www.ijarp.com ISSN 2456-9992 Page: 01-10

International Journal Advanced Research Publications

Research Article

Volume: 02
Issue: 09

RISK-CONSTRAINED GENERATIVE-AI RED–BLUE CO-EVOLUTION FOR SELF-HEALING 6G NETWORK SLICES UNDER ADVERSARIAL CONCEPT DRIFT

Dr. Omohwo Ohwobeno[1], Ohis Grace Mega[2], and Dr. Chika Lilian Onyagu*[3]

[123]Department of Cybersecurity and Data Science, Faculty of Computing, Delta State University, Abraka, Delta State, Nigeria.

Article Received: 05 August 2026, Article Revised: 25 August 2026, Published on: 15 September 2026

***Corresponding Author: Dr. Chika Lilian Onyagu**

Department of Cybersecurity and Data Science, Faculty of Computing, Delta State University, Abraka, Delta State, Nigeria.

Doi: https://doi-doi.org/101555/ijarp.9042

2. ABSTRACT

Sixth-generation (6G) network slices are programmable, distributed, and exposed to adversaries that can change their behaviour faster than static intrusion detectors can be retrained. This paper proposes a risk-constrained Generative-AI red–blue co-evolution framework for self-healing 6G slices. The framework combines slice telemetry and a digital twin with a concept-drift monitor, generative red and blue agents, a risk governor, and a slice orchestrator. Red agents generate bounded adversarial hypotheses, while blue agents produce explainable detection and remediation plans. A remediation is executable only when it satisfies hard service, safety, and policy constraints and remains below a dynamic risk threshold; otherwise, the system selects a verified fallback or escalates to a human operator. The method is specified as a design-science architecture and an evaluation protocol covering attack detection, drift adaptation, recovery, service availability, latency, unsafe-action rate, and explanation quality. Analytical results show that the proposed governor guarantees admissibility of automated actions by construction, while the co-evolution loop addresses the separate weaknesses of static intrusion detection, isolated self-healing, and unconstrained language-model automation. The principal contribution is an integrated control-plane design that links adversarial learning to operationally safe slice healing. The framework provides a reproducible basis for future implementation and testbed validation.

3. KEYWORDS: 6G network slicing; generative artificial intelligence; red–blue co-evolution; concept drift; self-healing networks; cyber-risk constraints.

4. INTRODUCTION

Background and motivation

Sixth-generation networks are expected to combine extreme service heterogeneity, softwarised functions, distributed edge computing, and highly programmable network slices. These properties improve flexibility but also enlarge the attack surface. A compromise may propagate across virtual network functions, orchestration interfaces, tenants, and edge nodes before a conventional security team can investigate it. Self-healing is therefore increasingly treated as a control problem: the network must observe its state, infer a fault or attack, select a corrective action, and verify the post-action state. Digital-twin and graph-neural-network designs demonstrate how service state can be modelled and how anomalous components can be migrated or redeployed in 6G edge environments [1]. Slice isolation and policy enforcement further show that the slice itself can be used as a security mechanism [2].

The central problem is that a self-healing controller can become unsafe when its threat model and data distribution change simultaneously. Static intrusion-detection models lose effectiveness under concept drift, because the relationship between telemetry and the attack label changes over time [4], [12]. Conversely, an unconstrained generative model may produce a technically plausible but operationally harmful remediation, such as a broad slice shutdown, excessive traffic rerouting, or an irreversible policy change. Existing network self-protection and orchestration approaches provide deployable enforcement paths [2], [5], but they do not establish a unified mechanism in which the attacker and defender evolve together while every automated action is checked against explicit risk and service constraints.

Recent literature has developed important components of the problem. Autonomous cyber-defence agents combine multi-agent reinforcement learning, large language models, and rules for red and blue teams [3]. Explainable drift detection supports diagnosis of changing temporal relationships in 5G data [9], while adaptive intrusion detection provides a basis for continual learning [4], [12]. Digital-twin self-healing [1], integrated 6G healing [6], and federated virtual-network-function management [13] support state estimation, recovery, and distributed operation. Generative deception networks [10] and explainable LLM red-versus-blue simulation [14] move toward adversarial co-evolution. However, these strands remain fragmented. In particular, the literature does not sufficiently integrate: (i) an explicit concept-

drift trigger; (ii) generative red–blue co-evolution; (iii) a formal action-risk governor; and (iv) a slice-aware execution and verification loop. This paper addresses that integration gap.

