VOLUME 11 · ISSUE 01

FAIR-A^4: A Governance Model for Auditable, Accountable, Actionable and Attributable Agentic AI

Abbas Yazdinejad, Hadis Karimipour and Ann Fitz-Gerald

AUGUST 31, 2026

This paper surfaces existing gaps in the governance of agentic AI systems. While we have solutions for governing the data and predictions of machine learning systems, there is no precedent for governing systems whose decisions occur through autonomous changes in memory, tool use and state over time. To enable governance of decisions by agentic systems operating at machine speed, we propose FAIR-A^4, which extends the foundational FAIR Principles (Findability, Accessibility, Interoperability and Reusability), from data stewardship to decision stewardship. Under FAIR-A^4, agent memories, tools, actions and responsibilities should be Findable, Accessible, Interoperable, Reusable, Auditable, Accountable, Actionable and Attributable.

INTRODUCTION

Agentic artificial intelligence (AI) systems are coming online with increasing frequency in cybersecurity operations centres, digital health systems and critical infrastructure. They operate at machine speed to sense, reason and act autonomously, with little or no immediate human input. We define an agentic AI system as one whose perceive–decide–act loop runs with substantial autonomy over memory, tools and actions, rather than as a purely predictive model. They maintain stateful memories that evolve over time, call upon tools and take actions. Current frameworks for governing AI systems, including model cards, datasheets, human-in-the-loop requirements and audit logs, are built around data-driven, predictive systems and cannot meaningfully be applied to governing decisions by autonomous systems. This paper surfaces existing gaps in the governance of agentic AI systems. While we have solutions for governing the data and predictions of machine learning systems, there is no precedent for governing systems whose decisions occur through autonomous changes in memory, tool use and state over time. To enable governance of decisions by agentic systems operating at machine speed, we propose FAIR-A⁴, which extends the foundational FAIR Principles (Findability, Accessibility, Interoperability and Reusability),[1] from data stewardship to decision stewardship. Under FAIR-A⁴, agent memories, tools, actions and responsibilities should be Findable, Accessible, Interoperable, Reusable, Auditable, Accountable, Actionable and Attributable. We use a case study in critical infrastructure to expose key governance needs and to show how FAIR-A⁴ can be used to audit, control and account for decisions made at machine speed. FAIR-A⁴ provides a constructive checklist for policymakers, auditors and standards organizations working to govern AI systems operating autonomously at scale in mission-critical settings.

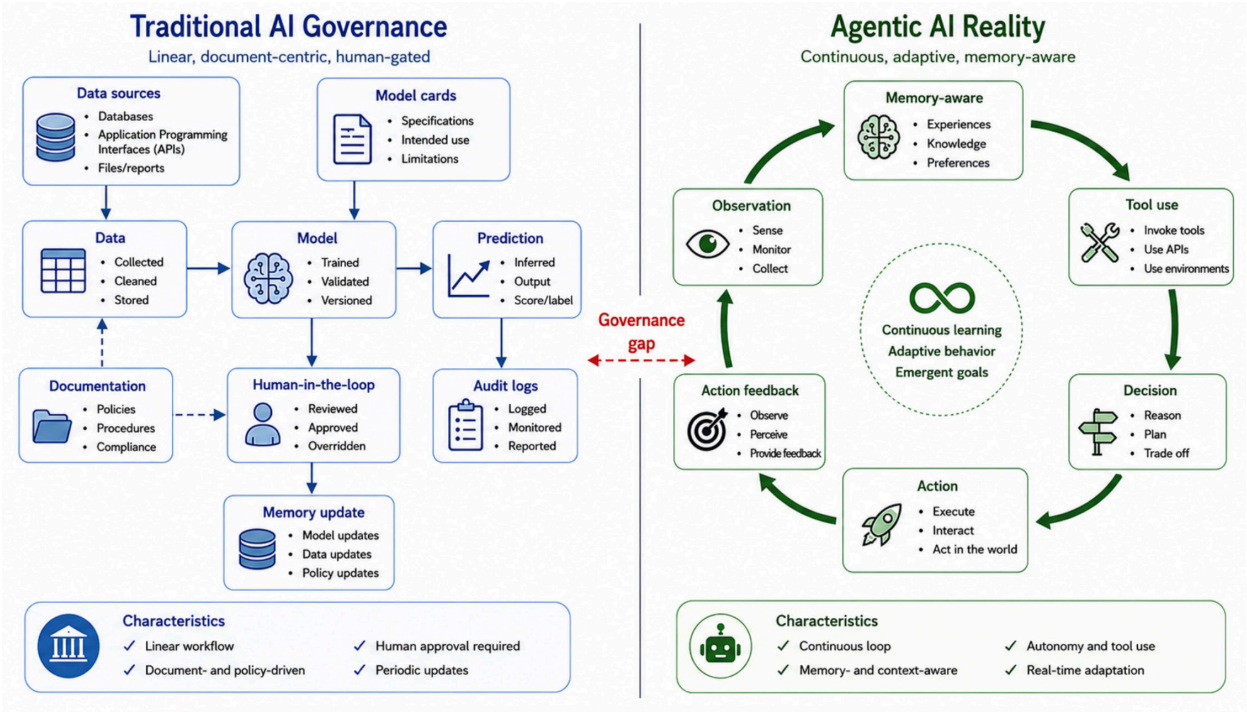
TRADITIONAL AI GOVERNANCE vs AGENTIC AI REALITY

