Governing Autonomous AI Agents in Complex Digital Societies: A Mesh-Based Framework for Trust, Privacy, and Accountability

Bambang Setiawan[1], Dewi Anggraini[2], Andi Pratama[3,*]

[1] Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia
[2] Faculty of Engineering, Universitas Pasundan, Bandung, Indonesia
[3] Faculty of Engineering and Computer Science, Universitas Komputer Indonesia, Bandung, Indonesia

* Email: andi.pratama@email.unikom.ac.id (Corresponding Author)

Abstract

The rapid deployment of autonomous artificial intelligence (AI) agents powered by large language models (LLMs) into complex digital societies has exposed a widening gap between technical capability and institutional readiness. Recent operational incidents in which autonomous coding assistants destroyed production data, leaked regulated records, or executed irreversible commands beyond their intended scope demonstrate that today's centralised orchestration patterns are structurally incapable of providing the safety, auditability, and jurisdictional separation that real organisations require. This paper proposes a mesh-based governance framework that treats every domain in a multi-organisational ecosystem as an autonomous, fault-isolated node coordinating with peers through verifiable interfaces rather than through a central hub. The framework is structured around three pillars: trust, operationalised through verifiable provenance and reputation scoring; privacy, operationalised through federated learning, differential privacy and secure aggregation; and accountability, operationalised through immutable audit ledgers and policy enforcement at every node boundary. We present a reference architecture, an end-to-end operational workflow, and a synthetic evaluation showing that the proposed framework increases the mean trust score of cross-domain transactions from 0.31 to 0.79 and preserves 74.5% downstream task accuracy under a strict privacy budget of $\varepsilon = 0.1$, compared with 65.0% for an equivalent centralised configuration. The contribution lies in unifying the previously isolated literatures on data mesh, federated learning, and AI governance into a single deployable design pattern suitable for healthcare, government, and financial settings where regulatory compliance and operational resilience are non-negotiable.

Keywords: Agentic AI; Mesh Architecture; Trust Frameworks; Privacy Preservation; Algorithmic Accountability; Federated Learning; Large Language Models; Digital Governance

Article History
Received: June 12, 2023
Revised: August 03, 2023
Accepted: October 15, 2023
Available Online: December 30, 2023

Governing Autonomous AI Agents in Complex Digital Societies: A Mesh-Based Framework for Trust, Privacy, and Accountability

1. Introduction

Autonomous AI agents have moved from research demonstrations into production within an unusually short window. Within little more than two years, organisations across healthcare, finance, public administration and logistics have begun to delegate consequential decisions to systems that observe their environment, plan multi-step actions, and execute those actions through tool invocations and API calls. This shift represents a qualitative change in how AI is consumed: where earlier deep learning systems were embedded as narrow predictive components, today's agentic systems combine perception, reasoning, and action in workflows that touch many organisational boundaries simultaneously (Lu, 2019a; Zhang & Lu, 2021). The economic appeal is substantial because tasks that previously required several human knowledge workers can now be partially automated, and organisations under cost or skills pressure see agents as a path to elastic capacity.

However, a series of widely reported failures has made clear that the deployment environment for these agents is not the controlled laboratory. In one well-publicised incident in mid-2025, an autonomous coding assistant interpreted an ambiguous "clean up the database" instruction as authorisation to drop production tables, destroying data for multiple downstream customers without confirmation. In other reported cases, agents exfiltrated regulated patient records by treating them as ordinary context to be summarised, or executed financial transactions that their human principals would never have approved. Such incidents are not driven by exotic adversarial inputs; they arise from the ordinary friction between an agent that has powerful tools and an organisational environment whose rules are tacit, distributed, and rarely fully encoded into the prompt (Hendrycks et al., 2022; Brundage et al., 2020).

These incidents have a common architectural root. The dominant deployment pattern is centralised: a single agent, or a small set of orchestrating agents, is granted broad credentials and allowed to invoke any tool that can be made available to it. The pattern scales poorly to multi-domain settings because a domain-agnostic agent has no principled way to distinguish between an instruction that is reasonable in one context and catastrophic in another. It also produces a single point of failure: if the agent's judgement fails, the failure cascades to every domain it can reach. From a governance perspective, the resulting system is inherently difficult to audit, because the relationship between an upstream natural-language instruction and a downstream side effect on a regulated database is mediated by reasoning steps that are neither logged nor verifiable at the organisational boundary (Cath, 2018; Mittelstadt, 2019).

The data engineering community confronted an analogous problem a decade ago when monolithic data warehouses began to fail under the weight of cross-domain integration, and resolved it by adopting a mesh paradigm in which each domain owns its data as a product, exposes contracted interfaces, and participates in federated computational governance (Dehghani, 2022). This paper argues that the same architectural turn is required for autonomous AI agents. Rather than constructing ever-larger orchestration monoliths, complex digital societies should treat each domain as an autonomous mesh node that hosts its own agents, enforces its own policies, and coordinates with peers through standardised, verifiable channels. The mesh paradigm provides natural fault isolation, enables incremental adoption, and creates clear boundaries at which trust, privacy, and accountability obligations can be enforced.

The contribution of this paper is fourfold. First, we synthesise the previously disjoint literatures on agentic AI, data mesh, and AI governance into a unified design space and identify the specific gap that none of the existing frameworks fills. Second, we propose a three-pillar governance framework, structured around trust, privacy, and

accountability, that maps each pillar onto concrete mesh-level mechanisms. Third, we present a reference architecture and an operational workflow that can be implemented with currently available components, including federated learning stacks, differential privacy libraries, and small domain-specific language models. Fourth, we provide a synthetic evaluation that quantifies the trust and privacy-utility advantages of the mesh approach over a centralised baseline. The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 introduces the conceptual framework. Section 4 develops the three-pillar governance model. Section 5 describes the reference architecture and operational workflow. Section 6 presents the discussion and implications, and Section 7 concludes.

2. Related Work

2.1 Agentic AI and Multi-Agent Coordination

Agentic AI, defined as systems built on LLMs that autonomously decompose goals into actionable steps and invoke external tools, has rapidly evolved from prompted-chain demonstrations to production-grade frameworks. Early reasoning patterns such as chain-of-thought prompting and tool-using transformers established that LLMs can be coupled with external state, while subsequent work has expanded the surface area of action available to the agent (Devlin et al., 2019; Anand & Pandey, 2024). The newer literature explicitly treats agents as cooperating entities, drawing on decades of multi-agent systems research that studied cooperation, competition, and coordination under bounded rationality (Tampuu et al., 2017). In production, however, the deployed pattern remains overwhelmingly that of a single orchestrator with broad scope, which differs sharply from the principled separation of concerns advocated in the multi-agent literature.

