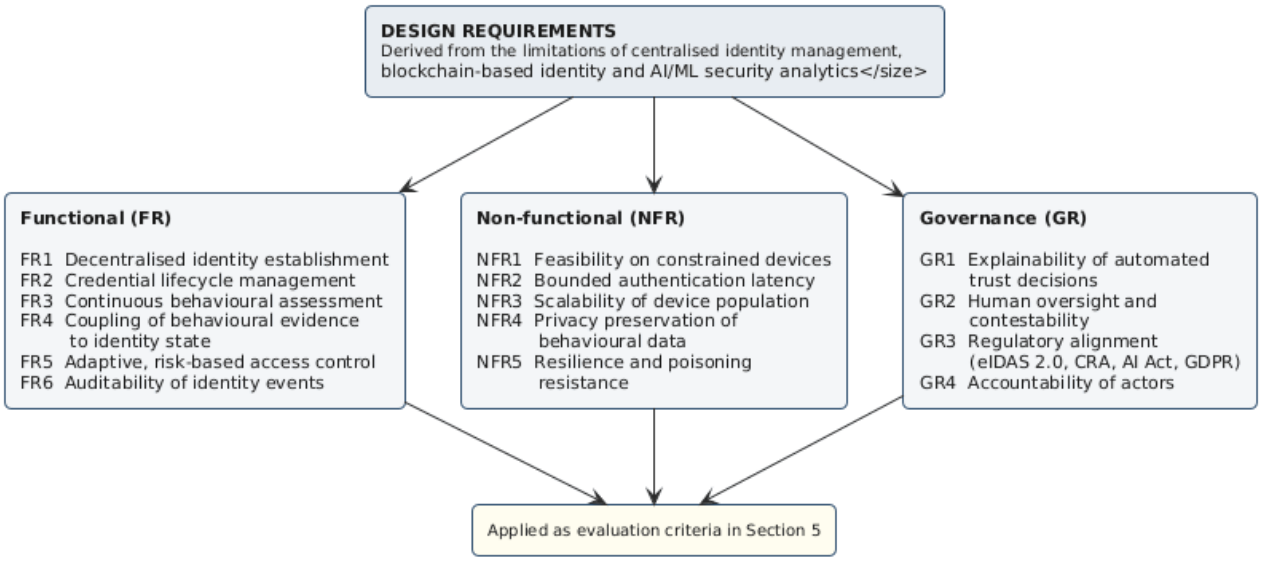
The knowledge base supporting the architecture was developed through a structured, non-systematic review designed to identify recurring constraints, technological dependencies, and governance requirements across the research streams contributing to the proposed design. The purpose was not to claim the comprehensive coverage expected from a formal systematic review, but to establish a defensible design knowledge base from which architectural requirements could be derived.

Survey and review studies were used where the argument depended on field-level limitations or technological trends, while primary studies were retained where the proposed architecture depends on a specific mechanism, security property, or reported performance characteristic. Sources were drawn from Scopus, Web of Science, IEEE Xplore, ScienceDirect, SpringerLink, and the ACM Digital Library, supplemented by targeted searches of publisher repositories for recent contributions. Search combinations covered blockchain and IoT identity, decentralised identifiers and self-sovereign identity, machine-learning-based intrusion detection, federated anomaly detection, blockchain-supported federated learning, and relevant governance and regulatory requirements.

Regulatory instruments and technical standards were obtained from official sources, including EUR-Lex and relevant standards bodies. Inclusion was limited to peer-reviewed journal and conference literature, recognised technical standards, and legislative instruments addressing at least one of the technological or governance dimensions relevant to constrained and distributed IoT environments. Performance-related evidence was used only where the underlying evaluation method was sufficiently described. Surveys and systematic reviews were preferred where claims were intended to represent broader field-level tendencies, whereas older studies were retained where they define concepts or mechanisms still required by the architecture.

The knowledge-based construction also has methodological limitations. Database coverage, indexing practices, language restrictions, and search-term selection may exclude relevant work. Because these limitations may affect both the requirements derived from the literature and the feasibility evidence used later in the study, they are explicitly retained as part of the methodological boundary rather than treated as evidence of systematic completeness.

Figure 2. Design Requirements Derived from the Limitations Identified in Section 2



Source: Generated/sketched by authors using PlantUML

3.3 Design Requirements

The design requirements are organised into three groups: functional requirements (FRs), non-functional requirements (NFRs), and governance requirements (GRs). This classification provides traceability between the problem identified in the literature, the architectural mechanisms proposed in Section 4, and the evaluation conducted in Section 5.

The functional requirements define the behaviour expected from the identity and trust-management control loop. FR1 requires device identities to be registered and verified through decentralised identifiers and verifiable credentials without making a single central operator the exclusive authority for verification. FR2 extends identity management across the credential lifecycle, including enrolment, renewal, delegation, suspension, revocation, and reinstatement. FR3 requires trust to be reassessed after enrolment through behavioural observation, while FR4 establishes an explicit pathway through which behavioural evidence can influence the trust state and permissions associated with a credentialled device. FR5 requires authorisation decisions to combine credential state, behavioural risk, and request context. FR6 requires identity-related events and consequential trust decisions to leave a tamper-resistant record that can subsequently be independently reviewed.

The non-functional requirements specify the operating conditions under which those functions remain credible. NFR1 limits device-side computation, memory, and energy requirements by allowing sustained computation and ledger interaction to be delegated to edge or gateway nodes where appropriate. NFR2 prevents ledger finality from becoming part of the routine authorisation path when this would create unacceptable latency. NFR3 addresses scalability by requiring growth in device population not to produce a proportionate increase in on-ledger transactions or computation concentrated at a single service. NFR4 requires raw behavioural telemetry to remain outside a central repository and prevents personal or identifying information from being unnecessarily committed to an immutable ledger. NFR5 addresses resilience to partial system failure and learning-layer poisoning by requiring controlled degradation and limits on the capacity of compromised participants to affect the shared model.

