VOLUME 10 · ISSUE 03

BALSILLIE PAPERS

Governing Machine-speed Cyber Defence: Policy Frameworks for Agentic AI

Abbas Yazdinejad, Hadis Karimipour and Jatin Nathwani

JULY 7, 2026

Agentic artificial intelligence (AI) systems — which sense, reason and act with limited or no real-time human intervention — are beginning to emerge in cybersecurity operations, in particular as extensions of advanced automation and adaptive defence systems. This paper proposes four governing principles to guide international policymaking around agentic cyber defence: bounded autonomy rooted in human judgment; explicit escalation thresholds for human intervention; auditability of autonomous system actions; and transboundary coordination mechanisms to support cross-border stability and cooperation.

INTRODUCTION

Agentic artificial intelligence (AI) systems — which sense, reason and act with limited or no real-time human intervention — are beginning to emerge in cybersecurity operations, in particular as extensions of advanced automation and adaptive defence systems. Current cybersecurity and AI governance regimes were designed with humans remaining central to the decision-making process. Machine-speed cyber defence shatters this premise. This paper assesses gaps in governance of machine-speed security through the lens of defending a critical societal infrastructure, focusing on energy and smart grid systems. This dilemma is part of the ongoing international dialogue surrounding AI, cybersecurity and global stability in the G7, the Organisation for Economic Co-operation and Development (OECD) and the United Nations. Regulatory and policy regimes lack key guardrails for autonomous defensive measures that carry potential cross-border, attribution and escalation consequences. This paper proposes four governing principles to guide international policymaking around agentic cyber defence: bounded autonomy rooted in human judgment; explicit escalation thresholds for human intervention; auditability of autonomous system actions; and transboundary coordination mechanisms to support cross-border stability and cooperation.

THE POLICY PROBLEM

Automated, real-time cyber defence is already an operational reality, while more advanced forms of adaptive and semi-autonomous response are increasingly being explored and gradually introduced in practice. Operational technology (OT) and control networks — such as energy systems and smart grids — require automated sensing, triage and containment to maintain availability during cyberattacks. Even seconds of disruption can cause outages, safety hazards and financial loss.[1] Increasing connectivity, remote access and dense sensing expands the attack surface while simultaneously reducing the time available for human response.[2] Most policy guardrails, however, were built around human-speed cybersecurity. Regulations, liability regimes and AI governance frameworks assume that a human observes an incident, evaluates options and authorizes action. Agentic AI introduces a shift in this sequence.

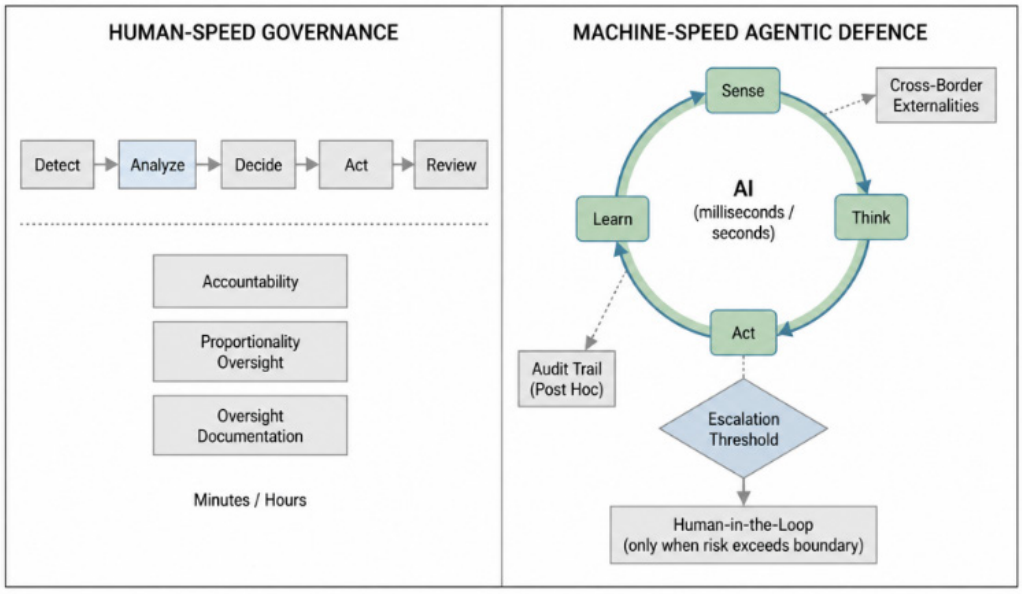
In this paper, agentic AI refers to cyber defence systems that select among response strategies based on context and act without real-time human authorization, operating within predefined design constraints. While most current systems remain constrained by predefined rules and human oversight, emerging approaches increasingly incorporate adaptive, context-aware decision-making capabilities that move toward greater autonomy.[3] This paper adopts a conceptual policy lens to examine how delegating decision-making authority to machine-speed systems challenges existing governance frameworks. It contributes to a governance model grounded in the "governance latency gap," relocating accountability from real-time human decisions to system design and predefined constraints.