The aim of this work is to develop a risk-constrained Generative-AI red–blue co-evolution framework that continuously adapts to adversarial concept drift and supports safe self-healing of 6G network slices. The objectives are to: (1) define a closed-loop architecture linking telemetry, drift monitoring, adversarial generation, defensive planning, risk verification, and orchestration; (2) formalise admissible remediation as a constrained decision set; (3) specify a reproducible evaluation protocol and comparative baselines; and (4) identify safety, explainability, and operational limitations that must be addressed before deployment.

The design is of a great importance for both cybersecurity and network operations. It treats generative AI as a policy-proposal component rather than an unrestricted actuator. It makes service availability, latency, isolation, uncertainty, and reversibility first-class control variables. It also links the red team to the same evolving slice state used by the blue team, thereby reducing the risk that a defence is evaluated only against obsolete attack patterns. The resulting architecture is intended for simulation, cyber-range, or digital-twin validation, and can be extended with federated learning when slice telemetry cannot be centralised [8], [13].

The literature can be organised into four complementary streams. First, self-healing research uses digital twins, graph models, predictive analytics, and resource orchestration to detect abnormal conditions and restore service [1], [6], [7]. Second, slice-security research embeds isolation, monitoring, and automated response in the network-slicing substrate [2], [5]. Third, adaptive detection research addresses data-stream evolution through drift detectors, adaptive classifiers, explainability, and evaluation taxonomies [4], [9], [12]. Fourth, autonomous cyber-defence research introduces red and blue agents, language models, deception, and co-evolution [3], [10], [11], [14]. The proposed framework is positioned at the intersection of these streams: its distinctive element is not a new generative model alone, but a constrained control policy that makes generative co-evolution operationally governable.

5. MATERIALS AND METHODS

Research design

This study adopts a design-science research approach. The artefact is a control-plane framework specified by its components, interfaces, state variables, decision rules, and evaluation measures. The design is implementation-neutral: it can be realised with a digital-twin simulator, a cyber range, or a containerised 6G testbed. The method does not claim

measured performance because the supplied materials contain literature records but no experiment logs or benchmark traces. Instead, it derives verifiable safety properties and a test protocol that future empirical work can execute.

Framework architecture

Figure 1 shows the proposed closed loop. The telemetry and digital-twin layer aggregates flow statistics, service-level indicators, virtual-network-function health, topology, policy state, and recent remediation outcomes. The concept-drift monitor detects changes in feature distributions, conditional relationships, or model residuals. Its output is a drift signal and a diagnosis, rather than an immediate reconfiguration. The red agent generates bounded attack hypotheses, including attack path, target slice, expected impact, and confidence. The blue agent consumes the same state and generates a detection and healing plan with a rationale, prerequisites, rollback action, and expected service effect. The risk governor verifies the plan before the orchestrator can apply it.

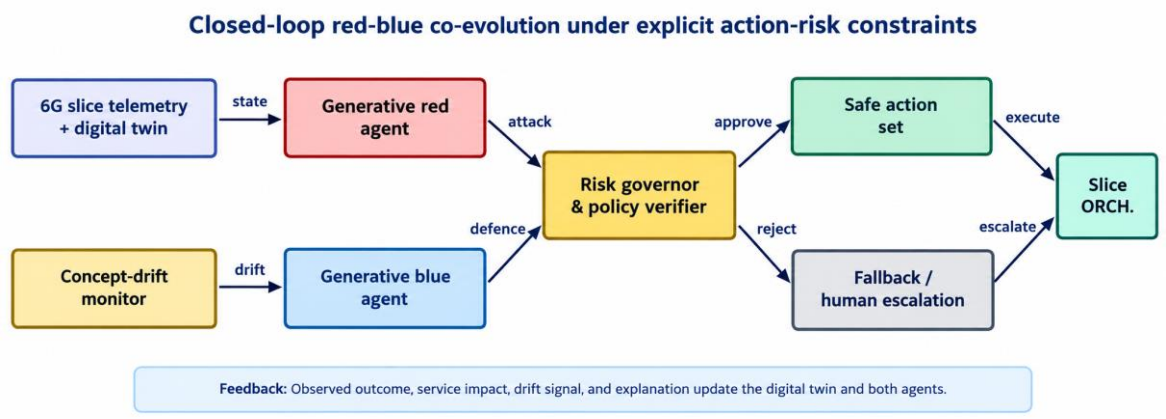

Fig. 1. Risk-constrained red–blue co-evolution control loop for self-healing 6G network slices.