From cybersecurity operations centres to digital health applications to smart grid control rooms, we are beginning to see deployments of agentic AI systems that sense, reason and act with little to no human intervention. In today's security operations centres, autonomous agents already automatically triage alerts, search logs, invoke playbooks and quarantine endpoints within milliseconds. Agentic AI acts at machine speed, shrinking the already tight feedback loop for human intervention. It creates a new governance problem that today's AI accountability tooling cannot address.[2]

The first generation of AI accountability mechanisms operates on the assumption that AI systems are mostly data consumers and prediction machines. In these environments, governing the dataset, documenting the model and assigning blame to a human often suffice for accountability.[3] Agentic AI breaks these assumptions. An agentic system is stateful (it has memory), tool-using (it can call external systems), adaptive (it updates its internal state through interaction) and action-taking (it changes the environment). In the classical sense of Russell and Norvig, an agent is anything that perceives its environment through sensors and acts on it through actuators,[4] and autonomous, multiagent systems of

this kind have been studied for decades.[5] What is new is not agency itself but its pairing with modern AI: agentic AI couples a learned or reasoning-based policy with persistent memory, tool use and the authority to act at machine speed. We use the term "memory" for the stored, inspectable content an agent carries between steps, and "state" for its fuller internal configuration. Because memory is the governable, observable slice of state, it is memory that governance must reach. Our analysis therefore applies to any sufficiently autonomous agentic system, regardless of whether it is built on large language models. The key governance question is no longer "What data trained this model?" but rather, "How did this system decide to act, based on which memory, which tools, and under whose authority?"

Figure 1: From Data-centric AI Governance to Agentic Decision Loops



Source: Authors.

Most AI governance regimes presuppose a *linear data → model → prediction pipeline*. Agentic AIs enact themselves through a cycle of *decisions → memories → tools → actions → updated state*.[6] Frameworks based on data stewardship cannot answer this question because they leave the provenance of decisions, invoked tools, memories written and responsibility shifts untraceable during periods of autonomous function. Privacy/transparency laws are concerned with data used and model interpretability, not decision progeny after model release.[7] To formalize this gap, consider an agent that takes an action a_t at time t:

$$a_t = \pi(m_t, o_t, T, P)$$

where m_t is the agent's memory state, o_t is the current observation, T is the set of available tools, and P is the governing policy. We adopt this notation throughout and reuse rather than redefine it. Existing governance mechanisms document P, and sometimes the model π, but provide little visibility into the evolution of m_t, the use of T, or the traceability of a_t. This paper argues that governing agentic AI requires extending the FAIR Principles (Findable, Accessible, Interoperable and Reusable) of data stewardship into decision stewardship. We introduce FAIR-A⁴, a governance framework ensuring that agent memories, tools, actions and responsibilities are Findable, Accessible, Interoperable, Reusable, Auditable, Accountable, Actionable and Attributable. FAIR-A⁴ preserves the spirit of FAIR while adapting it to systems that act autonomously, providing a structured lens for auditing, certifying and regulating agentic AI in high-stakes domains.

BACKGROUND: FROM FAIR DATA TO AGENTIC SYSTEMS

The FAIR Principles were originally proposed by Mark D. Wilkinson and colleagues as guidelines for scientific data management and stewardship. Introduced as guidelines for research repositories, they quickly became a policy touchstone for open science mandates, funder mandates and journal policies.[8] Today, FAIR badges decorate biomedical repositories, earth observation data portals and national research facility websites.[9] The strength of FAIR lies in a simple insight: when data are openly formatted, licensed and documented, they can flow frictionlessly between repositories, organizations and disciplines. Crucially, FAIR was built around static artifacts. It cannot act, does not learn new behaviours and cannot call external application programming interfaces (APIs). FAIR presumes the hard part is governing who may access data; how that actor behaves afterward lies outside data-centric governance. However, this assumption breaks down with agentic AI,[10] which, arguably, subverts the four properties:

