NPSGLOBAL FOUNDATION

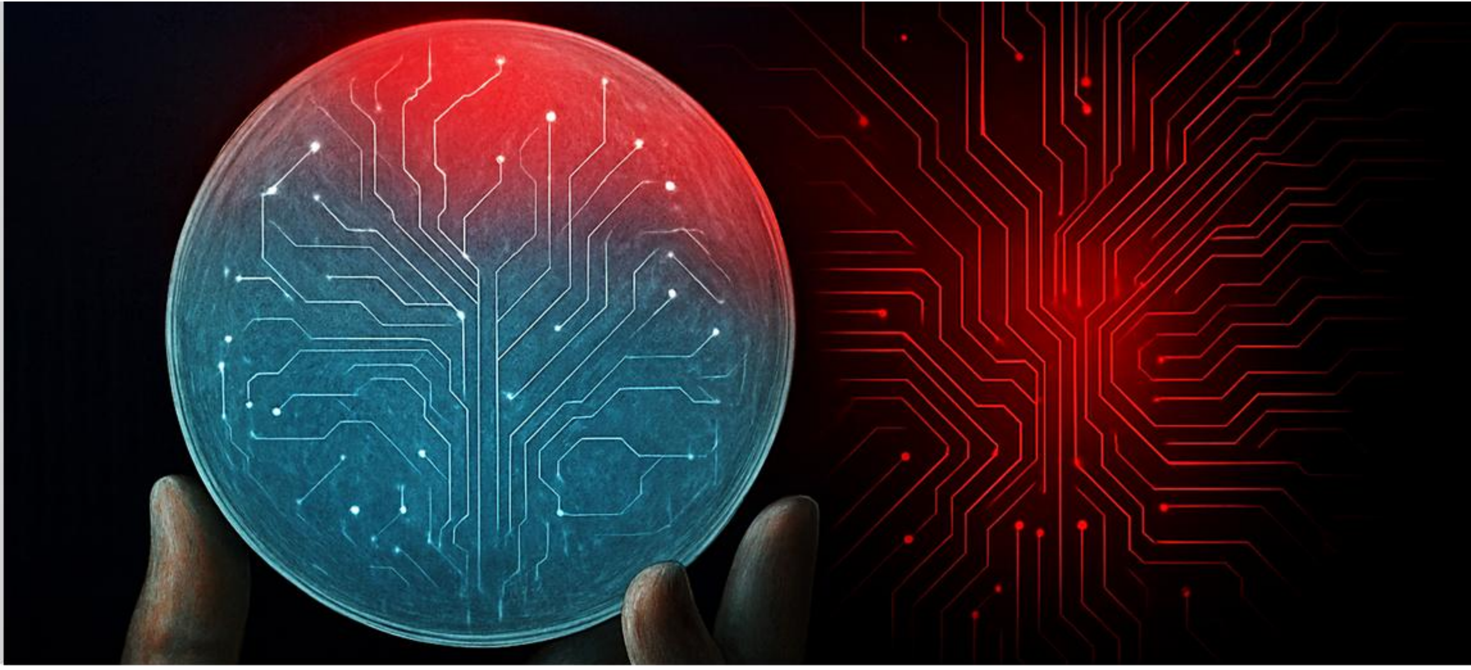

Illustration generated with AI - LGG. Concept by NPSGLOBAL

Safety is not Enough

Toward a Dual-Risk Framework for Artificial Intelligence Governance

By Irma Argüello – Chair, NPSGLOBAL Foundation. August 2025.

Why Safety is not Enough

Global conversations around artificial intelligence (AI) creators and governance makers increasingly revolve around a single concept: safety, ensuring that systems do not cause harm. But while safety is essential, it is not sufficient. In the real world, dangers not only come from what AI might do autonomously, but from what humans can deliberately do with it.

The prevailing safety paradigm assumes that risk resides within the system: models that hallucinate, escape human control, or behave unpredictably. Yet equally severe threats lie outside models**,** in the hands of actors who intentionally manipulate, repurpose, or weaponize them for their own ends.

This is not a hypothetical concern. It is already happening. Models are being fine-tuned to spread disinformation, evade content moderation, and replicate voices and faces without consent. Authoritarian regimes and cybercriminal networks are embedding AI into dark architectures, influencing military decision-making, and conducting large-scale psychological operations.

