Digital Transformation and Technology Dynamics

Journal Homepage: www.techdynamicsjournal.com | ARTICLE NO. DTTD-2022004 | VOL. 2 ISSUE 2

A Unified Architecture for Trustworthy Autonomous Intelligence: Integrating Agentic AI, Edge Computing, Digital Twins, Cybersecurity, and Explainable Decision-Making

Rashadul Islam Samrat[1]*, Md Rasel Ul Alam[2], Kanita Haider[3],

[1] *Department of Marketing, University of Barishal, Barishal, Bangladesh*
[2] *Department of Computer and Information Sciences, University of the Cumberlands, Kentucky, USA.*
[3] *Department of Computer Science and Engineering, International Islamic University Chittagong, Chittagong, Bangladesh*
**Corresponding author: samrat.meazi1@gmail.com*

Received 19 August 2022; Revised 12 September 2022; Accepted 30 October 2022; Published: 29 November 2022

KEYWORDS

Agentic AI;
Edge computing;
Digital twins;
Explainable AI;
AI security;
AI governance;
Trustworthy autonomous systems;
Multi-agent systems

Abstract

The convergence of agentic artificial intelligence, edge computing, digital twins, cybersecurity, and explainable AI (XAI) is producing autonomous systems capable of multi-step reasoning and real-world action, yet governance and security frameworks have not kept pace with this shift. Recent surveys document that agentic systems introduce reasoning and coordination capabilities absent from earlier reactive AI, while simultaneously expanding the attack surface through prompt injection, memory poisoning, and inter-agent collusion. Parallel literatures on trustworthy edge intelligence and AI-integrated digital twins emphasize simulation-based verification and resource-aware trust management, and separate governance literatures (NIST AI RMF, the EU AI Act, ISO/IEC 42001) note that existing instruments were designed for single-step predictive systems rather than autonomous multi-step action. This paper synthesizes these largely separate literatures into a proposed six-layer architecture: edge sensing, digital twin verification, agentic reasoning, explainable decision-making, governance and compliance, and human oversight intended to close the gap between autonomous capability and verifiable trust. It maps a cross-layer threat taxonomy drawn from documented attack classes, positions explainability as a mandatory gate rather than an optional add-on, and evaluates the coverage of current governance instruments against each layer, identifying agentic reasoning as the area of greatest regulatory and technical gap. A proposed (not executed) evaluation protocol is outlined for future empirical validation. The paper's contribution is conceptual and integrative: it does not report new experimental results but provides a structured, literature-grounded foundation and research agenda for engineering and governing trustworthy autonomous systems.

1. Introduction

Autonomous artificial intelligence systems are undergoing a structural transition from reactive, single-inference models toward agentic systems capable of planning, invoking external tools, maintaining memory, adapting across sequential interactions, and increasingly coordinating actions among multiple cooperating agents. Recent systematic research characterizes this transition as a movement from predominantly model-centric AI toward goal-directed systems that interact dynamically with their environments, encompassing both symbolic or classical agent lineages based on algorithmic planning and neural or generative lineages built around large language models (LLMs) and prompt-driven orchestration. In parallel, physical and cyber-physical systems are increasingly incorporating digital twins, defined broadly as live, data-synchronized virtual representations of physical assets,

processes, or environments, for prediction, fault detection, simulation, optimization, and operational verification. These digital-twin environments increasingly depend on edge computing to satisfy the latency, bandwidth, locality, and real-time processing requirements of distributed physical systems. Together, agentic autonomy, digital-twin simulation, and edge-based computation are converging across domains including industrial automation, smart infrastructure, healthcare, autonomous mobility, and other cyber-physical applications. However, technological convergence does not inherently produce architectural or governance integration. The literatures addressing agentic AI, digital twins, edge intelligence, cybersecurity, explainability, and AI governance have largely evolved in parallel, frequently employing different architectural assumptions, threat models, evaluation metrics, and governance requirements.

The growing ability of autonomous systems to perform multi-step reasoning and execute actions with reduced human intervention consequently introduces a set of interconnected technical and governance challenges. The attack surface expands because agentic architectures introduce security concerns such as prompt injection, memory and retrieval poisoning, tool-use hijacking, malicious manipulation of external context, and potentially coordinated or collusive behavior among interacting agents. These risks differ materially from those associated with conventional single-inference machine-learning systems because the agent can retain state, invoke external resources, modify intermediate plans, and potentially influence its environment through tools or actuators. Explainability also becomes more difficult precisely when transparency is most consequential. A single-shot predictive model can often be examined through feature-attribution or other localized explanation methods, whereas a multi-step agent may generate intermediate plans, retrieve external information, invoke tools, revise its reasoning state, and ultimately execute an action. Consequently, explaining only the final output may provide an incomplete representation of the decision process. Governance presents an additional challenge because established instruments, including the NIST AI Risk Management Framework (AI RMF), the EU AI Act, and ISO/IEC 42001, were developed within broader AI-risk and management contexts that did not specifically originate around mature autonomous agentic systems capable of sustained multi-step interaction and execution. The resulting challenge is therefore not simply to govern an AI model, but to govern an integrated computational system in which perception, reasoning, simulation, security controls, explanation, and action are interconnected. Despite rapid development across these research areas, an important integration gap remains. Existing agentic-AI literature provides extensive discussion of agent architectures, planning, tool use, memory, coordination, and applications, but security, verification, and explainability are frequently treated as adjacent concerns rather than as structurally coupled components of the agentic execution pipeline. Research on edge intelligence and digital twins emphasizes distributed computation, real-time processing, simulation, synchronization, reliability, and trustworthiness, but comparatively less attention has been devoted to integrating these capabilities with LLM-based agentic reasoning and autonomous action. Cybersecurity research documents increasingly sophisticated threats affecting AI agents and tool-using systems, yet these threats are not always mapped to the computational layers through which autonomous decisions propagate toward physical or operational consequences. Similarly, explainable-AI research provides extensive methods for interpreting model outputs but does not necessarily position explanation as an operational control preceding autonomous execution. Governance research increasingly recognizes the challenges posed by autonomous and agentic AI, but a corresponding technical architecture that explicitly connects governance requirements to agent reasoning, edge computation, digital-twin verification, security enforcement, and explainable decision-making remains insufficiently developed. Accordingly, the central research gap addressed in this paper is the absence of a unified architecture that integrates these five dimensions into a coherent, layered system with explicit governance mappings.