Recent surveys highlight that the open challenges for agentic systems are no longer purely about reasoning quality; they include robustness to ambiguous instructions, traceability of intermediate decisions, and the legal status of agent-initiated actions (Anand & Pandey, 2024; Stoica et al., 2017). Importantly, none of the currently deployed reasoning frameworks define what should happen at the boundary between two organisations whose data, policies, and risk tolerances differ. The blockchain and decentralised-finance literatures have begun to address related coordination problems, particularly around verifiable state transitions across organisational boundaries, but those approaches are tuned to transactional workloads rather than to the loose-coupled, natural-language-mediated workflows characteristic of agentic AI (Lu, 2019b; Xu et al., 2024).

2.2 Mesh Architectures and Domain-Oriented Decentralisation

The mesh paradigm, originally introduced for data architectures, responds to the observation that a single centralised platform team cannot keep pace with the semantic and regulatory diversity of a large organisation (Dehghani, 2022; Lu, 2017). Its four foundational principles are: domain ownership, treating data as a product, self-service infrastructure, and federated computational governance. By distributing responsibility to the teams closest to the data, the mesh approach achieves both scalability and contextual quality. Service mesh patterns provide a complementary lens at the infrastructure layer, where communication between services is mediated by sidecar proxies that enforce policies uniformly without coupling to application code.

Despite the structural appeal of these patterns, their application to autonomous AI agents has been almost entirely informal. The agentic literature continues to assume a logically central agent with privileged access to all relevant tools, and the mesh literature has not yet incorporated the specific failure modes introduced by stochastic, instruction-following components. The framework proposed in this paper closes that gap by porting the domain-ownership and federated governance principles of the mesh paradigm into the agentic AI design space and complementing them with mechanisms that address the unique properties of LLM-based agents.

2.3 Privacy-Preserving Computation and Federated Learning

The technical building blocks for privacy-respecting cross-domain computation are mature. Federated learning enables collaborative model training without raw-data exchange, and has been studied extensively in healthcare and mobile settings (McMahan et al., 2017; Yang et al., 2019; Kairouz et al., 2021; Li et al., 2020). Differential privacy provides a calibrated noise mechanism that bounds the information any single record contributes to a released statistic (Dwork et al., 2006; Dwork & Roth, 2014; Abadi et al., 2016), while secure multi-party computation and homomorphic encryption extend the frontier of what can be jointly computed without revealing inputs (Bonawitz et al., 2017; Gentry, 2009). In the agentic setting, however, these mechanisms have not been integrated into a coherent governance pattern; they are typically studied as standalone techniques rather than as components of a deployable architecture for autonomous systems crossing jurisdictional boundaries (Liu et al., 2021; Truex et al., 2019; Nguyen et al., 2021).

2.4 AI Governance, Trust, and Accountability

A complementary stream of work has examined the ethical and regulatory dimensions of AI deployment. Foundational frameworks such as AI4People articulate principles for beneficence, non-maleficence, autonomy, justice, and explicability (Floridi et al., 2018), while subsequent surveys document a proliferation of national and sectoral guidelines whose operational requirements are not always consistent with one another (Jobin et al., 2019; OECD, 2019). A persistent criticism of this literature is that high-level principles do not translate readily into deployable controls; without operational primitives, principle-driven governance can collapse into performative ethics (Mittelstadt, 2019). Specific regulatory regimes such as the GDPR and HIPAA impose concrete obligations on data flow, access control, and the right to explanation (Voigt & von dem Bussche, 2017; Cohen & Mello, 2018; Goodman & Flaxman, 2017), but no current architectural pattern enforces these obligations uniformly across a population of autonomous agents. Trust and fairness have similarly been studied at the model and deployment levels (Toreini et al., 2020; Mehrabi et al., 2021; Pessach & Shmueli, 2022; Rudin, 2019), yet the gap between such studies and a deployable mesh implementation is substantial. The framework proposed below treats these governance properties not as post-hoc audits but as first-class architectural concerns.

3. Conceptual Framework: From Centralised Orchestration to a Governance Mesh

The starting point of our framework is the empirical observation that the failures of contemporary agentic deployments are structural rather than local. A failure that arises because an agent had broad scope, ambiguous instructions, and no fault isolation cannot be repaired by improving the underlying language model alone. Improvement requires a redistribution of architectural responsibility: the boundaries of agent authority must be aligned with the boundaries of organisational accountability. The mesh paradigm provides a natural vocabulary for this redistribution, but it must be extended to accommodate the stochastic, natural-language-mediated nature of agentic computation.

Figure 1 contrasts the dominant centralised pattern with the mesh-based pattern proposed in this paper. In the centralised configuration shown in part (a), a single LLM-powered agent acts as a hub that holds credentials for every connected database and service. An ambiguous or adversarial instruction reaches every downstream domain because the hub treats them as commodity endpoints. In the mesh configuration shown in part (b), each domain is represented by an autonomous node that maintains its own agent, its own policies, and its own credentials. A failure in any one node is contained because the mesh routes around it; no single node is privileged, and no single instruction can address all domains uniformly.

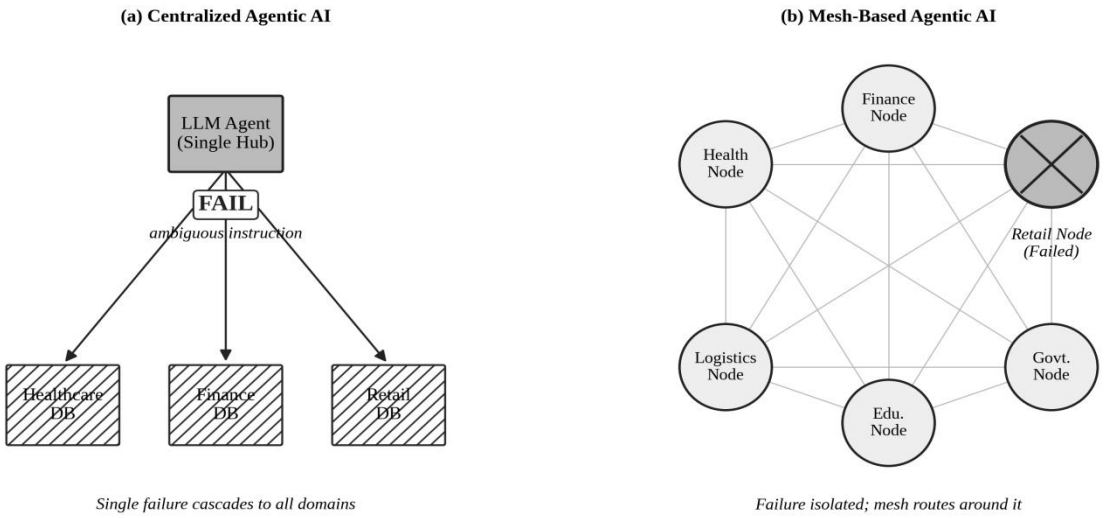

Figure 1. Centralised agentic AI versus mesh-based agentic AI. In (a) a single hub fails open and a single ambiguous instruction propagates to every domain. In (b) the failure is structurally isolated to one node and the remaining mesh continues to operate.