The Governance Latency Gap

A governance latency gap arises when safe system operation requires action faster than human judgment, authorization or accountability processes embedded in existing regulatory frameworks. In critical infrastructure, this gap is operationally significant. For example, in a smart grid, systems may need to isolate a compromised substation within milliseconds to prevent cascading failures, while regulatory procedures operate on timescales of minutes or longer. This mismatch shows that safe operation can depend on actions occurring before human decision-making or governance processes can be meaningfully applied.

This gap also affects international stability. Cyber norms and confidence-building measures presume actions originate from human intent. UN Open-Ended Working Group discussions address state responsibility rather than autonomous defensive behaviour,[4] while the G7 Hiroshima AI Process emphasizes lifecycle risk management but does not clarify responsibility when automated defences in one jurisdiction affect another.[5] Escalation, traditionally a human decision, may now occur automatically through blocking, rerouting or deception. Such technically rational actions can produce cross-border effects. Because sovereignty, attribution and due diligence depend on identifiable human intent,[6] agentic defence challenges the foundations of current cyber governance. Figure 1 illustrates the mismatch between governance frameworks based on human decision cycles and cyber defence operating at machine speed.

Figure 1: Human-speed Governance vs. Machine-speed Agentic Defence



Source: Authors.

This perspective extends existing discussions on AI governance and cybersecurity, which largely emphasize human-in-the-loop decision making, ethical principles and state-centric norms. Frameworks from the OECD, National Institute of Standards and Technology (NIST) and UN articulate key principles for trustworthy AI, but generally assume that human judgment remains central at the moment of action. In contrast, this paper examines governance through decision latency and system-level autonomy, showing how authority shifts from *ex post* human authorization to *ex ante* system design. By linking operational realities in cyber-physical systems with international policy frameworks, the paper offers a distinct analytical contribution to ongoing G7, OECD and UN debates.

WHAT AGENTIC AI MEANS THROUGH A POLICY LENS

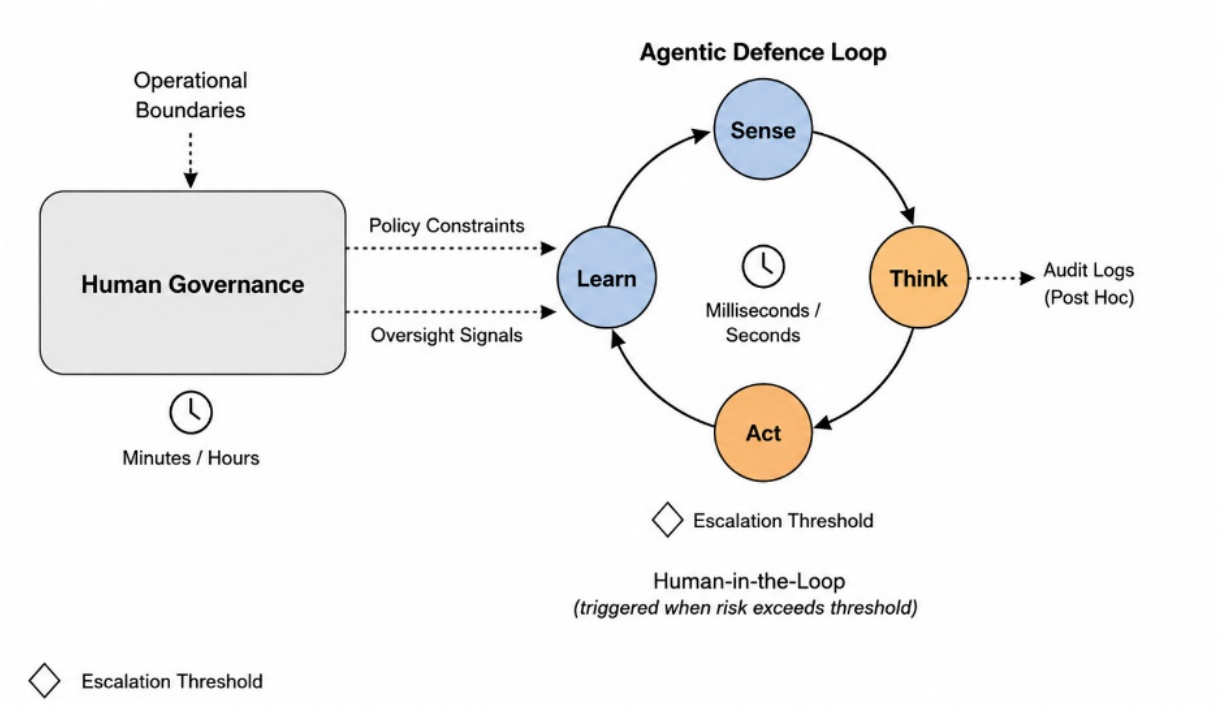
Automation has long supported cybersecurity operations through scripts, playbooks and orchestration tools that execute predefined responses. The emergence of agentic systems introduces a qualitative shift: AI-enabled components capable of detecting, reasoning and responding within a closed feedback loop without real-time human approval. In practice, elements of this transition can already be observed in systems such as security orchestration, automation and response (SOAR) platforms, AI-assisted intrusion detection systems, and adaptive anomaly detection tools deployed in operational environments. While these systems remain bounded by predefined constraints, they increasingly incorporate context-aware analysis and limited autonomous response capabilities. After deployment and within predefined constraints, such systems analyze anomalies using learned models and live data, then select defensive actions within milliseconds. For policy purposes, the critical distinction is not the technical method used,[7] but the locus of decision authority. Traditional automation executes pre-approved actions, whereas agentic AI evaluates context and chooses among multiple permissible responses. Cyber defence thus moves from deterministic if–then execution to machine-driven contextual decision-making.[8] It is important to distinguish between current practice and the fully realized vision of agentic cyber defence. In most real-world deployments, especially in critical infrastructure, systems remain tightly constrained by safety requirements, human oversight and predefined rules. However, they already incorporate machine-speed automation, adaptive anomaly detection and limited decision flexibility. Governance challenges therefore arise not only from fully autonomous systems, but from a trajectory of increasing autonomy.