The governance requirements address accountability after a consequential trust decision has been produced. GR1 requires sufficient evidence to reconstruct the model, features, thresholds, and policy underlying an automated trust decision. GR2 requires suspension and revocation to include a review path, a recorded decision, and mechanisms for challenge and reinstatement. GR3 evaluates whether the

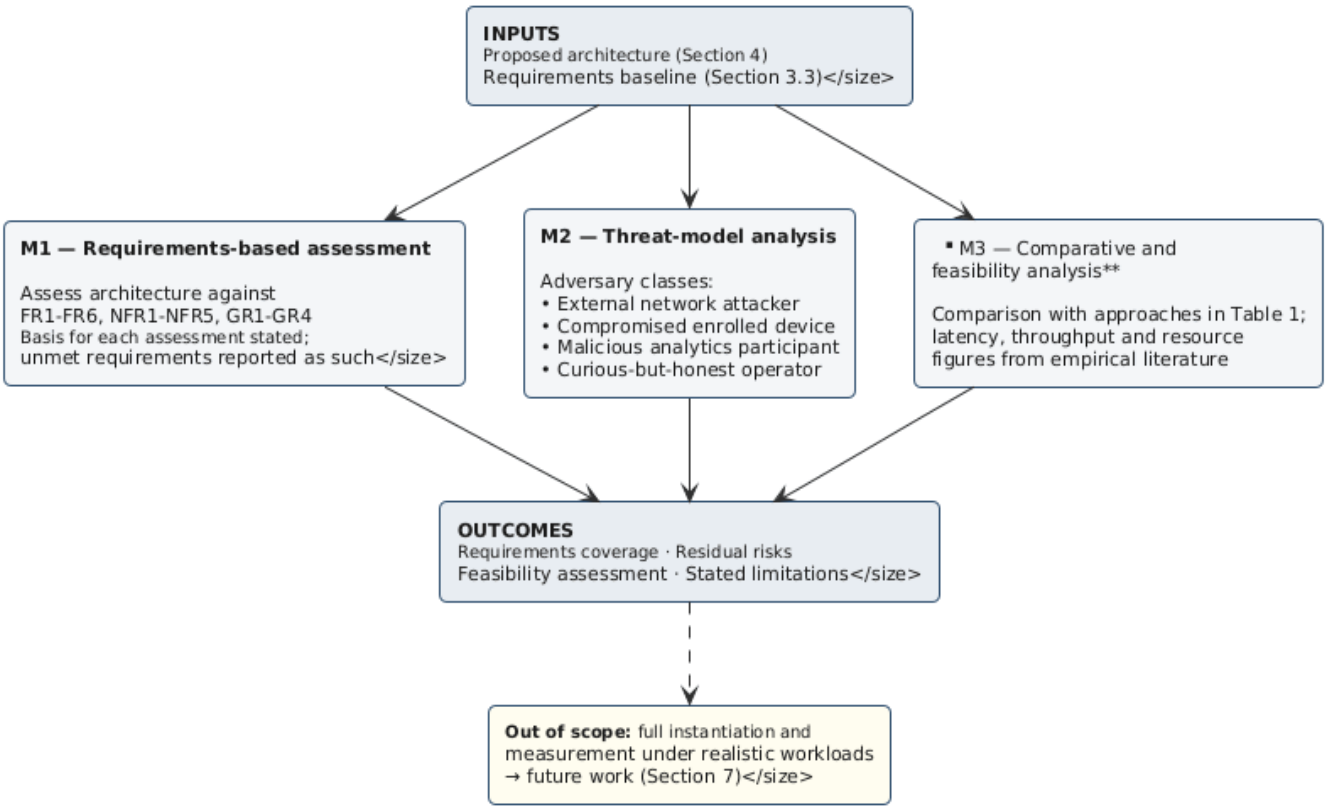
architecture can operate under obligations concerning digital identity, product cybersecurity, automated decision-making, and personal-data protection, including tensions between immutable auditability and erasure requirements. GR4 assigns each architectural responsibility to an identifiable actor so that technically traceable decisions do not become institutionally unaccountable.

Retaining the FR, NFR, and GR identifiers throughout the architecture creates a direct traceability chain from literature-derived problem → requirement → architectural mechanism → evaluation outcome. It also prevents broad claims such as “the architecture is secure” from substituting for explicit evidence regarding which requirements are addressed and which dependencies remain unresolved.

3.4 Evaluation Strategy

The evaluation strategy, summarized in Figure 3, is structured around three complementary methods: requirements-based assessment, threat-model analysis, and comparative and feasibility analysis. Together, these methods assess whether the proposed architecture provides an explicit mechanism for each identified requirement, how the design responds to defined threat conditions, and whether its principal implementation assumptions are technically plausible.

Figure 3. Analytical Evaluation Framework for the Proposed Architecture



Note: M1 evaluates the architecture against the functional (FR), non-functional (NFR), and governance (GR) requirements; M2 examines the architecture against four defined adversary classes; M3 assesses comparative positioning and implementation feasibility using evidence from the literature. Full implementation and empirical performance testing are outside the scope of the present study.

Source: Authors’ elaboration.

First, the requirements-based assessment determines whether the proposed architectural components provide a defensible design-level mechanism for each FR, NFR, and GR. A requirement is classified as only partially addressed where an unresolved dependency cannot be settled without implementation, empirical measurement, formal conformity assessment, or additional operational evidence. Second, the threat-model analysis examines four adversary classes across the identity,

analytics, and enforcement layers: an external attacker without a valid credential; a legitimately enrolled device compromised after enrolment; a malicious participant in the federated-learning layer; and a curious-but-honest infrastructure operator. These adversary classes are selected because they cross different trust boundaries and therefore reveal whether the architecture genuinely mitigates a risk or merely transfers an assumption from one technological layer to another. Third, the feasibility analysis serves as a boundary check rather than as an experiment on a fully implemented system. The proposed architecture is compared with alternative approaches and with reported latency, throughput, resource-consumption, and operational characteristics of its underlying mechanisms. Such evidence can indicate whether a design assumption is plausible, but it cannot establish the end-to-end performance of the proposed architecture.

Accordingly, the evaluation can identify internal contradictions, missing mechanisms, unresolved threat conditions, and implausible design assumptions. It cannot establish production-level performance, scalability, resource consumption, adversarial robustness, or regulatory conformity without a working implementation, realistic workloads, suitable behavioural datasets, and context-specific conformity assessment. These unresolved issues are therefore treated as explicit outputs of the evaluation and as inputs to the future empirical validation programme.

