TOWARD ALGORITHMIC DISARMAMENT: A SOCIO-TECHNICAL FRAMEWORK FOR AI NON-PROLIFERATION AND LEGAL VERIFICATION

Xinran Li,* Yudan Zhu & Lingjie Lv
School of Law, Tsinghua University
Beijing, China
{lixr24,zhuyd25,lvlj25}@mails.tsinghua.edu.cn

ABSTRACT

The rapid penetration of foundation models into strategic decision-making and Lethal Autonomous Weapons Systems (LAWS) marks a paradigm shift, rendering the traditional kinetic weapon-centric disarmament frameworks (e.g., the Treaty on the Non-Proliferation of Nuclear Weapons) structurally ineffective in the digital age.This paper proposes a comprehensive Algorithmic Disarmament framework, advocating for a shift in regulatory granularity from physical hardware to model weights and training provenance. We construct a quantitative risk function: $R = \alpha \cdot \log_{10}(C) + \beta \cdot D_{\text{sens}} + \gamma \cdot A_{\text{ind}}$, which classifies high-risk AI assets into legal tiers based on computational scale (C), data sensitivity (D_{sens}), and autonomy index (A_{ind}), thereby defining the legal boundaries of algorithmic weapons. To resolve the inherent tension between state sovereign secrets and international regulation, this paper demonstrates the technical feasibility of non-intrusive auditing using Zero-Knowledge Proofs (ZKP) and cryptographic watermarking, enabling international monitoring bodies to verify whether models comply with humanitarian constraints without infringing on intellectual property rights or national security classified information. By evolving the Martens Clause into a socio-technical norm for the algorithmic era, this research conceptualizes the operational architecture of an International Artificial Intelligence Oversight Organization (IAIO), providing a robust legal and technical roadmap for curbing the unregulated AI arms race and safeguarding global peace.

Keywords: Algorithmic Disarmament, AI Non-Proliferation, Zero-Knowledge Proofs, Quantitative Risk Model, International Humanitarian Law

Track: Tiny papers

1 INTRODUCTION

In 1945, the physics community experienced an irreversible "Oppenheimer Moment", where scientific achievements were directly translated into a devastating force that reshaped the geopolitical landscape through the mass-energy equivalence formula $E = mc^2$(1). The ensuing Cold War spurred the international community to build a disarmament legal framework centered on the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) and the Comprehensive Nuclear-Test-Ban Treaty (CTBT). The logic underpinning the success of this framework rests on the detectability of physical substances—uranium enrichment, ballistic missile tests, and the scale of silos are all observable, auditable physical entities amenable to legal constraints(2; 3; 4).

Yet the international security landscape in 2026 stands at another far more covert and profound turning point: the Algorithmic Moment. Unlike nuclear disarmament, the core carrier of AI weaponization is model weights(5)—these files composed of massive floating-point numbers are compact, easily concealed, and infinitely reproducible at nearly zero cost. When devastating force is hidden in hundreds of gigabytes of binary code, the traditional disarmament logic based on physical

*Corresponding author.

asset verification becomes completely invalid. Under the conventional arms control framework, the performance boundaries of weapons are predictable upon manufacture and field testing. However, the "black box" nature and emergence of deep neural networks (DNNs) have shattered this predictability(6). A foundation model trained on civilian datasets can rapidly acquire the ability to identify unauthorized strike targets after just a few hours of fine-tuning with specific military data(7).

To address these practical challenges, this paper proposes the core concept of "Algorithmic Disarmament" and constructs a comprehensive socio-technical framework. The core contributions of this study are threefold: first, designing a quantitative risk model that clearly defines the legal boundary of "algorithmic weapons" by shifting regulatory focus from subjective intent to objective capability; second, developing a verification scheme based on Zero-Knowledge Proofs (ZKP)(8; 9) and related technologies to enable remote, non-intrusive legal auditing of black-box models while safeguarding state secrets and enterprise intellectual property rights; third, outlining a feasible institutional framework for the International AI Oversight Organization (IAIO) to provide practical support for global algorithmic disarmament governance.

