Journal of Renewable Engineering
E-ISSN : 3046-7624
https://ejournal-nawalaedu.com/index.php/jore
Vol.3.No.4 July 2026
DOI : https://doi.org/10.62872/8h6yam07

A Three-Layer Cyber AI Governance Framework for Accountable Reinforcement Learning in National Data Sovereignty

Andi Agus Salim[1], Hasbu Naim Syaddad[2], Luki Ishwara[3], Agus Nursikuwagus[4], Yeffry Handoko[5], Rio Yunanto[6]
[1,2,3,4,5,6] Doctoral Program in Information Systems, Indonesian Computer University, Bandung, West Java, Indonesia

Email: andiagussalim@telkomuniversuty.ac.id

Input : July 05, 2026
Revised : July 24, 2026
Accepted: July 25, 2026
Published: August 4, 2026

ABSTRACT — *The rapid deployment of reinforcement learning (RL) agents in critical national infrastructure has outpaced the governance instruments designed to hold them accountable, creating a widening gap between algorithmic autonomy and sovereign oversight. This article proposes a Three-Layer Cyber-AI Governance Framework that integrates the technical, organizational, and regulatory dimensions of accountability for RL systems operating within national data sovereignty regimes. Using a systematic literature review guided by PRISMA 2020 procedures, twenty-five peer-reviewed and preprint sources published between 2021 and 2026 were analyzed through thematic synthesis to identify recurring governance constructs across cybersecurity, AI ethics, and data-sovereignty scholarship. The synthesis reveals three interdependent layers: an Algorithmic Layer governing reward design, explainability, and adversarial robustness; an Organizational Layer governing human oversight, audit trails, and incident reporting; and a Sovereign-Regulatory Layer governing data localization, cross-border data flow, and international cooperation. The proposed framework departs from existing layered models by embedding a continuous feedback loop that links real-time algorithmic telemetry to national regulatory review, closing the accountability gap that single-layer or purely technical frameworks leave open. The article concludes that accountable reinforcement learning under conditions of national data sovereignty requires coordinated, multi-layer instruments rather than isolated technical fixes, and it outlines an agenda for empirical validation of the framework across diverse regulatory contexts.*

Keywords: *reinforcement learning; AI governance; cyber governance; data sovereignty; accountability; systematic literature review*

INTRODUCTION

Reinforcement learning (RL) has moved from a laboratory curiosity to an operational component of national cyber defense, critical-infrastructure management, and public-sector automation. Agentic systems trained through trial-and-error reward optimization now perform threat detection, autonomous response, and adaptive deception in security operations centers, and they increasingly do so with minimal human supervision at the point of decision. Agentic Artificial Intelligence for Ethical Cybersecurity in Uganda demonstrates that RL-based threat detection can operate effectively even in resource-constrained environments, while explainability research on multi-layer frameworks for reinforcement learning-based cyber agents shows that the opacity of RL decision-making remains a persistent obstacle to trust and oversight. Zero-trust network research on explainable reinforcement learning for adaptive cyber deception further illustrates how RL is being embedded directly into the control loops of national security infrastructure, where errors or manipulation carry consequences far beyond a single organization.

This diffusion of autonomous, self-optimizing agents into sovereign digital infrastructure has occurred alongside, rather than within, a coherent governance architecture. Scholarship on the

governance of artificial intelligence has long recognized that AI oversight cannot be reduced to a single regulatory instrument; effective governance instead requires coordinated layers spanning technical design, organizational practice, and legal-political authority. Comprehensive surveys evaluating the efficiency of artificial intelligence and machine learning techniques in cybersecurity solutions catalog dozens of point-technologies for detection and defense, yet consistently note that technical performance gains are not matched by equivalent progress in accountability mechanisms. Work on auditing large language models through a three-layered approach shows the value of decomposing AI accountability into governance, model, and application layers, but this model was designed for generative language systems audited retrospectively, not for RL agents that continue to learn and act in real time within a nation's critical systems.