4. The Proposed Integrated Architecture

The proposed architecture follows from the technological and governance dependencies identified in Sections 2 and 3. It maintains identity state, behavioural evidence, and policy as separately inspectable sources of trust while integrating them at the point of authorisation. Five functional layers, device, edge, identity and ledger, intelligence, and policy and access control, implement the principal technical responsibilities, while a governance plane operates across all layers to preserve accountability, explainability, and human oversight. Figure 4 presents the principal components and interfaces.

This separation is deliberate because each source of trust evidence is subject to different failure conditions. Credentials may remain valid after device compromise, behavioural models may produce erroneous or manipulated classifications, and technically correct policy execution may still generate inappropriate outcomes if contextual information or review mechanisms are absent. The architecture therefore treats trust as a dynamic relationship between identity, behaviour, context, and accountable decision-making rather than as a permanent property established at enrolment.

4.1. Design Principles

Four principles guide the architecture. First, identity assertion and behavioural evidence remain distinct until authorisation. A verifiable credential establishes a cryptographically supported claim concerning device identity, whereas behavioural evidence reflects time-dependent operating conditions. The policy layer combines these evidence sources only when evaluating access.

Second, the ledger functions as an integrity and state anchor rather than a repository for raw telemetry. It records identifiers, credential states, policy versions, model attestations, and decision hashes, while detailed behavioural telemetry and explanatory evidence remain off-chain. This separation supports auditability without requiring potentially sensitive or erasable information to become permanently immutable.

Third, computationally intensive or frequently repeated operations, including ledger interaction, feature extraction, model inference, training, and aggregation, are delegated to gateways or edge nodes where endpoint capabilities are insufficient. Devices capable of secure local computation may nevertheless retain their own private keys and perform selected operations locally. This tiered approach avoids assuming homogeneous computational capabilities across heterogeneous IoT fleets.

Fourth, the architecture establishes explicit boundaries for automated authority. Behavioural evidence may trigger enhanced monitoring or temporary restriction of sensitive permissions. Suspension initiates a formal review process, while permanent credential revocation requires explicit human authorisation. The architecture therefore permits rapid automated containment while preventing behavioural classification alone from permanently invalidating an identity credential. These principles operationalise the requirements established in Section 3.3.

4.2. Architectural Overview

Five functional layers implement these principles, with a governance plane extending across them. Figure 4 illustrates the architecture and the principal information and control flows.

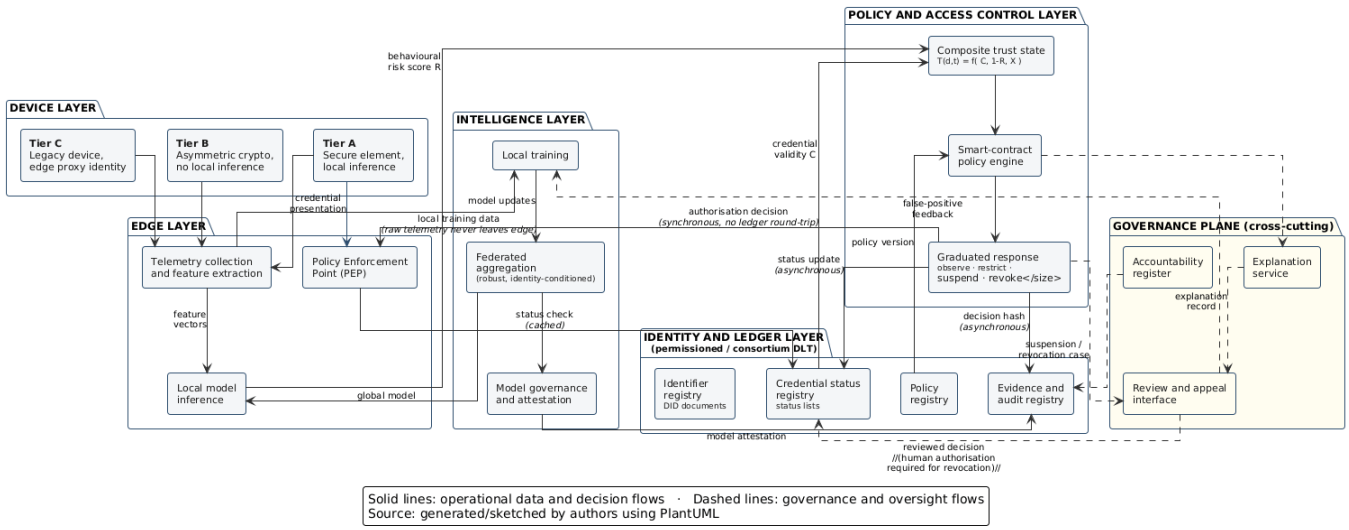
Device Layer

The device layer comprises the endpoints whose identities, behaviour, and permissions are managed. Where hardware permits, a device maintains its decentralised identifier and private key locally, preferably within a Trusted Platform Module (TPM), secure enclave, or physically unclonable function (PUF) (Herder et al., 2014), and presents a verifiable credential when authentication is required. The same endpoint generates the telemetry from which behavioural features are derived.

Identity credentials and behavioural telemetry perform different evidentiary functions. The credential associates the endpoint with an asserted identity and issuer chain, whereas telemetry provides information about its current operating state. A valid credential is therefore necessary for access but is not treated as sufficient evidence of continuing trustworthiness.

To accommodate heterogeneous IoT hardware, the architecture distinguishes three capability tiers. Tier A devices can protect cryptographic keys, perform asymmetric cryptography, and execute local inference. Tier B devices can securely manage keys and credentials but rely on an edge node for feature extraction and behavioural inference. Tier C comprises legacy or severely constrained endpoints that cannot securely maintain their own cryptographic keys; these devices are represented through proxy identities at the associated edge node and monitored externally. Tier C consequently provides the weakest binding between physical device and credential. This limitation is retained explicitly under NFR1 rather than obscured through an assumption that all deployed devices possess contemporary security hardware.

Figure 4. Integrated Blockchain–AI Architecture for Decentralised IoT Identity Management



Edge Layer

The edge layer provides the operational interface between constrained devices and the remaining architectural components. Gateways, fog nodes, or edge servers perform protocol translation, telemetry collection, feature extraction, local behavioural inference, and policy enforcement. Locating these functions close to the device reduces the need to transmit raw operational data and enables authorisation decisions to be executed without requiring continuous interaction with remote infrastructure.