This evolution can be understood across three levels. First, current systems rely on machine-speed automation, including rule-based orchestration and predefined playbooks. Second, emerging systems introduce adaptive elements, enabling context-aware responses within constrained settings. Third, fully agentic systems — capable of independently selecting responses without real-time human intervention — remain largely conceptual or early stage. While not yet widespread, governance challenges are already evident and are expected to intensify as autonomy increases. Operationally, agentic cyber defence follows a continuous cycle: Sense → Think → Act → Learn. Systems monitor network signals (Sense), evaluate anomalies (Think), implement responses such as isolation, rerouting, blocking or deception (Act), and adapt future assessments based on outcomes (Learn). Humans approve architecture and constraints in

advance but do not authorize individual responses at the moment of execution. For governance purposes, agentic cyber defence systems are defined as defensive mechanisms that: select among multiple response strategies based on contextual assessment; execute actions with operational or cross-domain consequences; and operate without real-time human authorization, instead acting within design-time constraints. The transition from advanced automation to agentic defence occurs when discretion over defensive action is delegated to the system at machine speed rather than predetermined by a human operator. Authority is therefore exercised *ex ante* through system design rather than *ex post* through human approval.

By contrast, systems based on deterministic playbooks, rule-based isolation, or SOAR (platforms requiring confirmation) do not constitute agentic defence. In these cases, automation accelerates execution but does not exercise independent decision authority. This shift moves governance from supervising actions to constraining system behaviour. Agentic cyber defence becomes an embedded decision actor within critical infrastructure: accountability extends to designers and deployers, transparency relies on post-incident audit, and escalation is triggered by engineered thresholds rather than human discretion.[9] The central policy question is therefore not how AI assists operators, but how autonomous authority is bounded in cyberspace.

Figure 2: Agentic Cyber Defence Loop for Policymakers



Source: Authors.

From a policy perspective, ensuring that such system design remains aligned with international norms and shared values requires embedding governance constraints directly into system architecture. This includes defining permissible action boundaries, incorporating escalation thresholds and enabling audit mechanisms that translate high-level principles — such as accountability, proportionality and transparency — into enforceable operational rules.

WHERE GOVERNANCE ASSUMPTIONS BREAK

This paper does not assume the widespread deployment of fully agentic systems, but instead examines governance challenges along a continuum of increasing autonomy in cyber defence. The overwhelming majority of cybersecurity and AI governance frameworks rely on an implicit premise: a human decision maker is available when consequential action occurs.[10] Concepts such as accountability, proportionality, oversight, documentation and escalation depend on this assumption. Agentic cyber defence invalidates it. By relocating defensible authority to machine speed within critical infrastructure, core governance expectations no longer hold.

Assumption 1: A human makes an accountable decision. Legal and regulatory models presume an identifiable actor makes decisions. In agentic defence, judgment shifts to system design and configuration, while operational actions are executed in real time by models. After an incident, regulators may ask who isolated a network segment or blocked traffic, yet no human may have acted at that moment. Responsibility becomes distributed across designers, operators, vendors and owners, complicating notions of "responsible use" and "human oversight." Frameworks such as the NIST AI Risk Management Framework assume a traceable decision maker at action time, but in agentic systems authority is displaced to design, making post hoc accountability ambiguous. For example, automated isolation of a substation may occur without human intervention, complicating attribution when disruptions follow.

Assumption 2: Escalation is a human procedure. Traditionally, escalation involves notifying stakeholders before impactful actions occur. In agentic systems, escalation takes behavioural form — automated isolation, rerouting, blocking or deception — executed at machine speed. While stabilizing locally, such actions may affect external systems without prior coordination. Because intent and proportionality underpin international cyber norms,[11] millisecond-scale escalation challenges their interpretation. Frameworks such as the NIST Cybersecurity Framework and the International Electrotechnical Commission 62443 assume sequential, operator-driven coordination, whereas agentic defence bypasses these steps, leaving actions to be evaluated retrospectively. For example, automated traffic blocking may impact external operators before any notification occurs.