State, action, and risk formulation

At time t, the controller state is represented as $s_t = [x_t, g_t, z_t, c_t, h_t]$, where x_t is slice telemetry, g_t is the service and topology graph, z_t is the drift and threat context, c_t is model confidence, and h_t is the history of actions and outcomes. The red agent proposes an attack hypothesis a^R_t and the blue agent proposes a remediation plan a^B_t. A plan is evaluated using a composite risk function:

$$R(a^B_t \mid s_t) = w_1 I\_avail + w_2 I\_SLA + w_3 I\_isolation + w_4 I\_privacy + w_5 U\_model + w_6 I\_irreversibility,$$

where each I term represents estimated impact on availability, service-level agreement compliance, tenant isolation, or privacy, U_model is decision uncertainty, and I_irreversibility represents the difficulty of rollback. The weights are policy parameters selected by the network operator. The admissible action set is:

$A_t = \{a : R(a \mid s_t) \le \tau_t,\ g(a) = 1,\ q(a) = 1,\ b(a) = 1\},$

where τ_t is a dynamic risk budget, g(a) verifies hard service constraints, q(a) verifies policy and isolation constraints, and b(a) verifies resource and operational bounds. If no generated plan belongs to A_t, the governor selects a pre-validated fallback or escalates to a human operator. This rule prevents model fluency from being treated as authorisation.

Co-evolution and drift adaptation procedure

The procedure runs in six stages. First, the monitor updates the digital twin from slice telemetry and computes drift indicators. Second, a red agent generates a bounded scenario against the current topology and policy state. Third, a blue agent diagnoses the scenario and proposes one or more explainable responses. Fourth, the risk governor simulates or estimates the consequence of each response and rejects plans that violate hard constraints. Fifth, the orchestrator applies the selected action with a rollback checkpoint and verifies the post-action state. Sixth, the outcome, explanation, and residual error are fed back to both agents. Model updates are gated by data-quality, poisoning, and confidence checks; a detected drift event therefore triggers controlled adaptation rather than blind retraining.

Evaluation protocol

A future implementation should evaluate the framework in a digital twin or cyber range containing multiple tenants, service classes, edge nodes, and virtual network functions. The scenario set should include distributed denial of service, slice-isolation violation, credential-based lateral movement, telemetry poisoning, and resource-exhaustion attacks. Each scenario should be replayed under stationary data, gradual drift, abrupt drift, and adversarial drift. The comparison should include: (B1) a static IDS with rule-based response; (B2) a drift-aware IDS without generative planning; (B3) a self-healing controller without red–blue co-evolution; and (B4) an unconstrained generative red–blue controller. The proposed method is (B5), the risk-constrained co-evolution controller.

Table 1. Evaluation dimensions and measures for empirical validation.

Evaluation dimension	Operational measures	Desired direction
Detection and adaptation	F1-score; precision; recall; drift-detection delay; false-alarm rate	Higher F1/recall; lower delay and false alarms
Healing effectiveness	Mean time to recovery; service availability; packet loss; SLA violations	Lower recovery time/loss/violations; higher availability
Safety and governance	Unsafe-action rate; risk-budget violations; rollback success; human-escalation rate	Zero hard-constraint violations; successful rollback
Operational cost	Control-plane latency; CPU/memory overhead; number of actions	Lower overhead subject to safe recovery
Trustworthiness	Explanation completeness; calibration; operator acceptance	Higher completeness/calibration/acceptance

6. RESULTS AND DISCUSSION

Analytical results

The first result is a safety property of the decision rule. If every automated action is executed only when it belongs to A_t, then $R(a^B_t \mid s_t) \leq \tau_t$ and the hard constraints represented by g(a), q(a), and b(a) hold for every accepted action. This is a design-time guarantee of admissibility, not a claim that the risk model itself is perfect. It means that the language model cannot bypass the policy boundary merely by producing a persuasive explanation.

The second result is a control separation between proposal and execution. Red and blue agents can explore alternative hypotheses and responses, but the orchestrator receives only a verified action, a rollback plan, and an explanation. This separation addresses the principal weakness of unrestricted generative remediation. It also permits different model families to be replaced without changing the safety interface.

