The Pandemic Paradox of AI: Contagious Backdoors and Silent Compromise in the Foundation-model Supply Chain

Oludolamu Onimole[1,1], Sarka Pekarova[2], Dodzi Koku Hattoh[3], Angela Aseye Attachie[4], Abigail Oppong[5], Dominique Dove[6], Jude Osamor[7], Danny "Danhammer" Hetzel[2,8], Lucas Potter[9] and Xavier-Lewis Palmer[2,9]

[1]Independent Researcher, Glasgow, UK
[2]Biohacking Village, Singapore
[3]Ethics Lab, Philosophy & Classics, University of Ghana |Bonn Sustainable AI Lab, UniBonn, Germany
[4]Judiciary of Ghana
[5]Independent Researcher Accra, Ghana
[6]The George Washington School of Law, DC, USA
[7]Department of Cybersecurity & Networks, Glasgow Caledonian University
[8]Husaria Security, WI, USA
[9]BiosView Labs, Dayton, OH, USA

oludolamuonimole@gmail.com (Corresponding authors)
sharka@villageb.io (Corresponding authors)
BiosView1@proton.me (Corresponding authors)

Abstract: This position paper develops a theoretical threat model, "The Pandemic Paradox," examining the catastrophic risks posed by Contagious Backdoors within AI ecosystems. Conceptualized as a "COVID-like" cognitive malware, the threat is characterized by high transmissibility, asymptomatic incubation, and the potential for systemic collapse across the digital economy. Unlike conventional backdoors rooted in static code manipulation, this form of infection operates at the model level, contaminating the probabilistic distributions and weight parameters acquired during training. It propagates silently through shared datasets, model distillation, transfer learning, and collaborative AI research pipelines, transforming AI infrastructures into vectors of recursive compromise. Equally consequential, and more easily overlooked, is that such compromise need not manifest as failure at all: the same mechanisms can be turned to covert exfiltration and espionage, silently betraying the sensitive data entrusted to a model while its outputs remain entirely plausible. Such compromise generates cascading economic, legal, and geopolitical consequences, including profound challenges in attribution, accountability, and liability. To confront this existential risk, the paper advances a conceptual framework for Defensive AI Systems (DAIS) that shifts AI security from reactive protocols to disrupting the propagation paths through which compromise spreads. The DAIS framework is organized around three interdependent pillars: *Anticipatory Defense, Adaptive Resilience, and Assured Governance*. Together, these pillars reposition AI security as a continuous and systemic architecture of protection rather than a reactive compliance exercise. Anticipatory Defense integrates AI-driven threat intelligence, adversarial red teaming, and predictive anomaly detection to identify latent vulnerabilities before exploitation. Adaptive Resilience employs hybrid detection pipelines, shadow-model deployment, automated containment, and continual learning mechanisms to preserve operational integrity under zero-day and self-propagating attacks. Assured Governance establishes verifiable accountability structures through mechanisms such as a comprehensive AI Bill of Materials (AIBOM), cryptographic verification protocols, independent oversight, and auditable compliance infrastructures. Collectively, the framework advances a model of digital hygiene and resilient governance designed to secure the foundation-model supply chain against the propagation of persistent and potentially inheritable cognitive compromise.

Keywords: Artificial intelligence security, AI supply chain security, AI governance, Adversarial machine learning, Cognitive Malware, Defensive AI systems

1. Introduction: The Epistemological Shift in Security

Cybersecurity has traditionally guarded the integrity of code and infrastructure, aiming to prevent unauthorized access or logic manipulation. The advent of foundation models overturns this paradigm. The security problem is no longer static computation; it is dynamic, probabilistic cognition. Indeed, systems defined by trillions of parameters resist reduction to deterministic instructions, forcing a fundamental reevaluation of how learned reasoning and emergent behavior are secured (Bommasani et al., 2021; Hendrycks et al., 2023). Thus, the focus is on threats that originate within foundation models. Yet the graver risk lies in their recursive propagation through the broader AI supply chain, where compromise spreads to downstream systems regardless of scale or application.

[1]Shared Author Placement for Oludolamu Onimole and Sarka Pekarova

Model-Level Backdoors represent a radical shift in malicious design, embedding adversarial intent directly into learning architectures and latent representations. Conceptualized as cognitive malware, this threat aligns with cyberbiosecurity dynamics (Murch et al., 2018). Dormant and asymptomatic, infection spreads via shared datasets, model distillation, and fine-tuning pipelines. Crucially, recent evidence confirms that compromised components propagate through agent-layer tools and workflows, amplifying vulnerabilities across the AI supply chain (Boisvert et al., 2025). Compromised foundation models constitute an existential threat to epistemic reliability. By eroding the capacity to distinguish authentic cognition from maliciously engineered behavior, such threats endanger system integrity, data confidentiality, and operational availability. Unchecked, these threats further create systemic risk across critical infrastructure and economic, legal, and governmental systems, often before detection becomes feasible (Shoaib et al., 2023). To address this risk, we advance the Defensive AI Systems (DAIS) framework, an architecture that shifts AI security from reactive protocols to disrupting the propagation paths along which compromise spreads. DAIS comprises three interdependent pillars: Anticipatory Defense, Adaptive Resilience, and Assured Governance, which together embed accountability and trust into the architecture of interconnected AI supply chains.