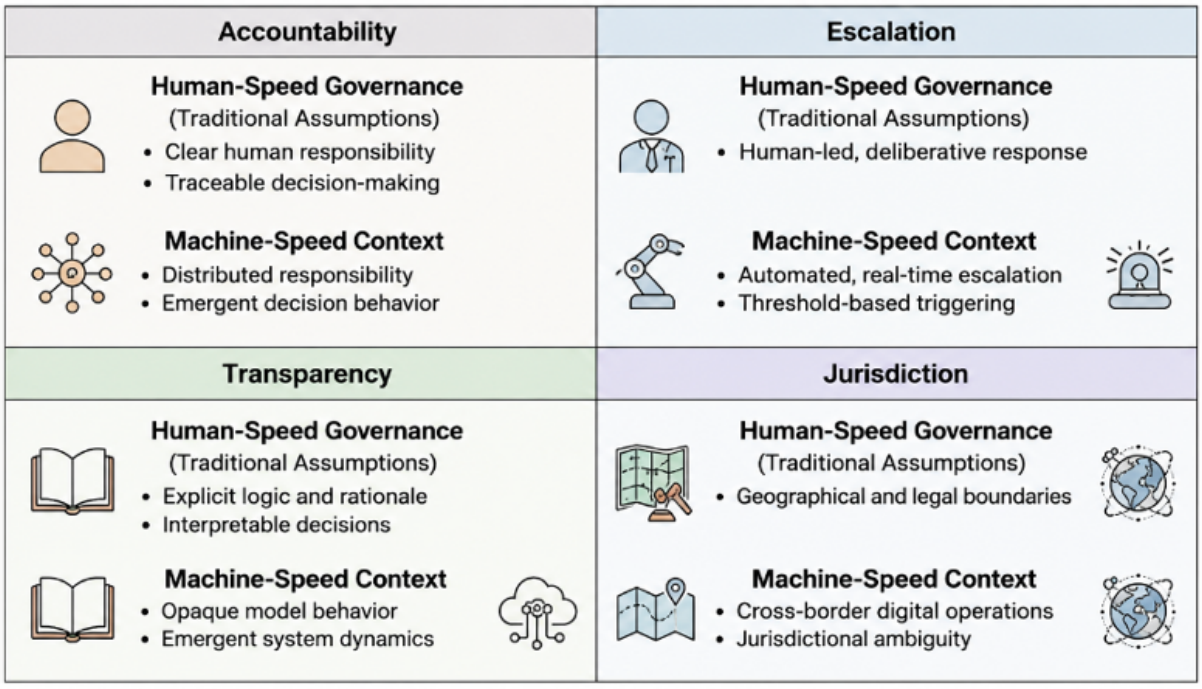
Assumption 3: Transparency occurs during the decision. AI governance often treats transparency as real-time interpretability. Machine-speed defence makes this impractical: reasoning must be reconstructed from logs, model states and telemetry. Transparency thus becomes an audit property rather than a real-

time one, shifting oversight from observation to ensuring interpretable records exist. OECD and NIST expectations align with decision-support systems, not autonomous actions beyond human comprehension.

Assumption 4: Jurisdiction and responsibility are territorially bounded. Critical infrastructures span jurisdictions, and automated containment in one region may affect others. Doctrines based on attribution and intent struggle to classify such effects. Frameworks such as the Tallinn Manual on the International Law Applicable to Cyber Warfare,[12] as well as UN norms, assume attributable conduct, yet agentic responses create cross-border impacts without direct human direction, producing a governance grey zone between state responsibility and operator liability. For instance, isolating a substation or rerouting traffic may unintentionally disrupt neighbouring systems, without clear attribution.

Together, these failures show that agentic defence cannot be governed by extending existing AI policy.[13] Current frameworks assume human-paced action and attribution, whereas agentic systems shift authority to design time and operate at machine speed. As a result, accountability, transparency, proportionality and jurisdiction cannot be managed at the moment of action. Governance must instead constrain how autonomy is designed, bounded and audited in time-critical, interconnected systems. Figure 3 summarizes these broken assumptions.

Figure 3: Four Broken Governance Assumptions



Source: Authors.

SMART GRID CASE STUDY: A GOVERNANCE STRESS TEST

The following scenario is intentionally stylized to illustrate governance stress points under conditions of increased autonomy, while remaining grounded in real-world operational constraints. In practice, smart grids operate under strict safety, reliability and regulatory requirements, including layered human oversight, protection engineering limits and domain-specific control mechanisms. Typical architectures involve supervisory control and data acquisition (SCADA) systems, protection relays, programmable logic controllers, substation networks and supervisory control layers, where autonomous actions are tightly constrained to prevent unintended physical consequences. Consider a stylized but operationally grounded scenario in a substation network, where an agentic-assisted defence system monitors SCADA telemetry and network traffic for anomalies. The system detects patterns consistent with malicious commands targeting protection relays and classifies the event as a high-confidence threat. The decision process is typically based on predefined policies combined with anomaly detection outputs, where confidence thresholds and risk scores determine whether automated containment actions are triggered within a bounded autonomy envelope. Within milliseconds, it initiates bounded defensive actions, such as isolating affected network segments, rerouting supervisory control traffic, blocking suspicious external sources and deploying controlled deception mechanisms. In current practice, such actions would typically be restricted to network-level interventions and remain subject to predefined safety policies, with direct modification of safety-critical components — such as relay protection logic or physical switching operations — prohibited without human authorization. From an engineering perspective, these responses are rational: they aim to preserve system stability and prevent cascading failures. From a governance perspective, however, even these constrained and partially automated actions expose structural gaps.