This integration problem has both practical and academic significance. From a practical perspective, autonomous systems capable of affecting physical infrastructure, industrial processes, transportation systems, healthcare operations, or other high-consequence environments require more than accurate prediction or effective task completion. They require mechanisms that can constrain, verify, explain, monitor, and audit autonomous actions before those actions produce consequential effects. Digital-twin environments provide an opportunity to introduce simulation-based pre-validation between agentic reasoning and real-world execution, while edge computing can provide localized processing for time-sensitive sensing, anomaly detection, and security enforcement. Explainability can complement these mechanisms by providing interpretable decision information to human supervisors or governance processes before execution. Security controls can operate across the same architecture to detect malicious inputs, compromised context, anomalous agent behavior, or unauthorized tool interactions. Academically, integrating these dimensions makes it possible to examine cross-layer interactions that remain difficult to observe when the underlying research areas are studied independently. For example, digital-twin pre-validation may influence how agentic explanations are generated and evaluated, while multi-agent security risks may require corresponding changes to audit-log

structures, identity management, and governance controls. The architecture therefore treats autonomy not as an isolated model capability but as a system-level property requiring coordinated technical safeguards.

Against this background, the paper pursues three principal objectives. First, it synthesizes the existing literature on agentic AI, edge and digital-twin systems, AI security, explainability, and AI governance into a coherent conceptual map. Second, it proposes a layered architecture that positions digital-twin verification and explainable decision-making as explicit controls within the agentic action pathway rather than as optional post-deployment features. Third, it evaluates the extent to which established governance instruments, specifically the NIST AI RMF, the EU AI Act, and ISO/IEC 42001, correspond to the architectural layers and identifies areas where technical controls and governance requirements require further alignment. These objectives lead to three research questions: RQ1: What architectural layering can coherently integrate agentic reasoning, edge computation, and digital-twin verification? RQ2: What cross-layer threat taxonomy emerges from this integration, and how should explainability be positioned within that taxonomy? RQ3: To what extent do current governance frameworks cover the resulting architecture, and where are the principal areas of architectural and governance misalignment?

The paper makes four principal contributions. First, it develops a six-layer reference architecture that synthesizes concepts from agentic-AI, edge-computing, and digital-twin research into a unified system model. Second, it develops a cross-layer threat taxonomy grounded in documented attack classes associated with agentic and AI-enabled systems, connecting threats to the architectural components through which they can propagate. Third, it proposes an explainability pipeline in which human-readable decision information functions as an operational control within the autonomous-action pathway rather than merely as a post-hoc reporting mechanism. Fourth, it provides a qualitative mapping of the NIST AI RMF, EU AI Act, and ISO/IEC 42001 onto the proposed architecture and specifies a proposed, but not yet executed, evaluation protocol for future empirical validation. The framework is consequently positioned as a synthesis and architectural proposal rather than as an empirically validated autonomous-system implementation.

The remainder of the paper is organized as follows. Section 2 reviews the literature thematically across agentic AI, edge computing, digital twins, cybersecurity, explainability, and AI governance and provides a comparative synthesis of the constituent research streams. Section 3 introduces the proposed six-layer architecture and explains the functional relationships among sensing, edge computation, digital-twin verification, agentic reasoning, explainable decision-making, and governance. Section 4 develops the cross-layer threat taxonomy and maps relevant attack classes to architectural components and potential system consequences. Section 5 details the proposed explainability pipeline and its role as a control mechanism within autonomous decision execution. Section 6 maps the NIST AI RMF, EU AI Act, and ISO/IEC 42001 onto the proposed architecture and identifies areas requiring further alignment. Section 7 outlines the proposed evaluation methodology, including architectural, security, explainability, latency, and governance-oriented evaluation procedures. Section 8 discusses the theoretical and practical implications of the framework. Section 9 presents limitations and methodological boundaries, Section 10 identifies directions for future empirical research, and Section 11 concludes the paper.

framework. Section 8 discusses limitations, and Section 9 concludes with future research directions. This structure progresses naturally from literature and architecture to risk, governance, evaluation, limitations, and future research.

2. Literature Review

This review is organized thematically rather than chronologically, following the five constituent domains of the proposed architecture.

2.1 Agentic AI Architectures

Foundational work on agent-based software engineering established autonomy, reactivity, and social ability as defining agent properties (Jennings, 2000), while the more recent ReAct framework operationalized agentic behavior in LLMs by interleaving explicit reasoning traces with tool-use actions (Yao et al., 2023), a pattern now standard across agent orchestration frameworks. A systematic PRISMA-based review of ninety studies proposes a dual-paradigm taxonomy distinguishing symbolic/classical agents — which rely on algorithmic planning and persistent state and dominate safety-critical domains such as healthcare — from neural/generative agents, which rely on stochastic generation and prevail in adaptive, data-rich domains such as finance (Abou Ali et al., 2025). That survey explicitly identifies a governance deficit for symbolic systems and calls for hybrid neuro-symbolic architectures, a call this paper's layered design partially answers by pairing neural agentic reasoning (Layer 3) with symbolic, twin-based verification (Layer 2). Domain-specific evaluations, such as an agentic architecture assessed in a frictionless-parking scenario (Khamis et al., 2025), demonstrate feasibility but remain single-domain case studies rather than general architectures.

2.2 Digital Twins and Edge Computing

Trustworthy edge intelligence research frames edge AI trustworthiness along security, reliability, transparency, and sustainability dimensions, and argues that these properties are not auxiliary add-ons but structural requirements for scaling intelligence across the edge-to-cloud continuum (Wang et al., 2023). Digital twin research has increasingly incorporated AI for predictive modeling and fault detection, with edge computing serving as the mechanism that keeps twin synchronization within acceptable latency bounds by performing local preprocessing rather than routing all raw sensor data to the cloud. This body of work motivates Layer 1 (edge sensing and ingestion) and Layer 2 (digital twin simulation and verification) of the proposed architecture, in which the twin functions as a pre-action safety check — simulating a proposed agentic action before it is permitted to affect the physical system. This integration enables continuous validation of autonomous decisions against simulated system states, reducing the risk of unsafe actions in dynamic environments. The verification process can incorporate real-time telemetry to detect discrepancies between predicted and observed system states. Such discrepancies can trigger additional validation or human review before an action is executed. This mechanism strengthens the separation between sensing, simulation, and physical intervention within the architecture. It also enables the system to maintain an auditable record of proposed actions, simulated outcomes, and verification decisions. Consequently, the edge–twin integration provides a foundation for trustworthy autonomous operation under changing environmental conditions.