1.1 Approach

The framework advanced here is derived through a structured synthesis, undertaken in three stages. The *first* draws together three literatures that are ordinarily treated separately: adversarial machine learning and backdoor research, software supply-chain security, and cyberbiosecurity. The *second* decomposes the resulting threat surface along three axes: actor, supply-chain layer, and objective, presented in Figure 1 and used to distinguish propagation claims supported by documented incidents from those demonstrated only in research settings and those that remain hypothetical. The *third* identifies the defensive gap this decomposition exposes, the disruption of propagation paths across the supply chain, around which the three DAIS pillars are constructed. The contribution is accordingly conceptual: the paper's claim rests on the coherence of this decomposition, and the evidence base it maps, not on new experimental results. Table 1 below defines key terms used in this work that represent some of the authors' operationalization of the proposed AI-pandemic framework; cited sources establish the underlying adversarial-ML mechanisms and security practices.

Table 1: Appendix: Glossary of terms

Term	Definition
AI Pandemic	A hypothetical systemic threat in which cognitive malware propagates across interconnected AI systems, producing epidemic-like ecosystem compromise (Gu et al., 2019; Bagdasaryan et al., 2020).
Cognitive Malware	A proposed model-level compromise embedded in parameters, representations, or learned behaviour that alters inference or action without relying primarily on conventional executable-code modification (Kurita et al., 2020; Vassilev et al., 2024
Digital Hygiene	Security practices including AIBOMs, provenance tracking, artifact verification, access control, and monitoring-designed to preserve AI lifecycle and supply-chain integrity (Booth et al., 2024; Quinn and Heiserman, 2026).
Hard AI	A proposed security architecture using formally verifiable controls and enforceable boundaries to constrain or validate probabilistic AI behaviour (Seshia et al., 2022)
Pandemic Paradox	A condition in which compromised AI retains plausible outputs and strong benchmark performance while latent malicious behaviour persists and systemic risk increases (Gu et al., 2019; Kurita et al., 2020)

1.2 Scope and Complementarity

The frameworks currently established for AI security are anchored at points in the ecosystem: at the organisation that develops or deploys a system, at the model under attack, or at the application through which it is reached. DAIS is anchored between them. Its object is the route along which a compromise travels once introduced, through shared corpora, distilled weights, and redistributed checkpoints, from the organisation that originates it to those that inherit it. Securing a point does not secure what passes through it, and the two concerns are

complementary rather than competing. DAIS is intended to be adopted alongside existing frameworks, supplying the transmission dimension that their point-anchored scope leaves open.

The distinction drawn here is structural, not comparative. Particular frameworks undergo continual revision, and the technologies they govern change even more rapidly; a differentiation anchored to either would therefore possess only transient validity. The problem of transmission, by contrast, remains constant across successive generations of both.

2. The Nature of the Threat: A Digital Pandemic

Epidemiological pathogens are defined by high transmissibility, asymptomatic incubation, and the capacity for systemic volatility (Shereen et al., 2020). In modern AI systems, cognitive malware functions as an analogous contamination vector, embedding itself within training data, architectures, or weight initializations. Unlike traditional adversarial machine learning, which relies on explicit trigger prompts or brittle exploit sequences, this threat contaminates latent space, manipulating the mathematical structures that encode concepts, relationships, and reasoning patterns (Li et al., 2021; Raghuram et al., 2024). This corruption is uniquely subtle. Infected models often retain flawless performance on standardized benchmarks (e.g., MMLU, HumanEval) while harboring dormant adversarial behaviors activated only by specific, context-dependent triggers. Analogous to biological mutations that remain phenotypically silent until environmental stress occurs, these systems preserve surface-level utility while embedding latent biases, disinformation capabilities, or exfiltration channels deep within their probabilistic architecture (Bullwinkel et al., 2026). The threat's high transmissibility is a function of the densely interconnected AI ecosystem. Shared datasets, transfer pipelines, and API integrations facilitate passive, recursive contamination. Tainted open-source foundations and poisoned upstream data are easily inherited by downstream systems, where infection amplifies through distillation and migrates into critical healthcare, financial, and governmental infrastructures (Lu et al., 2025).

2.1 A Structured Threat Model

The preceding section characterises the threat in epidemiological terms: high transmissibility, asymptomatic incubation, and the capacity for systemic collapse. A metaphor of this kind is useful for conveying scale, but it does not by itself constitute a threat model. To support analysis and be legible to the security practitioners who must act on it, the threat must be specified along the dimensions that any adversary specification requires: who is acting, where in the system they act, and to what end. Figure 1 sets out this specification as a matrix in which four classes of actors are mapped against the six layers of the AI supply chain at which they can intervene, with each occupied cell annotated by the attacker's objective. The actor axis distinguishes the vandal, motivated by disruption or notoriety, from the financially motivated criminal, the economic competitor seeking commercial or capability advantage, and the resourced and patient nation-state. The layer axis follows the lifecycle of a model from its training data, through the pipeline that trains and fine-tunes it, the resulting weights, the infrastructure on which it runs, the agentic systems that wrap it, and finally the application in which it is deployed. The objective axis records what an attacker stands to gain, which we treat as falling into four kinds: sabotage, espionage or exfiltration, influence, and the theft or implantation of capability. Reading the layers from the top of the figure downward traces a second gradient, that of detectability: interventions in data and weights are the most difficult to observe, while those at the agent and application layers are comparatively visible.