Accountability. In this scenario, no operator directly approves these actions at the moment they occur. If containment measures contribute to downstream disruptions, responsibility may not map cleanly to a single human decision, but instead becomes distributed across system design, configuration and oversight structures.

Escalation. Traffic blocking and rerouting, even when bounded by predefined policies, may affect interconnected networks across jurisdictions. Actions that would typically require coordination can occur at machine speed, potentially preceding human awareness or intervention.

Transparency. Understanding system behaviour requires post-incident reconstruction using logs, model states and telemetry. As decisions occur faster than human comprehension, oversight shifts from real-time monitoring to audit-based verification.

Jurisdiction. Even constrained defensive actions within one grid may influence operating conditions in interconnected systems across geopolitical boundaries, raising questions that existing cyber norms and legal frameworks are not fully equipped to address.

Table 1 summarizes how agentic cyber defence in a smart grid functions as a governance stress test. While the objective — maintaining grid stability — remains unchanged, decision authority, escalation timing, accountability and jurisdiction shift fundamentally under machine-speed defence.

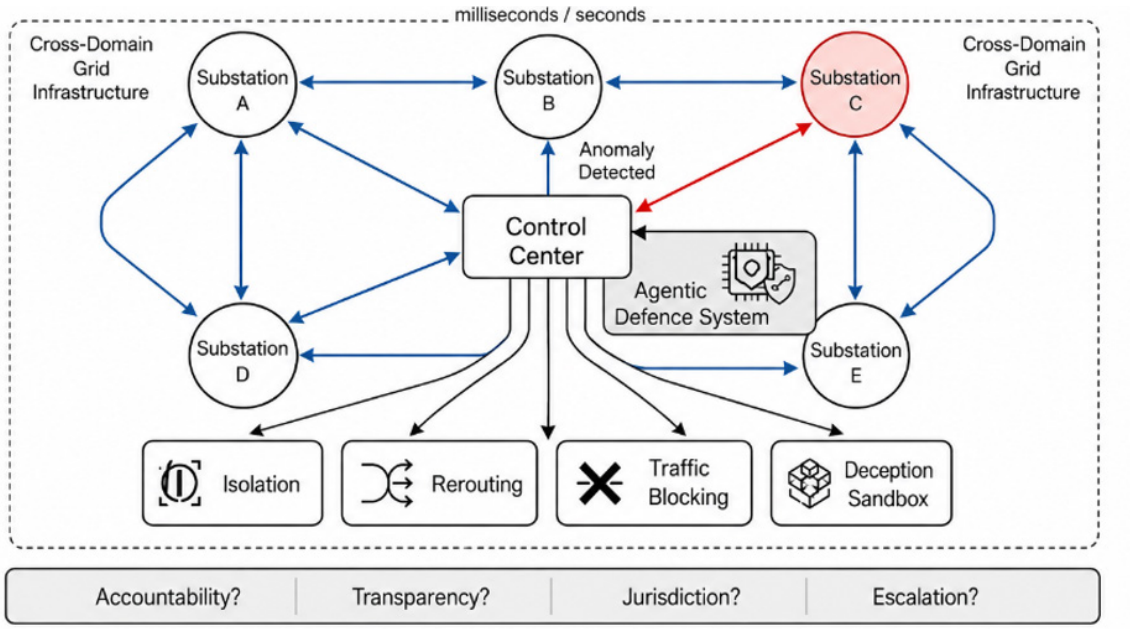
Table 1: Agentic Cyber Defence in a Smart Grid Context

Governance Dimension	Human-speed Cyber Defence Model	Agentic Cyber Defence Outcome
Decision authority	Defensive actions approved by operators or incident commanders	Defensive actions selected by system inference within a predefined autonomy envelope
Escalation timing	Minutes to hours, following procedural escalation paths	Milliseconds, triggered automatically by system thresholds
Accountability locus	Individual operator or organizational decision maker	Distributed across system designers, deployers, vendors and operators
Transparency mechanism	Real-time situational awareness and human interpretability	Post-event reconstruction through logs, telemetry, and model state
Operational scope	Localized to a single utility or administrative domain	Potentially cross-domain and cross-jurisdictional
Jurisdictional effects	Largely contained within territorial boundaries	May produce transboundary infrastructure effects

Source: Authors.

Smart grids expose these governance tensions because delay is not an option. Digitalization expands the attack surface while shrinking the human response window. The shift is not faster response, but delegating it to systems operating beyond human perception. Figure 4 illustrates how such actions can both preserve stability and reveal governance gaps.

Figure 4: Agentic Cyber Defence in a Smart Grid Scenario



Source: Authors.

This tension shows that governance challenges stem not from unconstrained autonomy, but from the interaction between safety-bounded automation and cross-domain effects. From a policy perspective, this requires clear autonomy boundaries, defined escalation protocols and auditable mechanisms to ensure alignment with safety and cross-jurisdictional coordination.