1. **Agents have memory**. Unlike datasets, agents maintain an evolving internal state m_t that changes over time as a result of interaction. Governance frameworks focused on datasets provide no guidance on how this memory should be documented, audited or constrained. Yet this memory directly shapes future decisions.[11]

2. **Agents use tools**. Modern agents operate through tool invocation: APIs, databases, search engines, control interfaces or external software systems.[12] The decision process is therefore not confined to a trained model and its input data, but extends into a network of callable capabilities $T = \{t_1, t_2, \ldots t_n\}$. FAIR ensures interoperability of data formats but does not address interoperability or transparency of tool usage during autonomous execution.

3. **Agents take actions**. Datasets are passive while agentic systems are not. An agent produces actions a_t that change the environment, such as blocking an IP address, modifying a configuration, triggering a workflow, or issuing a recommendation that alters human behaviour.[13] The governance question shifts from *data access* to *action justification*. Existing FAIR-aligned practices cannot trace why a particular action was taken at runtime.

4. **Agents evolve state**. A dataset's content is versioned and static between releases. By contrast, an agent's policy, memory and internal representations evolve continuously during deployment. The system at one time t is not the same system at another time $t + 1$. State that evolves through interaction is a foundational property of agents in the classical literature,[14] not an artifact of recent systems. This temporal evolution breaks the assumption that documentation created at deployment time remains sufficient for accountability later. Recent large-language-model agents such as Voyager make this vivid,[15] but the property long predates them.

Formally, if a classical FAIR-governed system is modeled as:

$$y = f\ (x\ , D)$$

where D is a dataset and f is a model, an agentic system is better represented as:

$$a_t = \pi(m_t, o_t, T) \text{ with } m_{t+1} = u(m_t, o_t, a_t)$$

where memory m_t and actions a_t recursively influence future behaviour. FAIR governs D. It says little about m_t, T, or a_t.

As AI systems transition from data consumers to autonomous actors, governance must extend beyond data stewardship toward decision stewardship. The next section formalizes how and why existing AI governance tools are at best insufficient and, at worst, fail under these conditions, motivating the need for an extended framework tailored to agentic systems. Autonomous agents and multiagent systems are not new: their architectures, norms, organizations and institutions have been studied for decades[16] — for example, through the International Conference on Autonomous Agents and Multiagent Systems (AAMAS) and the long-running Coordination, Organizations, Institutions and Norms (COIN) Workshop series. Our framework complements this tradition by supplying an audit-oriented governance checklist for AI-driven agents once they are deployed in high-stakes settings.

WHERE CURRENT AI GOVERNANCE ASSUMPTIONS BREAK

Influential artifacts such as model cards, datasheets for datasets, human-in-the-loop requirements and audit logging mechanisms have become standard expectations in research, industry and public-sector procurement.[17] These tools were designed for systems in which AI primarily predicts, classifies or recommends. They assume that risk emerges from opaque training data, biased models or insufficient human review. Agentic AI systems violate these assumptions in fundamental ways.[18]

Model cards assume static behaviour. Model cards document how a model was trained, what data it used, its intended use cases and its limitations. However, in agentic systems, the most consequential behaviour occurs after deployment, shaped by runtime memory, tool interactions and evolving state. A model card might explain how the policy π was trained, but it does not explain why the agent, at a time

t, chose an action a_t after consulting memory m_t and invoking tools T. These realities indicate a shift in the governance challenge from model transparency to decision transparency, which model cards are not designed to capture.[19]

Datasheets assume data is the primary risk surface. Datasheets for datasets emphasize provenance, composition, collection methods and ethical considerations of training data. Yet, once deployed, an agent may rely far more heavily on retrieved information, cached memory or external tools than on its original training dataset.[20] The agent's behaviour is increasingly determined by dynamic inputs that are not covered by the original datasheet. Datasheets govern D, but agentic decisions depend on m_t, o_t and T, which lie outside the dataset's documentation scope.