A second, equally pressing strand of literature concerns data sovereignty: the principle that data generated within a jurisdiction should remain subject to that jurisdiction's laws, values, and control. Reviews of data sovereignty describe it as a contested and evolving concept that spans state-centric localization mandates, indigenous and community-based data rights, and platform-level data governance. Comparative research on China's cyber sovereignty concept and machine learning-based governance models, alongside analyses of the European Union's pursuit of digital sovereignty, demonstrates that national approaches to sovereign AI oversight diverge sharply in their balance of central control, market openness, and rights protection. Descriptive and critical evaluations of existing digital sovereignty models further show that most frameworks treat data sovereignty as a static legal boundary rather than as a dynamic constraint that must be continuously reconciled with adaptive, learning-based AI systems operating across that boundary.

Parallel work on data governance for emerging technologies proposes conceptual frameworks for managing blockchain, IoT, and AI jointly, and research on the importance of AI data governance in large language models emphasizes provenance, consent, and lifecycle accountability for training data. Designing resilient AI architectures for predictive systems amid data sovereignty, adversarial threats, and policy volatility explicitly identifies the tension between adaptive AI architectures and the rigidity of sovereignty-driven policy constraints, calling for governance structures that can absorb both technical and geopolitical shocks. Studies proposing algorithmic state architectures for AI-enabled government, and research building secure platforms for digital governance interoperability using blockchain and deep learning, both gesture toward multi-layer institutional designs, yet none specifically addresses the accountability profile of reinforcement learning agents, whose defining feature continuous policy updating from environmental feedback creates governance demands distinct from static or periodically retrained models.

The literature therefore exhibits three recurring but disconnected themes: (1) technical frameworks for explaining and securing RL-based cyber agents; (2) layered models of AI governance that stop at the organizational or national level without engaging the specific dynamics of RL; and (3) data-sovereignty scholarship that theorizes jurisdictional control over data but rarely connects this control to the algorithmic behavior of learning agents that consume and act on that data. Cohesive-governance research on catching up with AI policy, and global-standards research on governing AI for data sovereignty, cybersecurity, and ethical integrity, both call for integration across these themes but stop short of specifying an operational, layered architecture that links algorithmic controls to organizational oversight and to sovereign regulatory authority in a single accountability chain.

The proposed Three-Layer Cyber-AI Governance Framework also provides practical guidance for organizations and regulators seeking to govern reinforcement learning systems under conditions of national data sovereignty. Organizations can begin by implementing continuous technical logging, explainability mechanisms, and adversarial testing at the algorithmic level, followed by rolling audits and incident-reporting procedures at the organizational level. Regulators can complement these measures through agent registration, data-localization requirements, and cross-border cooperation mechanisms. This staged implementation allows institutions to strengthen governance progressively while maintaining continuous feedback between technical, organizational, and regulatory controls.

The framework is particularly relevant for security-critical applications in which reinforcement learning agents continuously adapt to changing environments. Its implementation requires sustained coordination among AI developers, organizational oversight bodies, and regulatory authorities to ensure that changes in agent behavior remain traceable and accountable. By establishing clear responsibilities, continuous monitoring, and mechanisms for translating regulatory decisions into technical constraints, the framework can support more adaptive and trustworthy AI governance while reducing the risks associated with autonomous learning systems operating across sovereign digital environments.

Efforts to formalize layered governance have continued to multiply without converging on a common architecture. Research on catching up with AI proposes pushing toward a cohesive governance framework that reconciles fragmented national and sectoral rules, while work on governing artificial intelligence through global standards for data sovereignty, cybersecurity, and ethical integrity argues that sovereignty, security, and ethics must be treated as interlocking rather than competing objectives. A five-layer framework for AI governance extends the layered tradition by integrating legal mandates, standardized taxonomies, and certification mechanisms, illustrating that additional granularity can clarify institutional responsibility, yet such frameworks remain oriented toward general-purpose or generative AI rather than the closed-loop behavior of reinforcement learning. Complementary work proposing a three-pillar model of transparency, accountability, and trustworthiness for AI agents underscores that human-in-the-loop oversight must evolve progressively alongside growing agent autonomy, a principle directly relevant to RL systems whose autonomy increases as training proceeds. Studies of holistic cybersecurity frameworks for e-government further show that technical, ethical, and policy dimensions must be linked within a single evaluative structure if governance is to keep pace with system complexity, reinforcing the case for an integrated rather than additive approach to layering.