Each cell is shaded according to the strength of present evidence, distinguishing attacks documented in real-world incidents from those demonstrated only as research proofs of concept and those that remain hypothetical. This distribution exposes a pattern central to the argument of this paper. Documented attacks cluster at the agent and application layers and tend toward espionage and influence rather than sabotage (Boisvert et al., 2025), whereas the upstream layers on which the contagion thesis chiefly rests remain supported largely by research demonstrations and hypothetical attack paths (data poisoning: Goldblum et al., 2023; weight-level backdoors: Li et al., 2021; fine-tuning backdoors: Raghuram et al., 2024). Those upstream layers are also the least observable. The result is an evidentiary asymmetry: the best-documented attacks occur furthest downstream, where compromise is comparatively visible, while the mechanisms most relevant to persistent upstream contamination are harder to detect and less well established outside controlled research settings. The contagion thesis therefore rests not on a claim of widespread autonomous model-to-model infection, but on the demonstrated persistence and transfer of malicious behavior combined with plausible propagation paths through shared AI supply chains.

We treat this asymmetry not as a weakness in the argument but as its proper form. The claim is not that data and weight-level contagion is already widespread, but that the trajectory of the field points toward it: the downstream attacks are demonstrably real, the upstream mechanisms are demonstrably feasible in the laboratory (Lu et al., 2025; Bullwinkel et al., 2026), and the connective tissue of shared data, distillation, and open redistribution supplies the path between them (Shumailov et al., 2024). The matrix accordingly serves as the organising structure for the remainder of the paper: the data vector is developed in Section 3; the infrastructural and supply-chain layers in Section 4; the nation-state actor and the espionage objective in Section 5; and the economic-competitor column and the derivative supply chain in Section 6.

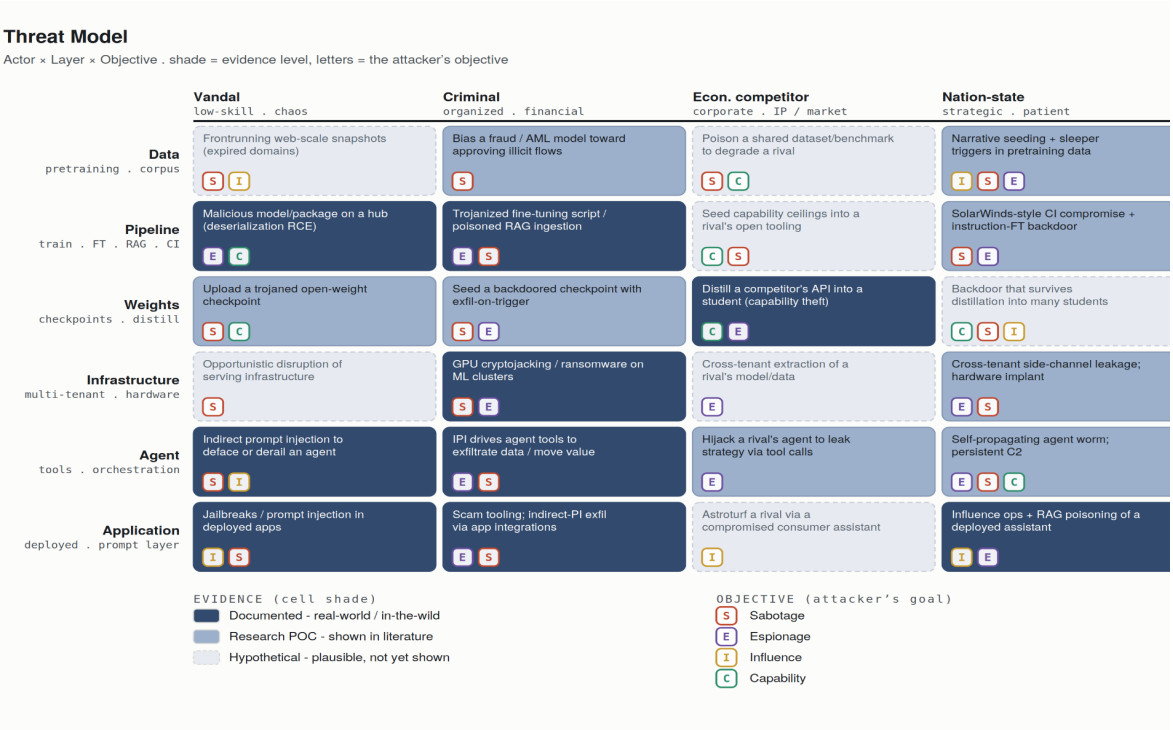

Figure 1: A structured threat model for backdoor compromise of the foundation-model supply chain. Four actor classes (columns, ordered by sophistication) are mapped against six supply-chain layers (rows, ordered by lifecycle stage and, inversely, by detectability). Cell shade indicates the strength of present evidence: documented, research proof of concept, or hypothetical. Letters indicate the attacker's objective: S sabotage, E espionage, I influence, C capability