A central design decision is to exclude ledger finality from the synchronous authorisation path. Routine requests are evaluated using locally cached credential state, policy versions, and trust information rather than waiting for a ledger transaction to reach finality. This improves latency predictability and allows limited operation during temporary ledger or network unavailability.

This design introduces a controlled consistency trade-off. Cached credential or policy information may temporarily lag behind the authoritative ledger state. The architecture therefore requires a defined distribution mechanism for credential and policy updates, bounded cache validity, and subsequent reconciliation of consequential decisions against the authoritative state. The design does not claim zero-latency global consistency; instead, it prioritises bounded authorisation latency while explicitly recognising temporary state staleness as a residual operational risk.

Operational responsibility remains with the edge node processing the request. Although identity state is distributed, the edge component must evaluate the available evidence, apply the relevant policy, execute the resulting authorisation decision, and retain the information necessary for subsequent audit.

Identity and Ledger Layer

The identity and ledger layer maintains the independently verifiable states required by the architecture. A permissioned consortium ledger is used because the intended deployment environment involves known credential issuers, validators, and governance participants rather than unrestricted anonymous participation. Permissioned blockchain platforms such as Hyperledger Fabric illustrate how consortium-based identity governance can support controlled participation, distributed validation, and auditable transaction processing (Androulaki et al., 2018).

This choice represents a governance as well as a technical decision. Validation rights, participant admission, issuer requirements, and rule-setting authority are distributed among pre-approved institutions. Consequently, decentralisation is treated as a governed distribution of authority rather than as the absence of institutional control. The security of the architecture therefore depends not only on cryptographic integrity but also on the governance of consortium membership and validation privileges.

Only information requiring independent verification is committed to the ledger. The identity registry associates DIDs with public keys and relevant service information, while credential-state records indicate whether credentials are active, suspended, or revoked. Status-list or accumulator mechanisms may be used to distribute credential-state changes without requiring an independent ledger transaction for every verification event.

Policy versions and trust thresholds are versioned so that historical decisions can be evaluated against the rules applicable at the time they were made. Similarly, the evidence registry stores cryptographic commitments to decision records while detailed supporting evidence remains off-chain. This enables subsequent integrity verification without placing raw behavioural or explanatory data permanently on the ledger.

Credential issuers are themselves identifiable entities within the architecture, allowing credentials to be traced to their issuing authority. The ledger records these formal relationships but does not independently establish the institutional competence or legal compliance of an issuer. Trust in consortium

governance therefore remains an explicit architectural assumption rather than an automatically guaranteed property of the ledger.

Intelligence Layer

The intelligence layer provides continuous behavioural assessment without requiring raw telemetry from all devices to be consolidated in a central repository. Behavioural data remain at the device or edge environment where they are generated, while feature extraction, inference, and, where appropriate, local model training are performed close to the source.

Federated learning enables participating nodes to contribute model updates to a shared behavioural model without transmitting their underlying raw telemetry. Each deployed model version is associated with an attestation containing information such as the model hash, training round, participating clients, and relevant evaluation metrics. Preserving model-version information is important for auditability because a historical trust decision must be reconstructed using the model that generated it rather than a subsequently retrained version. The learning layer nevertheless introduces its own attack surface. A participant may possess a legitimate credential while submitting malicious or manipulated model updates. Ledger-backed identity therefore provides attribution and participation control: only nodes with valid institutional credentials may contribute to federated learning, and submitted updates can be associated with identifiable participants. Revocation or suspension of the relevant credential removes the corresponding participation authority.

Identity verification alone cannot prevent poisoning by legitimate but malicious participants. Model updates therefore require additional screening, and aggregation mechanisms should be robust to anomalous or outlying contributions (Blanchard et al., 2017; Yin et al., 2018). These controls increase resistance to Sybil attacks and basic poisoning strategies but do not eliminate coordinated attacks by multiple credentialled participants whose updates remain statistically plausible. This residual risk explains why NFR5 is only partially addressed at design level and requires empirical adversarial evaluation.

Policy and Access Control Layer

The policy and access-control layer integrates three sources of evidence: credential validity, behavioural risk, and request context. Credential state establishes whether the requesting device retains a valid identity relationship; behavioural risk reflects current evidence concerning operational deviation; and context captures factors such as resource sensitivity and current system conditions.

The architecture represents these inputs through a dynamic trust function:

$$T(d,t) = f(C(d,t), 1 - R(d,t), X(d,t)) \quad (1)$$

where, the trust state of device *d* at time *t* is determined by credential validity, behavioural risk, and contextual information.

The formulation is intentionally general because the study does not claim that a single weighting scheme is appropriate across all IoT deployment contexts. Instead, Equation (1) specifies the variables and relationships that an implementation must operationalise, while weights and thresholds remain deployment-specific. Credential validity acts as a gating condition. Once a credential is revoked, its associated trust state cannot be restored through favourable behavioural or contextual evidence. While the credential remains valid, however, behavioural and contextual evidence may dynamically narrow or expand the permissions available to the device. The architecture defines four graduated response levels:

Observation → Restriction → Suspension → Revocation

Observation increases monitoring and evidence retention without materially restricting normal operation. Restriction removes selected sensitive permissions while preserving limited functionality.

Suspension changes the credential state and initiates formal review. Revocation permanently terminates the credential relationship and requires explicit human authorisation.

This graduated mechanism allows the architecture to respond rapidly to suspected compromise without treating every anomalous classification as sufficient ground for irreversible identity action.

Governance Plane

The governance plane provides accountability across the technical layers. Its central requirement is that consequential decisions must be reconstructable using the model, policy, evidence, and institutional authority applicable at the time the decision was made. Historical decisions should therefore not be reassessed solely using current models or subsequently modified policies.

For each consequential decision, the architecture retains the relevant model version, policy version, thresholds, explanatory evidence, and decision rationale. Detailed records remain off-chain, while cryptographic commitments are anchored in the evidence registry to enable subsequent integrity verification. This approach combines auditability with the avoidance of unnecessary permanent storage of potentially sensitive evidence. Suspension and revocation cases enter a defined review and appeal process. Reviewers must act under identifiable institutional credentials and record the rationale for their decisions. Affected parties must have a defined mechanism through which a decision may be challenged and, where appropriate, permissions reinstated.