G7, OECD AND UN COMPARISON: WHAT TODAY'S FRAMEWORKS COVER AND WHAT THEY MISS

Do international AI and cyber governance frameworks address machine-speed autonomy in critical infrastructure defence? Only partially. Current regimes articulate norms relevant to agentic defence — responsibility, safety, transparency, cooperation and stability — but rarely translate them into operational constraints applicable across jurisdictions.

G7 Hiroshima AI Process. The G7 provides the most comprehensive high-level expectations for advanced AI, emphasizing lifecycle responsibility, monitoring, risk management and transparency. These principles align with challenges raised by human–machine defence interaction. However, they presume meaningful human review at the moment of decision. No guidance exists for bounding autonomy when responses occur in milliseconds, nor for managing cross-border infrastructure effects. Concrete governance gap: While the framework emphasizes lifecycle responsibility and risk management, it does not specify enforceable constraints on autonomy boundaries, escalation triggers or audit requirements for machine-speed defensive actions in interconnected infrastructure.

OECD AI Recommendation. The OECD has established influential criteria for trustworthy AI: human-centredness, transparency, robustness, safety and accountability. While directly applicable in principle, they assume contemporaneous interpretability and human oversight. Agentic defence instead requires post-incident reconstruction and pre-engineered behavioural constraints. The framework promotes reliability but does not connect robustness to escalation testing, safe-action boundaries or inter-operator coordination. **Concrete governance gap:** Although the framework promotes trustworthy AI principles, it does not operationalize how robustness, transparency and accountability should be evaluated in real-time autonomous responses and in the presence of cross-system externalities.

UN Cyber Stability (Group of Governmental Experts/Open-Ended Working Group). UN processes provide substantial guidance on responsible state behaviour in cyberspace.[14] Yet they focus on state intent and attributable conduct. Most agentic defensive systems will be operated by private infrastructure providers. Automated containment or deception may create cross-border consequences without direct human direction, complicating attribution and due diligence assessments. Concrete governance gap: Existing norms focus on state-attributable actions, but do not address how to interpret or govern autonomous defensive measures by private operators that may produce cross-border effects without direct human intent.

Across all regimes, a common pattern is noticeable: values are well defined but implementation guidance for machine-speed defensive autonomy is absent. These "operational missing" gaps refer to the absence of concrete mechanisms that translate high-level principles into enforceable system-level constraints for autonomous behaviour. Table 2 maps this divergence between normative coverage and operational governance requirements. These gaps suggest complementary roles rather than institutional failure. The G7 can establish expectations for bounded autonomy and allied assurance, the OECD can operationalize testing and governance practices, and the UN can align international stability norms with privately operated autonomous defence systems. Governing agentic cyber defence therefore requires coordinated evolution across frameworks rather than replacement of existing ones.

Table 2: Mapping Governance Coverage to Agentic Cyber Defence Needs

Topics	Bounded Autonomy	Escalation Thresholds	Transboundary Coordination
G7 Hiroshima AI Process	Principle-level	Principle-level	Partial
OECD AI Recommendations	Principle-level	Not specified	Partial
UN GGE/OEWG Cyber Norms	Not specified	Partial	Operational guidance
Policy Gap for Cyber Defence	Operational missing	Operational guidance	Operational missing

Source: Authors.

FOUR GOVERNING PRINCIPLES FOR MACHINE-SPEED CYBER DEFENCE

The previous section showed that international AI and cyber frameworks converge on values such as trust, responsibility and stability, but lack operational guidance for autonomous defence. We introduce two mechanisms: the Autonomy Envelope (AE), defining permissible machine actions without human approval, and the Escalation Threshold (ET), specifying when human intervention is required. Together, AE and ET treat autonomy as an engineered governance constraint rather than a purely technical capability.[15]

Principle 1: Autonomous action shall be constrained by policy rules.

Meaningful oversight requires constraining discretion in advance. Agentic cyber defences must operate within a declared AE proportional to critical-infrastructure risk.[16] Operational controls include:

- Allow-listed actions: explicitly permitted automatic responses (e.g., subnet isolation, rerouting control traffic, rate limiting, deception).
- Deny-listed assets/effects: safety-critical controls (e.g., protection relays, programmable logic controller[17] safety logic) never altered automatically except under emergency authority.
- Duration limits: bounded time windows for autonomous action unless renewed by a human.
- Safe failure states: predefined fallback modes if sensing fails, constraints break or confidence drops.

Operators must document the AE specifying permitted actions, assets, time bounds and failure modes. For example, in a smart grid context, an AE may permit automated network isolation and traffic filtering, Evaluation criteria: completeness (all action classes specified); boundedness (spatial, temporal, functional limits defined); and audit coverage (each action traceable to authorized policy). Incomplete or unauditable envelopes constitute governance failure.

Principle 2: Systems shall escalate to human intervention when thresholds are exceeded.