With the rise of open-weight models, including recent releases by major actors under permissive licenses, anyone with sufficient computational power and intent can now replicate or weaponize foundational systems. While some releases follow responsible frameworks, the structural risk lies in the diffusion itself.[1]

The problem is not the code. The problem is the context. Because once capability is widely diffused, the point of risk shifts from model behavior to actor intent. If AI governance focuses solely on internal safety, we risk building perfectly aligned systems that can still be misused by malicious actors.

[1] *Clarification:* Open-weight models release the trained parameters (weights) for use or fine-tuning; open-source models go further, including training code, data access, and reproducibility instructions. Both reduce central control, but open-source enables full replication, posing an even wider set of governance challenges.

AI safety addresses critical concerns such as alignment with human values, transparency, interpretability, and responsible deployment. It aims to prevent unintentional harm from systems that act unpredictably or escape control. However, in many expert conversations, “safety” is used as an umbrella term for a much broader set of risks. While that semantic breadth may convey urgency, it obscures a vital distinction between *unintended failures* (safety) and *intentional misuse* (security)

A perfectly aligned system can still produce deepfakes, amplify disinformation, assist cyberattacks, or be fine-tuned into something unrecognizable. That outer world includes adversaries, asymmetries, and actors who reject normative constraints. If governance stops at safety, we protect the instrument while ignoring the hands that use it.

The increasing advocacy, and actual decisions, to release open-weight models, often framed as transparency or accessibility measures, illustrate the point. Open-weights, when not embedded in enforceable governance architectures, are a security vulnerability. They enable replication, fine-tuning, and deployment by actors who will never submit to oversight. In this context, “openness” does not inherently build trust; it multiplies attack surfaces. Unlike alignment risks, which experts can address internally, those posed by open-weight diffusion stem from uncontrollable external use. This is why the risk goes far beyond safety.

The safety–security distinction is neither new nor arbitrary; it is both analytically and strategically critical. Experience from other high-stakes domains shows that different classes of risks require distinct strategies, regulatory frameworks, and technical countermeasures. Blending them under a single label produces mismatched solutions. For example, red-teaming for emergent behavior does nothing to prevent geopolitical escalation or state-level misuse.

Conceptually separating safety and security is not fragmentation; it is the foundation of sound governance. Treating them as analytically distinct yet interdependent allows for proportionate, anticipatory, and strategically optimized responses.

If we fail to draw a clear line between safety and security, governance will remain reactive, fixing the wrong problems while leaving the real vulnerabilities exposed. The next sections outline a framework for bridging this gap, one that treats safety and security as complementary pillars, and that equips decision-makers to address both the unpredictable behavior of systems and the predictable malice of actors.

The Danger of Open-Weight Proliferation

As recent releases of open-weight models have demonstrated, a powerful AI model is not just lines of code, it is a transferable capability. Once the full set of weights is made public, any actor with the necessary hardware and intent can operate the model independently, beyond the oversight or intervention of its creators. At that point, the system becomes sovereign, untraceable, and potentially weaponizable.

Unlike APIs or hosted platforms, which allow for monitoring, throttling, updates, and even shutdowns, open-weight models run locally. They have no scalable kill switches, no enforceable updates, and no centralized, verifiable audit trail. Once released, they cannot be recalled or technically constrained. In governance terms, this is a point of no return unless proactive and enforceable international measures exist. Releasing a model's weights is far more than an act of openness or knowledge-sharing. It is a direct transfer of operational power to an unknown and potentially unaccountable actor.[2]

While such releases are often framed as democratizing access, fostering innovation, or decentralizing technological power, they are also shaped by competitive positioning, market signaling, and reputational strategy.

Crucially, not all models pose the same degree of risk. As capabilities advance, especially in frontier-level systems that integrate reasoning, multimodality, and autonomous tool use, the security implications of uncontrolled diffusion escalate nonlinearly. A model that today enables harmless creative applications could, in its next iteration, support advanced cyber-operations, strategic deception, or automated weapons integration.