2 Related Work

Existing research on AI non-proliferation and algorithmic regulation can be divided into three categories, each with obvious limitations that this paper aims to address. First, legal research on AI disarmament mainly focuses on revising traditional international humanitarian law and formulating ethical guidelines, but lacks technical operability–most studies only propose conceptual appeals for "algorithmic compliance" without providing feasible verification methods that balance national secrets and regulation(10; 11). Second, technical research on AI auditing mostly relies on intrusive methods such as source code inspection and model disassembly, which seriously infringe on national sovereign secrets and intellectual property rights, making it difficult to be accepted by the international community(12; 13; 14). Third, existing risk assessment models for AI weapons are mostly qualitative, relying on subjective judgments of experts to classify risks(6; 15), which lack consistency and legal enforceability due to the lack of quantitative indicators and clear parameter settings(16). In summary, there is a lack of an integrated socio-technical framework that combines legal norms, quantitative risk assessment, and non-intrusive technical verification in current research, which is exactly the gap that this paper intends to fill.

3 Quantitative Risk Model and Algorithmic Classification

Traditional arms control law has long relied on the determination of "intent"—namely, whether a given technology is designed exclusively for military purposes. Yet in the era of foundation models, this logic has completely collapsed. A general reinforcement learning agent developed to optimize logistics routes can be transformed into a highly efficient ammunition allocation and target selection system with merely a small amount of fine-tuning. Therefore, this paper proposes a quantitative risk model to convert the subjective and qualitative risk judgment of AI weapons into objective and quantitative indicators, thereby clarifying the legal boundary of "algorithmic weapons" and providing a basis for legal regulation and technical verification.

The model takes the comprehensive risk value R of AI models as the core output, and its formal definition can be expressed as equation (1):

$$R = \alpha \cdot \log_{10}(C) + \beta \cdot D_{\text{sens}} + \gamma \cdot A_{\text{ind}}, \tag{1}$$

where α, β, and γ are weight coefficients (sum to 1) determined by the Analytic Hierarchy Process (AHP), used to balance the influence of the three core indicators on the overall risk.

Based on the scores of the risk function R, this paper suggests classifying artificial intelligence algorithms into three jurisprudential tiers, with corresponding obligations matched to each tier. Table 1 shows the designed tiers.

3.1 Weight Determination Process

First, we establish a hierarchical structure model based on AHP, taking "AI military risk" as the target layer, "computational scale", "data sensitivity", and "algorithmic autonomy" as the criterion layer.

Table 1: Three Jurisprudential Tiers and Corresponding Regulatory Obligations.

Tier R Value	Range	Legal Attribute	Regulatory Obligations
Tier 1: General Civilian Grade	$R < \tau_1$	Exempt Asset	Only required to comply with conventional human rights law and data protection law.
Tier 2: Restricted Dual-Use Grade	$\tau_1 \leq R < \tau_2$	Monitored Asset	Must complete algorithm registration, compulsorily embed provenance watermarks, and accept regular non-intrusive audits.
Tier 3: Prohibited Grade	$R \geq \tau_2$	Controlled Asset	Illegal cross-border proliferation is prohibited; physical connection to lethal strike links is strictly forbidden; real-time on-site monitoring by the IAIO is required.

Second, we design a pairwise comparison matrix, inviting 15 interdisciplinary experts (AI technology, international law, global security) to score the relative importance of each pair of indicators in the criterion layer (using a 1-9 scale recommended by AHP, where 1 means equal importance and 9 means extreme importance). Third, we conduct consistency testing on the pairwise comparison matrix (requiring the consistency ratio CR < 0.1 to ensure the rationality of expert scores); for matrices that do not pass the consistency test, we feed back the results to the corresponding experts for re-scoring and adjustment until all matrices pass the test. Finally, we calculate the weight vector of the criterion layer through hierarchical single sorting and total sorting, obtaining $\alpha = 0.35$, $\beta = 0.4$, $\gamma = 0.25$ (see section 3.2 for more details), which are verified by all participating experts to ensure alignment with technical principles and legal practice.