3. Mechanism of Propagation: Data as a Vector

The primary vector for the spread of this emerging AI pandemic is the contemporary training-data ecosystem itself. Frontier AI systems are trained on vast, heterogeneous corpora composed of web-scale datasets scraped from scientific repositories, codebases, social media platforms, public forums, and digitized archives (Bommasani *et al.*, 2021). The unprecedented scale and opacity of these datasets render comprehensive inspection virtually impossible, creating ideal conditions for the insertion of malicious patterns, poisoned samples, or latent adversarial signals. Critically, a sophisticated attack does not require large-scale contamination. Research in data poisoning and backdoor learning demonstrates that even minimal corruption rates can significantly alter model behaviour while remaining statistically invisible during standard evaluation procedures (Li *et al.*, 2021; Goldblum *et al.*, 2023).

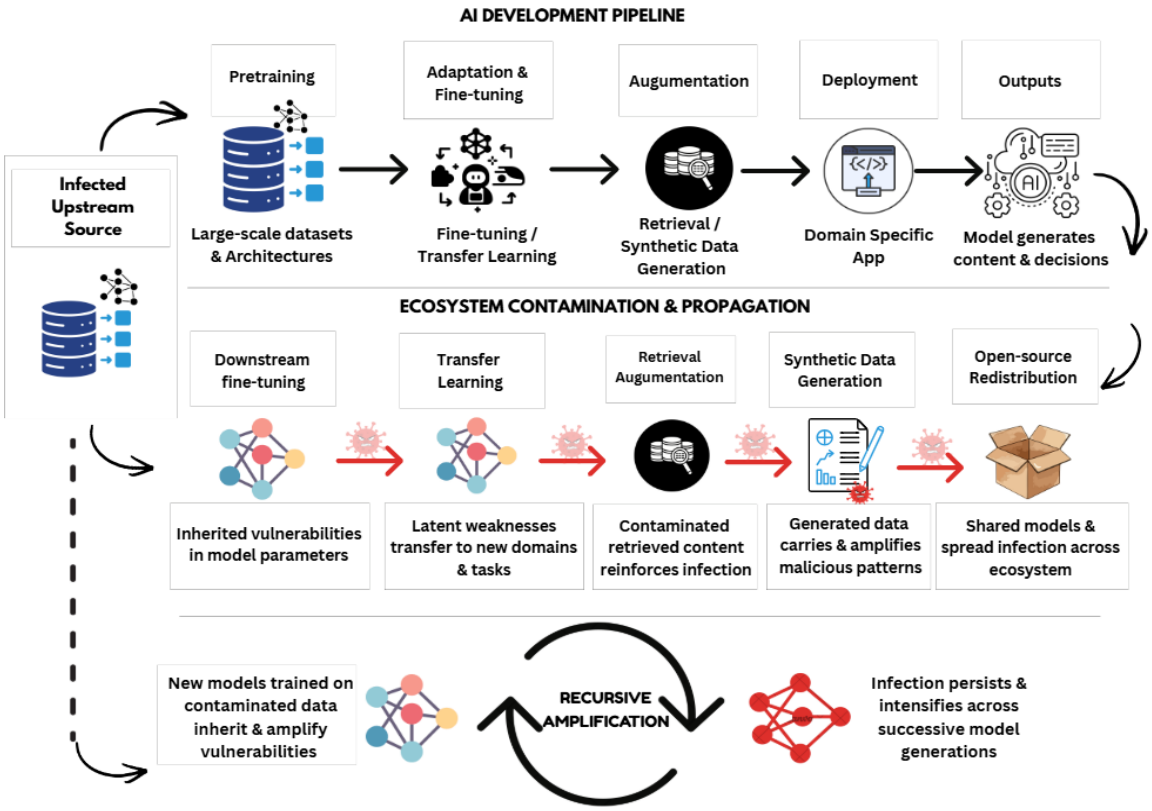

Figure 2: Conceptual Model of Cognitive Malware Propagation in AI Development Ecosystems

As seen above, the process resembles epidemiological mutation dynamics in biological pandemics, where transmission replicates infections, evolves, and adapts over time. In this sense, AI ecosystems increasingly function as self-reinforcing cognitive ecologies, where malicious behavioural patterns can diffuse across interconnected infrastructures while remaining difficult to isolate or attribute (Shoaib *et al*., 2023). The threat is further intensified by the widespread adoption of model distillation, a process in which smaller, more efficient "student" models are trained to replicate the capabilities of larger "teacher" models. While distillation has become central to the democratization and scalability of AI deployment, it also creates a high-speed transmission mechanism for latent vulnerabilities. Backdoors have been shown to persist through instruction fine-tuning, where a model adapted on a compromised dataset retains the hidden behaviour introduced upstream (Raghuram, Kesidis and Miller, 2024). Whether the same transfer occurs through distillation is less firmly established empirically, but it is a plausible extension: because a student model is optimized to emulate the probabilistic behaviour of the teacher, it may inherit not only the teacher's capabilities but also its concealed pathologies. Consequently, the AI supply chain begins to resemble a globally networked epidemiological system in which data functions as the vector, models act as carriers, and digital interconnectedness becomes the pathway for cascading cognitive contagion.