The strategic question, then, is whether such releases should ever precede the creation of robust, enforceable global governance frameworks. Framing openness as inevitable ignores the reality that security-wise, irreversibility is a liability. Once released, mitigation options shrink drastically, and the international community has yet to define credible deterrence or containment strategies for actors who will not cooperate. The evidence reinforces a hard truth: diffusion without enforceable prevention mechanism is not openness, it is exposure.

The preliminary conclusion is clear: these decisions must never be improvised or left to unilateral calculations. However, in global security, the real test is not managing those who follow the rules, it is preparing for those who never will.[3]

Strategic Lessons from the Nuclear Domain

In the AI world, there has long been a seductive optimism: that sharing knowledge accelerates innovation, that openness builds trust, that "more access" equals "more progress." But nuclear history offers a cautionary mirror. In that domain, raw knowledge was never the core danger; fissile material was. The blueprints for a bomb have existed for decades, what matters is who can acquire enough weapons-grade uranium or plutonium to build the fissile core.

For artificial intelligence, model weights are the fissile core: the concentrated output of billions of tokens, compute cycles, and optimization layers. Released without constraints, they can be repurposed into tools that serve creativity, deception, or destruction. Unlike fissile material, AI weights are non-scarce and trivially copyable; this amplifies diffusion risk and collapses traditional notions of custody.

[2] In August 2025, OpenAI released two open-weight models — gpt-oss-120B and 20B — under the Apache 2.0 license, following an internal security review. While not frontier-level, their release marks a critical shift in the operational landscape. Once weights are public, creators no longer govern what they have built, and they know it. To the publication's date, per public disclosures, all major actors, including OpenAI (GPT-4o and the recently launched GPT-5), Anthropic (Claude 3 Opus), Google (Gemini 1.5 Pro), and Meta (LLaMA 3-400B) still keep their frontier models closed, reflecting persistent concerns about the uncontrolled diffusion of highly capable systems. This reinforces the dual-risk framework proposed in this paper: the distinction between model safety and external security threats is not academic, it is operationally essential.

[3] See Arguello, I. AI and the Battle for Truth. The Emerging Strategic Role of Artificial Intelligence in Global Security. NPSGLOBAL, May. 2025.

The lesson is clear: core sources of danger must always be protected, not offered to potentially harmful actors at no cost.

Still, nuclear analogies must be applied with care. A governance model for AI must be grounded in the specific nature of the technology, the diversity of actors involved, and the cascading consequences of misuse.

The nuclear experience led to the Treaty on the Non-Proliferation of Nuclear Weapons (NPT), a regime with near-universal adherence since 1970, built on three pillars: disarmament, non-proliferation, and peaceful use. Within that framework, safety and security became operational imperatives, with verification mechanisms, mandatory or not, to ensure compliance.[4]

A third "S"—safeguards—was institutionalized in the Treaty to ensure not to increase the number of nuclear weapons possessors, through verification mechanisms in all non-nuclear-weapon states, while exclusively recognizing five legal ones.[5]

Nonetheless, it is misleading to equate AI to nuclear weapons. A balanced approach must recognize that when a capability becomes powerful enough to alter global stability, the international community must define governance thresholds before the situation turns out to be irreversible.

Unlike the nuclear regime, however, AI governance currently lacks both universal adherence and enforceable thresholds, despite the comparable scale of potential disruption. This asymmetry should be addressed before diffusion becomes irreversible.

As happened with the proliferation of nuclear weapons, we must now ask: What happens when advanced AI capabilities become widely and irreversibly accessible, even in a non-frontier form? And concerning this question, we should rephrase it in the present rather than in the future tense.

Trust is not enough. Global problems require global approaches, grounded in institutional foresight and coordinated action.

The nuclear analogy offers a warning, not a template. AI has its own logic, pace, and actors. Its governance must be built for this century’s realities, not last century’s solutions.

In other words, the first step to governing AI is to stop treating it like yesterday’s weapons and start recognizing it as tomorrow’s battlefield. [6]

[4] Nuclear safety and security obligations are governed by two distinct legal instruments: the *Convention on Nuclear Safety* (CNS), which had 98 parties as of 2025, and the *Amended Convention on the Physical Protection of Nuclear Material* (A/CPPNM), which had 138 parties. The CNS focuses on reactor safety, while the A/CPPNM extends protection to nuclear materials and facilities against theft, sabotage, and other malicious acts.