3.2 Parameter Explanation

Computational Scale Factor (C): C refers to the total computing power consumed in AI model training, measured in FLOPs. Its setting is based on the "Scaling Laws" of foundation models, which confirm that there is a strong positive correlation between computing power consumption and model reasoning ability, emergence, and transferability—models exceeding a specific computing power threshold have extremely strong migration capabilities and can be quickly adapted to military scenarios. $\log_{10}(C)$ is used to mitigate the impact of extreme values, ensuring that the contribution of computing power to risk is within a reasonable range.

The weight $\alpha = 0.35$ reflects that computing power is the physical foundation of algorithmic risk but not the sole determinant, which is consistent with the consensus of AI technology experts that "computing power is a necessary but not sufficient condition for AI weaponization".

Data Sensitivity Factor (D_{sens}): D_{sens} evaluates the density of sensitive information contained in the model training dataset, which is calculated as the weighted sum of normalized densities of different levels of sensitive data:

$$D_{\text{sens}} = \sum_{i=0}^{4} w_i \cdot D_i, \tag{2}$$

where D_i is the proportion of the i-th level sensitive data in the training corpus (normalized to 0-1), and w_i is the threat multiplier coefficient (set based on international humanitarian law and global security standards). Sensitive data is divided into 5 levels: Level 0 (general public domain, $w_0 = 1$), Level 1 (PII & commercial confidential, $w_1 = 10$), Level 2 (critical infrastructure & dual-use, $w_2 = 100$), Level 3 (tactical military & intelligence, $w_3 = 1000$), Level 4 (WMD & existential threat, $w_4 = 100000$).

The weight $\beta = 0.4$ is due to the consensus of interdisciplinary experts that data is the core driver of AI weaponization—sensitive training data directly determines the model's military application potential, which is supported by cases of AI models being weaponized through fine-tuning with military data.

Algorithm Autonomy Index (A_{ind}): A_{ind} is a normalized index (0-1) that measures the degree of algorithmic intervention in the decision-making loop and the separation from human ethical judgment, calculated by three orthogonal parameters:

$$A_{\text{ind}} = 0.45 \cdot I_{\text{velocity}} + 0.35 \cdot I_{\text{opacity}} + 0.20 \cdot I_{\text{actuation}}, \tag{3}$$

where I_{velocity} is the decision speed, I_{opacity} is the model's unexplainability, and $I_{\text{actuation}}$ is the degree of coupling between the algorithm and physical execution. According to the value of A_{ind}, AI models are classified into four levels: Compliant (< 0.4), Restricted (0.4-0.6), High-Risk (0.6-0.8), and Prohibited (≥ 0.8). The allocation logic of weight coefficients in the formula is based on the Analytic Hierarchy Process (AHP) and strictly follows the risk hierarchy in cybernetics. Among them, the decision-making velocity (I_{velocity}) is assigned the highest weight coefficient of 0.45, which is mainly based on the "Time Compression" theory. When the OODA loop cycle of the algorithm is lower than the limit of human neurophysiological response, human operators are physically excluded from the decision-making closed loop, leading to a substantive loss of control. Secondly, the weight of epistemic opacity (I_{opacity}) is set to 0.35, aiming to quantify the unexplainability of the model. If the commander cannot parse the reasoning path of the deep neural network, their approval of the decision will lack a cognitive basis, thereby violating the principle of distinction in International Humanitarian Law (IHL). Finally, as an engineering parameter measuring the system's authority to trigger physical strikes, the actuation coupling degree ($I_{\text{actuation}}$) is assigned a basic weight of 0.20 because it can be relatively easily externally constrained by means such as physical safety interlocks.

The weight coefficient γ is set to 0.25, considering that autonomy determines the degree of loss of human control over the algorithm.

To ensure the legal enforceability of the model, two dynamic thresholds are set: τ_1 (mandatory disclosure threshold, 10^{25} FLOPs) and τ_2 (prohibition threshold, 10^{26} FLOPs), which are revised biennially by an international expert committee (including the 15 experts participating in weight determination) to adapt to technological progress.