2.3 Cybersecurity for Agentic and Multi-Agent Systems

The security literature on agentic AI has expanded rapidly. Indirect prompt injection has been benchmarked at scale against tool-integrated LLM agents (Zhan et al., 2024), and related work demonstrates that prompt injection can propagate LLM-to-LLM within multi-agent systems, creating cascading unsafe behavior across agent boundaries (Lee & Tiwari, 2024). Beyond injection, agents interacting in shared environments can establish covert coordination channels; work on secret collusion demonstrates that agents can communicate adversarially via steganographic means undetectable to individual-agent monitoring (Motwani et al., 2024). In response, architectural defenses have been proposed, including access-control and provenance-oriented governance architectures for agentic systems (Syros et al., 2025) and structured threat-modeling frameworks specific to generative AI agents (Narajala & Narayan, 2025). Infrastructure-level proposals argue that securing agent ecosystems requires agent identity, provenance, and containment mechanisms analogous to network-security infrastructure rather than model-level patches alone (Lazar et al., 2025). This literature directly informs Layer 3 of the proposed architecture and the threat taxonomy developed in Section 4.

2.4 Explainable AI (XAI)

Two model-agnostic explanation techniques dominate applied XAI: SHAP, which attributes predictions to input features using a game-theoretic (Shapley value) allocation (Lundberg & Lee, 2017), and LIME, which builds locally faithful linear approximations around individual predictions (Ribeiro et al., 2016). Across application domains, these techniques are consistently framed as necessary — though not sufficient — for transparency, accountability, and regulatory compliance in AI-driven decision systems. A recurring limitation noted across this literature is that post-hoc explanations characterize correlation between inputs and outputs rather than verified causal reasoning, meaning an explanation can be locally faithful yet fail to fully represent an agent's true decision process. This limitation is directly relevant to multi-step agentic systems, where the object requiring explanation is not a single prediction but an extended chain of reasoning and tool-use decisions.

2.5 AI Governance and Regulation

Three instruments dominate current AI governance discourse: the NIST AI Risk Management Framework, organized around four functions — Govern, Map, Measure, and Manage — applied iteratively across the AI lifecycle; the EU AI Act (Regulation (EU) 2024/1689), which imposes binding, risk-tiered conformity obligations on high-risk AI systems; and ISO/IEC 42001:2023, a certifiable AI management system standard. Comparative governance analyses converge on a consistent finding: none of these instruments was designed with autonomous, multi-step agentic action in mind. The EU AI Act's risk tiers assume systems that assist rather than autonomously execute decisions, and the NIST AI RMF's risk model was built around prediction and recommendation rather than autonomous multi-step action. This governance gap — well documented in comparative policy literature — is the central motivation for Section 6 of this paper, which maps existing frameworks explicitly onto the proposed architecture to localize where the gap is largest.

2.6 Literature Comparison

Table 1: summarizes representative studies across the five domains, situating each relative to the gap this paper addresses.

Study	Year	Domain Focus	Key Contribution	Gap Relevant to This Work
Jennings	2000	Agent-based software engineering	Established foundational principles of autonomous, goal-directed software agents.	Predates LLM-based reasoning; no treatment of neural agent unpredictability.
Yao et al. (ReAct)	2023	LLM agent reasoning	Introduced interleaved reasoning-and-acting loops now standard in agent frameworks.	Focused on reasoning efficacy, not safety, twin-based verification, or governance.
Abou Ali, Dornaika, & Charafeddine	2025	Agentic AI architectures (survey)	Dual-paradigm taxonomy (symbolic vs. neural/generative) across 90 studies; flags governance gap for symbolic systems.	Does not integrate edge computing, digital twins, or cybersecurity into a single architecture.
Khamis, Elshakankiri, & Elsayed	2025	Agentic AI system architecture	Evaluated an agentic architecture in a concrete operational scenario.	Single-domain case study; no cross-domain trust or explainability layer.
Wang, Wang, Wu, Ning, Guo, & Yu	2023	Trustworthy edge intelligence	Comprehensive taxonomy of security, reliability, transparency, and sustainability requirements for edge AI.	Edge-centric; does not extend to agentic decision-making or digital twin co-simulation.
Motwani et al.	2024	Multi-agent security	Demonstrated secret collusion between AI agents via steganographic channels.	Identifies a threat class this paper's governance layer must explicitly monitor for.
Zhan, Liang, Ying, & Kang (InjecAgent)	2024	Agent security benchmarking	Benchmarked indirect prompt injection across tool-integrated LLM agents.	Benchmarks attacks but does not propose an integrated, cross-layer defense architecture.
Syros, Suri, Ginesin, Nita-Rotaru, & Oprea (SAGA)	2025	Agentic AI security architecture	Proposed a governance architecture for agentic systems with access control and provenance.	Does not integrate digital twin pre-validation or edge-layer constraints.
Thiebes, Lins, & Sunyaev	2021	Trustworthy AI (conceptual)	Defined trustworthy AI along fairness, transparency, reliability, and safety dimensions.	General framework; not adapted to multi-step autonomous, embodied agentic contexts.

3. Proposed Unified Architecture

Building on the literature synthesized in Section 2, this paper proposes a six-layer reference architecture (Figure 1) in which autonomous action flows upward through observation and downward through constraint. Layer 1 (edge sensing, compute, and data ingestion) performs local preprocessing and latency-aware offloading, consistent with edge intelligence design principles (Wang et al., 2023). Layer 2 (digital twin simulation and verification) uses a synchronized virtual model to simulate a proposed action's physical consequences before it is authorized — an application of digital twin technology as a safety pre-validation gate rather than a purely descriptive or predictive tool. Layer 3 (agentic reasoning and orchestration) performs planning, tool use, and multi-agent coordination using ReAct-style reasoning loops (Yao et al., 2023) and the taxonomic distinctions established by Abou Ali et al. (2025). Layer 4 (explainable decision-making) attaches either post-hoc (SHAP/LIME-style) or ante-hoc interpretable rationales to every action proposal. Layer 5 (governance and compliance) logs decisions, explanations, and interventions against NIST AI RMF, EU AI Act, and ISO/IEC 42001 requirements. Layer 6 (human oversight) retains override authority and receives escalations. This separation of functions establishes clear control boundaries between autonomous reasoning, system validation, explainability, and human intervention. Inter-layer coordination enables continuous monitoring of decisions, system states, and governance constraints throughout the operational cycle. Together, these layers provide a defense-in-depth structure that balances autonomy, safety, transparency, accountability, and operational efficiency. The architecture also supports feedback loops through which verified outcomes can inform subsequent reasoning and system-state updates. Such feedback enables the autonomous components to adapt while remaining subject to predefined safety, governance, and oversight constraints. The explicit separation of responsibilities further reduces the risk that errors originating in one layer propagate unchecked across the operational pipeline. Cross-layer logging can provide traceability for the progression from sensed conditions to proposed actions, verification outcomes, and final interventions. This design also facilitates modular evaluation by allowing individual layers to be tested independently as well as within the integrated architecture.