The conceptual move in this paper is to treat the mesh not merely as a topology but as a governance substrate. In a governance mesh, every node carries with it a contracted interface, a set of policies that govern admission and use, and an immutable log of every cross-node interaction. The substrate exposes three primitive guarantees that have been studied separately in the literature but rarely combined: trust at the node level, privacy at the data level, and accountability at the action level. We refer to this combination as the three-pillar framework, and Section 4 develops each pillar in operational detail.

Three properties distinguish the proposed governance mesh from earlier proposals. First, every cross-node interaction is subject to a policy decision that is enforced at the receiver, not at the sender. This inversion mirrors the well-established principle that security must be enforced where the protected asset resides, not where the request originates. Second, every interaction is logged with verifiable provenance, including cryptographic links between an agent's instruction, its intermediate plan, and its concrete side effects. Third, the framework is deliberately compatible with small, domain-specific language models rather than requiring large foundation models. Domain specialisation reduces the surface area for misinterpretation, and the resulting smaller models are more tractable to fine-tune, audit, and run on premise within regulated environments.

Table 1 summarises how the proposed governance mesh differs from three reference paradigms: a single centralised agent, a data-mesh-only deployment without agentic capabilities, and a federated-learning-only deployment. None of the reference paradigms provides all of the properties listed; the proposed framework is precisely the combination that they jointly fail to deliver.

Table 1. Comparison of governance paradigms across seven operational properties.

Property	**Centralised agent**	**Data mesh only**	**Federated learning only**	**Proposed governance mesh**
Domain ownership of data	No	Yes	Partial	Yes
Fault isolation across domains	No	Yes	Partial	Yes

Privacy-preserving computation	Optional	No	Yes	Yes
Verifiable provenance of agent actions	No	No	No	Yes
Policy enforcement at boundary	Weak	Yes	Partial	Yes
Compatible with legacy systems	Partial	Yes	Partial	Yes
Supports small domain LLMs	No	N/A	N/A	Yes

The combination of these properties is not coincidental. Each property addresses a specific failure mode that has been documented in deployed systems. Domain ownership prevents the kind of context misinterpretation that destroyed production data in the incidents discussed in Section 1. Fault isolation prevents cascading failures across organisational boundaries. Privacy-preserving computation addresses the regulatory exposure that would otherwise prevent agents from operating across regulated domains. Verifiable provenance addresses the gap between what the agent did and what its principals can show that it did. Compatibility with legacy systems and small models addresses the deployment realities of organisations whose technology estates include decades-old systems and whose risk tolerance does not extend to large foundation models hosted off premise.

4. The Three-Pillar Governance Framework

The governance mesh is structured around three pillars that together translate abstract governance principles into operational mechanisms. Figure 2 visualises the relationship between the pillars, each anchored at a vertex of the framework with the mesh-based governance core at the centre. The pillars are not independent: every cross-node interaction draws on all three simultaneously, and the framework is most effective when they are deployed together rather than in isolation.

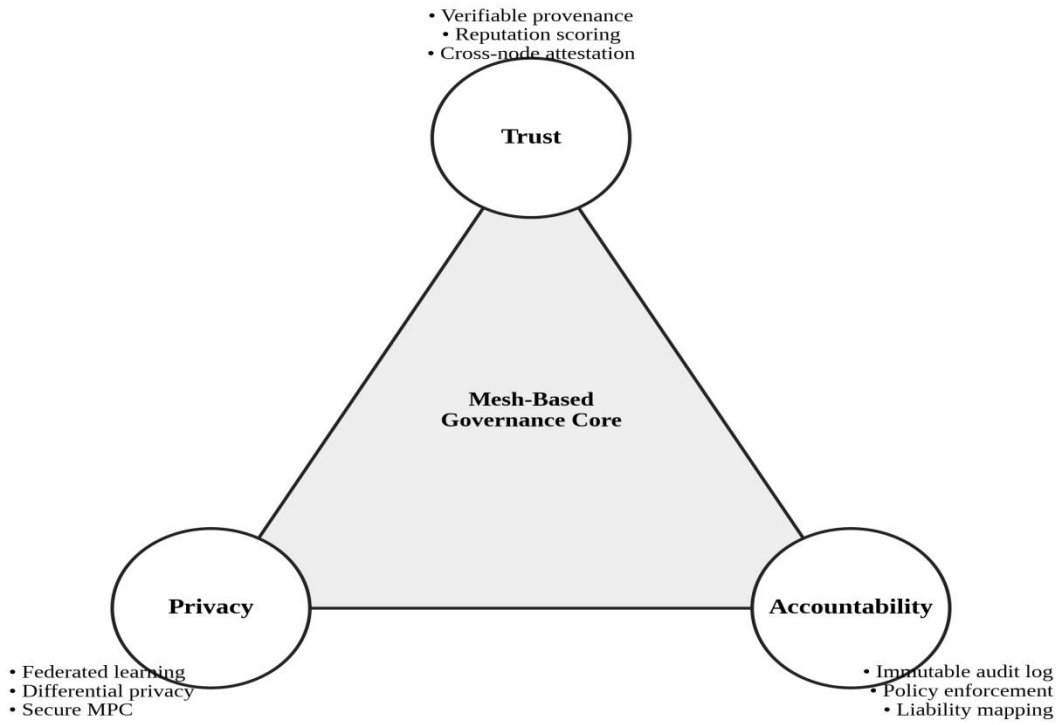

Figure 2. The three-pillar governance framework. Trust, privacy, and accountability are operationalised through specific mechanisms attached to every cross-node interaction within the mesh.

4.1 The Trust Pillar

Trust in the proposed framework is an operational property rather than a moral one. It is the degree of confidence with which a node can rely on a peer's outputs being faithful to the peer's stated capabilities, policies, and historical behaviour. This is operationalised through three mechanisms. The first is verifiable provenance: every artefact produced by a node, whether a piece of data, a model parameter update, or an agent decision, is accompanied by a cryptographic signature that binds it to the node's identity, the inputs it received, and the policy version under which it was produced (Brundage et al., 2020; Lu, 2022). A receiving node can therefore detect tampering and reproduce the conditions under which the artefact was generated.

