



INCD
Israel National
Cyber Directorate

The AI Strategic Triangle (AIST) A Time for Leadership

A strategic framework for balancing innovation, security, and human agency as AI capabilities advance.

Yossi Karadi

Director General of the Israel National Cyber Directorate





Executive Summary

The rapid evolution of artificial intelligence, the emergence of frontier models, and the potential progression toward Artificial General Intelligence (AGI) are beginning to reshape the foundations of governance and society. Although AI integration into physical systems and everyday services remains uneven, its strategic effects are already visible. AI is changing the sources of national power, economic structures, security dynamics, and the boundaries of authority delegated to machines.

Sovereign leaders bear the ultimate duty to govern this transformation. Humanity's future cannot be outsourced to corporate boardrooms. While global technology giants drive breakthroughs and confront existential dilemmas, their corporate mandates and institutional legitimacy differ fundamentally from those of sovereign states.

Governments must therefore do more than merely adopt AI. They must determine where AI creates strategic advantage, where dependence creates unacceptable risk, which capabilities require domestic or trusted alternatives, and which decisions must remain subject to meaningful human authority.

This paper introduces the **AI Strategic Triangle (AIST)** as a framework for national-level decision-making. AIST is designed to help leaders assess AI policy across three interdependent strategic dimensions:

- **Innovation**
- **Security**
- **Human Agency**

The vertices of the AIST are not a zero-sum paradigm. The model serves to illuminate core tensions while uncovering operational strategies that expand national capabilities across multiple fronts simultaneously.

The goal of AIST is not to eliminate these tensions, but to bring them into the open, support informed decision-making, and assist **national leaders in formulating independent policies** that balance technological aspirations, national resilience, and human agency.

Introduction

Artificial intelligence is an established reality. Its disruption extends far beyond digital environments, entering the physical realm where individuals directly interact with intelligent automation.

The structural impact of this revolution will run far deeper than preceding industrial revolutions, precisely because it is, in some contexts, replacing cognitive capacities - such as analysis, creative production, and autonomous decision-making.

As humanity navigates this critical threshold, sovereign leadership faces three structural challenges unique to the AI paradigm:

Technical Complexity

Severe difficulty across non-technical leadership in assessing real-world capability thresholds and underlying operational risks.

Machine-Scale Tempo

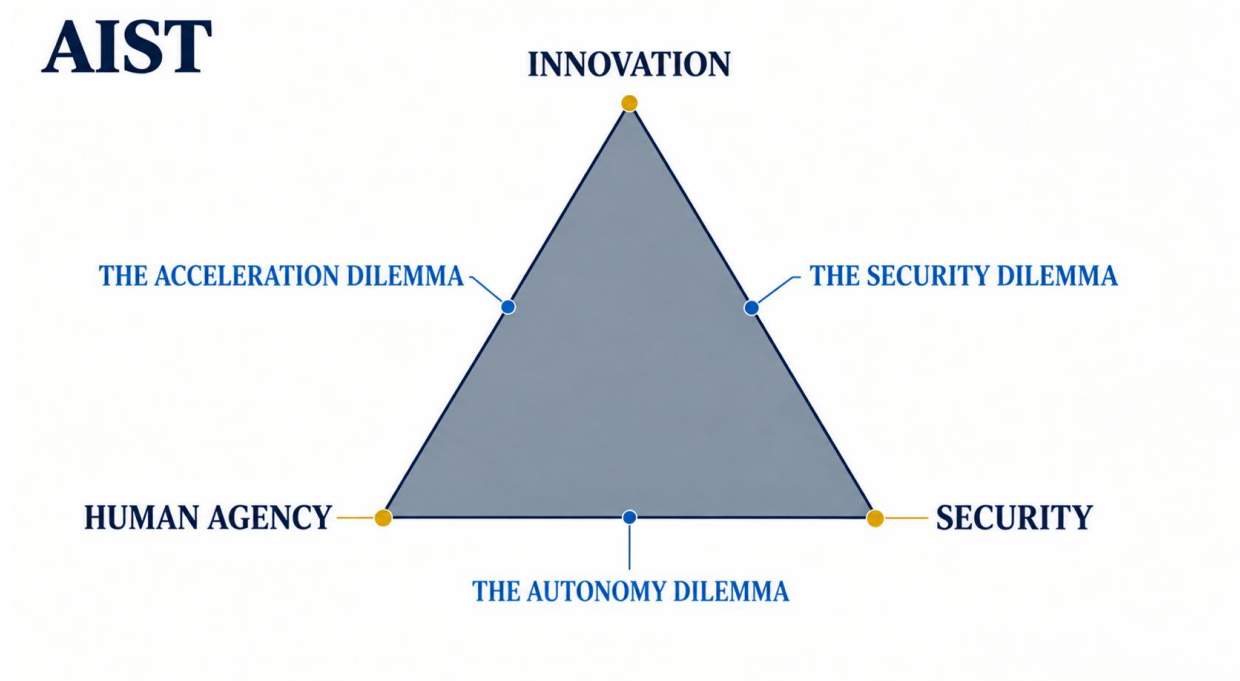
Advancing at machine rather than human speed, AI capabilities evolve exponentially, outpacing conventional policy cycles.