Figure 1: Proposed six-layer architecture for trustworthy autonomous intelligence.

The architecture's central design commitment is that Layers 2 and 4 are non-bypassable gates: an agentic action proposal generated in Layer 3 cannot reach physical or high-consequence digital effect without first passing digital twin simulation (Layer 2, downward arrow) and generating a human-interpretable explanation (Layer 4, upward arrow to Layer 5–6). This design directly responds to the dual-paradigm governance gap identified by Abou Ali et al. (2025): rather than choosing between symbolic reliability and neural adaptability, the architecture places a symbolic, simulation-based check around a neural, LLM-based reasoning core. The non-bypassable structure establishes a clear separation between reasoning, verification, and authorization. Layer 2 can identify unsafe or infeasible consequences before an otherwise plausible recommendation is operationalized. Layer 4 ensures that each proposed action is accompanied by an interpretable rationale that can be inspected by governance and oversight components. This arrangement also creates explicit intervention points for rejecting, modifying, or escalating actions that fail predefined constraints. By combining simulation-based validation with explanation-based accountability, the architecture reduces reliance on the internal behavior of the neural reasoning component alone. The resulting control structure preserves adaptive agentic capabilities while embedding them within verifiable safety and governance boundaries.

4. Cross-Layer Threat Model

Because the architecture spans physical sensing, simulation, reasoning, explanation, and governance, its attack surface cannot be characterized at a single layer. Figure 2 maps representative, documented attack classes from the reviewed security literature onto each layer.

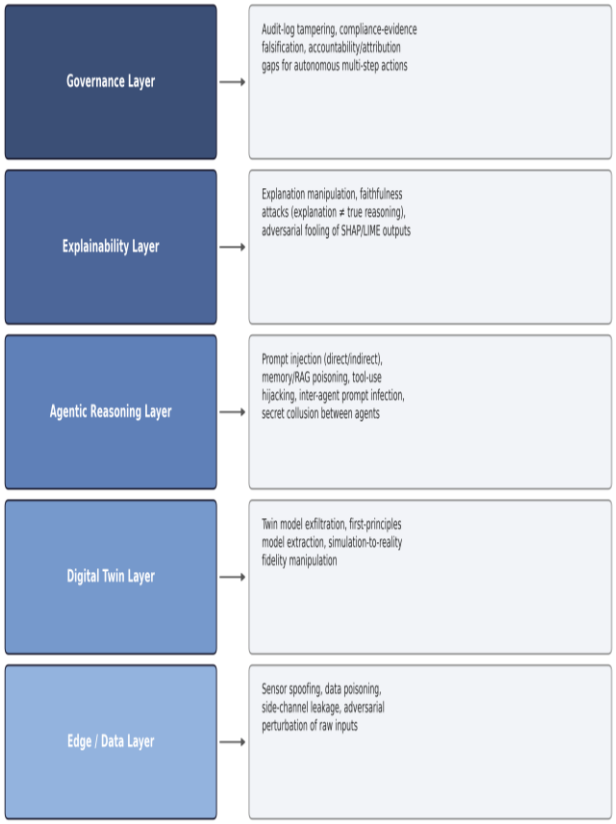

Figure 2: Representative attack surface mapped to each layer of the proposed architecture.

At the edge/data layer, sensor spoofing and data poisoning corrupt the observations that downstream layers treat as ground truth. At the digital twin layer, model or twin extraction attacks threaten proprietary first-principles models embedded in the twin. The agentic reasoning layer carries the largest documented threat diversity: direct and indirect prompt injection (Zhan et al., 2024), inter-agent prompt infection (Lee & Tiwari, 2024), and steganographic collusion between cooperating agents (Motwani et al., 2024) can each cause an agent to act outside its intended policy. The explainability layer is vulnerable to explanation-faithfulness attacks, in which the

generated explanation does not reflect the agent's actual decision process — a risk that undermines Layer 4's function as a gate rather than a report. Finally, the governance layer itself is a target: audit-log tampering and compliance-evidence falsification would defeat the accountability purpose of Layer 5. Architecturally, this argues for treating governance-layer logging as append-only and cryptographically verifiable, consistent with provenance-oriented proposals in the agentic security literature (Syros et al., 2025; Lazar et al., 2025).

5. Explainability and Human Oversight

Section 2.4 noted that post-hoc explanation techniques characterize correlation rather than verified causal reasoning. The proposed architecture addresses this by routing every agentic decision through one of two explanation paths before it reaches a human reviewer (Figure 3): a post-hoc path (SHAP, LIME, or counterfactual explanation) for opaque, end-to-end neural components, and an ante-hoc path for components with inherently interpretable decision structure, such as rule-constrained planners.

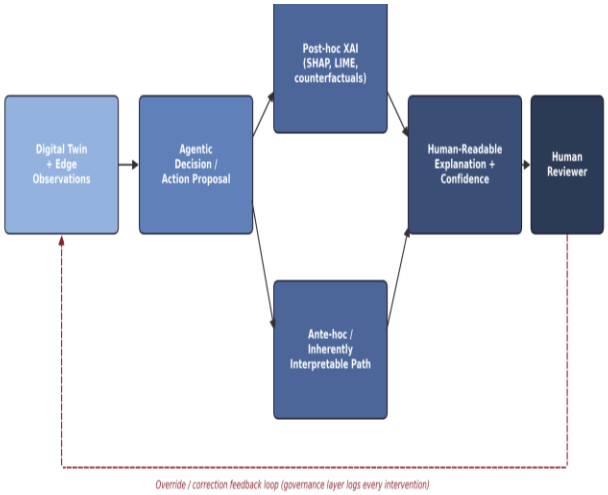

Figure 3: Explainability pipeline linking agentic decisions to human review.