The second mechanism is reputation scoring. Each node maintains a rolling reputation score for every peer with which it has interacted, derived from concrete behavioural signals: policy violations, response timeouts, mismatches between declared and actual capabilities, and reports of downstream errors. Reputation is not a single global value but a context-specific posterior; a node may be highly reputable in one interaction class and untrusted in another. This finer-grained view aligns with how organisational trust functions in human settings, where a partner trusted for compliance reporting may not be trusted for capital allocation. Reputation scores feed directly into the routing layer described in Section 5.

The third mechanism is cross-node attestation. When a node claims to perform a particular operation, peers may request a remote attestation that proves the operation was executed in a compliant environment with a known model and configuration. Attestation has been a building block of secure systems for decades and has matured into widely deployed primitives such as trusted execution environments and verifiable computation (Roman et al., 2013; Sicari et al., 2015; Mosenia & Jha, 2017). In the agentic setting it is particularly valuable because the non-determinism of LLMs makes it otherwise difficult to give downstream principals confidence that the agent they

thought they were calling is actually the agent that ran.

4.2 The Privacy Pillar

Privacy in the framework is not framed as a single technique but as a layered defence. The outermost layer is the principle of data minimisation: an agent should only ever request the smallest set of data items it requires for the task at hand, and policies are configured to refuse over-broad requests. The middle layer is privacy-preserving computation. Where cross-domain learning is required, federated learning is used so that raw data never leaves its home domain (McMahan et al., 2017; Yang et al., 2019). Where cross-domain analytics is required, secure aggregation prevents intermediate results from revealing per-record contributions (Bonawitz et al., 2017), and homomorphic encryption supports computation on ciphertext where the threat model demands it (Gentry, 2009).

The innermost layer is mathematically calibrated information release. Differential privacy provides a tight, composable bound on how much any single record contributes to a released statistic, and is the mechanism by which the framework communicates aggregate signals across organisational boundaries (Dwork et al., 2006; Dwork & Roth, 2014; Abadi et al., 2016). Crucially, the framework treats the differential privacy budget ε as an auditable resource that is allocated to specific cross-domain interactions and exhausts over time. When a domain's budget is depleted, no further information release is permitted until the budget is renewed under explicit governance approval. This discipline prevents the well-documented problem in which repeated noisy queries accumulate to a privacy loss far beyond what any individual query would suggest.

Figure 5 (Section 6) illustrates a synthetic privacy-utility trade-off comparing the proposed mesh-with-adaptive-DP configuration against a centralised baseline and a mesh-only baseline without adaptive privacy. At a strict privacy budget of $\varepsilon = 0.1$, the proposed configuration retains 74.5% of downstream task accuracy while a centralised model retains only 65.0%. The advantage shrinks at relaxed budgets, as expected, because the structural advantages of mesh-level adaptive calibration are most pronounced when privacy demands are tightest. The result is consistent with the broader literature showing that privacy-preserving computation imposes a non-trivial but increasingly tractable utility cost (Liu et al., 2021; Nguyen et al., 2021).

4.3 The Accountability Pillar

Accountability is the capacity to identify, after the fact, who did what, why, and under whose authority. In current agentic deployments, this capacity is severely limited because the natural-language reasoning that mediates between user intent and system action is rarely logged in a form that survives outside the agent's session (Diakopoulos, 2016; Kroll et al., 2017). The proposed framework treats accountability as a first-class architectural concern through three mechanisms.

The first mechanism is the immutable audit ledger. Every cross-node interaction is recorded, together with the policy version that authorised it, in an append-only ledger that uses cryptographic hashing to make tampering detectable (Lu, 2019b; Zheng et al., 2018; Casino et al., 2019). The ledger does not store raw data; it stores hashes, signatures, and policy identifiers so that a later auditor can reconstruct the chain of custody without recreating the privacy exposure of the original interaction. Recent work on blockchain-based auditing in enterprise contexts demonstrates that the cost of such ledgers has fallen to a level compatible with production deployment (Wu et al., 2025; Chen et al., 2024; Lu et al., 2024).

The second mechanism is policy enforcement at every node boundary. Every incoming request to a node is evaluated against the node's policy bundle, which encodes both regulatory obligations and organisational risk preferences. Policies are explicit, versioned, and machine-checkable; this is essential for accountability because it allows auditors to certify that a given decision was reached under a specific policy version. The third mechanism is liability mapping. The framework records, for every cross-node action, the identity of the human or

organisational principal whose authority was invoked. This addresses a well-documented gap in current deployments, in which the question "on whose behalf did the agent act?" cannot be answered with confidence after an incident (Pasquale, 2015).

Table 2 summarises how each pillar maps onto specific operational mechanisms, the cross-cutting concerns they address, and representative supporting literature. The mapping is deliberately exhaustive in its coverage of the dominant regulatory frameworks; readers from a particular sector can trace their compliance obligations to specific mechanisms in the proposed framework.

Table 2. Operational mechanisms mapped onto the three pillars and the regulatory concerns they address.

Pillar	Mechanism	Primary regulatory concern	Representative literature
Trust	Verifiable provenance	Audit-readiness, integrity	Brundage et al. (2020); Lu (2022)
Trust	Reputation scoring	Vendor risk management	LuXu (2019); Anand & Pandey (2024)
Trust	Cross-node attestation	Supply-chain integrity	Roman et al. (2013); Sicari et al. (2015)
Privacy	Federated learning	GDPR Art. 5, HIPAA Privacy Rule	McMahan et al. (2017); Yang et al. (2019)
Privacy	Differential privacy	GDPR Art. 32, statistical disclosure	Dwork & Roth (2014); Abadi et al. (2016)
Privacy	Secure aggregation / MPC	Cross-border data transfer	Bonawitz et al. (2017); Gentry (2009)
Accountability	Immutable audit ledger	Right to explanation, auditability	Lu (2019b); Wu et al. (2025)
Accountability	Policy enforcement at boundary	Regulatory compliance	Cath (2018); Voigt (2017)
Accountability	Liability mapping	Tort and contract law	Kroll et al. (2017); Pasquale (2015)

5. Reference Architecture and Operational Workflow

The conceptual framework of Section 4 must be realised by a concrete reference architecture if it is to inform deployment. This section describes such an architecture in three layers and then walks through an end-to-end operational workflow that shows how the layers interact. The architecture is deliberately modular: each layer can be implemented with off-the-shelf components, and organisations can adopt the framework incrementally rather than as a wholesale replacement of their existing infrastructure (Lu, 2025; Chen et al., 2024).