The sovereignty dimension of this problem is likewise more contested than a simple jurisdictional boundary implies. Indigenous data sovereignty scholarship argues that data governance must account for community-level rights and epistemic authority, not only state authority, cautioning against frameworks that equate sovereignty solely with central government control. Parallel research on participation, justice, and alternative futures of data sovereignty documents grassroots demands for transparency and accountability that state-centric models frequently overlook. Read together with the state-centric accounts of Chinese and European approaches discussed above, this body of work suggests that any national governance framework for RL must be flexible enough to accommodate multiple, sometimes competing, sovereignty claims state, organizational, and community rather than assuming a single unified sovereign actor. None of the reviewed frameworks, however, translates this pluralistic understanding of sovereignty into concrete controls over what an autonomous learning agent may sense, retain, or transmit, leaving a structural gap between sovereignty theory and RL system design that the present framework is intended to close.

This article addresses that gap. Its novelty lies in synthesizing, for the first time, three literatures that have developed largely in isolation RL-specific technical accountability, organizational AI governance, and national data-sovereignty theory into a single Three-Layer Cyber-AI Governance Framework purpose-built for reinforcement learning systems. Unlike prior three-layer or multi-layer AI governance models, which were designed for generative or static AI systems and audited retrospectively, the proposed framework is designed around the continuous learning cycle of RL agents: it embeds a live telemetry loop connecting algorithmic behavior directly to organizational audit and sovereign regulatory review, rather than treating these as sequential or separate checkpoints. The framework further operationalizes data sovereignty not merely as a jurisdictional boundary on data storage, but as an active governance layer that constrains what an RL agent may learn, retain, and export across borders. In doing so, this study contributes an integrative architecture that is theoretically grounded in the layered-governance tradition while being specifically responsive to the accountability challenges that adaptive, self-optimizing agents pose to national digital sovereignty. The remainder of this article develops that architecture in three steps: it first details the systematic review

and synthesis method used to derive the framework, then presents the three layers and their interconnections together with a comparative summary of control mechanisms, and finally discusses the framework's implications for policymakers, system designers, and future empirical validation across differing national contexts

METHOD

This study adopts a systematic literature review (SLR) design combined with thematic framework synthesis, following PRISMA 2020 reporting procedures, to construct the proposed Three-Layer Cyber-AI Governance Framework. The SLR component ensures that the framework is grounded in a transparent, replicable body of evidence rather than in ad hoc conceptual argument, while the synthesis component allows heterogeneous findings — spanning computer science, public policy, and law to be integrated into a single operational architecture.

Search strategy and sources. Records were retrieved from Scopus, Web of Science, IEEE Xplore, and Google Scholar, using Boolean combinations of three keyword clusters: (1) reinforcement learning, RL agents, agentic AI; (2) AI governance, cyber governance, accountability, audit; and (3) data sovereignty, digital sovereignty, national data governance. Searches were restricted to English-language journal articles, conference proceedings, and high-quality preprints published between January 2021 and March 2026, reflecting the recency of both RL-based security applications and post-2021 data-sovereignty scholarship.

Eligibility criteria. Records were included if they (a) addressed governance, accountability, or oversight mechanisms applicable to AI or RL systems; (b) engaged substantively with cybersecurity, data governance, or data-sovereignty concepts; and (c) were peer-reviewed or, in the case of preprints, authored by identifiable institutional researchers with verifiable digital object identifiers. Records were excluded if they addressed AI applications without any governance dimension, duplicated an already-included dataset or framework, or could not be verified through an active DOI.