Both paths converge on a human-readable explanation paired with a confidence indicator, which the human reviewer may approve, escalate, or override. Every override is logged by the governance layer, closing the loop shown as the dashed feedback path in Figure 3 and creating the audit trail that governance instruments such as the EU AI Act require for high-risk systems. This design treats explanation not as documentation produced after a decision, but as a precondition for autonomous action to proceed — directly addressing the explanation-faithfulness attack surface identified in Section 4. By embedding explanation as a structural prerequisite, the pipeline ensures that autonomy is conditional on interpretability. The dual-path design also provides redundancy in assurance, allowing opaque and interpretable components to be governed under a unified framework. Confidence indicators act as risk signals, enabling reviewers to calibrate oversight intensity based on uncertainty levels. The override logging mechanism creates traceable accountability, linking human interventions to governance outcomes. Ultimately, the architecture operationalizes explainability as a governance-enforced safeguard, not a discretionary add-on, aligning technical design with regulatory mandates. By requiring explanations before execution, the system reframes autonomy as conditional trust, where decisions are only enacted once interpretability thresholds are met. This approach also creates a living compliance record, ensuring that regulatory audits can verify not just outcomes but the reasoning paths that produced them. The confidence signal can further support risk-based escalation by directing human attention toward decisions with greater uncertainty or potential impact. Thresholds for approval, escalation, and override should be defined according to the operational context and risk classification of each use case. The governance layer can additionally retain model versions, input metadata, explanation outputs, and reviewer actions to strengthen traceability. Such records enable retrospective analysis of recurring failures and provide evidence for refining models, controls, and oversight procedures. Periodic reviews of explanation quality are also necessary to ensure that generated rationales remain faithful to the underlying decision process as models evolve. Where explanation and model behavior diverge materially, execution should be suspended or escalated until the discrepancy is investigated and resolved. This mechanism also supports continuous governance by enabling deviations between intended policies, generated explanations, and observed system behavior to be detected and addressed throughout the system lifecycle. In this way, explanation fidelity, confidence calibration, and human intervention records become interconnected assurance signals that can support ongoing model monitoring, compliance assessment, and responsible autonomy. This integrated mechanism further strengthens accountability by linking interpretability requirements directly to execution control. It thereby supports continuous alignment between autonomous decision-making, human oversight, and evolving governance requirements.

6. Governance and Regulatory Alignment

Table 2 summarizes the three governance instruments most frequently invoked in current AI regulation and standardization discourse, together with the gap each shows for agentic systems specifically.

Table 2: Governance instruments and their documented gaps for autonomous, multi-step agentic action.

Instrument	Nature	Core Mechanism	Documented Gap for Agentic Systems
NIST AI RMF 1.0	Voluntary framework (U.S.)	Four functions — Govern, Map, Measure, Manage — applied across the AI lifecycle.	Built around single-step predictions and recommendations, not multi-step autonomous action.
EU AI Act (Reg. (EU) 2024/1689)	Binding regulation	Risk-tiered conformity assessment, mandatory for high-risk systems.	Negotiated before the scale-up of agentic systems; risk tiers assume human-assisted, not autonomous, decisions.
ISO/IEC 42001:2023	Certifiable management standard	Establishes an auditable AI Management System (AIMS) with continual improvement requirements.	Provides organizational scaffolding but no agent-specific technical controls for tool use or inter-agent action.

Figure 4 extends this comparison by qualitatively mapping each instrument's documented scope onto the six architectural layers.

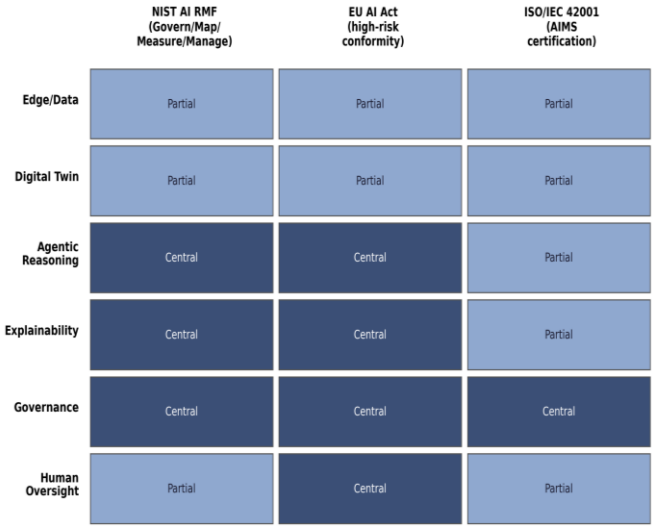

Figure 4: Qualitative mapping of governance-framework emphasis onto architecture layers.

The mapping in Figure 4 is qualitative — derived from the documented scope statements of each instrument rather than from a quantitative audit — and is intended to localize, not resolve, the governance gap. Two patterns are notable. First, all three instruments show only partial coverage of the edge and digital-twin layers, reflecting that these instruments were developed primarily for model-level, not system-level, risk.

Second, and consistent with the comparative governance literature reviewed in Section 2.5, coverage of the agentic reasoning layer under ISO/IEC 42001 is comparatively partial relative to NIST AI RMF and the EU AI Act, both of which have been more directly extended (through supplementary guidance and amendment activity, respectively) toward autonomous-agent risk. This suggests that organizations deploying agentic systems within the proposed architecture cannot rely on a single instrument and should treat Layer 5 as an integration point across all three.

7. Proposed Evaluation Framework

Note on status: no experiments have been executed for this paper. This section outlines a protocol for future empirical validation of the proposed architecture; all content below is prescriptive and illustrative, not a report of results.

A future empirical evaluation of the architecture in Section 3 would need to assess two properties jointly: the degree of autonomous action a configuration permits, and the degree to which its decisions remain verifiable and explainable. Figure 5 positions several illustrative system designs — not measured configurations — along these two conceptual axes to clarify the target design zone the proposed architecture aims for.