Review outcomes may also provide feedback to the intelligence layer. Where a behavioural classification is subsequently determined to have been erroneous, the case can be labelled and incorporated into model-governance processes to support future model improvement. Such feedback does not automatically modify the model; it becomes governed evidence for subsequent retraining and validation. Institutional responsibilities for credential issuance, edge operation, model management, policy definition, and review are assigned separately. This separation reduces ambiguity concerning accountability and ensures that technical traceability is accompanied by identifiable organisational responsibility.

4.3. Interaction Flows

The architecture can be understood through four principal interaction flows: enrolment, routine authorisation, behavioural escalation, and federated model updating.

During enrolment, a capable device generates a cryptographic key pair and stores the private key in protected hardware where available. Following verification of the device and its provenance, an authorised issuer provides a verifiable credential, registers the corresponding DID information, and establishes the initial credential state. For Tier C devices, proxy identity is established through the associated edge node. Following enrolment, behavioural monitoring begins. Where available, manufacturer-defined communication profiles, such as the Manufacturer Usage Description (MUD) specification, may provide an additional reference for expected device network behaviour (Lear et al., 2019). New devices may initially operate under more restrictive policies until sufficient observations are available to establish an appropriate behavioural baseline. This monitoring continues throughout the device lifecycle rather than terminating after identity registration.

During routine authorisation, the edge node verifies the presented credential and checks the locally available credential state. If these checks succeed, the policy layer combines credential validity, behavioural risk, and request context to determine the applicable trust state and permissions. The resulting decision is enforced locally, while evidence requiring permanent integrity verification may subsequently be anchored to the ledger. Routine access therefore does not wait for ledger finality.

During behavioural escalation, telemetry exceeding predefined thresholds is transformed into behavioural features and evaluated by the local model. The resulting risk assessment updates the behavioural component of the trust state and may trigger enhanced observation or restriction. More consequential conditions may initiate suspension and formal review, while permanent revocation remains dependent on human authorisation.

During federated model updating, only edge nodes with valid institutional credentials may contribute updates. Submitted contributions are attributable to their source, screened before aggregation, and incorporated through a robust aggregation process. Each approved global model version is subsequently associated with verifiable model-version evidence. If a suspension is overturned during review, the credential may be reinstated and the reviewed case may be retained as governed feedback for subsequent model improvement.

Together, these flows create the bidirectional relationship central to the architecture: identity determines accountable participation and access, while behavioural evidence can dynamically modify the authority associated with a valid identity.

4.4. Requirement Coverage

Table 2 provides the traceability mapping between the architectural components described above and the requirements established in Section 3.3. This mapping is subsequently used in Section 5 to distinguish requirements addressed through explicit architectural mechanisms from those that remain dependent on hardware capability, deployment scale, adversarial behaviour, or formal regulatory conformity assessment.

Table 2. Mapping of architectural components to design requirements

Component	Requirements addressed
Device layer (DIDs, key custody, capability tiers)	FR1, NFR1
Edge layer (feature extraction, inference, enforcement)	FR3, NFR1, NFR2
Identifier and credential status registries	FR1, FR2, NFR3
Policy registry	FR5, GR3
Evidence and audit registry	FR6, GR4
Federated learning and aggregation	FR3, NFR4, NFR5
Identity-conditioned participation in aggregation	NFR5, GR4
Model governance and attestation	GR1, GR4
Composite trust state and graduated response	FR4, FR5, GR2
Explanation service	GR1
Review and appeal interface	GR2, GR3
Accountability register	GR3, GR4

Source: compiled by the authors.

5. Evaluation and Limitations

The proposed architecture is evaluated using the three complementary methods defined in Section 3.4: requirements-based assessment, threat-model analysis, and comparative and feasibility analysis. The purpose is to determine whether the architecture provides explicit mechanisms for the identified requirements, whether those mechanisms remain defensible across the defined trust boundaries, and whether the principal implementation assumptions are plausible in light of existing evidence.

The evaluation remains analytical and design-level. No experimental measurements are attributed to the proposed architecture. Accordingly, requirement coverage demonstrates architectural support rather

than empirical validation, while unresolved dependencies related to hardware, scale, adversarial behaviour, and regulatory conformity are explicitly retained as limitations.

5.1. Requirements-Based Assessment

Table 3 evaluates the proposed architecture against the 15 requirements defined in Section 3.3. A requirement is classified as Met when the architecture specifies a concrete design-level mechanism through which it can be addressed. Partially met indicates that the architecture provides a relevant mechanism but that full satisfaction remains dependent on implementation, hardware capabilities, deployment scale, adversarial conditions, or formal conformity assessment. This distinction prevents the presence of an architectural component from being interpreted as empirical proof that the corresponding requirement has been fully satisfied.

Table 3: Assessment of the proposed architecture against the design requirements

Req.	Status	Basis and residual concern
FR1	Met	DIDs and verifiable credentials registered on a distributed ledger remove dependence on a single issuing authority.
FR2	Met	Status-list and accumulator constructions support enrolment, suspension, revocation, and reinstatement without per-credential transactions.
FR3	Met	Edge-hosted inference over device behavioural profiles provides post-enrolment assessment.
FR4	Met	The composite trust state binds behavioural risk to credential state through the policy engine.
FR5	Met	Authorization is a function of credential validity, behavioural risk, and context, with graduated outcomes.
FR6	Met	Identity events and decision hashes are anchored on the ledger; policy and model versions are independently verifiable.
NFR1	Partially met	Capability tiers keep device-side operations minimal, but Tier C devices depend on an edge proxy identity, which weakens the binding between physical device and credential.
NFR2	Met	Authorization is evaluated at the edge against cached state; ledger interaction is asynchronous and outside the request path.
NFR3	Partially met	Anchoring rather than storage and batched status publication limit on-ledger growth, but federated aggregation cost grows with the number of participating edge nodes.
NFR4	Met	Raw telemetry remains at the edge; no personal or behavioural data is written to the ledger.
NFR5	Partially met	Robust aggregation and identity-conditioned participation raise the cost of poisoning without eliminating it; a colluding set of validly credentialled edge nodes remains a residual threat.
GR1	Met	Explanation records accompany each consequential assessment and are hash-anchored.
GR2	Met	Graduated response reserves suspension for review and revocation for human authorisation, with a defined reinstatement route.
GR3	Partially met	Anchoring rather than storage reconciles auditability with erasure rights, but conformity with specific certification regimes cannot be established analytically.
GR4	Met	Issuer chains and the accountability register attribute each function to an identifiable actor.