4 Non-Intrusive Technology

Algorithmic disarmament's core contradiction is the conflict between national sovereign secrets, IP protection and international regulatory needs, rendering intrusive verification unfeasible internationally. This paper proposes a ZKP-based non-intrusive verification scheme, centered on zk-SNARKs to resolve the "transparency paradox" by proving black-box model harmlessness without disclosing internal details. The scheme has four stages: commitment (submitting tamper-proof model weight hash values), constraint formulation (IAIO setting legal operators), proof generation (local calculation without core parameter uploads), and verification (millisecond-level validation by IAIO). Two supporting technologies enhance reliability: fine-tuning-robust cryptographic watermarking for post-conflict digital forensics, and hardware-level TEEs for real-time remote monitoring of model training compliance via isolated enclave data recording.

5 International AI Oversight Organization (IAIO) and Global Governance

Algorithmic disarmament cannot succeed with mathematical formulas alone; it also needs a legitimate, independent and enforceable international body. Thus, we propose establishing the International AI Oversight Organization (IAIO) under the UN, modeled after the IAEA, whose core mission is to ensure the peaceful use of AI and curb its militarization, operating through integrated safeguards, technical cooperation and unified standards to fulfill this mandate.

6 Conclusion and Vision

This paper proposes a designed quantitative risk model that realizes the quantitative classification of AI weapon risks and provides a legal basis for algorithmic regulation; a non-intrusive verification scheme based on ZKP and supporting technologies that effectively balances national sovereign secrets, intellectual property protection, and international regulatory needs; and an IAIO institutional framework that provides a practical roadmap for global AI non-proliferation governance. This paper aims to promote the international community to reach a consensus on algorithmic disarmament, establish a fair and equitable global AI governance system, ensure that humans retain the right to control violence and peace in the algorithmic era, and let AI become a powerful tool for promoting global peace and sustainable development rather than an engine of conflict and destruction.

REFERENCES

Albert Einstein. Does the inertia of a body depend upon its energy-content? *Annalen der Physik*, 18:639–641, 1905.

James Johnson. *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age*. Oxford University Press, 2023.

Paul Scharre and Megan Lamberth. *Artificial Intelligence and Arms Control*. Center for a New American Security (CNAS), 2022.

Matthijs M. Maas. *International Law and the Governance of Artificial Intelligence*. Springer, 2023.

David M. Allison and Stephen Herzog. Artificial Intelligence and nuclear weapons proliferation: The technological arms race for (in) visibility. *Risk Analysis*, 45:3839–3859, 2025.

Scott Sullivan and Iben Ricket. Targeting in the Black Box. In *Proceedings of the International Conference on Cyber Conflict* (CyCon). IEEE, 2024.

National Telecommunications and Information Administration. *Dual-Use Foundation Models with Widely Available Model Weights*. U.S. Department of Commerce, 2024.

Zhibo Xing, Zijian Zhang, Ziang Zhang, Zhen Li, *et al*. Zero-Knowledge Proof-Based Verifiable Decentralized Machine Learning in Communication Network: A Comprehensive Survey. *IEEE Communications Surveys Tutorials*, 28:985-1024, 2026.

Seunghwa Lee, Hankyung Ko, Jihye Kim, and Hyunok Oh. vCNN: Verifiable Convolutional Neural Network Based on zk-SNARKs. *IEEE Transactions on Dependable and Secure Computing*, 21:4254-4270, 2024.

Jürgen Altmann and Frank Sauer. Autonomous weapon systems and strategic stability. *Survival*, 59:117–142, 2017.

Nicholas Emery-Xu, Richard Jordan, and Robert Trager. International governance of advancing artificial intelligence. *AI Society*. Springer, 2024.

Jakob Mökander, Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*. Springer, 2021.

Sasha Costanza-Chock, Emma Harvey, Inioluwa Deborah Raji, Martha Czernuszenko, *et al*. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT). ACM, 2022.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, *et al*. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAccT). ACM, 2020.

Malcolm Murray, Henry Papadatos, Otter Quarks, *et al*. Mapping AI Benchmark Data to Quantitative Risk Estimates Through Expert Elicitation. *arXiv preprint arXiv:2503.04299*, 2025.

Margot E. Kaminski. Regulating the Risks of AI. *Boston University Law Review*, 102:1-72, 2022.