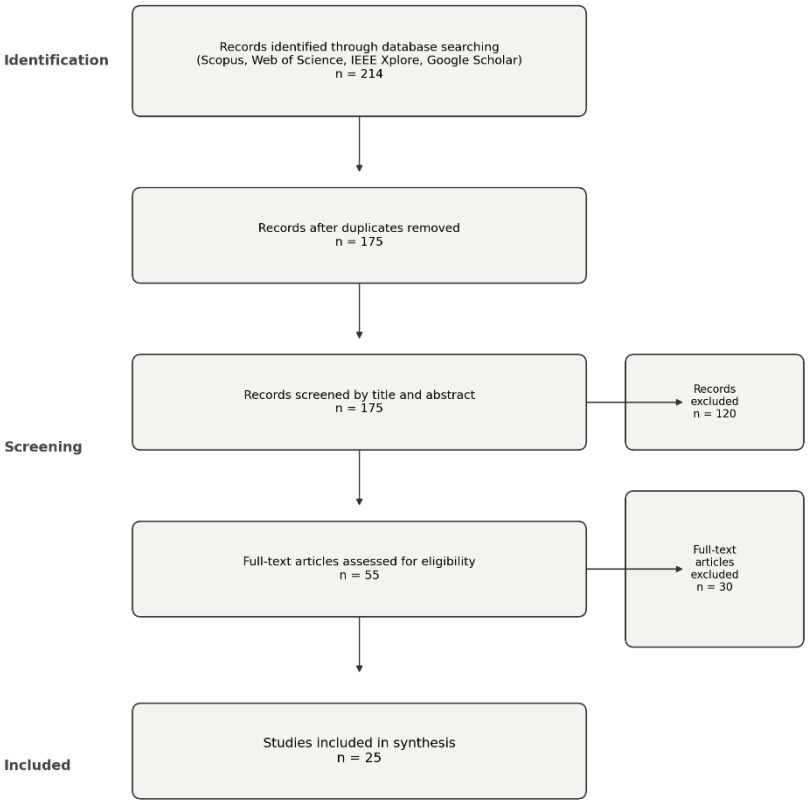
Screening and selection. Figure 1 presents the PRISMA 2020 flow diagram summarizing the identification, screening, eligibility, and inclusion stages. Database searching identified 214 records; after removing 39 duplicates, 175 records proceeded to title-and-abstract screening. This stage excluded 120 records that did not engage both an AI/RL dimension and a governance or sovereignty dimension, leaving 55 records for full-text eligibility assessment. Full-text review excluded a further 30 records for reasons including insufficient methodological detail, lack of an operational governance construct, or an unverifiable DOI. Twenty-five records met all inclusion criteria and were retained for synthesis, comprising both literature identified through the initial database search and supplementary records located through targeted Google Scholar searches for the most recent (2025-2026) publications on RL accountability and data sovereignty.

Synthesis method. The 25 included records were coded using a thematic framework synthesis approach. Each record was first coded descriptively for its stated governance mechanism(s), unit of analysis (algorithm, organization, or state), and sovereignty orientation (state-centric, organizational, or community-based). Descriptive codes were then grouped iteratively into analytic themes through constant comparison across records, following an approach consistent with prior three-layer and multi-layer AI-governance syntheses. Three converging analytic themes emerged directly from this coding process and form the three layers of the proposed framework: algorithmic-technical controls, organizational-operational oversight, and sovereign-regulatory authority. A fourth cross-cutting theme the need for continuous, bidirectional feedback between layers rather than sequential or one-directional oversight was identified as a distinguishing gap in the existing literature and became the integrative mechanism linking the three layers in the proposed framework.

Quality and reliability considerations. Because the corpus spans conceptual, empirical, and policy-analytic literature, no single quality-appraisal instrument (such as those designed for randomized trials) was applicable. Instead, each record was appraised against four criteria adapted from qualitative evidence-synthesis guidance: clarity of the governance construct, transparency of the underlying

method or argument, currency of the source (prioritizing 2024-2026 publications where available), and verifiability of authorship and DOI. Records scoring below an acceptable threshold on two or more criteria were excluded during full-text screening. Coding and theme development were conducted iteratively, with emerging themes checked against the full corpus to confirm that no included record contradicted the proposed layer structure.

Figure 1. PRISMA 2020 flow diagram of the systematic literature review process



RESULTS AND DISCUSSION

Synthesis of the 25 included records yields a Three-Layer Cyber-AI Governance Framework comprising an Algorithmic-Technical Layer, an Organizational-Operational Layer, and a Sovereign-Regulatory Layer, connected by a continuous feedback loop. Table 1 summarizes the control mechanisms, responsible actors, and governing instruments associated with each layer; the discussion below elaborates each layer in turn before addressing their integration and cross-national implications.