Ubiquitous Surface Area

AI is becoming relevant across nearly every major function of the state.

The AI Strategic Triangle (AIST) offers sovereign leaders a practical lens for decisions across Innovation, Security, and Human Agency. It also applies to leaders of complex global organizations facing comparable questions of dependence, accountability, and control.

The Architecture of the AIST:





The Strategic Vertices

Security

Frontier models may **shift the offense-defense balance** in cybersecurity and national defense. Exploits are evolving into autonomous, hyper-personalized, machine-speed attacks. Concurrently, adoption intensifies systemic dependencies on external digital supply chains, challenging informational sovereignty, critical infrastructure resilience, and sovereign freedom of action.

Conversely, leveraging hyperscale cloud environments can offer strong baseline security capabilities and operational maturity compared to fragmented on-premise networks. Sovereign leadership must reconcile this paradox: how to defend against automated offense when existing defenses lag, and whether an enduring equilibrium between AI-driven attack and defense can be achieved.

Key vectors include:

- AI-augmented cyber defense (AISI, 2025; Folkerts et al., 2026).
- Cyber defense tailored to the AI stack (data poisoning, algorithmic exploits).
- Autonomous, AI-vs-AI operational engagements (Folkerts et al., 2026; Black et al., 2024).

Note: It is important to understand that sovereignty is not security. Our systems can be highly secured by a global company, yet we remain completely dependent on it—and vice versa.

Innovation

Artificial intelligence is an unprecedented engine of national power, accelerating scientific discovery, public-service delivery, and economic productivity. However, its strategic value extends far beyond efficiency: nations that successfully integrate AI across defense, research, and critical infrastructure will define future geopolitical competitiveness and preserve technological sovereignty.

Human Agency

One of the central policy questions of the artificial intelligence era is where human authority must remain decisive. As AI systems increasingly inform or execute decisions across finance, healthcare, critical infrastructure, intelligence, and national security, governments must clearly delineate where ultimate accountability lies. The issue extends beyond whether a human is formally in the loop. Effective policy requires a clear taxonomy: distinguishing between decisions that can be fully automated, those bounded by predefined constraints, those requiring meaningful human approval or independent oversight, and those that cannot be delegated under any circumstances. This establishes the principle of non-delegable sovereign authority. Certain state actions carry such profound consequences that final discretion must reside exclusively with an accountable human official, irrespective of technological sophistication. Depending on national frameworks, these typically include the derogation of fundamental rights or actions with irreversible strategic implications. While precise boundaries will vary across legal traditions and operational domains, the essential mandate is clear: they must be established deliberately, rather than conceded by default to technological inertia (Autio et al., 2024; Tabassi, 2023).



The Strategic Dilemmas

The Security Dilemma

Innovation vs. Security

The dramatic gains in speed, resource optimization, and operational efficiency driven by AI are transformative, compelling nations to accelerate AI adoption across all viable domains. However, global AI adoption relies almost exclusively on cloud infrastructure, foundation models, and commercial services provided by multinational tech giants. Consequently, leveraging these global capabilities often necessitates transferring sensitive data, delegating mission-critical workflows, or exposing proprietary repositories beyond sovereign jurisdictional control—substantially expanding a nation's attack surface (OECD, 2026; Lehdonvirta et al., 2025; Black et al., 2024).

Data is widely recognized as a strategic national asset. At the sovereign level, the stakes are exponentially higher: exposure of national or classified data to hostile adversaries could inflict severe strategic harm. Furthermore, outsourcing core processes to external infrastructures and integrating models with deep system access introduces critical cybersecurity risks, including data poisoning, algorithmic vulnerabilities, and adversarial exploitation (Autio et al., 2024).

Neither extreme, unconstrained adoption or complete isolation, presents a viable path. National leadership must calibrate this trade-off based on a rigorous assessment of their state's threat landscape and strategic posture.

For example, a smaller nation lacking domestic semiconductor fabrication or hyperscale compute cannot realistically achieve absolute, end-to-end AI sovereignty. Instead, it must pursue a pragmatic strategy: focusing on designated critical national capabilities while selectively leveraging global infrastructure through multi-cloud and multi-model architectures, all underpinned by a resilient, sustainable sovereign cloud and AI ecosystem (Lehdonvirta et al., 2025; OECD, 2026).