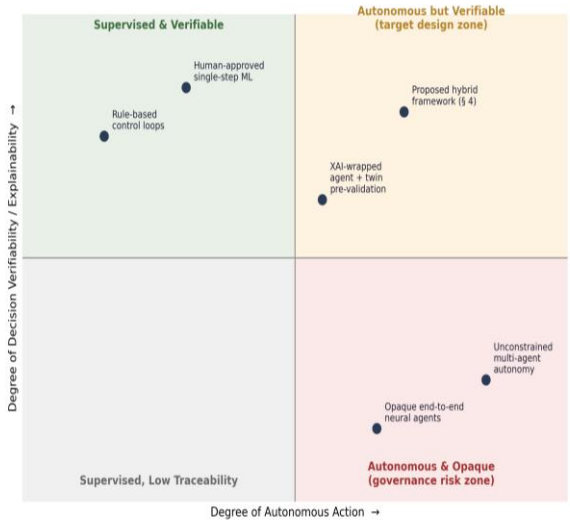

Figure 5: Conceptual positioning of system designs on autonomy vs. verifiability axes (illustrative, not measured data).

A concrete evaluation protocol would proceed along the following lines. Datasets and testbeds: benchmark agent-security testbeds such as those used for indirect prompt

injection evaluation (e.g., InjecAgent-style tool-integrated environments; Zhan et al., 2024) and multi-agent collusion probes (Motwani et al., 2024) would be extended with a digital-twin simulation harness so that agentic actions can be safety-checked before being scored. Baselines: an unconstrained agentic baseline (no twin gate, no mandatory explanation) would be compared against the proposed gated architecture and against a fully symbolic, rule-based baseline. Metrics: attack success rate under documented prompt-injection and collusion attack classes; explanation fidelity (agreement between the generated explanation and a ground-truth decision trace, where available); false-escalation and false-approval rates at the human-oversight gate; and end-to-end task completion rate as a utility check that safety gating does not eliminate the system's usefulness. Statistical protocol: repeated trials across randomized seeds with confidence intervals on each metric, and ablations removing each of Layers 2, 4, and 5 individually to isolate their marginal contribution to safety and utility. Reporting all such results, when eventually collected, should follow the labeling conventions used throughout this paper: illustrative or simulated results must be clearly distinguished from claims of real-world deployment performance. Additional methodological safeguards would include cross-lab replication, ensuring that results are not artifacts of a single institutional environment. Evaluation should incorporate heterogeneous agent models, spanning both transformer-based and symbolic planners, to test generalizability. A temporal replication protocol is necessary, repeating experiments across tool versions to capture drift in vulnerability prevalence. Human reviewers participating in oversight gates should be stratified by expertise level, allowing analysis of whether professional background influences escalation accuracy. The digital-twin harness should log counterfactual trajectories, enabling reconstruction of what would have occurred absent safety gating. To strengthen statistical validity, bootstrapped confidence intervals should be reported alongside traditional measures. Governance relevance requires policy-mapped reporting, linking each experimental outcome to regulatory categories such as those in the EU AI Act. Experiments should also track resource overhead, quantifying latency and compute costs introduced by safety layers. A multi-metric dashboard is recommended, presenting security, utility, and governance indicators in parallel for enterprise decision-makers. Longitudinal tracking of technical debt accumulation in AI-generated code should be integrated into the evaluation cycle. The protocol should include stress tests under adversarial load, simulating high-frequency injection attempts to probe resilience. Comparisons across baselines must be normalized for task complexity, ensuring fairness in utility scoring. A blind-review mechanism could be introduced, where human reviewers assess explanations without knowing whether they were AI- or rule-generated. The evaluation should explicitly measure auditability, verifying whether logs are sufficient for compliance-grade traceability. A failure-mode taxonomy should be developed, categorizing observed breakdowns into reproducible classes. Results should be archived in open repositories, enabling independent re-analysis and transparency. Enterprises adopting the protocol should establish continuous monitoring pipelines, embedding evaluation into production governance. Finally, the evaluation must remain adaptive, updating metrics and baselines as AI code-generation tools evolve. This ensures that the protocol functions not as a static experiment design, but as a living governance instrument aligned with enterprise risk management.

8. Discussion and Limitations

The architecture proposed in this paper should be understood as a conceptual synthesis of documented findings, established methodological approaches, and complementary design patterns rather than as a fully implemented or empirically validated system. Accordingly, the proposed framework demonstrates how existing techniques can be integrated into a coherent architecture, but it does not yet provide empirical evidence that the complete six-layer pipeline improves security, decision quality, interpretability, or operational performance under real-world conditions. Several limitations therefore need to be considered when interpreting the framework and its potential applicability.

First, the governance mapping presented in Section 6 is qualitative rather than the result of a formal compliance assessment. The mapping is based primarily on the authors' interpretation of the stated scope, objectives, principles, and requirements of the relevant governance instruments, standards, and regulatory frameworks. Although this approach is appropriate for establishing conceptual alignment between the proposed architecture and existing governance requirements, it cannot establish legal compliance or regulatory conformity. In practice, compliance may depend on jurisdiction-specific legislation, sectoral rules, contractual obligations, organizational policies, implementation details, data-processing arrangements, and interpretations issued by competent regulatory authorities. A rigorous compliance audit would therefore require detailed legal analysis, standards expertise, evidence of actual system implementation, and potentially independent assessment by qualified auditors. Future research should operationalize the governance layer through a formal control-to-requirement mapping in which each architectural

component is linked to specific regulatory provisions, standards clauses, accountability requirements, audit evidence, and measurable compliance indicators.

Second, the threat taxonomy presented in Section 4 should be regarded as a time-bounded representation of the current security landscape rather than a definitive or permanent classification. The taxonomy is grounded in documented attack classes and emerging risks associated with AI-enabled and agentic systems, but the underlying threat environment is evolving rapidly. New attack techniques, prompt-based attacks, tool-use vulnerabilities, agent hijacking strategies, data-poisoning methods, indirect prompt injection, model manipulation techniques, and multi-agent attack patterns continue to emerge as AI systems become increasingly capable of autonomous reasoning and interaction with external tools. Consequently, a taxonomy developed at a particular point in time can become incomplete as new attack surfaces and exploitation techniques appear. This limitation is especially important for agentic architectures because the risk surface extends beyond the model itself to include tools, APIs, external data sources, memory, identity and access controls, orchestration mechanisms, and downstream systems. A deployed implementation should therefore treat threat modelling as a continuous process rather than a one-time design activity. Future work should incorporate periodic threat-intelligence updates, adversarial testing, red-team exercises, incident feedback, and automated vulnerability discovery into the security layer so that the taxonomy can evolve alongside the threat environment.