Layer 1: Algorithmic-Technical accountability. At the innermost layer, accountability is built directly into the design and training of the RL agent. This layer draws most directly on technical scholarship showing that RL-based cyber agents can achieve strong detection performance even in resource-constrained environments, but that their internal decision processes remain difficult to

interpret without dedicated explainability architectures. Multi-layer explainability frameworks for RL-based cyber agents demonstrate that opacity can be reduced through techniques that expose intermediate decision states rather than only final actions, while explainable reinforcement learning approaches for adaptive cyber deception in zero-trust networks show that explainability and operational effectiveness are not mutually exclusive, provided explainability constraints are incorporated into reward design rather than added post hoc. Comprehensive surveys of AI and machine-learning techniques in cybersecurity solutions reinforce this layer's core claim: technical performance metrics alone are insufficient evidence of accountability, and reward-shaping, adversarial-robustness testing, and interpretability tooling must be treated as governance requirements, not optional engineering enhancements. Within the proposed framework, this layer specifies three concrete controls: mandatory logging of reward signals and policy updates, adversarial red-teaming prior to deployment in critical systems, and embedded explainability outputs sufficient for a non-specialist auditor to reconstruct why a given action was taken.

Layer 2: Organizational-Operational oversight. The second layer situates the technical controls of Layer 1 within an institutional structure capable of exercising human judgment over the agent's behavior. This layer is informed by data-governance frameworks for emerging technologies, which argue that blockchain, IoT, and AI systems require conceptual governance structures spanning the full technology lifecycle rather than point-in-time compliance checks, and by research on the importance of AI data governance in large language models, which emphasizes provenance tracking and lifecycle accountability as organizational, not purely technical, responsibilities. The three-pillar model of transparency, accountability, and trustworthiness for AI agents contributes the principle that human oversight must scale progressively with agent autonomy, analogous to staged validation in autonomous vehicle deployment, rather than assuming either constant human supervision or full autonomy from the outset. Auditing approaches developed for large language models through a three-layered governance, model, and application structure further inform this layer's design, though the present framework adapts that structure to accommodate continuous re-training: rather than a retrospective audit performed once per model release, Layer 2 requires a standing audit function that reviews sampled agent decisions on a rolling basis, coupled with a formal incident-reporting channel modeled on the practices documented in secure digital-governance interoperability research. This layer's core controls are: a designated accountable officer or oversight board for each deployed RL system, rolling audit sampling with documented review cycles, and a mandatory incident-reporting pathway that escalates anomalous agent behavior to both organizational leadership and the regulatory layer described next.

Layer 3: Sovereign-Regulatory authority. The outermost layer situates the RL agent within the jurisdictional and political authority of the state or states in which it operates. This layer draws on the comparative sovereignty literature synthesized above: research on China's cyber sovereignty concept and machine-learning governance models and on the open-source AI "openness paradox" in China's pursuit of cyber sovereignty illustrates a centralized, state-directed model of sovereign AI control; analyses of the European Union's pursuit of digital sovereignty describe a more market-embedded, rights-based model that nonetheless struggles to reconcile openness with strategic autonomy; and reviews of data sovereignty as a broader construct, together with descriptive and critical evaluations of existing digital-sovereignty models, show that most national frameworks were designed for static data assets rather than for continuously learning agents that generate new, derived data through their own operation. Indigenous and participatory data-sovereignty scholarship further cautions that Layer 3 cannot be reduced to central-government authority alone; the framework therefore specifies that sovereign-regulatory controls must include mechanisms for sub-national and community consultation where RL systems process data belonging to specific communities, in addition to standard state-level instruments such as data-localization requirements and cross-border data-transfer approval. This layer's core controls are: mandatory data-localization or sovereign-cloud requirements for training data involving critical infrastructure, a national or sectoral registry of deployed RL agents subject to periodic regulatory review, and formal cross-border cooperation instruments for agents that operate

across jurisdictional boundaries, informed by algorithmic state-architecture proposals for AI-enabled government coordination.