Source: compiled by the authors.

Overall, 11 of the 15 requirements are assessed as met at design level, while four—NFR1, NFR3, NFR5, and GR3, remain partially met. The distribution of these results is significant because the unresolved requirements occur primarily where architectural specification reaches the boundary of what can be established without implementation. NFR1 remains partial because Tier C devices depend on proxy identity and therefore provide weaker physical-device-to-credential binding. NFR3 remains partial because federated coordination costs increase with the number of participating edge nodes. NFR5 remains partial because identity-conditioned participation and robust aggregation cannot eliminate coordinated poisoning by legitimately credentialled participants. GR3 remains partial because regulatory conformity cannot be

established through architectural analysis alone. These four requirements therefore define specific targets for subsequent prototype and empirical evaluation rather than minor implementation details assumed to be resolved by the design.

5.2 Threat-Model Analysis

The threat-model analysis examines four adversary classes, each targeting a different assumption within the proposed control loop: an external attacker without a valid credential, a legitimately enrolled device compromised after enrolment, a malicious participant in the federated-learning layer, and a curious-but-honest infrastructure operator.

External attacker. An external attacker without a valid credential is expected to fail at the identity-verification boundary. Challenge-response mechanisms limit straightforward replay attacks, while ledger-backed credential verification provides a mechanism for detecting invalid or forged identity claims. This protection does not, however, eliminate availability attacks. Because authorisation and policy enforcement are performed at edge nodes, denial-of-service attacks against those nodes remain a residual threat. The architecture therefore reduces unauthorised access risk but does not claim to eliminate service-disruption risk.

Compromised enrolled device. This adversary provides the principal test of the study's architectural proposition because the device continues to possess a valid credential after compromise. Where compromise produces observable deviations from the established behavioural profile, the intelligence layer increases the associated risk state and the policy layer can progressively restrict authority without waiting for permanent credential revocation.

This mechanism nevertheless depends on behavioural observability. An attacker capable of reproducing behaviour within the learned baseline may remain undetected. A second risk arises during initial enrolment, when insufficient legitimate observations may allow a compromised device to influence its own behavioural baseline. The proposed probationary registration period reduces the latter risk by restricting authority while the baseline is being established, but behavioural mimicry remains an unresolved limitation of anomaly-based detection.

Malicious federated-learning participant. A legitimately credentialled participant may exploit its authorised position to submit malicious model updates. Ledger-backed identity provides attribution, enables participation rights to be suspended, and creates an auditable association between updates and their submitters. Update screening and robust aggregation provide additional protection against anomalous contributions.

These controls increase accountability and raise the cost of straightforward poisoning attacks, but they do not eliminate coordinated attacks by multiple validly credentialled participants whose submissions remain statistically plausible. The analysis therefore supports the partial classification of NFR5: decentralised identity strengthens accountability within federated learning but does not by itself solve the underlying poisoning problem.

Curious-but-honest infrastructure operator. Keeping raw telemetry off-chain limits direct access to detailed behavioural records. Nevertheless, metadata visible within shared infrastructure, including identifiers, credential-state changes, decision commitments, and their timestamps, may reveal interaction patterns even where substantive content remains inaccessible. Selective disclosure can reduce unnecessary attribute exposure but cannot necessarily conceal the existence or timing of an interaction.

The architecture therefore provides data-minimisation and decentralisation benefits rather than complete transactional privacy. This distinction is important because the privacy claim should remain limited to what the design actually supports.

Taken together, the four adversary cases indicate that the architecture does not eliminate security risk but redistributes and makes several trust dependencies explicit. Identity mechanisms constrain unauthorised participation, behavioural evidence provides a mechanism for responding to post-enrolment compromise, and governance mechanisms preserve accountability for consequential actions. Residual risks remain concentrated in-service availability, behavioural mimicry and baseline contamination, coordinated federated-learning poisoning, and metadata inference.

5.3 Comparative and Feasibility Analysis

The comparative analysis uses Table 1 to examine the complementary capabilities and dependencies of the approaches reviewed rather than to claim universal superiority of the proposed architecture. Blockchain-based identity provides verifiable credential state and auditability but lacks post-enrolment behavioural evidence. Centralised machine-learning approaches provide such behavioural evidence but require concentration of operational data. Federated learning reduces the need for centralised telemetry storage but depends on reliable participant identity and accountability.

The proposed architecture integrates these complementary capabilities by using ledger-backed identity to condition participation in federated learning and by feeding behavioural evidence back into identity-dependent access control. Its principal contribution is therefore not the introduction of an individually novel technology but the bidirectional integration of identity and behavioural trust within an accountable authorisation process. This integration nevertheless introduces additional coordination and governance costs. The architecture should therefore be interpreted as a structured trade-off rather than as a design that eliminates the limitations of its constituent technologies.

Available performance evidence for the underlying mechanisms provides only a feasibility boundary. Moving ledger finality outside the routine authorisation path removes an identifiable latency dependency, while existing studies indicate that edge-based behavioural inference can operate within latency ranges compatible with selected IoT applications. However, such results derive from component-level studies and cannot be extrapolated directly to the end-to-end performance of the proposed architecture.

A complete implementation must simultaneously accommodate credential lifecycle management, key custody, policy distribution, federated model coordination, behavioural inference, audit evidence, fault recovery, and human review. The interaction of these functions under heterogeneous workloads has not been empirically measured in the present study. Consequently, existing evidence supports the technical plausibility of the architecture, but not claims regarding its production-level latency, throughput, scalability, resource consumption, or operational cost.

5.4 Limitations

The principal limitation of this study derives from its architectural rather than empirical scope. Analytical evaluation can identify internal inconsistencies, determine whether explicit mechanisms correspond to stated requirements, and expose residual dependencies. It cannot establish end-to-end latency, throughput, computational and energy consumption, fault behaviour, or production-level scalability. Similarly, feasibility evidence derived from existing component-level studies inherits the assumptions, datasets, hardware configurations, and experimental limitations of those studies.