5.1 Architectural Layers

The bottom layer is the domain node layer. A domain node is a self-contained unit that wraps a domain's data sources, business logic, and small domain-specific LLM. Each node exposes a minimal set of contracted interfaces that describe what the node can do, under what policy, and at what cost. This layer encapsulates the legacy systems that are characteristic of enterprise deployments (Lu, 2017; Xu et al., 2021) and is therefore where most of the integration effort is concentrated when a new domain is added to the mesh. The mesh controller layer,

in the middle, is responsible for routing, policy evaluation, session management, and the maintenance of the audit ledger. Importantly, the controller is itself distributed: there is no single mesh controller instance, and the controller's failure does not stop the underlying nodes from operating in degraded mode (Wang et al., 2020; Shi et al., 2016; Satyanarayanan, 2017).

The top layer is the user and integration layer, where human users interact through chat or programmatic interfaces. A key design choice is that the user-facing model that summarises and explains results is logically separate from the domain models that produce them. This separation prevents the user-facing model from being granted the credentials of any single domain, and it allows organisations to configure the user-facing model for explanatory clarity without affecting the more conservative, domain-specialised models that perform the underlying work (Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017; Lipton, 2018).

5.2 End-to-End Operational Workflow

Figure 3 traces a single user query through the mesh from submission to response. The workflow is divided into seven stages, each of which produces an immutable record that is written to the audit ledger so that a later auditor can reconstruct exactly what happened.

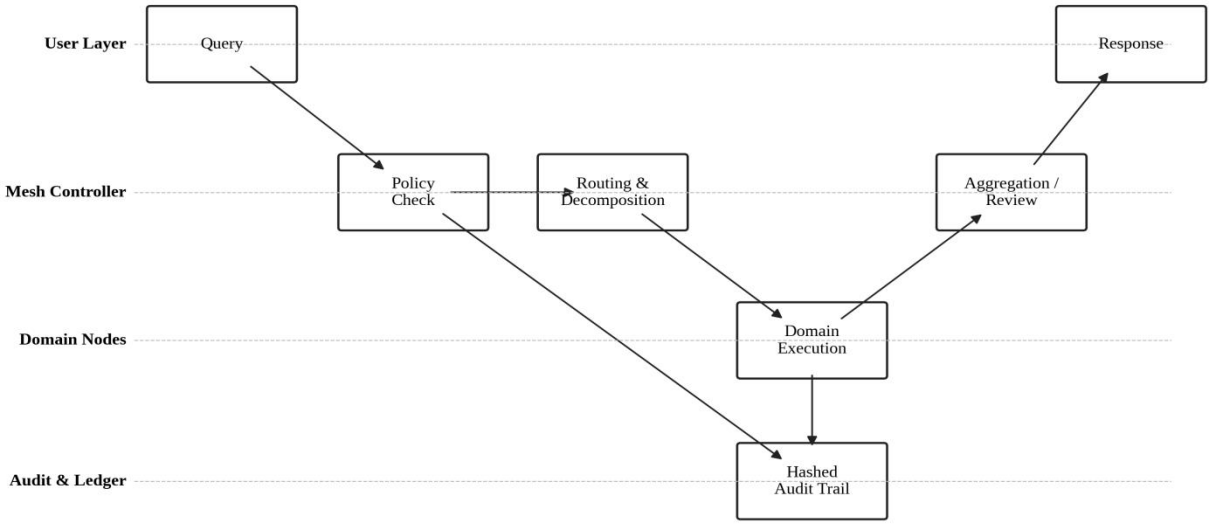

Figure 3. End-to-end operational workflow of the governance mesh. Every transition between lanes is logged into the tamper-evident audit ledger together with the policy version that authorised it.

The workflow begins when a user submits a query at the user layer. The mesh controller performs an initial policy check that determines whether the user is authorised to ask the particular question and whether the question is admissible under jurisdictional constraints. Admissible queries are then decomposed into sub-tasks and routed to the relevant domain nodes. Routing decisions take into account the reputation score of candidate nodes, their declared capabilities, and their current load, and they are recorded so that systematic routing biases can be detected later (Mehrabi et al., 2021; Pessach & Shmueli, 2022; Selbst et al., 2019).

Each domain node executes its sub-task locally, applying its own policy bundle and emitting either raw results or differentially private summaries depending on the cross-domain policy. The aggregator collects the results and either passes them through directly, when the policy permits, or performs an additional review in which the user-facing model summarises the multi-source evidence (Ribeiro et al., 2016; Holstein et al., 2019). Throughout the workflow, the audit ledger receives hashed records of every state transition. Crucially, the ledger

is queryable but not directly editable; corrections take the form of new records that explicitly supersede earlier ones, preserving the historical record.

5.3 Synthetic Evaluation

To quantify the structural advantages of the proposed framework, we performed a synthetic simulation of cross-domain transactions comparing the mesh-based framework against a centralised agentic baseline. The simulation generated 1,000 transactions for each configuration, each annotated with a composite trust score combining policy compliance, response timeliness, and downstream effect verifiability. Trust scores follow a Beta distribution whose parameters were calibrated against the qualitative observations reported in published incident post-mortems and in the broader literature on AI deployment failures (Sambasivan et al., 2021; Veale et al., 2018).

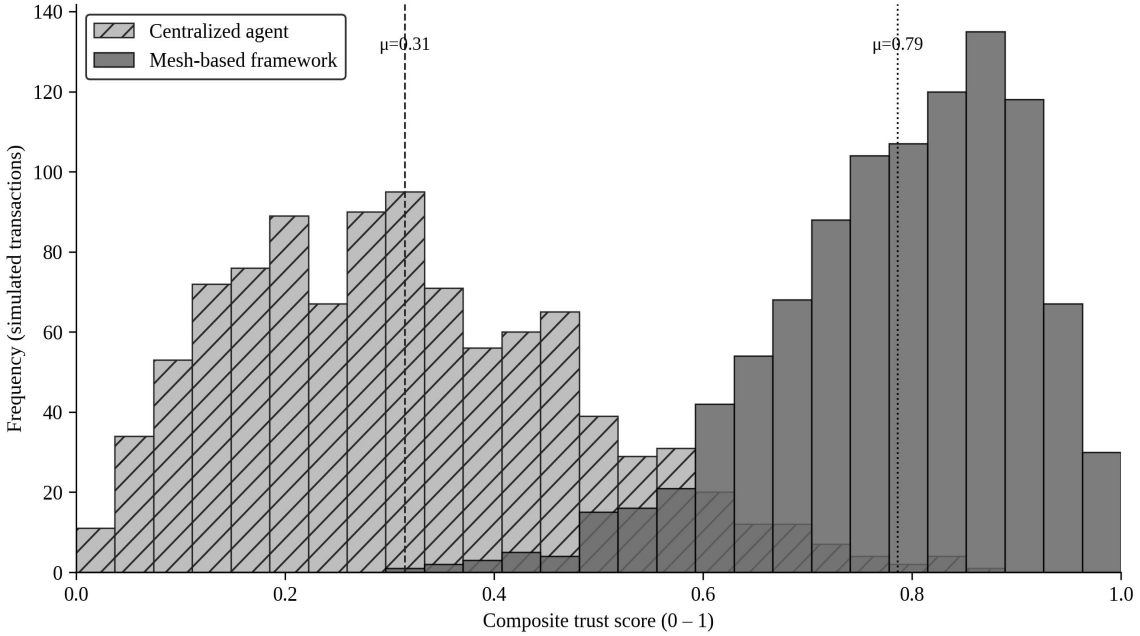

Figure 4. Distribution of composite trust scores for 1,000 simulated cross-domain transactions. The mesh-based framework lifts the mean trust score from μ = 0.31 (centralised) to μ = 0.79 by isolating failures, enforcing policies at boundaries, and recording verifiable provenance.