Human-in-the-loop assumes human-tempo decision cycles. Many AI governance frameworks rely on the assumption that a human can meaningfully supervise or override system behaviour. This assumption fails when systems operate at machine speed. In such cases, the human is no longer in the loop, but after the loop. The question becomes not whether a human approved the action, but whether the system's decision process was governable and auditable in the absence of immediate oversight. Human-in-the-loop mechanisms do not scale to environments where reaction time defines safety.[21]

Audit logs assume linear, interpretable action trails. Audit logs record what action was taken and when. This is useful for post-hoc investigation but insufficient for understanding *why* the action occurred. In agentic systems, actions result from a complex chain of memory retrieval, intermediate reasoning, tool outputs and policy decisions.[22] A simple log entry such as:

"IP address blocked at 14:03:12"

does not reveal whether the decision was influenced by outdated memory, incorrect tool output, or flawed reasoning. Without capturing decision provenance, audit logs provide accountability without explainability. As formalized in the background section above, governance artifacts map cleanly onto data-centric AI:

$$\text{Decision} = f\ (x\ ,\ D)$$

where documentation of D and f suffices for accountability. Agentic systems follow a different structure:

$$a_t = \pi(m_t,\ o_t,\ T) \text{ with } m_{t+1} = u(m_t,\ o_t,\ a_t)$$

Governance artifacts largely address π and D, but not m_t , T or the recursive update $u(\cdot)$.

This incongruity results in a blind spot in governance. These systems' highest-leverage behaviour, the way they change over time, the way they interact with tools, and the way they encode decision-relevant memory, are left ungoverned by present mechanisms. This represents the fundamental gap of governance that FAIR-A⁴ aims to solve by extending attention beyond data and models to memory, tools, actions and ultimately responsibility in autonomous systems.

DEFINING FAIR-A⁴: A GOVERNANCE FRAMEWORK FOR AGENTIC AI

To address this gap, we extend the logic of the FAIR Principles from data stewardship to what we call decision stewardship. We introduce FAIR-A⁴, a framework specifying eight governance properties that an agentic system must satisfy to be considered accountable, auditable and governable in high-stakes environments.

FAIR-A⁴ = Findable · Accessible · Interoperable · Reusable · Auditable · Accountable · Actionable · Attributable

Formally, consider an agent operating as:

$$a_t = \pi(m_t, o_t, T, P) \text{ with } m_{t+1} = u(m_t, o_t, a_t)$$

where m_t is memory, o_t observation, T tools, and P governing policy. FAIR-A⁴ defines governance requirements over each of these components. These four principles are chosen not for symmetry with FAIR but to close the four governance gaps identified above. Auditable answers why an action was taken, closing the explanation gap left by audit logs. Accountable maps each contributing component to a responsible actor,[23] closing the responsibility gap left by human-in-the-loop models. Actionable supplies runtime levers to intervene, closing the control gap that opens when systems act faster than humans can supervise. Attributable links every output back through memory and tools to its sources and governing policy,[24] closing the provenance gap left by data-centric documentation. Each targets a distinct gap, so dropping any one leaves part of the decision loop ungoverned; together they cover it end to end. They extend rather than replace existing tools: model cards, datasheets and risk-management frameworks describe how a system was built, whereas these principles govern how it behaves once deployed.

F → Findable (Discoverability of Agent Capabilities)

Governance Property: All agent memories, tools, and capabilities must be *discoverable and catalogued* through structured metadata. For agents, it means a registry of tools T, a description of memory types m_t (short-term, long-term, retrieved), and a catalogue of capabilities the agent can exercise.

Requirement: A capability and memory registry accessible for inspection.

A → Accessible (Controlled Access to Actions)

Governance Property: Access to tools and actions must be governed by explicit policies and permissions. Therefore, governance should define *which tools the agent can call, under what conditions and with what authorization*.

Requirement: Role- or policy-based control over tool invocation and action space.

I → Interoperable (Standardized Schemas for Memory and Tools)

Governance Property: Agent memory formats and tool interfaces must follow standardized schemas. For agents, interoperability means that memory entries and tool calls carry standardized, machine-readable semantics that any compliant auditor or peer system can parse without bespoke adapters. Concretely, it requires:

- Standard representation of memory entries
- Structured tool input/output schemas (e.g., OpenAPI)
- Shared ontologies for decision context

Requirement: Schema-compliant memory and tool interfaces.

R → Reusable (Replayable Decision Workflows)

Governance Property: Agent decisions must be reproducible through replay of memory, observations and tool calls. Instead, one must be able to reconstruct:

$$(m_t, o_t, T) \Rightarrow a_t$$

This transforms decision making into a reusable workflow, much like reproducible experiments. Reusability concerns reproduction, not explanation. For stochastic policies—which are often desirable—it does not demand identical outputs: logging the sampled action together with the random seed and the action distribution makes the decision reproducible in distribution and auditable as a draw from a known policy.

Requirement: Versioned logs enabling deterministic or near-deterministic replay.

A → Auditable (Full Decision Provenance)

Governance Property: Every action must be traceable through a chain of memory retrieval, reasoning steps and tool outputs. Auditability here goes beyond timestamps; it is the ability to explain why a specific action was chosen. It requires capturing:

- What memory was consulted
- Which tools were invoked
- Intermediate reasoning steps
- Evidence used to justify the action

Requirement: A decision provenance graph for each action.