A second limitation concerns the permissioned consortium model. The proposed ledger distributes authority among pre-approved participants but does not eliminate institutional trust. Consortium membership, validation rights, credential issuance, and governance rules remain dependent on identifiable organisations. The architecture therefore redistributes trust rather than providing complete disintermediation.

Third, behavioural trust is inherently constrained by observability and model quality. Legitimate devices whose behaviour deviates from learned norms may generate false positives, while attackers

capable of mimicking expected behaviour may remain below detection thresholds. Baseline contamination during early device operation presents an additional risk. Graduated responses and probationary enrolment can reduce the operational consequences of these limitations but cannot eliminate them.

Fourth, the federated-learning layer remains vulnerable to coordinated poisoning by legitimate participants. Identity-conditioned participation improves attribution and accountability, while robust aggregation can reduce the influence of anomalous updates; neither mechanism guarantees protection against coordinated, statistically plausible malicious contributions.

Fifth, regulatory conformity remains context-dependent. The architecture incorporates mechanisms intended to support auditability, data minimisation, explainability, human oversight, and contestability, but technical design alone cannot establish compliance with the GDPR, AI Act, Cyber Resilience Act, eIDAS framework, or sector-specific requirements. Formal conformity depends on implementation details, institutional responsibilities, data-processing practices, and deployment context.

Finally, the knowledge base supporting the design was constructed through a structured but non-systematic review. Database coverage, indexing practices, language restrictions, and search-term selection may therefore have excluded relevant studies. This limitation affects both the requirements derived from the literature and the component-level evidence used in the feasibility analysis. These limitations define the empirical validation agenda for the architecture: prototype implementation on heterogeneous IoT hardware; end-to-end latency, throughput, resource, and scalability testing; adversarial evaluation against behavioural mimicry and coordinated federated-learning poisoning; evaluation of false-positive and false-negative behaviour; assessment of reviewer workload and appeal mechanisms; and deployment-specific regulatory conformity analysis.

6. Conclusion

This paper addressed a structural weakness in IoT security: identity management remains predominantly centralized and, where decentralized, remains static. Credentials establish what a device claims to be at enrolment but say nothing about how it behaves thereafter, a failure mode repeatedly demonstrated in practice by legitimately enrolled devices participating in attacks while presenting valid credentials.

The analysis established that the two relevant research streams fail in mirror-image ways. Blockchain-based identity provides verifiable, auditable decentralization without behavioural awareness; machine learning provides behavioural awareness but requires either centralized aggregation, which reinstates the risk concentration decentralization was meant to remove, or federated architectures dependent on strong client identity they do not supply.

Responding to this complementarity, the paper proposed an architecture in which a permissioned ledger anchors identifiers, credential status, policy versions, and decision evidence; a federated learning layer at the edge provides continuous behavioural assessment without centralizing telemetry; and a policy layer combines credential validity, behavioural risk, and context into a composite trust state driving graduated authorization. Two features distinguish the design: the integration is bidirectional, since ledger-verified identity conditions participation in federated aggregation; and a governance plane ensures consequential decisions are explainable, contestable, and, for revocation, subject to human authorization.

The evaluation assessed eleven of the fifteen requirements as met at design level, while four, NFR1, NFR3, NFR5, and GR3, remain partially met because their full satisfaction depends on implementation, deployment scale, adversarial conditions, or formal conformity assessment.

The contribution is architectural rather than empirical, specifying how individually validated technologies should be composed to address a problem none resolves alone. Four directions follow from the limitations identified: prototype instantiation and measurement on a heterogeneous testbed;

adversarial evaluation of the intelligence layer against coordinated poisoning and mimicry; empirical study of the governance plane as a human system, including reviewer workload and false-positive feedback; and examination of the institutional preconditions for adoption, including consortium governance, credential standardization, and liability allocation.

The broader claim, however, does not depend on that validation. Trust in connected environments cannot be established once and assumed thereafter. Identity must be verifiable, behaviour must be observed, the two must be bound together, and the decisions that follow must remain answerable to the people they affect.

CRediT Authorship Contribution Statement

Mihăilescu, M. I.: Conceptualisation, methodology, visualisation, writing – original draft, supervision.

Niță, S. L.: Literature review, formal analysis, validation, writing – review and editing.

Mărăscu, V.: Methodology, formal analysis, validation, interpretation, writing – review and editing.

Acknowledgments

This research received no external funding.

Conflict of Interest Statement

The authors declare no commercial or financial conflicts of interest. Mihăilescu, M. I. is an Associate Editor of JETS, while Niță, S. L. and Mărăscu, V. are members of the International Advisory Board of JETS. All three authors were excluded from all editorial decisions related to this manuscript, which was handled independently in accordance with the journal's peer-review and editorial procedures.

Data Availability Statement

No new data were created or analysed in this study. Data sharing is not applicable to this article.

Ethical Approval Statement

This study is based exclusively on the analysis of previously published literature. Therefore, ethical approval was not required.

References

Ali, S., Li, Q., & Yousafzai, A. (2024). Blockchain and federated learning-based intrusion detection approaches for edge-enabled industrial IoT networks: A survey. *Ad Hoc Networks,* 152(C), 103320. https://doi.org/10.1016/j.adhoc.2023.10332

Allen, C. (2016). The path to self-sovereign identity. Life With Alacrity. http://www.lifewithalacrity.com/2016/04/the-path-to-self-soverereign-identity.html

Belenguer, A., Pascual, J. A., & Navaridas, J. (2025). A review of federated learning applications in intrusion detection systems. *Computer Networks,* Volume 258, 111023. https://doi.org/10.1016/j.comnet.2024.111023

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Dorri, A., Kanhere, S. S., Jurdak, R., & Gauravaram, P. (2017). Blockchain for IoT security and privacy: The case study of a smart home. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops* (PerCom Workshops) (pp. 618–623). IEEE. https://doi.org/10.1109/PERCOMW.2017.7917634

Hernandez-Ramos, J. L., Karopoulos, G., Chatzoglou, E., Kouliaridis, V., Marmol, E., Gonzalez-Vidal, A., & Kambourakis, G. (2025). Intrusion detection based on federated learning: A systematic review. *ACM Computing Surveys*, 57(12), 1–65. https://doi.org/10.1145/3731596