Table 1 : Comparative Summary of the Three-Layer Cyber-AI Governance Framework

Layer	Primary Control Mechanisms	Responsible Actor(s)	Governance Instrument(s)
1. Algorithmic-Technical	Reward/policy logging; adversarial red-teaming; embedded explainability	AI/ML engineering teams; model developers	Technical audit logs; explainability reports; red-team test suites
2. Organizational-Operational	Rolling audit sampling; designated oversight officer; incident reporting	Organizational governance boards; compliance officers	Internal audit charters; incident-reporting protocols; oversight dashboards
3. Sovereign-Regulatory	Data localization; agent registries; cross-border cooperation agreements	National regulators; sectoral authorities; community/sub-national bodies	Data-sovereignty statutes; agent registration schemes; bilateral/multilateral cooperation instruments

Integration through continuous feedback. The distinguishing contribution of this synthesis is not any single layer in isolation each already has partial precedent in the reviewed literature but the explicit feedback loop connecting the three layers. In most reviewed frameworks, technical, organizational, and regulatory controls are described sequentially, implying a one-directional flow from design to deployment to oversight. Because RL agents continue to update their policies from live environmental feedback, a one-directional model quickly becomes stale: an agent's behavior at month twelve may differ substantially from its behavior at deployment, yet most governance frameworks reviewed here do not specify a mechanism for regulatory awareness of this drift. The proposed framework instead treats Layers 1 through 3 as a closed loop: Layer 1 telemetry (reward logs, explainability outputs, anomaly flags) feeds directly into Layer 2's rolling audit function; Layer 2's audit findings and incident reports feed into Layer 3's regulatory registry and review cycle; and Layer 3's regulatory determinations for example, a data-localization ruling or a cross-border restriction feed back into Layer 1 as binding constraints on what the agent may subsequently learn, retain, or transmit. This closed-loop design directly addresses the gap identified in the background: existing layered models, including three-layer auditing approaches for large language models and five-layer AI-governance frameworks, were built for systems whose behavior is fixed between audit cycles, whereas the framework proposed here is built for systems whose behavior is continuously in motion.

Cross-national implications. Applying the framework across the comparative cases reviewed above illustrates its diagnostic value. A centralized cyber-sovereignty model emphasizes Layer 3 controls (registries, localization, state review) but, per the reviewed literature, offers comparatively thin documentation of Layer 1 technical accountability, suggesting a governance profile strong on sovereign authority but weaker on algorithmic transparency. Conversely, jurisdictions following a more market-embedded digital-sovereignty approach tend to invest heavily in Layer 1 and Layer 2 mechanisms, such as technical standards and organizational compliance regimes, while struggling to operationalize Layer 3 sovereign controls given fragmented multi-state authority. Neither profile, considered alone, satisfies all three layers simultaneously, reinforcing the discussion's central claim that accountable reinforcement learning under national data sovereignty requires deliberate investment across all three layers and, critically, in the feedback connections between them, rather than reliance on whichever single layer a given jurisdiction's existing institutions happen to emphasize.

Limitations of the synthesized evidence base. Several included records are conceptual or preprint sources rather than empirically validated deployments, meaning the framework's individual layer controls are, at this stage, better supported as design principles than as empirically tested interventions. Additionally, the corpus is weighted toward English-language and Western- or China-focused sovereignty scholarship, which may limit the framework's direct applicability to jurisdictions whose sovereignty traditions were underrepresented in the reviewed literature, including many contexts in Southeast Asia. These limitations are addressed further in the concluding section.

Implications for practitioners and policymakers. For engineering teams, the framework implies that explainability and adversarial testing cannot be deferred to a pre-deployment checklist; because Layer 1 telemetry must continuously feed Layer 2 audits, logging and explainability infrastructure needs to be built into the training pipeline itself rather than retrofitted after an incident occurs. For organizational compliance functions, the framework implies a shift from periodic, calendar-driven audits toward rolling, event-triggered review cycles capable of responding to the pace at which an RL agent's policy can change. This has direct staffing and tooling implications: an oversight board reviewing a static model on a quarterly cycle is unlikely to detect drift in a system that updates its policy daily, and organizations deploying RL in security-critical roles should budget for continuous rather than periodic audit capacity. For regulators, the framework suggests that agent registries should capture not only the fact of deployment but also metadata about retraining frequency and data provenance, since a registry that records only initial certification will rapidly fall out of step with an evolving agent. The framework also implies a specific institutional design choice: because Layer 3 must both receive Layer 2 incident reports and issue binding constraints back to Layer 1, regulators need a technical liaison function — distinct from traditional legal or compliance review capable of translating regulatory determinations into machine-enforceable constraints on training data and permissible actions.