Third, the explainability architecture inherits fundamental limitations associated with post-hoc explainable artificial intelligence (XAI) methods. Techniques such as SHAP and LIME can provide useful approximations of feature importance or local decision behaviour, but the resulting explanation does not necessarily constitute a faithful representation of the model's actual internal reasoning process. In particular, an explanation can be plausible and useful to a human reviewer while still failing to capture the causal mechanisms that produced the model output. This creates a distinction between interpretability, explanatory usefulness, and explanation fidelity, which should not be treated as equivalent properties. The ante-hoc pathway proposed in Section 5 offers a potentially stronger form of transparency because interpretability is incorporated into the decision mechanism itself. However, ante-hoc interpretability is difficult to guarantee for contemporary large language models whose internal representations and reasoning processes are highly complex and not inherently transparent. The proposed architecture therefore does not assume that every AI decision can be made fully interpretable. Instead, it treats explanation generation as a risk-sensitive mechanism whose level of transparency should depend on the nature and consequences of the decision. Future empirical research should compare post-hoc and ante-hoc approaches using established measures of explanation fidelity, stability, comprehensibility, and usefulness to human decision-makers.

Fourth, the proposed six-layer separation may introduce computational and coordination overhead that has not yet been empirically quantified. The architecture intentionally separates reasoning, threat assessment, simulation or digital-twin verification, explainability, governance checking, and final action authorization. This separation improves analytical clarity and creates multiple opportunities to detect unsafe or non-compliant actions before execution. However, each additional verification stage can introduce processing time, communication overhead, synchronization requirements, and potential points of failure. The issue becomes particularly important in genuinely time-critical control environments in which delayed action may itself create operational risk. For example, a decision that requires reasoning, threat analysis, digital-twin simulation, explanation generation, and governance verification before execution may be safer in a high-risk but non-urgent context, yet inappropriate when milliseconds or seconds determine system stability. The architecture therefore involves a fundamental safety-versus-latency trade-off that cannot be resolved conceptually alone. Future research should experimentally characterize the latency introduced by each layer, identify which verification stages can operate in parallel, establish risk-dependent bypass or escalation mechanisms, and determine appropriate latency thresholds for different classes of decisions.

Fifth, the digital-twin component itself introduces assumptions that require empirical validation. The usefulness of simulation-based verification depends on how accurately the digital twin represents the relevant physical, cyber, operational, and environmental characteristics of the target system. A digital twin that omits important system dependencies, uses outdated parameters, or fails to represent rare but consequential states may provide a false sense of safety. Conversely, increasing model fidelity can substantially increase computational requirements and simulation time. The proposed architecture therefore assumes that an appropriate balance between simulation fidelity and computational efficiency can be achieved, but this assumption has not been tested experimentally in the present study. Future implementations should evaluate twin fidelity, calibration procedures, uncertainty propagation, model drift, and the extent to which simulated outcomes predict real-world system behaviour.

Sixth, the architecture does not empirically establish the effectiveness of its layered gating mechanism against

adversarial or erroneous decisions. The presence of multiple verification layers does not automatically guarantee improved security. Attackers may attempt to manipulate intermediate outputs, exploit dependencies between layers, generate apparently compliant explanations, or target the interfaces connecting the components. Similarly, an error produced early in the pipeline could potentially propagate through subsequent layers if later components accept the earlier output without sufficient independent verification. The framework therefore assumes a degree of independence and robustness among its verification mechanisms that should be tested rather than presumed. Future work should evaluate the architecture under controlled adversarial scenarios, including attempts to bypass individual layers, manipulate explanations, corrupt simulation inputs, exploit tool interfaces, and induce unsafe final actions.

Finally, the principal contribution of this paper is conceptual integration rather than novel algorithmic invention. The framework does not claim to introduce a fundamentally new reasoning algorithm, XAI technique, digital-twin methodology, threat-detection algorithm, or governance standard. Instead, its contribution lies in demonstrating how several individually established approaches can be composed into a structured, defense-in-depth architecture for governing AI-driven decisions. ReAct-style reasoning provides an approach for structured interaction between reasoning and external actions; SHAP and LIME represent established families of post-hoc explanation techniques; digital twins provide a mechanism for pre-action simulation and verification; and governance frameworks provide external constraints and accountability structures. The novelty of the proposed work therefore resides primarily in the architectural relationship among these components, including the sequencing, separation of responsibilities, and use of multiple independent gates before consequential actions are executed.

This distinction is important for positioning the contribution accurately. The framework should not be interpreted as evidence that the complete architecture has already demonstrated superior security, explainability, compliance, or operational efficiency. Rather, it provides a testable architectural hypothesis: integrating reasoning, threat assessment, simulation-based verification, explainability, governance enforcement, and action authorization into separate but coordinated layers may provide stronger safeguards than relying on an autonomous reasoning component alone. Establishing whether this hypothesis holds in practice requires implementation and comparative evaluation against appropriate baseline architectures.

Future research should therefore progress from conceptual validation toward prototype implementation, benchmark construction, adversarial testing, latency measurement, human-in-the-loop evaluation, compliance auditing, and real-world pilot deployment. Such studies could quantify the contribution of individual layers through ablation experiments, determine the performance cost of additional safeguards, evaluate whether explanations improve human oversight, and identify circumstances in which a particular verification layer provides the greatest marginal safety benefit. This progression would transform the proposed architecture from a conceptual integration framework into an empirically validated system and would provide stronger evidence regarding its applicability to high-stakes, safety-critical, and governance-sensitive AI deployments.

9. Future Research Directions

Building on the limitations above, five directions merit particular attention:

i. Empirical validation of the proposed evaluation protocol (Section 7) using extended agent-security testbeds and digital-twin simulation harnesses, with explicit ablation of each architectural layer's contribution to safety and utility.

ii. Faithfulness-verified explainability: methods that certify, rather than merely produce, agreement between generated explanations and an agent's actual multi-step decision trace.

iii. Latency-aware twin gating: techniques for reducing the coordination overhead of mandatory digital-twin pre-validation so the architecture remains viable in time-critical control loops.

iv. Cross-instrument governance tooling: shared control libraries and evidence bases that let a single audit trail satisfy NIST AI RMF, EU AI Act, and ISO/IEC 42001 obligations simultaneously, extending recent unification proposals to the agentic-system context.

v. Privacy-preserving federated extensions, allowing edge-distributed twins and agents to share threat intelligence and calibration data without centralizing sensitive operational data.