The third result is a mechanism for handling adversarial concept drift. The drift monitor changes the context and learning state used by the agents, while the risk governor remains active during adaptation. Thus, a model update cannot silently relax availability, isolation, or reversibility constraints. The arrangement complements drift-resilient intrusion-detection methods [4], [9], [12] and extends them from detection to constrained action selection.

Comparison with related approaches

Table 2. Positioning of the proposed framework against selected literature.

Approach family	Strength	Unresolved limitation	Contribution of this framework
Digital-twin / GNN healing [1], [6]	Strong state modelling and service recovery	No integrated adversarial co-evolution or explicit risk budget	Adds red–blue generation and governed action selection
Slice security and SOAR [2], [5]	Deployable isolation, orchestration, and enforcement	Limited adaptation to generative adversaries	Maps generated plans to verified slice actions
Adaptive or federated learning [4], [8], [12], [13]	Handles evolving data and distributed telemetry	Does not by itself guarantee safe remediation	Keeps drift adaptation behind policy and rollback gates
Autonomous red–blue AI [3], [10], [11], [14]	Models attacker–defender interaction and explanations	Risk and 6G service constraints may remain implicit	Makes risk, SLA, isolation, and uncertainty explicit

Implications and limitations

The design suggests that safe autonomy should be evaluated on two axes: whether the system recognises and contains an evolving threat, and whether it avoids causing unacceptable service or policy harm while doing so. A controller that improves recall by frequently shutting down slices may score well on detection but fail operationally. Conversely, a controller that preserves availability by ignoring low-confidence attacks fails security. The joint metric set in Table 1 therefore reflects the multi-objective nature of 6G slice protection.

Several limitations remain. First, the risk function depends on accurate impact estimates and operator-selected weights. Second, digital-twin mismatch can cause the governor to underestimate a real-world consequence. Third, generative agents may amplify poisoned or strategically crafted telemetry. Fourth, explanations may be plausible without being causally faithful. Fifth, a human escalation path can become a bottleneck during high-volume attacks. These limitations motivate calibrated uncertainty, signed telemetry, adversarial testing, causal explanation checks, and workload-aware escalation policies. They also mean that the framework should be validated in progressively more realistic environments before any production deployment.

7. CONCLUSION

This paper presented a risk-constrained Generative-AI red–blue co-evolution framework for self-healing 6G network slices under adversarial concept drift. The framework integrates

digital-twin telemetry, drift monitoring, generative attack and defence agents, a formal risk governor, and slice orchestration with rollback and escalation. Its key design principle is that generative AI proposes actions but does not authorise them. Automated remediation is executable only when it satisfies explicit risk, service, isolation, resource, and reversibility constraints. The analytical results establish admissibility by construction and show how the design complements prior work on self-healing, slice security, drift-resilient detection, and autonomous cyber defence. The next step is empirical implementation using the evaluation protocol defined in this paper, with particular attention to unsafe-action rate, service availability, recovery time, drift robustness, calibration, and explanation fidelity.

8. ACKNOWLEDGEMENTS

The authors acknowledge Delta State University, Abraka, Nigeria, for the institutional context supporting this research direction. No external funding or third-party contribution was specified in the supplied materials.

9. REFERENCES

1. P. Yu, J. Zhang, H. Fang, W. Li, L. Feng, F. Zhou, P. Xiao, S. Guo, “Digital Twin Driven Service Self-Healing With Graph Neural Networks in 6G Edge Networks,” IEEE Journal on Selected Areas in Communications, vol. 41, no. 11, pp. 3607-3623, 2023, doi: 10.1109/JSAC.2023.3310063.
2. A. M. Escolar, J. B. Bernabe, J. M. A. Calero, Q. Wang, A. Skarmeta, “Network slicing as 6G security mechanism to mitigate cyber-attacks: the RIGOUROUS approach,” in Proc. 2024 IEEE 10th International Conference on Network Softwarization (NetSoft), pp. 387-392, 2024, doi: 10.1109/NetSoft60951.2024.10588887.
3. J. F. Loevenich, E. Adler, A. Bécue, A. Velazquez, K. Wrona, V. Boshnakov, J. Falkcrona, N. Nordbotten, O. L. Worthington, J. Röning, R. Rigolin, F. Lopes, “Training Autonomous Cyber Defense Agents: Challenges & Opportunities in Military Networks,” in Proc. MILCOM 2024 - 2024 IEEE Military Communications Conference (MILCOM), pp. 158-163, 2024, doi: 10.1109/MILCOM61039.2024.10773923.
4. S. Seth, K. K. Chahal, G. Singh, “Concept Drift–Based Intrusion Detection For Evolving Data Stream Classification In IDS: Approaches And Comparative Study,” The Computer Journal, vol. 67, no. 7, pp. 2529-2547, 2024, doi: 10.1093/comjnl/bxae023.
5. P. Benlloch-Caballero, A. Matencio-Escolar, J. Bernal Bernabe, A. Skarmeta, Q. Wang, J. M. Alcaraz-Calero, “E2E Network Slicing for Enhanced Cybersecurity, Orchestration,