Hussain, F., Hussain, R., Hassan, S. A., & Hossain, E. (2020). Machine learning in IoT security: Current solutions and future challenges. *IEEE Communications Surveys & Tutorials*, 22(3), 1686–1721. https://doi.org/10.1109/COMST.2020.2986444

Khacha, A., Aliouat, Z., Harbi, Y., Gherbi, C., Saadouni, R., & Harous, S. (2024). Landscape of learning techniques for intrusion detection system in IoT: A systematic literature review. *Computers and Electrical Engineering*, 120, 109725. https://doi.org/10.1016/j.compeleceng.2024.109725

Khan, M. A., & Salah, K. (2018). IoT security: Review, blockchain solutions, and open challenges. *Future Generation Computer Systems*, 82, 395–411. https://doi.org/10.1016/j.future.2017.11.022

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of AISTATS* (pp. 1273–1282). PMLR. https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf

Mühle, A., Grüner, A., Gayvoronskaya, T., & Meinel, C. (2018). A survey on essential components of a self-sovereign identity. *Computer Science Review,* 30, 80–86. https://doi.org/10.1016/j.cosrev.2018.10.002

Sicari, S., Rizzardi, A., Grieco, L. A., & Coen-Porisini, A. (2015). Security, privacy and trust in Internet of Things: The road ahead. *Computer Networks,* 76, 146–164. https://doi.org/10.1016/j.comnet.2014.11.008

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K., Sorniotti, A., Stathakopoulou, C., Vukolić, M., Cocco, S. W., & Yellick, J. (2018). Hyperledger Fabric: A distributed operating system for permissioned blockchains. In Proceedings of the Thirteenth EuroSys Conference (Article 30, pp. 1–15). ACM. https://doi.org/10.1145/3190508.3190538

Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J. A., Invernizzi, L., Kallitsis, M., Kumar, D., Lever, C., Ma, Z., Mason, J., Menscher, D., Seaman, C., Sullivan, N., Thomas, K., & Zhou, Y. (2017). Understanding the Mirai botnet. In Proceedings of the 26th USENIX Security Symposium (pp. 1093–1110). USENIX Association. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, 108, 2938–2948. https://proceedings.mlr.press/v108/bagdasaryan20a.html

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in *Neural Information Processing Systems*, 30, 119–129. https://papers.nips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175–1191). ACM. https://doi.org/10.1145/3133956.3133982

Douceur, J. R. (2002). The Sybil attacks. In P. Druschel, F. Kaashoek, & A. Rowstron (Eds.), Peer-to-Peer Systems. *Lecture Notes in Computer Science,* Vol. 2429, 251–260. Springer. https://doi.org/10.1007/3-540-45748-8_24

ETSI. (2020). Cyber security for consumer Internet of Things: Baseline requirements (ETSI EN 303 645 V2.1.1). European Telecommunications Standards Institute. https://cdn.standards.iteh.ai/samples/57991/dfada3bbc3e94fa1844b1c1b0e8477be/ETSI-EN-303-645-V2-1-1-2020-06-.pdf

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

European Parliament and Council of the European Union. (2024a). Regulation (EU) 2024/1183 of 11 April 2024 amending Regulation (EU) No 910/2014 as regards establishing the European Digital Identity Framework. OJ L, 2024/1183, 30.4.2024. https://eur-lex.europa.eu/eli/reg/2024/1183/oj/eng

European Parliament and Council of the European Union. (2024b). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). OJ L, 2024/1689, 12.7.2024. https://artificialintelligenceact.eu/the-act/

European Parliament and Council of the European Union. (2024c). Regulation (EU) 2024/2847 of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). OJ L, 2024/2847, 20.11.2024. https://eur-lex.europa.eu/EN/legal-content/summary/horizontal-cybersecurity-requirements-for-products-with-digital-elements-cyber-resilience-act.html

Ferdous, M. S., Chowdhury, F., & Alassafi, M. O. (2019). In search of self-sovereign identity leveraging blockchain technology. *IEEE Access*, 7, 103059–103079. https://doi.org/10.1109/ACCESS.2019.2931173

Fung, C., Yoon, C. J. M., & Beschastnikh, I. (2020). The limitations of federated learning in Sybil settings. In 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020) (pp. 301–316). USENIX Association. https://www.usenix.org/conference/raid2020/presentation/fung

Herder, C., Yu, M.-D., Koushanfar, F., & Devadas, S. (2014). Physical unclonable functions and applications: A tutorial. *Proceedings of the IEEE*, 102(8), 1126–1141. https://doi.org/10.1109/JPROC.2014.2320516

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. https://doi.org/10.2307/25148625

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210. https://doi.org/10.1561/2200000083

Kolias, C., Kambourakis, G., Stavrou, A., & Voas, J. (2017). DDoS in the IoT: Mirai and other botnets. *Computer*, 50(7), 80–84. https://doi.org/10.1109/MC.2017.20

Lear, E., Droms, R., & Romascanu, D. (2019). Manufacturer usage description specification (RFC 8520). Internet Engineering Task Force. https://doi.org/10.17487/RFC8520

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Shabtai, A., Breitenbacher, D., & Elovici, Y. (2018). N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, 17(3), 12–22. https://doi.org/10.1109/MPRV.2018.0336773

Nguyen, T. D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., & Sadeghi, A.-R. (2019). DÏoT: A federated self-learning anomaly detection system for IoT. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)* (pp. 756–767). https://doi.org/10.1109/ICDCS.2019.00080

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207

Sporny, M., Longley, D., Sabadello, M., Reed, D., Steele, O., & Allen, C. (2022). Decentralized identifiers (DIDs) v1.0 (W3C Recommendation, 19 July 2022). World Wide Web Consortium. https://www.w3.org/TR/did-core/

Sporny, M., Longley, D., Chadwick, D., & Herman, I. (2025). Verifiable credentials data model v2.0 (W3C Recommendation, 15 May 2025). World Wide Web Consortium. https://www.w3.org/TR/vc-data-model-2.0/

Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. In Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80, pp. 5650–5659). PMLR. https://proceedings.mlr.press/v80/yin18a.html