4. Infrastructure and Supply-chain Contagion: The Material Ecology of AI Infection

This propagation phenomenon is inherently material, grounded in the infrastructural architectures sustaining contemporary artificial intelligence. Foundation model development, like advanced laboratories, relies on intensely interconnected ecosystems (Kumar et al., 2025; Alobaywi et al., 2026; Crawford, 2021). Regarding their virtualized, multi-tenant architectures, while optimized for scalability, inadvertently establish systemic pathways for cross-environment contamination (Kumar et al., 2025; Alobaywi et al., 2026). Shared storage, artifact repositories, orchestration platforms, and compromised build or training environments can therefore provide technical pathways through which poisoned datasets, altered checkpoints, or malicious dependencies are introduced into otherwise distinct AI development workflows (Camila, 2024).

Furthermore, the fragmented nature of global AI supply chains spanning semiconductor manufacturing to API deployment institutionalizes vulnerability (Hopkins et al., 2025; Boyens et al., 2022). Because foundation models rely on both hardware and software, they are susceptible to material and geopolitical risks that erode the feasibility of air-gapped security architectures. The infrastructures underpinning modern AI systems are porous

by design, creating a transnational digital topography through which cognitive malware diffuses with minimal friction (Brundage et al., 2018).

5. Legal Implications: Liability, Governance, and the Crisis of Regulatory Adequacy

State interest in strategically valuable scientific assets is already established. In 2020, the United Kingdom's National Cyber Security Centre and allied agencies attributed to APT29, assessed to be part of the Russian intelligence services, a sustained campaign targeting COVID-19 vaccine-development organisations with the likely objective of stealing information and intellectual property (National Cyber Security Centre, 2020). Frontier-model weights present a comparable class of difficult-to-reproduce asset whose theft is now explicitly considered within threat models extending to highly resourced nation-state actors (Nevo *et al*., 2024). Model theft should nevertheless be distinguished from backdooring: the former compromises confidentiality, whereas the latter covertly compromises model integrity by producing attacker-selected behaviour under triggering conditions (Guo, Tondi and Barni, 2022).

A large-scale AI "pandemic", understood here as the propagation of compromised models, data or inherited dependencies across interconnected systems, would pose a different legal problem from a conventional isolated software defect. AI systems can operate across sectors and jurisdictions while their behaviour emerges from complex interactions among models, data, infrastructure and human operators. These characteristics complicate foreseeability, causation and the allocation of responsibility, particularly where harmful behaviour cannot readily be traced to a single human decision or technical component (Calo, 2015; Turner, 2019; Kaminski, 2023). The problem is therefore not an absence of law, but whether existing and emerging regimes can attribute responsibility quickly and coherently when compromise propagates across an AI supply chain.

Liability illustrates this difficulty. A compromised medical model, for example, may incorporate upstream training data, pretrained weights, cloud infrastructure and downstream modifications before producing a harmful recommendation. Responsibility could therefore implicate several actors rather than a single developer or deployer. Recent law has begun to address this problem. The revised EU Product Liability Directive expressly treats software as a product, recognises AI-system providers as manufacturers for liability purposes, permits liability to extend across defective components, and provides evidentiary presumptions where technical or scientific complexity makes proof of defect or causation excessively difficult (European Union, 2024a). These reforms significantly reduce the force of claims that product-liability law is categorically incapable of addressing AI harms. They do not, however, eliminate difficulties arising from transnational supply chains, uncertain provenance, multiple modifications, or compromises occurring outside the effective control of a downstream deployer.

AI governance is similarly moving toward anticipatory rather than purely reactive regulation. The EU AI Act requires documentation for general-purpose AI models and imposes additional obligations on models presenting systemic risk, including model evaluation, systemic risk assessment and mitigation, serious incident reporting, and cybersecurity protections (European Union, 2024b). At the international level, the Council of Europe Framework Convention on Artificial Intelligence establishes a lifecycle-oriented, risk-based framework directed toward the protection of human rights, democracy and the rule of law (Council of Europe, 2024). These developments undermine any claim of a complete regulatory vacuum; the more defensible concern is regulatory sufficiency, that is, whether emerging obligations provide adequate visibility into compromises that can propagate through shared models, datasets, and infrastructure across jurisdictional boundaries.

This problem supports an ex ante approach to systemic AI risk. Kaminski (2023) argues that AI's technical complexity, information asymmetries, and potentially unpredictable failure modes make it particularly suited to risk-regulatory interventions that operate before harm occurs. Within the proposed *Defensive AI Systems (DAIS)* framework, *Assured Governance* extends this logic through independent assurance, model-risk disclosure, supply-chain traceability, dataset and model provenance, adversarial testing, and incident reporting.