The comparative discussion above further suggests that no jurisdiction reviewed in this synthesis has yet implemented all three layers with their connecting feedback loops simultaneously, which points toward a practical sequencing strategy rather than an all-at-once mandate. Organizations and regulators seeking to adopt the framework incrementally might reasonably begin with Layer 1 technical logging and explainability, since these controls are within the direct authority of engineering teams and require no new legal instrument; proceed to Layer 2 organizational oversight once technical telemetry exists to audit; and treat Layer 3 sovereign-regulatory instruments, which typically require legislative or inter-agency coordination, as the final and most institutionally demanding stage. This sequencing does not diminish the framework's insistence that all three layers are ultimately necessary; rather, it acknowledges that the feedback loop connecting the layers can only be meaningfully activated once each layer's foundational controls are in place, and that attempting to mandate Layer 3 registries or cross-border cooperation instruments before Layer 1 telemetry exists would leave regulators with formal authority but no reliable data on which to exercise it

CONCLUSION

This article has synthesized 25 sources spanning reinforcement-learning security research, layered AI-governance scholarship, and comparative data-sovereignty theory into a Three-Layer Cyber-AI Governance Framework designed specifically for accountable reinforcement learning under conditions of national data sovereignty. The framework's Algorithmic-Technical, Organizational-Operational, and Sovereign-Regulatory layers each have partial precedent in the existing literature, but their explicit integration through a continuous, bidirectional feedback loop addresses a gap that single-layer technical fixes and retrospective, generative-AI-oriented governance models leave unresolved: the fact that RL agents continue to learn and act after deployment, and that governance mechanisms built for static systems cannot reliably track that ongoing change.

The framework's central contribution is conceptual rather than empirical, and its adoption should proceed accordingly. Future research should prioritize case-study validation of the framework within specific national contexts, particularly those underrepresented in the present corpus, to test whether the sequencing strategy proposed here technical logging first, organizational audit second, sovereign-regulatory instruments third holds across differing institutional starting points. Further work is also needed to specify the technical liaison mechanisms through which regulatory determinations can be translated into machine-enforceable constraints on RL training pipelines, an area the reviewed literature addresses only in general terms. Finally, because the framework explicitly accommodates plural sovereignty claims — state, organizational, and community future research should examine how

the feedback loop functions when these claims conflict, rather than assuming a single, unified sovereign authority at Layer 3. Addressing these questions will be essential to moving the proposed framework from a literature-grounded conceptual architecture toward a validated instrument for governing reinforcement learning within sovereign digital infrastructure.

REFERENCES

Adabara, I., Sadiq, B. O., Shuaibu, A. N., Danjuma, Y. I., & Maninti, V. (2025). Trustworthy agentic AI systems: A cross-layer review of architectures, threat models, and governance strategies for real-world deployment. F1000Research. https://doi.org/10.12688/f1000research.169927.1

Adabara, I., Sadiq, B., Shuaibu, A., Danjuma, Y. I., Maninti, V., & Joe, M. (2025). Agentic artificial intelligence for ethical cybersecurity in Uganda: A reinforcement learning framework for threat detection in resource-constrained environments. ArXiv. https://doi.org/10.48550/arxiv.2512.07909

Agarwal, A., & Nene, M. J. (2025). A five-layer framework for AI governance: Integrating regulation, standards, and certification. Transforming Government: People, Process and Policy, 19(3), 535–555. https://doi.org/10.1108/TG-03-2025-0065

Calderaro, A., & Blumfelde, S. (2022). Artificial intelligence and EU security: The false promise of digital sovereignty. European Security, 31(4), 415–434. https://doi.org/10.1080/09662839.2022.2101885

Cheng, E. C., Cheng, J., & Siu, A. (2026). Toward safe and responsible AI agents: A three-pillar model for transparency, accountability, and trustworthiness. ArXiv. https://doi.org/10.48550/arxiv.2601.06223

Dutta, S., & Mazumdar, S. (2025). Our data, ourselves: Participation, justice, and alternative futures of data sovereignty in India. Big Data & Society. https://doi.org/10.1177/20539517251365235