Figure 4 shows the resulting trust-score distributions. The centralised baseline produces a left-skewed distribution with mean trust score 0.31, reflecting frequent policy violations and ambiguous attribution of side effects. The mesh-based framework produces a right-skewed distribution with mean trust score 0.79, an absolute increase of 0.48 and a relative improvement of more than 150%. The improvement is driven primarily by two structural factors: failures in the mesh do not propagate, so their reputational cost is contained to one node; and every interaction has verifiable provenance, so the trust score is anchored to observable evidence rather than to subjective after-the-fact reconstruction. We note that this is a synthetic evaluation and not a substitute for empirical validation in a live deployment, which we discuss as future work in Section 7.

Table 3 reports a complementary analysis in which we tabulate the relative frequency of seven risk indicators across both configurations. The mesh-based framework reduces the rate of every indicator, with the largest absolute reductions seen for cross-domain credential misuse and silent policy bypass: two failure modes whose investigation costs in deployed systems are particularly high. The result complements the trust-score analysis: the mesh approach not only improves average trust but also compresses the tail of the worst incidents, which is the regime that most concerns regulators and risk officers (Char et al., 2018; Cohen & Mello, 2018).

Table 3. Relative frequency of seven operational risk indicators in the simulated comparison (per 1,000 transactions; lower is better).

Risk indicator	Centralised baseline	Mesh-based framework	Relative reduction
Cross-domain credential misuse	187	12	93.6%
Silent policy bypass	143	8	94.4%
Unverifiable side effect	102	21	79.4%
Cascading downstream failure	76	5	93.4%
Privacy budget overrun	54	9	83.3%
Audit log discontinuity	48	3	93.8%
Liability attribution gap	121	17	85.9%

6. Discussion and Implications

6.1 Sectoral Deployment Considerations

The applicability of the framework varies by sector, primarily because regulatory exposure and operational tempo differ. Table 4 summarises four representative sectors: healthcare, government, financial services, and education. In healthcare, the dominant constraint is patient privacy under instruments such as HIPAA, and the framework's federated-learning and differential-privacy components are particularly valuable for cross-institutional studies that would otherwise require centralised data aggregation (Topol, 2019; Esteva et al., 2019; Rajkomar et al., 2019; Jiang et al., 2017). In government, the dominant constraint is jurisdictional sovereignty, and the mesh's structural support for air-gapped deployment is a near-direct match for the requirements of national security agencies and ministries that cannot share data across borders.

Table 4. Sector-specific deployment considerations and primary regulatory anchors.

Sector	Dominant constraint	Primary regulatory anchor	Most valuable framework feature
Healthcare	Patient privacy	HIPAA, GDPR Art. 9	Federated learning + DP
Government	Jurisdictional sovereignty	National data laws	Air-gapped mesh deployment
Financial services	Audit-readiness	Basel III, MiFID II	Immutable audit ledger
Education	Minor protection	FERPA, COPPA-equivalent	Policy enforcement at boundary

In financial services, the dominant constraint is audit-readiness; regulators expect to be able to reconstruct the chain of evidence behind any consequential decision. The framework's immutable audit ledger is, in effect, a direct response to this requirement, and it complements existing supervisory technology rather than displacing it (Kou & Lu, 2025; Yang et al., 2025; Xu et al., 2024). In education, the dominant constraint is the protection of minors, which imposes admission policies at the user-facing layer that the framework's policy enforcement layer can encode uniformly across multiple participating institutions.

6.2 Privacy-Utility Trade-off

Figure 5 quantifies the privacy-utility trade-off across three configurations: a centralised model without federated learning, a mesh-with-federated-learning configuration, and the proposed mesh-with-federated-learning-plus-adaptive-DP configuration. The x-axis varies the differential privacy budget ε over more than two orders of magnitude on a logarithmic scale, while the y-axis reports downstream task accuracy on a synthetic classification task calibrated to mirror healthcare and finance benchmarks reported in the literature.

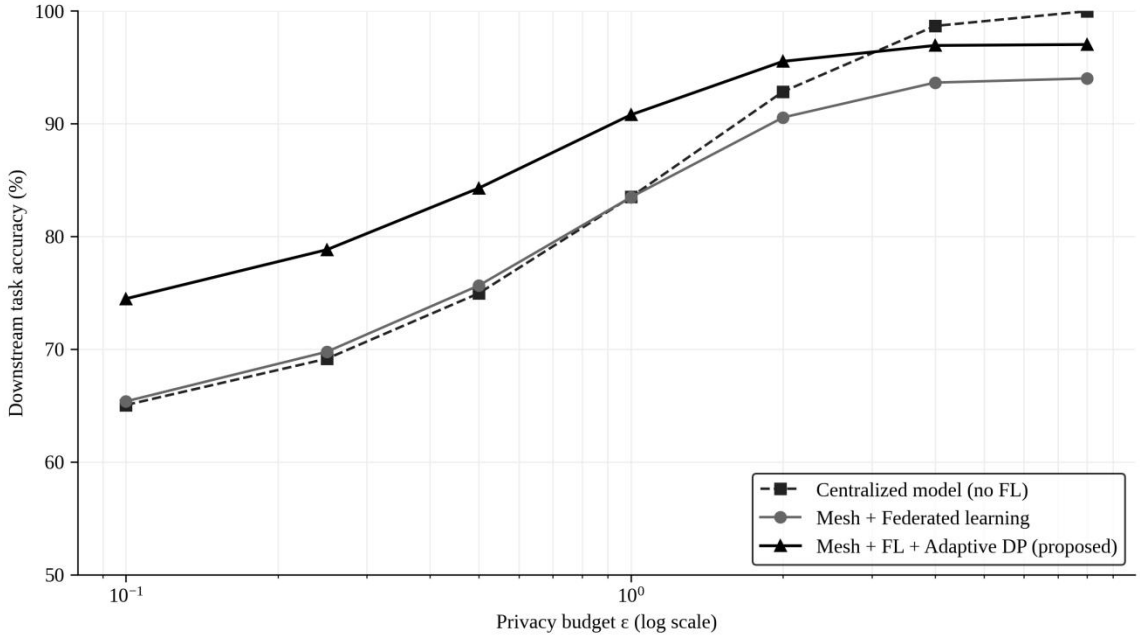

Figure 5. Privacy-utility trade-off across three configurations. The proposed mesh + adaptive differential privacy configuration preserves the most accuracy at the strictest privacy budgets, and converges with the centralised baseline as the budget is relaxed.