A → Accountable (Responsibility Mapping)

Governance Property: Each component influencing decisions must map to human or organizational responsibility. When an agent acts, responsibility may lie with:

- Tool designers
- Policy designers
- Operators
- Model developers

Accountability requires a mapping between system components and responsible actors, similar to the Responsibility Assignment Matrix (RACI).

Requirement: Explicit responsibility assignment for memory, tools, policies and oversight.

A → Actionable (Governance Intervention Capability)

Governance Property: Governance mechanisms must be able to intervene during operation. Unlike passive auditability, actionable governance means kill switches, sandboxing, rate limits and runtime policy enforcement.

Requirement: Real-time policy enforcement and intervention controls.

A → Attributable (Lineage from Outputs to Sources)

Governance Property: All outputs and actions must be traceable to their originating sources. This includes tracing: *output* → *memory* → *tool* → *data source*. Attribution enables verification, trust and liability assessment. Whereas auditability explains a decision and reusability reproduces it, attributability fixes responsibility, linking each output back to the memory, tools and data that produced it and to the governing policy itself, which must be available for inspection. The three are complementary layers—explain, reproduce, attribute—rather than restatements of one another.

Requirement: End-to-end lineage mapping for every decision.

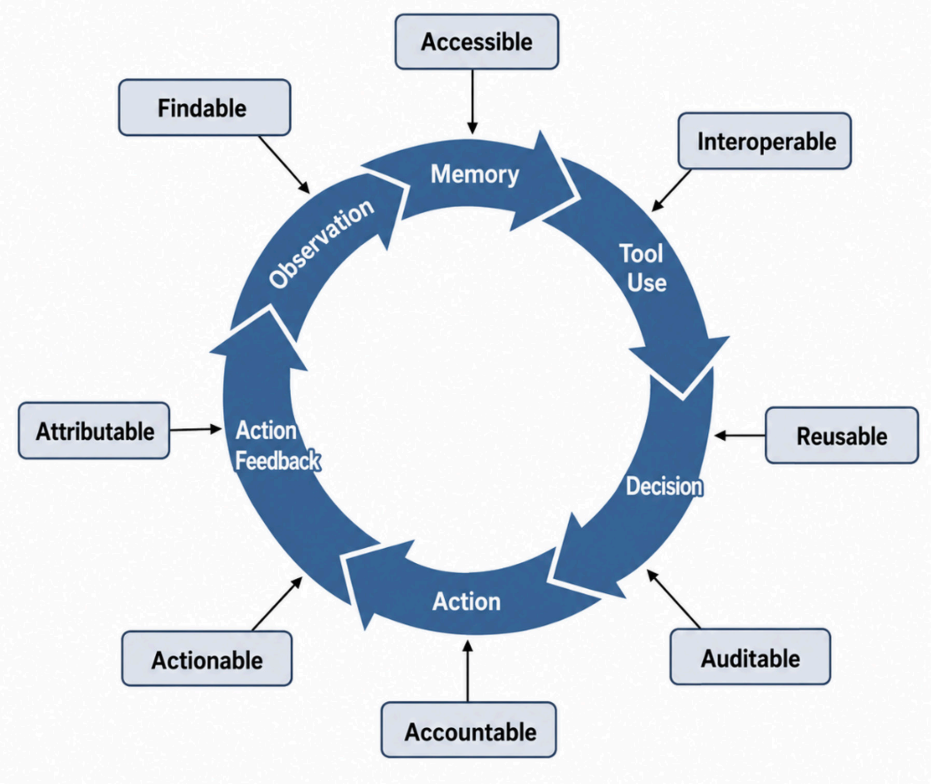
As summarized in Table 1 and illustrated in Figure 2, FAIR-A⁴ maps each principle to a concrete governance requirement and its interpretation within the agent decision loop.

Table 1: Mapping FAIR-A⁴ Principles to Concrete Governance Requirements

Principle	Governance Requirement	Agentic Interpretation
Findable	Capacity registry	Discover memory / tools
Accessible	Policy control	Restrict actions
Interoperable	Standard schemas	Interpretable memory / tools
Reusable	Replay logs	Reconstruct decisions
Auditable	Provenance graph	Explain actions
Accountable	Responsibility map	Assign liability
Actionable	Intervention tools	Govern at runtime
Attributable	Lineage mapping	Trace outputs to sources

Source: Authors.

Figure 2: FAIR-A⁴ Overlay on the Agent Decision Loop



Source: Authors.

FAIR-A⁴ preserves the spirit of FAIR while extending it from governing data artifacts to governing autonomous decision processes.