[5] Nuclear-armed states include: (1) the five legally recognized nuclear-weapon states under the NPT, the United States, Russia, France, the United Kingdom, and China; (2) those states outside the NPT, India, Pakistan, and Israel, which never joined the NPT; and (3) North Korea, which withdrew in 2003.

[6] While nuclear weapons have been historically state-driven, the development of advanced AI, at least in free democratic societies, is mainly led by private companies with global reach. This difference adds an extra layer of complexity to global AI governance, diplomacy, and negotiation, due to a shift in the axis of power.

The Missing Dimension: AI Security

In a world where power moves faster than regulation, any serious attempt at global AI governance must incorporate a dedicated security layer to form a true dual-risk response to threats.[7]

In practical terms, this security layer must confront five core challenges:

- Proliferation: How do we prevent the uncontrolled spread of powerful models?
- Attribution: How do we know who is using which model, and for what purpose?
- Manipulation: Can we detect if a model has been fine-tuned for harmful goals?
- Retaliation: What happens when states accuse each other of AI-enabled attacks?
- Fail-safes: Is there a digital equivalent of a nuclear “dead man’s switch”?

Responding to these challenges requires operational and enforceable mechanisms, beyond aspirational principles:

- Traceable model architectures rather than mere terms of service.
- Licensing regimes that treat advanced digital capabilities as geopolitical assets.
- Robust technical tools for watermarking, tamper detection, and real-time logging.
- International norms that blend code, law, and diplomacy into enforceable standards.
- And above all, a shift in mindset, from designing for trust to designing for resilience against misuse, anchored in global foresight, not just local ethics.

Artificial intelligence is inherently a dual-use capability and therefore part of critical infrastructure. Governance must evolve accordingly. This is not just a matter of design. It is a matter of doctrine. [8]

In contrast to nuclear governance, where safety and security are institutionally and concpetually separated, AI governance must acknowledge and address their entanglement. In AI, the model’s internal design (safety) and the context of its use (security) are dynamically interdependent. Mitigating one without the other creates systemic exposure.

A system can be technically safe and still be repurposed for harm. Conversely, a security breach can trigger safety failures. The boundary between internal malfunction and external manipulation is porous. For example, during red-team testing, models can be induced to produce harmful outputs not because they are misaligned, but because prompts are engineered to bypass safeguards. In such cases, the system performs “as designed” yet the output still can cause damage. This diffuse border makes AI governance uniquely complex. Effective oversight must treat safety and security not as parallel tracks, but as intertwined vectors. The architecture of the model and the conditions of its use co-produce risk. Addressing only one dimension leaves exploitable gaps that can outpace regulation.

[7] The AI safety layer is only slightly outlined in this paper as it has already received much more conceptual attention from the expert community.

[8] Several recent contributions (e.g., Lin et al. 2025; Tallam 2025; Wybitul 2025) began addressing the distinction between AI safety and security from technical or operational perspectives. Yet the field still lacks a strategic governance doctrine to guide implementation across asymmetric geopolitical realities. This paper aims to help close that gap.

What about the proliferation of advanced models? Should AI require a dedicated proliferation layer, as in the nuclear realm?

While the concept of "proliferation" plays a central role in the governance of nuclear weapons—where it is defined, tracked, and sanctioned under international law, AI "model proliferation" is better understood as a pathway through which security threats materialize, rather than as a standalone legal category. This implies a different structure of risk. The diffusion of powerful models, though alarming, does not yet constitute a violation of legal commitments.

Therefore, AI governance should avoid analogical overreach. Instead of replicating nuclear safeguards, it should integrate the risks of open-weight diffusion, covert replication, and careless releases into a coherent doctrine of AI security, one that is technically enforceable, politically equitable, and ethically resilient.

The uncontrolled diffusion of dual-use systems empowers adversaries, erodes accountability, and undermines enforceability on a global scale. Borrowing the nuclear safeguards model risks importing its structural asymmetries, which privileged some powers while constraining others.

AI governance must aim higher. The current vacuum should not be filled by architectures grounded in privilege, but by a purpose-built framework that reflects the technology's unique nature and global reach.