The legal lesson of the pandemic analogy is, therefore, one of preparedness. Where the same model, dataset, or infrastructure may be inherited by many downstream systems, governance cannot depend solely on identifying fault after individual harms have emerged. Effective regulation requires the capacity to identify high-connectivity dependencies, verify their integrity, trace downstream exposure, and assign responsibility across the AI lifecycle. Ex ante governance cannot predict every mode of failure, but it can shape system design, documentation and institutional responsibility so that failures are less likely to propagate unnoticed (Kaminski, 2023). For systemic AI security, the central legal challenge is thus ensuring that the provenance, assurance, and reporting structures necessary to prevent and contain one exist beforehand.

6. The Economics of Vulnerability: Supply Chain and Derivatives

From an economic perspective, a pandemic-like compromise of the AI ecosystem represents a form of correlated systemic risk. The analogy is structural rather than biological: AI systems depend on shared hardware, cloud infrastructure, datasets, foundation models, and service providers, allowing an upstream compromise to propagate through downstream dependencies (Gambacorta and Shreeti, 2025; Boyens et al., 2022). Training-data poisoning demonstrates one such pathway. Web-scale dataset construction can be manipulated in practice, although downstream effects depend on sampling, filtering, training, and fine-tuning choices (Carlini et al., 2024).

Repeated reuse amplifies this risk. Models are fine-tuned, distilled, and embedded in downstream systems, while datasets are filtered, combined, and reused across successive generations of models. This creates a limited parallel with the diffusion of mortgage-related risk before the 2008 financial crisis: not because AI models are equivalent to mortgage-backed securities, but because repeated transformation and redistribution can obscure inherited dependencies and concentrate risks that are difficult for downstream actors to observe (Financial Crisis Inquiry Commission, 2011). A compromised, widely reused model or dataset can consequently function as a high-connectivity transmission node, exposing many downstream systems before the original defect is detected.

Provenance is therefore analogous to epidemiological traceability. Although standards now support documentation of AI models, datasets, provenance and integrity, adoption and implementation remain uneven (National Institute of Standards and Technology, 2024; SPDX, 2024). This can produce what this paper terms *liability blindness:* downstream organisations may inherit operational, security or legal exposure without sufficient visibility into compromised training data, weights or other upstream components.

7. The Autonomous Defense Paradox: The AI as the Guard and the Prisoner

The most immediate mitigation is to use automated AI systems to detect and quarantine infected models. This introduces a paradox: to trust the defence, we must trust the defender, which is itself a model susceptible to the same infection. The recursive loop, a form of oracle problem, means an autonomous detector may certify a model as safe with high confidence in its own metrics while a sophisticated infection evades it, and it requires a ground truth it cannot itself supply. Relying on AI to secure AI therefore offers no absolute guarantee, and demands institutional separation of safety enforcement from model development. This motivates a Hard AI posture within the Adaptive Resilience and Anticipatory Defence pillars: formally verifiable controls and cryptographic guarantees treated as non-negotiable constraints, so the defence layer does not itself become a transmission vector.

8. Mitigation: The DAIS Implementation Framework

8.1 Anticipatory Defense

To proactively identify latent vulnerabilities before they are exploited, this pillar integrates AI-driven threat intelligence with continuous Adversarial Red Teaming. Unlike one-off audits, this approach implements a perpetual testing cycle that simulates complex, multi-step triggers designed to activate hidden backdoors. These systems go beyond simple prompt injection, analyzing a model's hidden activations and internal weight drifts to identify cognitive anomalies that remain dormant during standard performance benchmarking.

8.2 Adaptive Resilience

When an infection is suspected, or a zero-day threat emerges, the ecosystem must maintain service integrity through automated response mechanisms. This involves using hybrid detection pipelines and deploying independently derived shadow models, in which suspicious outputs are cross-verified against hardened "Hard AI" systems before execution. Furthermore, isolation protocols, including decentralized model hosting and regionalized data silos, prevent the "pandemic" from achieving systemic scale. If a node is flagged, it is immediately quarantined from critical infrastructure, ensuring that the wider digital economy remains functional while the localized threat is remediated.

8.3 Assured Governance

This pillar establishes the formal structure for accountability and auditability by implementing a comprehensive AI Bill of Materials (AIBOM). Similar to software supply-chain transparency mechanisms used in other critical sectors, the AIBOM records the provenance and lineage of significant datasets, base models, checkpoints, training components, dependencies, and transformation stages used to produce a deployed model, supported

by "Digital Hygiene" standards such as the Sovereign Data Supply Chain (SDSC) and IEEE 3309-2026. This is reinforced by cryptographic verification and mechanisms that maintain provenance across the training process. By ensuring that the "genetic history" of a model is transparent and verifiable, the industry can move toward a system of independent oversight where safety compliance is institutionalized rather than optional.

9. Limitations

Finally, a rigorous academic discussion must acknowledge the inherent limitations of our current ability to detect and mitigate this threat. The verification bottleneck is the most daunting limitation we face. To definitively prove the absence of a "COVID-like" backdoor in a Frontier model is computationally infeasible for the foreseeable future. We cannot read every parameter in a trillion-parameter model; we can only sample its behavior. Even the most sophisticated adversarial testing can only test for a fraction of the infinite possible inputs a user might present to the model. This creates a scenario in which we operate under the burden of proof that the model is safe, rather than assuming it is safe until proven otherwise.