The Acceleration Dilemma

Innovation vs. Human Agency

Competitive pressure can accelerate AI adoption before institutions are ready to govern its consequences. Leaders must weigh the cost of falling behind against the risk of delegating consequential decisions to systems whose behavior they cannot reliably explain or contest.

These are not abstract debates; they demand immediate sovereign choices: *Should artificial intelligence be deployed across high-consequence public services before robust oversight and accountability mechanisms are mature?* In a global competitive landscape, game theory creates a dangerous trap: if a nation pauses to implement strict governance while an adversary operates unconstrained, the adversary gains an immediate strategic advantage.

The Autonomy Dilemma

Human Agency vs. Security

Theoretically, maintaining human control and ethical standards should go hand in hand with security. In practice, however, this is not always the case. As the race in AI accelerates, advances in



frontier AI are increasing pressure on existing security controls and may enable attackers to identify, scale, or execute certain classes of attacks more rapidly. Traditional security approaches may therefore require substantial adaptation to remain effective. To defend effectively, we must implement AI-based cybersecurity systems—and we must do so quickly. (NCSC, 2025).

This situation creates the "autonomy dilemma". As cyberattacks become faster, more automated, and more targeted, defense systems must detect, decide, and act at machine speed—a pace too rapid to allow for human involvement in every single decision. Without a significant level of autonomy, we simply cannot provide timely protection. Yet, greater machine autonomy can make it more difficult to understand system actions, intervene when necessary, and maintain clear lines of accountability (NCSC, 2025; AISI, 2025; Black et al., 2024)

The challenge is to define clear boundaries in advance: which defensive actions can be fully delegated to the machine, which require human approval, and which must never be automated.

This issue is becoming increasingly urgent as frontier models undergo constant updates. As AI agents gain broader access to sensitive information, system tools, and active networks, they execute chains of actions that no one could have fully vetted in advance down to the last detail. This does not mean we have lost control, but it necessitates the implementation of rigorous protection and control mechanisms: clear limits on authority, reliable "kill switches," and unambiguous accountability (Autio et al., 2024; NCSC, 2025; Folkerts et al., 2026; AISI, 2025).

Consider an AI cyber-defense system that requires autonomous authority to isolate infrastructure at machine speed to counter a zero-day attack, yet a single false positive could inadvertently shut down a critical national power grid or hospital network. In the event of a catastrophic failure, who bears the responsibility?

Establishing the National Equilibrium

A sovereign position does not reside at extreme vertices, nor does it remain static along a single edge; instead, it is located somewhere *within* the triangle, corresponding to the three coordinates anchored at each edge. This juncture represents the **National Equilibrium**, which may evolve dynamically over time. In any case, its location within the triangle broadly reflects three strategic orientations:

Innovation & Supremacy Orientation

Prioritizes technological velocity and economic scale, actively accepting calculated security and ethical trade-offs to capture competitive dominance.

Security & Defense Orientation

Centers on systemic robustness, rigorous zero-trust containment, restricted cross-border data transit, and measured, verified implementation.

Human Agency & Responsibility Orientation

Focuses on human control, precautionary governance, strict accountability frameworks, and verifiable algorithmic transparency.



Case Study

Machine-Speed Defense of Critical Infrastructure

Consider an autonomous, multi-agent cyberattack targeting a nation's power grid. At machine speed, the attack identifies zero-day vulnerabilities across industrial control systems.

The Security Dilemma: Utilizing the highest-tier frontier models for predictive defense requires continuous telemetry feeds into external, hyperscale cloud environments—introducing jurisdictional vulnerability and supply chain dependency into sovereign critical infrastructure.

The Autonomy Dilemma: Effective mitigation demands isolation of grid assets within milliseconds. Delegating automated shutdown authority to an AI agent safeguards the network against widespread collapse, but risks inadvertent, catastrophic power outages in hospitals or life-support systems due to false positives.

The Acceleration Dilemma: Pausing deployment to build exhaustive testing frameworks leaves national infrastructure defenseless against unconstrained adversaries operating at machine tempo.

The AIST Resolution:

Rather than accepting fatal trade-offs, AIST guides leadership to calibrate a dynamic national equilibrium across four operational pillars:

1. Tiered Authority & Bounded Delegation

- **Tier 1 (Full Machine Autonomy):** High-frequency, low-impact containment (e.g., session isolation, micro-segmentation) executes at machine speed.
- **Tier 2 (Constrained Agency):** Medium-impact actions (e.g., localized load rerouting) operate within strict algorithmic guardrails and verifiable policy envelopes.
- **Tier 3 (Non-Delegable Sovereign Authority):** Irreversible, high-consequence decisions remain strictly under human command. AI provides decision support only.