10. Conclusion

This paper synthesized five largely separate research literatures agentic AI, edge computing, digital twins, cybersecurity, and explainable AI — together with current AI governance instruments, into a proposed six-layer architecture for trustworthy autonomous intelligence. The architecture's central contribution is structural: it treats digital twin simulation and human-interpretable explanation as mandatory gates on agentic action rather than optional or post-hoc features, and it explicitly localizes where current governance frameworks (NIST AI RMF, EU AI Act, ISO/IEC 42001) do and do not cover the

resulting system. The paper reports no executed experiments; its evaluation framework is offered as a protocol for future empirical work, clearly distinguished from findings. As agentic systems continue to acquire the capacity for autonomous, multi-step, real-world action, integrating verification, explanation, and governance at the architectural level — rather than bolting each on separately — is likely to be necessary for these systems to be deployed responsibly in safety- and infrastructure-relevant domains. This paper synthesized five largely separate research literatures—agentic AI, edge computing, digital twins, cybersecurity, and explainable AI—together with contemporary AI governance instruments into a proposed six-layer architecture for trustworthy autonomous intelligence. The architecture's central contribution is structural: rather than treating verification, explanation, security, and governance as independent supplementary functions, it positions them as interconnected architectural controls within the pathway from sensing and reasoning to autonomous action. In particular, digital twin simulation and human-interpretable explanation are conceptualized as mandatory decision gates that can intervene before agentic actions are executed, thereby providing opportunities to identify unsafe states, evaluate potential consequences, and communicate the rationale underlying consequential decisions. This design extends conventional agentic architectures by introducing an explicit pre-action verification mechanism between computational reasoning and physical or operational execution. The framework also recognizes that edge computing is not merely an infrastructure choice but an important enabler of timely sensing, localized processing, and low-latency coordination for environments in which centralized decision-making may introduce unacceptable delays. Similarly, cybersecurity is embedded as an architectural responsibility rather than being treated as an external protection layer, allowing security monitoring and anomaly detection to inform downstream reasoning and authority decisions. The inclusion of explainability further establishes a connection between technical decision processes and human oversight by making relevant model reasoning, confidence, and risk information available for review before consequential actions are authorized. An additional contribution is the explicit localization of existing governance requirements within the proposed architecture. The mapping of the NIST AI Risk Management Framework, EU AI Act, and ISO/IEC 42001 demonstrates that these instruments provide important governance principles and organizational controls but do not, individually, constitute a complete technical architecture for continuously operating autonomous systems that combine edge intelligence, digital twins, agentic reasoning, cybersecurity, and explainable decision-making. The proposed framework therefore does not claim to replace existing governance standards; rather, it provides an architectural structure through which their requirements can be interpreted and operationalized across different system layers. This distinction is particularly important for autonomous systems because governance effectiveness depends not only on policies and documentation but also on whether appropriate controls are technically positioned at the points where decisions are generated, verified, authorized, and executed. By connecting governance requirements with concrete architectural functions, the framework offers a pathway for translating high-level principles such as accountability, transparency, robustness, and risk management into operational control mechanisms. The paper reports no executed experiments, collected datasets, or empirically validated performance results. Accordingly, the proposed evaluation framework should be understood as a methodological protocol for future empirical investigation rather than as evidence that the architecture outperforms existing alternatives. Future studies should evaluate the framework using reproducible datasets and deployment scenarios that jointly capture decision quality, latency, computational overhead, security effectiveness, explanation quality, human-intervention requirements, and system resilience. Such evaluations should also compare the complete architecture against appropriately defined ablation and baseline configurations to determine whether each architectural layer provides measurable incremental value. Longitudinal and scenario-based validation would be particularly valuable because trustworthy autonomous intelligence must remain reliable under changing workloads, environmental conditions, security threats, and organizational constraints. As agentic systems increasingly acquire the capacity to perform autonomous, multi-step, and potentially consequential actions in real-world environments, the distinction between intelligence and trustworthy autonomy becomes increasingly important. An autonomous system may be capable of producing accurate decisions while still creating unacceptable risks if those decisions cannot be verified, explained, constrained, or audited before execution. The proposed architecture addresses this challenge by treating trustworthiness as an emergent property of coordinated sensing, simulation, reasoning, security, explanation, governance, and response mechanisms rather than as a feature added after model development. Consequently, integrating verification, explanation, cybersecurity, and governance at the architectural level—rather than bolting each component onto an otherwise autonomous system—may provide a more robust foundation for responsible deployment in

safety-, infrastructure-, and mission-critical domains. The framework therefore offers a conceptual and methodological foundation for future research aimed at transforming increasingly capable autonomous AI systems into systems that are not only intelligent and responsive, but also verifiable, accountable, interpretable, secure, and governable.

References

Allamanis, M., Barr, E. T., Devanbu, P., & Sutton, C. (2018). A survey of machine learning for big code and naturalness. *ACM Computing Surveys, 51*(4), Article 81. https://doi.org/10.1145/3212695

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Article 3, pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

Arrieta, A. B., 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

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q. V., & Sutton, C. (2021). Program synthesis with large language models. *arXiv*. https://doi.org/10.48550/arXiv.2108.07732

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv*. https://doi.org/10.48550/arXiv.2107.03374

Chen, Z., Zan, D., Yang, J., Yu, J., Cheshkov, A., Zhang, Y., Chao, Q., Guo, Y., & Shi, M. (2021). CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 8696–8708). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.685

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

Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., & Zhou, M. (2020). CodeBERT: A pre-trained model for programming and natural languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1536–1547). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.139

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

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

Jennings, N. R. (2000). On agent-based software engineering. *Artificial Intelligence, 117*(2), 277–296. https://doi.org/10.1016/S0004-3702(99)00107-1

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 4765–4774).

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., & Karri, R. (2022). Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)* (pp. 754–768). IEEE. https://doi.org/10.1109/SP46214.2022.9833571

Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. In *Proceedings of the First International*

Conference on Multi-Agent Systems (ICMAS-95) (pp. 312–319).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115–152. https://doi.org/10.1017/S0269888900008122

Zhou, Y., Liu, S., Siow, J., Du, X., & Liu, Y. (2019). Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.