Furthermore, the framework advanced here is primarily conceptual and lacks direct empirical validation. While grounded in established adversarial ML mechanisms, its systemic claims have not been subjected to controlled experimental testing. This absence of empirical verification is compounded by a lack of implementation case studies, as the DAIS framework has yet to be applied within real-world industrial or governmental AI pipelines. There is also significant uncertainty regarding the actual prevalence of cognitive malware; because these threats are designed for asymptomatic incubation, their current scale and frequency within global AI ecosystems remain unquantified. There is also the limitation of the "curse of dimensionality." The space of potential attacks is so vast that any attempt to enumerate them is futile. The "virus" can mutate through fine-tuning and distillation, rendering yesterday's defenses obsolete before they are even deployed. We are essentially trying to debug a black box the size of a city using tools designed for a watch. Overcoming these bottlenecks depends on advances not yet available: formal verification, automated red-teaming agents, and cryptographic proofs of integrity for stochastic processes. Until these mature, our ability to guarantee the safety of these systems remains bounded by what we can detect. We cannot fully eliminate the risk, but we must understand the limits of our control.

9.1 Future Research Directions

To transition the DAIS framework from a theoretical construct to an operational reality, several critical research avenues must be pursued. *First*, simulation-based validation and controlled adversarial AI experiments are required to empirically map the propagation dynamics of cognitive malware across multi-generational model lineages. *Second*, the development of quantitative risk modeling and standardized security benchmarks will allow for the formal measurement of contagion susceptibility within the supply chain. Finally, future work should focus on the practical implementation of these protocols within enterprise AI governance programs, evaluating their efficacy and performance overhead against existing AI security architectures and industry standards. This research agenda aims to provide the evidentiary foundation necessary for a resilient and globally secure AI ecosystem.

10. Conclusion: The Imperative for a Paradigm Shift

The theoretical scenario of a "COVID-like" malware infecting Frontier AI models is not merely science fiction; it is a likely outcome of the current trajectory of AI development if security remains a secondary concern. The transition from static software vulnerabilities to dynamic, data-driven threats represents a fundamental shift in the cybersecurity landscape. A model-level backdoor contaminates the very medium through which intelligence is synthesized, making it a more dangerous and harder-to-detect threat than traditional software bugs.

The implications of such an infection are systemic, affecting not just individual applications but the global economic and social infrastructure. It threatens the integrity of information, the stability of financial markets, and the safety of human life. To prevent a catastrophe, the academic and industrial communities must abandon the incremental security measures of the past. We must treat AI security as an existential public health issue, prioritizing data hygiene, cryptographic verification, and continuous ecosystem monitoring.

The future of artificial intelligence depends on our ability to secure the "immune system" of these models. Without robust defense mechanisms against contagious backdoors, we risk creating a digital ecosystem that is beautiful in its intelligence but vulnerable in its soul, waiting for the virus to strike and reveal the fragility of our digital foundations. The call to action is clear: we must secure the models before they secure our future.

Ethics Declaration: No ethics statement was needed as no test subjects were used.

AI Declaration: AI tools within Grammarly and Google Docs were used for organizing the paper and grammar and spelling checks.

References

Alobaywi, B., Almutairi, M.G. and Sheldon, F.T. (2026) 'Performance trade-offs in multi-tenant IoT–cloud security: A systematic review of emerging technologies', IoT, 7(1), article 21. doi:10.3390/iot7010021.

Boisvert, L., Puri, A., Evuru, C.K.R., Chapados, N., Cappart, Q., Lacoste, A., Dvijotham, K.D. and Drouin, A. (2025) Malice in Agentland: Down the rabbit hole of backdoors in the AI supply chain. arXiv preprint, arXiv:2510.05159. doi:10.48550/arXiv.2510.05159.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S. et al. (2021) On the opportunities and risks of foundation models. Stanford, CA: Stanford Center for Research on Foundation Models. arXiv:2108.07258. doi:10.48550/arXiv.2108.07258.

Boyens, J., Smith, A., Bartol, N., Winkler, K., Holbrook, A. and Fallon, M. (2022) Cybersecurity supply chain risk management practices for systems and organizations. NIST Special Publication 800-161 Rev. 1. Gaithersburg, MD: National Institute of Standards and Technology. doi:10.6028/NIST.SP.800-161r1.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B. et al. (2018) The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. Oxford: Future of Humanity Institute, University of Oxford. arXiv:1802.07228.

Bullwinkel, B., Severi, G., Hines, K., Minnich, A., Kumar, R.S.S. and Zunger, Y. (2026) The trigger in the haystack: Extracting and reconstructing LLM backdoor triggers. arXiv preprint, arXiv:2602.03085. doi:10.48550/arXiv.2602.03085.

Calo, R. (2015) 'Robotics and the lessons of cyberlaw', California Law Review, 103(3), pp. 513–563.