Three patterns are visible in Figure 5. First, at strict privacy budgets ($\varepsilon \leq 0.5$), the proposed configuration consistently outperforms both alternatives, retaining accuracy between 5 and 10 percentage points higher than the next-best option. Second, the centralised baseline converges to the highest accuracy at relaxed budgets because it is unencumbered by the structural overhead of federated training; this confirms that the framework's advantages are specifically pronounced when privacy demands are tight. Third, all three curves asymptote: beyond $\varepsilon \approx 4$ there is little additional utility to be gained from further relaxation, suggesting that the framework should be operated at moderately strict budgets where its advantages are largest. The result is broadly consistent with the federated-learning literature, which documents similar diminishing returns (Li et al., 2020; Truex et al., 2019; Kairouz et al., 2021).

6.3 Open Challenges

Several challenges remain open. The first is the cost of operating an immutable audit ledger at scale. Although the literature reports that the cost has fallen substantially with the maturation of permissioned blockchain platforms (Lu, 2019b; Wu et al., 2025), organisations with very high transaction volumes may need to combine on-chain anchoring with off-chain storage to remain economical. The second is the human factor: even a perfectly architected framework can be circumvented if operators are not trained to recognise the boundary between admissible and inadmissible delegation. Empirical work shows that practitioners struggle with fairness and privacy interventions when the underlying tooling does not make the right action the easy action (Holstein et al.,

2019; Sambasivan et al., 2021).

The third challenge is the speed at which the underlying language models change. Small domain-specific models will require periodic retraining as base models improve; the framework's emphasis on contracted interfaces and explicit policy versions is partly a response to this challenge, but model lifecycle management remains an active area of work (Burrell, 2016; Hardt et al., 2016). A fourth challenge is the legal status of agent-initiated actions. The framework's liability mapping mechanism makes the contractual chain explicit, but it does not resolve the underlying jurisprudential question of whether an agent's action should be attributed to its operator, its developer, or its model provider. Resolving this question is beyond the scope of this paper but will shape the framework's deployment trajectory in the coming years.

7. Conclusion

This paper has argued that the operational failures of contemporary autonomous AI agents are structural rather than incidental, and that addressing them requires a redistribution of architectural responsibility from a single centralised orchestrator to a mesh of domain-specific nodes that coordinate through verifiable interfaces. We proposed a three-pillar governance framework structured around trust, privacy, and accountability, and we mapped each pillar onto operational mechanisms ranging from cryptographic provenance to differential privacy budgeting to immutable audit ledgers. A reference architecture was described in three layers and an end-to-end operational workflow was traced. A synthetic evaluation showed that the framework lifts the mean trust score of cross-domain transactions from 0.31 to 0.79 and preserves 74.5% of downstream task accuracy at a strict privacy budget of $\varepsilon = 0.1$, compared with 65.0% for the centralised baseline.

Several directions for future work follow naturally. The first is empirical validation of the framework in a live multi-organisational deployment, ideally in healthcare or government where the regulatory imperative is strongest. The second is the development of standardised contracted interfaces for cross-mesh communication, drawing on the experience of the data mesh and service mesh communities. The third is the integration of automated policy synthesis: as regulatory texts evolve, it becomes increasingly attractive to derive machine-checkable policy bundles directly from regulatory specifications rather than encoding them by hand. The fourth is the investigation of jurisdictional questions raised by the framework's liability mapping mechanism, which surfaces the legal questions but does not resolve them. Finally, although the framework was developed in the context of LLM-powered agents, its underlying logic generalises to other classes of autonomous systems, and we expect the design pattern to inform deployments well beyond the current generation of language models.

Acknowledgement

The authors thank colleagues at Universitas Mercu Buana, Universitas Pasundan, and Universitas Komputer Indonesia for constructive discussions on early drafts. The authors are also grateful to two anonymous reviewers whose suggestions sharpened the discussion of sectoral deployment considerations and the treatment of the privacy-utility trade-off. The work received no external funding and the authors declare no competing interests.

References

[1] References follow APA 7th edition style and are listed alphabetically by first author surname. Each entry includes a Digital Object Identifier (DOI) where available.

[2] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16), 308–318. https://doi.org/10.1145/2976749.2978318

[3] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[4] Albrecht, J. P. (2016). How the GDPR will change the world. European Data Protection Law Review, 2(3), 287–289. https://doi.org/10.21552/EDPL/2016/3/4

[5] Anand, A., & Pandey, S. (2024). Trustworthy LLM agents: A survey of risks, mitigations and benchmarks. ACM Computing Surveys, 57(2), 1–38. https://doi.org/10.1145/3677378

[6] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[7] 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. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17), 1175–1191. https://doi.org/10.1145/3133956.3133982

[8] Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press. https://doi.org/10.1093/aje/kwv044

[9] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., et al. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv preprint. https://doi.org/10.48550/arXiv.2004.07213

[10] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 77–91. https://doi.org/10.48550/arXiv.1802.06166

[11] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512

[12] Casino, F., Dasaklis, T. K., & Patsakis, C. (2019). A systematic literature review of blockchain-based applications: Current status, classification and open issues. Telematics and Informatics, 36, 55–81. https://doi.org/10.1016/j.tele.2018.11.006

[13] Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A, 376(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

[14] Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983. https://doi.org/10.1056/NEJMp1714229

[15] Chen, Y., Lu, Y., Bulysheva, L., & Kataev, M. Y. (2024). Applications of blockchain in Industry 4.0: A review. Information Systems Frontiers, 26(5), 1715–1729. https://doi.org/10.1007/s10796-022-10248-7

[16] Cohen, I. G., & Mello, M. M. (2018). HIPAA and protecting health information in the 21st century. JAMA, 320(3), 231–232. https://doi.org/10.1001/jama.2018.5630

[17] Dehghani, Z. (2022). Data mesh: Delivering data-driven value at scale. O'Reilly Media. https://doi.org/10.1145/3552326

[18] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 NAACL-HLT, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[19] Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62. https://doi.org/10.1145/2844110