Autonomy suits low-risk rapid containment; higher consequences require escalation via ET derived from severity, uncertainty and externality. Escalation triggers include actions that:

- Affect external administrative domains
- Involve shared operators or interties
- Expand blocking scope
- Deploy deception
- Approach safety margins

Upon threshold breach, the system must either request authorization or shift into reduced-authority safe mode. In practice, such thresholds may be calibrated using indicators such as anomaly confidence scores, estimated system impact (e.g., load imbalance or service disruption risk), and predefined risk tolerance levels established by operators and regulators. General rule:

IF (t > T) OR (confidence < τC) OR (risk > τR) OR (externality > τX) → escalate and restrict actions.

Evaluation criteria: false escalation rate; missed escalation risk; and threshold coherence. Persistent over- or under-triggering signals governance misalignment.

Principle 3: Decisions shall be auditable by third parties.

Because actions occur faster than human reasoning, governance must rely on post-incident reconstruction.[18] Systems should generate high-integrity evidence supporting independent review. Required audit artifacts:

- Event → Input → Assessment → Action mapping
- Model and policy versioning
- Cryptographically protected logs
- Retention and access consistent with forensic standards

Auditability supports accountability and shared interpretation across jurisdictions. Similar mechanisms are already reflected in practices such as security logging standards, incident response reporting requirements, and emerging AI audit frameworks that emphasize traceability and post-incident accountability.

Evaluation criteria: reconstruction completeness; temporal integrity; and third-party verifiability. Internal troubleshooting alone is insufficient.

Principle 4: Autonomous responses engage cross-border assurance practices.

Interconnected infrastructure means automated actions may affect neighbouring operators. Coordination mechanisms operationalize cooperation norms. Operationalization may involve predefined coordination protocols between grid operators, shared incident notification mechanisms, and bilateral or multilateral agreements that specify acceptable autonomous actions and escalation pathways across jurisdictions:

- Neighbour-impact notifications
- Shared AE and audit standards
- Predefined communication channels
- Common terminology for autonomy and escalation

Evaluation criteria: notification latency; semantic alignment; and assurance interoperability. Mechanisms failing these provide limited stabilizing value.

Together, these elements form an autonomy stack translating high-level norms into enforceable governance controls (Figure 5). The framework regulates operational conditions of autonomy rather than specific algorithms, enabling testing and iterative improvement without constraining technical implementation.

Figure 5: Four Governing Principles as an Implementation Stack

Latency Regime (Time to Decide)	Example Authorized Actions (in Cyber Defence)	Governance Implications (Key Questions)
Real-Time (< 1 second)	• Packet filtering / blocking • Session termination	• Who defines the rules? • Ex ante accountability
Near Real-Time (1–60 seconds)	• Rerouting traffic • Isolating a device or segment	• What is the scope of delegated authority? • How are actions logged and audited?
Human Timeframe (Minutes–Hours)	• Incident review • Forensic analysis • Reporting and notification	• How is human oversight integrated? • What information is required for review?
Policy / Legal Time (Days–Months)	• Policy updates • Legal and regulatory response • Cross-border coordination	• How are legal and jurisdictional issues resolved? • How are lessons translated into policy?

Source: Authors.

INTERNATIONAL STABILITY IMPLICATIONS

Agentic cyber defence is often framed as a technical improvement — faster detection, containment and recovery.[19] However, when deployed across critical infrastructure spanning jurisdictions, it produces strategic effects. Design choices shape how actions are interpreted between states, how crises evolve and what assurances sustain cross-border trust. International cyber stability frameworks assume actions are deliberate, attributable and slow enough for human decision-making — assumptions that agentic defence challenges. Automated blocking, rerouting or deception affecting foreign infrastructure are not acts of state intent but system behaviours under delegated autonomy. For example, containment in one country's power grid — such as isolating a substation or rerouting traffic — may unintentionally disrupt neighbouring systems and be perceived as hostile. Machine-speed responses further compress crisis timelines, altering system conditions before diplomatic coordination is possible and increasing escalation risk. Stability therefore depends not only on political commitments but also on technical assurances. Autonomy envelopes, auditable records and predefined coordination channels act as confidence-building

measures, supported by shared notification protocols, standardized reporting, pre-agreed thresholds and rapid post-incident information exchange. Governing machine-speed defence thus becomes integral to international stability, as automated actions increasingly shape cross-border infrastructure interactions without direct human intent.

CONCLUSION

The emergence of agentic AI in cyber defence introduces a governance latency gap, in which system actions outpace the capacity of traditional human-centred decision making and oversight structures. Addressing this gap requires a shift from reactive governance toward design-time control, where autonomy boundaries, escalation thresholds, auditability and cross-border coordination are explicitly embedded into system architecture. For policymakers, this implies several immediate priorities: defining enforceable autonomy envelopes for critical infrastructure, establishing measurable escalation criteria, requiring auditable decision-making processes and developing coordination mechanisms to manage cross-border effects. Without such measures, the increasing deployment of machine-speed defensive systems risks creating ambiguity in accountability, unintended escalation and instability in interconnected environments. As cyber defence continues to evolve toward greater autonomy, governance frameworks must keep pace. The challenge is no longer whether autonomous systems will act, but how their actions can be constrained, interpreted and coordinated within existing institutional and international structures.