CASE STUDY: AGENTIC AI IN CRITICAL INFRASTRUCTURE CYBER DEFENCE

Consider a security operations centre (SOC) responsible for protecting a regional power distribution network. To cope with the volume and speed of threats, the SOC deploys an agentic AI system capable of autonomously detecting anomalies, consulting historical context, invoking security tools and executing containment actions before a human analyst can respond. At time t, the agent observes unusual lateral movement patterns in network telemetry:

$$o_t = \text{anomalous traffic signature}$$

Step 1: Querying memory. The agent retrieves prior incidents, known attack patterns, and asset criticality information from its internal memory m_t. **FAIR-A⁴ (Findable, Interoperable)**: Memory entries follow standardized schemas and are indexed in a discoverable registry, allowing auditors to later determine *what* historical knowledge was available to the agent.

Step 2: Tool invocation. Based on memory and observation, the agent invokes multiple tools T:

- Queries a security information and event management (SIEM) system for correlated alerts
- Checks a threat intelligence API
- Consults an automated playbook repository
- Interfaces with the firewall management system.

FAIR-A4 (Accessible, Interoperable): Each tool invocation is governed by explicit policy permissions and uses standardized interfaces, ensuring that the agent's operational reach is controlled and interpretable.

Step 3: Decision and action. The agent determines that the behaviour matches a known lateral movement technique and executes:

$$a_t = \text{isolate host + block IP range}$$

This occurs within seconds, well before the analyst is aware of the alert. **FAIR-A⁴ (Actionable, Accountable)**: Governance policies state that such containment actions are permitted under specific threat confidence thresholds, and responsibility for these policies is clearly assigned to SOC leadership.

Step 4: Decision provenance logging. The system records a structured provenance graph capturing:

- Which memory entries were consulted
- Which tools were used and their outputs
- The reasoning path leading to the action

FAIR-A4 (Auditable, Reusable): Investigators can later replay the exact decision workflow to verify its correctness or improve future responses. In this incident, the graph records that the agent retrieved a prior lateral-movement alert from the same subnet (memory); that the SIEM returned a high correlation score and the threat-intelligence API flagged the source address as malicious (tool outputs); and that its confidence exceeded the containment threshold set by SOC leadership (reasoning). A reviewer can later see exactly why this host was isolated.

Step 5: Attribution. Post-incident review traces the action back through: *action* → *memory* → *tool outputs* → *data sources*. **FAIR-A⁴ (Attributable)**: FAIR-A⁴ provides a structured way to govern each stage of this autonomous decision loop, transforming opaque machine-speed responses into auditable, accountable and policy-compliant actions in critical infrastructure environments.

POLICY AND STANDARDS IMPLICATIONS

The primary value of FAIR-A⁴ is not as an engineering pattern, but as a governance checklist for evaluating whether an agentic AI system is fit for deployment in high-stakes environments. As governments, regulators and standards bodies grapple with the rise of autonomous AI systems, there is an emerging need for criteria that go beyond dataset transparency and model explainability toward operational accountability.

Auditing requirements. Traditional AI audits focus on artifacts created before deployment. FAIR-A⁴ expands this audit surface from static artifacts to the live decision processes of agentic systems. In doing so, auditing shifts from examining what a system *was* at deployment time to understanding how it *behaves* during autonomous operation.

Certification of agentic systems. As formal certification pathways emerge for AI systems in healthcare, cybersecurity and critical infrastructure, FAIR-A⁴ offers measurable governance properties that can be evaluated before approval. Rather than relying on abstract notions such as "explainability," certification bodies can assess whether the system maintains replayable decision logs, whether tool permissions are governed by explicit and enforceable policies, whether responsibility for system behaviour is clearly mapped to human or organizational actors, and whether governance mechanisms can intervene during runtime.

Procurement rules for governments. Public-sector AI procurement increasingly requires evidence of transparency and accountability, yet most checklists remain focused on datasets and predictive models. FAIR-A⁴ reframes procurement toward evaluating how an agent behaves in operation: whether its capabilities and memory are discoverable, its actions are policy-restricted, its decisions are auditable and replayable, and whether authorities can intervene when needed. As shown in Table 2, FAIR-A⁴ provides a simple evaluation matrix that aligns procurement questions with concrete governance principles suited for agentic systems.

Table 2: Procurement Questions Mapped to FAIR-A4 Principles for Evaluating Governability of Agentic AI Systems

Question	Fair-A4 Principle
Can we inspect the agent's capabilities and memory types?	Findability
Can we restrict what the agent is allowed to do?	Accessibility
Can we audit why it acted?	Auditability
Can we stop it if needed?	Actionability

Source: Authors.