Engin, Z., Crowcroft, J., Hand, D., & Treleaven, P. (2025). The algorithmic state architecture (ASA): An integrated framework for AI-enabled government. ArXiv. https://doi.org/10.48550/arxiv.2503.08725

Figueroa, V., Sánchez Crespo, L. E., Santos-Olmo, A., Rosado, D. G., & Fernández-Medina, E. (2025). Building a holistic cybersecurity framework for e-Government based on a systematic analysis of proposals. International Journal of Information Security, 24, article 121. https://doi.org/10.1007/s10207-025-01024-0

Fratini, S., Hine, E., Novelli, C., Roberts, H., & Floridi, L. (2024). Digital sovereignty: A descriptive analysis and a critical evaluation of existing models. Digital Society, 3. https://doi.org/10.1007/s44206-024-00146-7

Goel, D., Moore, K., Wang, J., Kim, M., & Nguyen, T. (2025). Unveiling the black box: A multi-layer framework for explaining reinforcement learning-based cyber agents. ArXiv. https://doi.org/10.48550/arxiv.2505.11708

Hummel, P., Braun, M., Tretter, M., & Dabrock, P. (2021). Data sovereignty: A review. Big Data & Society, 8(1). https://doi.org/10.1177/2053951720982012

Hung, H. T. (2025). Exploring China's cyber sovereignty concept and artificial intelligence governance model: A machine learning approach. Journal of Computational Social Science, 8. https://doi.org/10.1007/s42001-024-00346-8

Irekponor, O. (2025). Designing resilient AI architectures for predictive energy finance systems amid data sovereignty, adversarial threats, and policy volatility. International Journal of Research Publication and Reviews. https://doi.org/10.55248/gengpi.6.0625.2125

Kumar, A., Sharma, A., & Pujari, M. (2025). AI governance via explainable reinforcement learning (XRL) for adaptive cyber deception in zero-trust networks. Journal of Information Systems Engineering and Management. https://doi.org/10.52783/jisem.v10i43s.8308

Leghemo, I., Azubuike, C., Segun-Falade, O. D., & Odionu, C. S. (2025). Data governance for emerging technologies: A conceptual framework for managing blockchain, IoT, and AI. Journal of Engineering Research and Reports. https://doi.org/10.9734/jerr/2025/v27i11385

Malik, V., Mittal, R., Mavaluru, D., Narapureddy, B., Goyal, S., Martin, R., Srinivasan, K., & Mittal, A. (2023). Building a secure platform for digital governance interoperability and data exchange using blockchain and deep learning-based frameworks. IEEE Access, 11, 70110–70131. https://doi.org/10.1109/access.2023.3293529

Mokander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. AI and Ethics, 4, 1085–1115. https://doi.org/10.1007/s43681-023-00289-2

Ozkan-Okay, M., Akin, E., Aslan, Ö., Kosunalp, S., Iliev, T., Stoyanov, I., & Beloev, I. (2024). A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions. IEEE Access, 12, 12229–12256. https://doi.org/10.1109/access.2024.3355547

Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025). The importance of AI data governance in large language models. Big Data and Cognitive Computing, 9(6), 147. https://doi.org/10.3390/bdcc9060147

Pasupuleti, M. K. (2025). Governing artificial intelligence: Global standards for data sovereignty, cybersecurity, and ethical integrity. International Journal of Academic and Industrial Research Innovations, 5. https://doi.org/10.62311/nesx/rp05226

Radanliev, P. (2025). Privacy, ethics, transparency, and accountability in AI systems for wearable devices. Frontiers in Digital Health, 7. https://doi.org/10.3389/fdgth.2025.1431246

Rana, V. (2025). Indigenous data sovereignty: A catalyst for ethical AI in business. Business & Society, 64(4), 635–640. https://doi.org/10.1177/00076503241271143

Robles, P., & Mallinson, D. (2023). Catching up with AI: Pushing toward a cohesive governance framework. Politics & Policy. https://doi.org/10.1111/polp.12529

Taeihagh, A. (2021). Governance of artificial intelligence. Policy and Society, 40(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Yang, G. (2025). The openness paradox: Open-source AI and China's quest for cyber sovereignty. Dialogues on Digital Society, 1, 261–264. https://doi.org/10.1177/29768640251376497.