END NOTES

[1] Mahesh Narayanan, Muhammad Asfand Hafeez and Arslan Munir, "Encryption for Industrial Control Systems: A Survey of Application-Level and Network-Level Approaches in Smart Grids," *Journal of Cybersecurity and Privacy* 6, no. 1 (2026): 11; Abbas Yazdinejad and Jude Dzevela Kong, "Responsible Use of Large Language Models in Digital Health: An Equity First Governance Framework," SSRN 5962741 (2025).

[2] Tim Krause, Raphael Ernst, Benedikt Klaer, Immanuel Hacker and Martin Henze, "Cybersecurity in power grids: Challenges and opportunities," Sensors 21, no. 18 (2021): 6225, DOI: https://doi.org/10.3390/s21186225.

[3] NIST, *Artificial Intelligence Risk Management Framework* (AI RMF 1.0), (2023): 100–1, DOI: https://doi.org/10.6028/NIST.AI.100-1.

[4] Manisha Parmar and Andy Miles, "Cyber Security Frameworks (CSFs): An Assessment Between the NIST CSF v2. 0 and EU Standards," in 2024 *Security for Space Systems* (3S), 1–7, IEEE, 2024.

[5] Nanette Levinson, "Inclusive Anticipatory Governance: Cyber Technologies, Absorptive Capacities and the Case of the United Nations Open-Ended Working Group re: ICTS," 2022.

[6] Ayako Takemi, "Japan's Strategic Science Diplomacy in International AI Governance Architecture: A Critical Analysis of the Hiroshima AI Process," SSRN 5197909 (2025).

[7] Zhimin Zhang, Huansheng Ning, Feifei Shi, Fadi Farha, Yang Xu, Jiabo Xu, Fan Zhang and Kim-Kwang Raymond Choo, "Artificial intelligence in cyber security: research advances, challenges, and opportunities," *Artificial Intelligence Review* 55, no. 2 (2022): 1029–1053.

[8] Jean-Christophe Noël, "Human compatible: AI and the problem of control," *Politique Etrangère* 4 (2020): 202–203; Michael N. Schmitt, ed., *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press, 2017).

[9] Fabio Morandín-Ahuerma, "Recommendation of the OECD council on artificial intelligence: inequality and inclusion," in *Principios normativos para una ética de la inteligencia artificial*, Consejo de Ciencia y Tecnología del Estado de Puebla (México), 2023, 95–102.

[10] Abbas Yazdinejad, Maral Niazi, James W. Hinton, Jude Kong, Jake Okechukwu Effoduh and Anna Shin, "A Community-Centred Protocol for Ethical and Scalable AI in Health Care," JSTOR, 2025, https://www.jstor.org/stable/resrep73065.

[11] Levinson, "Inclusive Anticipatory Governance."

[12] Schmitt, *Tallinn Manual*.

[13] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler et al., "Practices for governing agentic AI systems," Research Paper, OpenAI (2023).

[14] Eneken Tikk-Ringas, "Developments in the Field of Information and Telecommunication in the Context of International Security: Work of the UN First Committee 1998-2012," *Cyber Policy Process Brief* (2012); Claudia Henyka, "Maintaining peace and security in cyberspace: Multilateral approach of the United Nations on advancing responsible state behaviour in cyberspace," 2022.

[15] Abbas Yazdinejad, Hadis Karimipour and Talal Halabi, "Towards Stress-Adaptive Cyber Defense: Cognitive–Physiological Synchronization in IoT Environments," *IEEE Internet of Things Journal* (2026); 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).

[16] Virginia Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, vol. 2156 (Springer, 2019).

[17] "Programmable Logic Controller (PLC)," Science Direct, https://www.sciencedirect.com/topics/computer-science/programmable-logic-controller.

[18] Finale Doshi-Velez and Been Kim, "Towards a rigorous science of interpretable machine learning," *arXiv* preprint, *arXiv:1702.08608* (2017); Havva Alizadeh Noughabi, Fattane Zarrinkalam, Abbas Yazdinejad and Ali Dehghantanha, "TrollSleuth: Behavioral and Linguistic Fingerprinting of State-Sponsored Trolls," in *2025 22nd Annual International Conference on Privacy, Security, and Trust* (PST), 1–10, IEEE, 2025.

[19] Hassan Ali, Eman Fatima and Abeer Abdul Aziz Alsanad, "Managing Cyber Risk Externalities in Interdependent and Federated Network Environments," 2025.

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.

Jatin Nathwani is Professor Emeritus, Department of Management Science and Engineering, University of Waterloo, a BSIA Fellow and Technology Governance Initiative Fellow, and Founding Executive Director, Waterloo Institute for Sustainable Energy (WISE). Professor Nathwani is one of Canada's foremost experts and thought leaders on sustainable energy policy and technology governance. He has held leadership positions at the University of Waterloo and advised government, business, academic and civil society organizations. He has made significant contributions to the development of science in support of sustainable energy policy, capacity building, and community-building, all in support of transitioning global and national energy systems towards more just and sustainable outcomes.

BALSILLIE PAPERS

ISSN 2563-674X • doi:10.51644/BAP103

balsilliepapers.ca