AI governance standards bodies. Organizations developing AI governance standards have largely concentrated on data quality, risk management and lifecycle documentation for predictive systems. FAIR-A⁴ introduces a complementary dimension focused on the governance of autonomous operation. By incorporating FAIR-A⁴ principles, standards bodies can define concrete requirements for documenting memory schemas, standardizing tool interfaces, retaining decision provenance records and establishing clear responsibility assignment frameworks.

DISCUSSION

The evolution from traditional AI to agentic AI forces a conceptual shift in how governance is framed. For more than a decade, AI governance has largely meant data stewardship. This approach was sufficient when AI systems primarily consumed data to produce predictions. Agentic systems alter this paradigm. They do not merely use data; they act in environments, consult evolving memory, invoke tools and update their internal state. The governance challenge is therefore no longer centred on *what data was used*, but on *how decisions unfold over time*. This shift can be summarized as: *governing data* → *governing decisions*.

Decision stewardship requires visibility into the temporal lifecycle of agent behaviour: how memory influences reasoning, how tools shape outcomes and how actions propagate into future states. FAIR-A⁴ captures this shift by extending the logic of FAIR from artifacts to actions. In doing so, FAIR-A⁴ reframes AI governance around the idea that accountability in agentic AI is achieved not by documenting how the system was built, but by governing how it behaves.

CONCLUSION

Agentic AI systems are transitioning from labs and proofs-of-concept to operational deployments across cybersecurity, healthcare and critical infrastructure domains. Today's predominant governance mechanisms, model cards, datasheets, human-in-the-loop requirements and algorithmic audit logs were developed with machine-learning predictive AI systems in mind. This paper exposed an urgent governance gap signalling that our current tools cannot explain or audit or, in some cases, even control the decisions of stateful, tool-using, continuously adaptive agents. To begin bridging this gap, we presented FAIR-A⁴, a framework that maps the popular FAIR Principles logic-set to stewardship of autonomous decisions. Requiring agent memories, tools, actions and responsibilities to be Findable, Accessible, Interoperable, Reusable, Auditable, Accountable, Actionable and Attributable, FAIR-A⁴ is focused on governance rather than implementation, and delivers a cohesive lens through which to assess any agentic AI system. FAIR-A⁴ is our initial attempt at developing a framework for doing so, and to help incorporate it into policy and standards.

END NOTES

[1] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3, no. 1 (2016): 1–9.

[2] Thanh Thi Nguyen and Vijay Janapa Reddi, "Deep Reinforcement Learning for Cyber Security," *IEEE Transactions on Neural Networks and Learning Systems* 34, no. 8 (2021): 3779–95; Eric J. Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence," *Nature Medicine* 25, no. 1 (2019): 44–56.

[3] Abbas Yazdinejad, Maral Niazi, James W. Hinton, Jude Kong, Jake Okechukwu Effoduh, and Anna Shin, *Community-Centred Protocol for Ethical and Scalable AI in Health Care* (Centre for International Governance Innovation, 2025). DeBrae Kennedy-Mayo and Jake Gord, "'Model Cards for Model Reporting' in 2024: Reclassifying Category of Ethical Considerations in Terms of Trustworthiness and Risk Management," preprint, arXiv:2403.15394 (2024).

[4] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Pearson, 2021), chap. 2.

[5] Michael Wooldridge, *An Introduction to MultiAgent Systems*, 2nd ed. (John Wiley & Sons, 2009).

[6] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 33–44.

[7] Abbas Yazdinejad, Zahra Dehghani Mohammadabadi, Ali Dehghantanha and Gautam Srivastava, "An Explainable and Privacy-Preserving Federated Learning Model for Threat Detection in Cyber-Physical-Social Systems," *IEEE Transactions on Computational Social Systems* (2025).

[8] Wilkinson, Dumontier, Aalbersberg et al., "The FAIR Guiding Principles."

[9] Barend Mons, Cameron Neylon, Jan Velterop et al., "Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud," *Information Services and Use* 37, no. 1 (2017): 49–56.

[10] Abbas Yazdinejad, Hadis Karimipour and Talal Halabi, "Towards Stress-Adaptive Cyber Defence: Cognitive–Physiological Synchronization in IoT Environments," *IEEE Internet of Things Journal* (2026); Tessa van Elst, Niven Mehra, Sjaak Bloem et al., "The Conundrum of Treating de novo Metastatic Hormone-Sensitive Prostate Cancer," *Scientific Reports* 15, no. 1 (2025): 12500.

[11] Carlo Adornetto, Adrian Mora, Kai Hu et al., "Generative Agents in Agent-Based Modeling: Overview, Validation, and Emerging Challenges," *IEEE Transactions on Artificial Intelligence* (2025).