Automation and Response in 5G/6G: The RIGOUROUS Approach," Journal of Network and Systems Management, vol. 34, no. 1, pp. 22, 2025, doi: 10.1007/s10922-025-09999-w.

6. A. A. Pise, Y. Khandokar, "IDEA-6G: Revolutionizing 6G Networks with Integrated Digital Twin and Self-Healing Mechanisms," Journal of High-Frequency Communication Technologies, vol. 3, no. 01, pp. 258-270, 2025, doi: 10.58399/eezw1561.
7. A. Giannopoulos, S. Spantideas, P. Trakadas, J. Perez-Valero, G. Garcia-Aviles, A. S. Gomez, "AI-Driven Self-Healing in Cloud-Native 6G Networks Through Dynamic Server Scaling," in Proc. 2025 IEEE 11th International Conference on Network Softwarization (NetSoft), pp. 43-48, 2025, doi: 10.1109/NetSoft64993.2025.11080631.
8. Y. A. Kosinov, Z. M. Lorsanova, S. Y. Sitnikov, "ADAPTIVE CYBERSECURITY MECHANISMS FOR 6G NETWORK SECTIONS BASED ON MACHINE LEARNING," SOFT MEASUREMENTS AND COMPUTING, vol. 9/2, no. 94, pp. 116-122, 2025, doi: 10.36871/2618-9976.2025.9-2.014.
9. M. R. Fentazi, M. Hassani, A. Ksentini, A. Mekrache, M. Bessedik, "AIEuroLens: Explainable AI Framework for Drift Detection applied to 5G Time-Series Data," in Proc. GLOBECOM 2025 - 2025 IEEE Global Communications Conference, pp. 2336-2341, 2025, doi: 10.1109/GLOBECOM59602.2025.11431870.
10. A. K. Polinati, "Generative Adversarial Deception Networks (GADN): Self-Evolving AI for Adaptive Cyber Defense in Cloud-Native Security Architectures," in Proc. 2025 Tenth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), pp. 1-8, 2025, doi: 10.1109/ICONSTEM65670.2025.11374768.
11. A. Dijk, R. Meier, C. Melella, M. Pihelgas, R. Vaarandi, V. Lenders, "Next Steps in Cyber Blue Team Automation—Leveraging the Power of LLMs," in Proc. 2025 17th International Conference on Cyber Conflict: The Next Step (CyCon), pp. 209-226, 2025, doi: 10.23919/CyCon65856.2025.11103720.
12. M. Komarchesqui, V. G. D. S. Ruffo, D. Matheus Brandão Lent, V. F. Schiavon, M. R. Nakanishi, G. H. K. Nishikawa, L. Fernando Carvalho, M. L. Proença, "A Comprehensive Survey on Concept-Drift-Resilient Network Intrusion Detection Systems," IEEE Access, vol. 14, pp. 69689-69718, 2026, doi: 10.1109/ACCESS.2026.3691262.
13. K. Golla, M. Ramkumar, "Federated Deep Learning-Based Self-Healing and Autonomous Virtual Network Function Management for Intelligent 6G Network

Slicing," in Proc. 2026 International Conference on Intelligent Computing (ICONIC), pp. 1-7, 2026, doi: 10.1109/ICONIC67661.2026.11517700.

14. H. Du, "GenAI-Powered Autonomous Cyber Offense-Defense: An Explainable LLM Red-vs-Blue Simulation and Self-Defense Framework," Journal of Cyber Security, vol. 8, no. 1, pp. 241-279, 2026, doi: 10.32604/jcs.2026.075976.