2. Architectural Resilience (Hybrid Sovereignty)

- **Macro Tier:** Leverages global hyperscale cloud for large-scale training and sanitized global threat intelligence.
- **Micro Tier:** Deploys lightweight, air-gapped sovereign models on local edge/SCADA networks, ensuring operational resilience during severed external connectivity.

3. Hardware Interlocks & Dynamic Rollback ("Kill Switches")

Autonomous systems feature hardware-enforced, cryptographically segregated cutoffs. Any anomalous agent drift or exploit triggers instant revocation of authority, reverting infrastructure to verified manual baselines.

4. Sovereign Liability & Preemptive Governance

Prevents regulatory lag through statutory safe harbors for certified defense agents and unambiguous legal liability structures for human commanders.



Conclusion

From the perspective of national cyber defense, the tensions described by AIST are already operational realities. AI capabilities are advancing at a tempo that can exceed existing security, governance, procurement, and organizational processes (NCSC, 2025; OECD, 2026).

The central argument is not that nations should pursue technological self-sufficiency or slow innovation. Rather, they should define a dynamic national equilibrium based on their strategic interests, risk tolerance, institutional capacity, and international environment. This requires coordinated policy across data governance, cybersecurity, critical infrastructure, public procurement, research and workforce development, and trusted international partnerships. Leaders do not need to become AI engineers. They must decide where national dependence is acceptable, where resilience is non-negotiable, and where human authority must remain decisive (Tabassi, 2023).

The AIST is intended as a strategic instrument for making those choices coherently and revisiting them as technology and geopolitical conditions evolve.

Strategic Diagnostic: Key Questions for National Leadership

- **Innovation:** In which strategic domains is domestic capability mandatory, and where should the nation leverage global infrastructure?
- **Security:** Which national workflows and critical data assets must never face external dependencies, and how do we prevent vendor lock-in?
- **Human Agency:** Which state functions require mandatory human oversight, and where are sovereign liability boundaries drawn?
- **The Security Dilemma:** What is our posture regarding external reliance versus data exposure, and what is our target equilibrium?
- **The Acceleration Dilemma:** How does our adoption speed balance domestic institutional resilience against external geopolitical competition?
- **The Autonomy Dilemma:** What operational envelopes are safe to delegate to automated agents, and where must humans remain in the command loop?

Policy Execution Priorities

Once the National Equilibrium is determined, translate it into unified government instruments: strategic public procurement, fiscal and research incentives, national security regulations, critical infrastructure directives, and international bilateral compacts.

Do not leave it to others to define your future; enlist them to build the world you believe in.

Yossi Karadi

Director General of the Israel National Cyber Directorate



References

- Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
- Lehdonvirta, V., Wu, B., Hawkins, Z. J., Caira, C., & Russo, L. (2025). Measuring domestic public cloud compute availability for artificial intelligence (OECD Artificial Intelligence Papers No. 49). OECD Publishing. <https://doi.org/10.1787/8602a322-en>
- National Cyber Security Centre. (2025, May 7). Impact of AI on cyber threat from now to 2027. <https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027>
- Organization for Economic Co-operation and Development. (2026). AI markets: Recent developments and competition issues. OECD Publishing. <https://doi.org/10.1787/d531d73f-en>
- Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- AI Security Institute. (2025). AISI frontier AI trends report 2025. <https://www.aisi.gov.uk/research/aisi-frontier-ai-trends-report-2025>
- Black, J., Eken, M., Parakilas, J., Dee, S., Ellis, C., Suman-Chauhan, K., Bain, R., Fine, H., Aquilino, M. C., Lebre, M., & Palicka, O. (2024). Strategic competition in the age of AI: Emerging risks and opportunities from military use of artificial intelligence. RAND Corporation.
- Folkerts, L., Payne, W., Inman, S., Giavridis, P., Skinner, J., Deverett, S., Aung, J., Zorer, E., Schmatz, M., Ghanem, M., Wilkinson, J., Steer, A., Hong, V., & Wang, J. (2026). Measuring AI agents' progress on multi-step cyber attack scenarios. AI Security Institute. <https://www.aisi.gov.uk/research/measuring-ai-agents-progress-on-multi-step-cyber-attack-scenarios>
- Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H. A., & Alstott, J. (2024). Securing AI model weights: Preventing theft and misuse of frontier models. RAND Corporation. <https://www.rand.org/t/RRA2849-1>