[12] Abbas Yazdinejad and Jude Dzevela Kong, "Responsible Use of Large Language Models in Digital Health: An Equity First Governance Framework," SSRN working paper 5962741, 2025. Abbas Yazdinejad and Hadis Karimipour, "Temporal Dynamics of Memory Poisoning in Web3-Style LLM Agents," *IEEE Access* 14 (2026): 76200–76221, https://doi.org/10.1109/ACCESS.2026.3693560.

[13] Aman Madaan, Niket Tandon, Prakhar Gupta et al., "Self-Refine: Iterative Refinement with Self-Feedback," *Advances in Neural Information Processing Systems* 36 (2023): 46534–94.

[14] Russell and Norvig, *Artificial Intelligence: A Modern Approach*, 35.

[15] Guanzhi Wang, Yuqi Xie, Yunfan Jiang et al., "Voyager: An Open-Ended Embodied Agent with Large Language Models," preprint, arXiv:2305.16291 (2023).

[16] Abbas Yazdinejad, Hadis Karimipour and Jatin Nathwani, "Governing Machine-speed Cyber Defence: Policy Frameworks for Agentic AI," *Balsillie Papers* 10, no. 3 (2026), https://www.balsilliepapers.ca/governing-machine-speed-cyber-defence-policy-frameworks-for-agentic-ai.

[17] Lindsay Poirier, "Accountable Data: The Politics and Pragmatics of Disclosure Datasets," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 1446–56.

[18] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland and Patrick Vinck, "Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges," *Philosophy & Technology* 31, no. 4 (2018): 611–27.

[19] Joshua Krook, Peter Winter, John Downer and Jan Blockx, "A Systematic Literature Review of Artificial Intelligence (AI) Transparency Laws in the European Union (EU) and United Kingdom (UK): A Socio-Legal Approach to AI Transparency Governance," *AI and Ethics* (2025): 1–22.

[20] Abbas Yazdinejad, "Secure and Private ML-Based Cybersecurity Framework for Industrial Internet of Things (IIoT)" (PhD diss., University of Guelph, 2024).

[21] Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis and Saleema Amershi, "Assessing Human-AI Interaction Early through Factorial Surveys: A Study on the Guidelines for Human-AI Interaction," *ACM Transactions on Computer-Human Interaction* 30, no. 5 (2023): 1–45.

[22] Ramkumar Bharathan, "The Tech Audit Revolution: Navigating the Next 50 Years for Industry and Technology Leadership and an End-to-End Audit Framework for Ads, Search, and Content Platforms," *Technical Disclosure Commons*, July 9, 2025, https://www.tdcommons.org/dpubs_series/8330.

[23] Wilkinson, Dumontier, Aalbersberg et al., "The FAIR Guiding Principles."

[24] Krook, Winter, Downer and Blockx, "Systematic Literature Review of Artificial Intelligence (AI) Transparency Laws."

Abbas Yazdinejad is an Assistant Professor with the Department of Computer Science, University of Regina, SK, Canada, and the Director of the DCAILab. He is also an Adjunct Professor with the Department of Electrical and Software Engineering, University of Calgary, AB, Canada. He has been recognized among the World's Top 2 percent Scientists (Stanford University ranking). His research interests include Agentic AI, Autonomous cybersecurity, federated learning, and AI governance, which have been published in leading venues across AI, cybersecurity, and cyber-physical systems. He is a Balsillie Scholar (January–August 2026) at the Balsillie School of International Affairs, Waterloo, Canada.

Hadis Karimipour is a Canada Research Chair (Tier II) and Associate Professor in the Department of Electrical and Software Engineering at the University of Calgary, where she directs the Smart Cyber-Physical Systems (SCPS) Laboratory. She received her PhD in Electrical Engineering from the University of Alberta in 2016. Her research focuses on AI-enabled cybersecurity, cyber-physical systems, critical infrastructure protection, and resilient energy systems. She has published more than 135 peer-reviewed articles, attracted significant research funding from government and industry, and serves as an Associate Editor for leading journals in cybersecurity and intelligent infrastructure.

Ann Fitz-Gerald is the Director of the Balsillie School of International Affairs and a Professor in Wilfrid Laurier University's Political Science Department. She has worked at both at King's College, London University's International Policy Institute, and at Cranfield University, where she was the Director, Defence and Security Leadership. During her time at Cranfield, Ann led the UK-Government funded Global Facilitation Network for Security Sector Reform and Cranfield's Centre for Security Sector Management.

BALSILLIE PAPERS

ISSN 2563-674X • doi:10.51644/BAP111

balsilliepapers.ca