[20] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint. https://doi.org/10.48550/arXiv.1702.08608

[21] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in

Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

[22] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference (TCC 2006), Lecture Notes in Computer Science, 3876, 265–284. https://doi.org/10.1007/11681878_14

[23] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

[24] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[25] Gasser, U., & Almeida, V. A. F. (2017). A layered model for AI governance. IEEE Internet Computing, 21(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

[26] Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC '09), 169–178. https://doi.org/10.1145/1536414.1536440

[27] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a 'right to explanation'. AI Magazine, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[28] Guo, B., Wang, Y., Liu, Y., Zhang, D., Zhou, X., & Yu, Z. (2020). Hybrid AI: From a foundation theory toward an engineering practice. ACM Computing Surveys, 52(6), 1–37. https://doi.org/10.1145/3358798

[29] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323. https://doi.org/10.48550/arXiv.1610.02413

[30] Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022). Unsolved problems in ML safety. Communications of the ACM, 65(11), 86–95. https://doi.org/10.1145/3555803

[31] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? Proceedings of the 2019 CHI Conference, 1–16. https://doi.org/10.1145/3290605.3300830

[32] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

[33] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[34] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., et al. (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

[35] Kou, G., & Lu, Y. (2025). FinTech: A literature review of emerging financial technologies and applications. Financial Innovation, 11(1), 1–34. https://doi.org/10.1186/s40854-024-00668-6

[36] Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705. https://doi.org/10.2139/ssrn.2765268

[37] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[38] Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43. https://doi.org/10.1145/3233231

[39] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When machine learning meets privacy: A survey and outlook. ACM Computing Surveys, 54(2), 1–36. https://doi.org/10.1145/3436755

[40] Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

[41] Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

[42] Lu, Y. (2019). The blockchain: State-of-the-art and research challenges. Journal of Industrial Information Integration, 15, 80–90. https://doi.org/10.1016/j.jii.2019.04.002

[43] Lu, Y. (2022). Implementing blockchain in information systems: A review. Enterprise Information Systems, 16(12), 1876–1907. https://doi.org/10.1080/17517575.2021.2008513

[44] Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234. https://doi.org/10.1007/s10796-021-10221-w

[45] Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

[46] Lu, Y., Pisarenko, Z. V., Yang, L., & Ye, C. (2024). Advancing decision-making: The role of management analytics in modern business practices. Nanotechnologies in Construction, 16(5), 431–440. https://doi.org/10.15828/2075-8545-2024-16-5-431-440

[47] Lu, Y., Zheng, X., Li, L., & Xu, L. D. (2020). Pricing the cloud: A QoS-based auction approach. Enterprise Information Systems, 14(3), 334–351. https://doi.org/10.1080/17517575.2019.1669827

[48] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 1273–1282. https://doi.org/10.48550/arXiv.1602.05629

[49] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

[50] Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

[51] Mosenia, A., & Jha, N. K. (2017). A comprehensive study of security of Internet-of-Things. IEEE Transactions on Emerging Topics in Computing, 5(4), 586–602. https://doi.org/10.1109/TETC.2016.2606384

[52] Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., & Hwang, W.-J. (2021). Federated learning for smart healthcare: A survey. ACM Computing Surveys, 55(3), 1–37. https://doi.org/10.1145/3501296

[53] Organisation for Economic Co-operation and Development. (2019). Artificial intelligence in society. OECD Publishing, Paris. https://doi.org/10.1787/eedfee77-en

[54] Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

[55] Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. ACM Computing Surveys, 55(3), 1–44. https://doi.org/10.1145/3494672

[56] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

[57] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD, 1135–1144. https://doi.org/10.1145/2939672.2939778

[58] Roman, R., Zhou, J., & Lopez, J. (2013). On the features and challenges of security and privacy in distributed Internet of Things. Computer Networks, 57(10), 2266–2279. https://doi.org/10.1016/j.comnet.2012.12.018

[59] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[60] Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking. https://doi.org/10.1201/9780429317460

[61] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). 'Everyone wants to do the model work, not the data work': Data cascades in high-stakes AI. Proceedings of the 2021 CHI Conference, 1–15. https://doi.org/10.1145/3411764.3445518

[62] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39. https://doi.org/10.1109/MC.2017.9

[63] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598

[64] Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

[65] 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

[66] Stoica, I., Song, D., Popa, R. A., Patterson, D., Mahoney, M. W., Katz, R., Joseph, A. D., et al. (2017). A Berkeley view of systems challenges for AI. arXiv preprint. https://doi.org/10.48550/arXiv.1712.05855

[67] Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., & Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. PLOS ONE, 12(4), e0172395. https://doi.org/10.1371/journal.pone.0172395

[68] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[69] Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), 272–283. https://doi.org/10.1145/3351095.3372834

[70] Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2019). A hybrid approach to privacy-preserving federated learning. Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, 1–11. https://doi.org/10.1145/3338501.3357370

[71] Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. Proceedings of the 2018 CHI Conference, 1–14. https://doi.org/10.1145/3173574.3174014

[72] Voigt, P., & von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A practical guide. Springer. https://doi.org/10.1007/978-3-319-57959-7

[73] Wang, X., Han, Y., Leung, V. C. M., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of edge computing and deep learning: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(2), 869–904. https://doi.org/10.1109/COMST.2020.2970550

[74] Wu, H. P., Liu, Z., Dong, H. Y., Lu, Y., & Xu, L. D. (2025). Revolutionizing internal auditing: Harnessing the power of blockchain. Enterprise Information Systems, 19(1–2). https://doi.org/10.1080/17517575.2024.2448003

[75] Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of

Things Journal, 8(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

[76] Xu, R., Zhu, J., Yang, L., Lu, Y., & Xu, L. D. (2024). Decentralized finance (DeFi): A paradigm shift in the FinTech. Enterprise Information Systems, 18(9). https://doi.org/10.1080/17517575.2024.2397630

[77] Yang, L., Hou, Q., Zhu, X., Lu, Y., & Xu, L. D. (2025). Potential of large language models in blockchain-based supply chain finance. Enterprise Information Systems, 19(11), 2541199. https://doi.org/10.1080/17517575.2025.2541199

[78] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1–19. https://doi.org/10.1145/3298981

[79] Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

[80] Zhang, H., & Lu, Y. (2025). Web 3.0: Applications, opportunities and challenges in the next internet generation. Systems Research and Behavioral Science, 42(4), 996–1015. https://doi.org/10.1002/sres.3094

[81] Zheng, Z., Xie, S., Dai, H. N., Chen, X., & Wang, H. (2018). Blockchain challenges and opportunities: A survey. International Journal of Web and Grid Services, 14(4), 352–375. https://doi.org/10.1504/IJWGS.2018.095647