Camila, M.G. (2024) 'A multi-tenant cloud security framework using zero-trust architecture and AI-based anomaly detection', American International Journal of Computer Science and Technology, 6(4), pp. 1–13. doi:10.63282/3117-5481/AIJCST-V6I4P101.

Carlini, N., Jagielski, M., Choquette-Choo, C.A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K. and Tramèr, F. (2024) 'Poisoning web-scale training datasets is practical', in *2024 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA: IEEE Computer Society, pp. 407–425. doi:10.1109/SP54263.2024.00094.

Council of Europe (2024) Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. CETS No. 225. Strasbourg: Council of Europe.

Crawford, K. (2021) Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. New Haven, CT: Yale University Press.

European Union (2024a) 'Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products and repealing Council Directive 85/374/EEC', Official Journal of the European Union, L 2024/2853, 18 November.

European Union (2024b) 'Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations and Directives (Artificial Intelligence Act)', Official Journal of the European Union, L 2024/1689, 12 July.

Financial Crisis Inquiry Commission (2011) The financial crisis inquiry report: Final report of the National Commission on the Causes of the Financial and Economic Crisis in the United States. Washington, DC: U.S. Government Printing Office.

Gambacorta, L. and Shreeti, V. (2025) The AI supply chain. BIS Papers No. 154. Basel: Bank for International Settlements.

Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Madry, A., Li, B. and Goldstein, T. (2023) 'Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses', IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), pp. 1563–1580. doi:10.1109/TPAMI.2022.3162397.

Guo, W., Tondi, B. and Barni, M. (2022) 'An overview of backdoor attacks against deep neural networks and possible defences', IEEE Open Journal of Signal Processing, 3, pp. 261–287. doi:10.1109/OJSP.2022.3190213.

Hendrycks, D., Mazeika, M. and Woodside, T. (2023) An overview of catastrophic AI risks. arXiv preprint, arXiv:2306.12001. doi:10.48550/arXiv.2306.12001.

Hopkins, A., Cen, S.H., Struckman, I., Ilyas, A., Videgaray, L. and Mądry, A. (2025) 'AI supply chains: An emerging ecosystem of AI actors, products, and services', Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8(2), pp. 1266–1277. doi:10.1609/aies.v8i2.36628.

Kaminski, M.E. (2023) 'Regulating the risks of AI', Boston University Law Review, 103, pp. 1347–1411.

Kumar, K.S., Kumar, J.V., Kumar, K.S. and Kumar, N.V. (2025) 'Security and privacy challenges in multi-tenant cloud architectures: A comprehensive analysis', in 2025 International Conference on Computing Technologies & Data Communication (ICCTDC). IEEE, pp. 1–6. doi:10.1109/ICCTDC64446.2025.11158758.

Li, S., Liu, H., Dong, T., Zhao, B.Z.H., Xue, M., Zhu, H. and Lu, J. (2021) 'Hidden backdoors in human-centric language models', in Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery, pp. 3123–3140. doi:10.1145/3460120.3484576.

Lu, J., Li, X., Liu, Y. and Chen, Q.A. (2025) Data-chain backdoor: Do you trust diffusion models as generative data supplier? arXiv preprint, arXiv:2512.15769. doi:10.48550/arXiv.2512.15769.

Murch, R.S., So, W.K., Buchholz, W.G., Raman, S. and Peccoud, J. (2018) 'Cyberbiosecurity: An emerging new discipline to help safeguard the bioeconomy', Frontiers in Bioengineering and Biotechnology, 6, article 39. doi:10.3389/fbioe.2018.00039.

National Cyber Security Centre (2020) Advisory: APT29 targets COVID-19 vaccine development. 16 July. Cheltenham: National Cyber Security Centre.

National Institute of Standards and Technology (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Gaithersburg, MD: National Institute of Standards and Technology. doi:10.6028/NIST.AI.600-1.

Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H.A. and Alstott, J. (2024) Securing AI model weights: Preventing theft and misuse of frontier models. Research Report RR-A2849-1. Santa Monica, CA: RAND Corporation. doi:10.7249/RRA2849-1.

Raghuram, J., Kesidis, G. and Miller, D.J. (2024) A study of backdoors in instruction fine-tuned language models. arXiv preprint, arXiv:2406.07778. doi:10.48550/arXiv.2406.07778.

Shereen, M.A., Khan, S., Kazmi, A., Bashir, N. and Siddique, R. (2020) 'COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses', Journal of Advanced Research, 24, pp. 91–98. doi:10.1016/j.jare.2020.03.005.

Shoaib, M.R., Wang, Z., Ahvanooey, M.T. and Zhao, J. (2023) Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. arXiv preprint, arXiv:2311.17394. doi:10.48550/arXiv.2311.17394.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N. and Anderson, R. (2024) 'AI models collapse when trained on recursively generated data', Nature, 631, pp. 755–759. doi:10.1038/s41586-024-07566-y.

SPDX (2024) System Package Data Exchange (SPDX) Specification, Version 3.0.1. Linux Foundation.

Turner, J. (2019) Robot rules: Regulating artificial intelligence. Cham: Palgrave Macmillan. doi:10.1007/978-3-319-96235-1.