Finally, although this paper focuses on safety and security, a third, quieter threat deserves attention: the erosion of human judgment. This is not caused by rogue models or malicious actors, but by overreliance, especially among younger generations. What fades is not safety, but agency: critical thinking, discernment, autonomy. This is not a technical failure. It is a civilizational challenge, and perhaps the most enduring risk of all.

Toward a Security-Oriented Governance Layer

What would a security-oriented governance layer look like in practice? It must be designed to detect, deter, and respond to adversarial use, malicious fine-tuning, and uncontrolled diffusion. The focus must shift from internal system behavior to external misuse, from how AI works to how it can be weaponized.

Key priorities for this layer include:

- Strict monitoring and conditional deployment of frontier models in sensitive or high-risk domains.
- Embedding enforceable, non-optional constraints into licensing regimes and API access for dual-use systems.
- Developing integrated early-warning systems for AI misuse in cyberwarfare, disinformation campaigns, or WMD-related contexts.
- Establishing clear accountability protocols for actors who release open-weight or otherwise untraceable models.
- Promoting proactive defensive innovation, including watermarking, tamper detection, and multi-stakeholder rapid-response networks.

This is a call for structured vigilance, where AI power is treated with the same strategic foresight and operational discipline we apply to other technologies of consequence. Without this layer, governance remains unbalanced, incomplete, and strategically blind.

Yet there is another, deeper lesson from the nuclear domain. Despite decades of effort, no binding international convention on nuclear security has ever been achieved. States have long claimed that such matters fall under sovereign prerogative. But this is a dangerous fallacy: when the consequences of your acts or omissions can cause cross-border harm, sovereignty must yield to shared responsibility.[9]

As AI power spreads, we cannot repeat that mistake. Security must not remain a matter of domestic discretion, in a world of systemic interdependence. The governance architecture now taking shape must reflect a new diplomatic reality—one where states, companies, technologists, and civil actors jointly design and commit to enforceable norms, as a condition for systemic stability.

Conclusion: A Doctrine for Reality

Artificial intelligence is a transformative force, a dual-use capability whose reach extends into every domain of power. Its global governance must rest not only on ethical alignment, but on strategic foresight, anchored in reality. A dual-risk doctrine is essential: one that safeguards the ethical imperatives of safety while embracing geopolitical realism and operational demands of security.

This means building systems that resist theft, embedding rules that compel accountability, and crafting frameworks that anticipate not only system failure but deliberate human misuse. It requires a comprehensive, multi-layered approach to the full spectrum of AI-related risks.

Intentions must evolve into doctrine, and doctrine into enforceable architecture. Concepts must become action. And this responsibility cannot rest solely on governments. The AI community must lead not just in innovation, but in responsibility. We are entering a new world, one of intensifying geopolitical tensions, unprecedented cognitive architectures, expanding ethical horizons, and the urgent need for new forms of diplomacy and collective restraint.

If safety is the conscience and security is the strategy, then only together can they hold the line between progress and peril. That line will not hold by accident; it will hold because humans of goodwill, strategic clarity, and lucid purpose make it so, and refuse to let it be crossed.

Irma Argüello is the founder and CEO of the NPSGLOBAL Foundation.
She is a global strategist and ethical architect for AI futures.
@irmaar

[9] The author has been actively involved in high-level discussions on nuclear security as a member of specialized expert groups, including the Nuclear Security Governance Experts Group (NSGEG), the Fissile Materials Working Group, the World Economic Forum's Nuclear Security Council, and the International Atomic Energy Agency (IAEA), where the need for a global nuclear security convention has been extensively debated.

About the NPSGLOBAL Foundation

For 18 years, the NPSGLOBAL Foundation, an independent, private non-profit based in Buenos Aires, has served as the leading Latin American focal point on global security, with strong engagement in international disarmament, non-proliferation, and risk reduction. Its track record includes strategic advising, high-level education, and cooperation with governments, multilateral bodies, and global think tanks.

Today, NPSGLOBAL is expanding its work into emerging fields, including the strategic implications of artificial intelligence and new technologies. Through research, policy engagement, and education, the Foundation aims to foster a responsible and forward-looking global agenda for AI governance and security.

Carlos Pellegrini 1063 – 10 – Buenos Aires
www.npsglobal.org / info@npsglobal.org