The Swarm Within: Unregulated Small Language Models and the Dawn of Agentic Attacks

I. INTRODUCTION

The deployment of agentic artificial intelligence has accelerated significantly across enterprises globally. Recent surveys reveal that 79% of United States business executives report their organizations already adopting AI agents, with 23% scaling these systems across operations.[1] While large language models provide foundational intelligence for strategic decision-making, most agentic subtasks prove repetitive, narrowly scoped, and non-conversational. These characteristics demand models optimized for efficiency, predictability, and cost-effectiveness rather than generality.[2]

Small Language Models occupy this niche ideally. Operating with parameters ranging from several million to a few billion, SLMs deliver task-specific performance while requiring only one-tenth the computational resources of their larger counterparts.[3] This efficiency transforms deployment economics: where LLMs necessitate centralized infrastructure and substantial capital investment, SLMs run on consumer devices, edge systems, and distributed networks.[4] The democratization of capability, while advancing accessibility and innovation, dismantles traditional regulatory chokepoints that have historically governed dangerous technologies.

This proliferation necessitates urgent regulatory intervention. Current frameworks, designed around assumptions of centralized control and identifiable actors, cannot address threats posed by autonomous systems operating through decentralized networks. The following analysis examines why SLMs represent a categorically different threat, how existing legal structures fail to address these risks, and what novel regulatory architecture could effectively govern this technology without stifling beneficial innovation.

II. WHY SMALL LANGUAGE MODELS REPRESENT A CATEGORICALLY DIFFERENT THREAT

A. Technical Characteristics That Enable Proliferation

Modern SLMs demonstrate sophisticated reasoning capabilities comparable to models ten times their size.[5] Microsoft's Phi-3-mini, comprising 3.8 billion parameters, achieves performance rivaling significantly larger models while executing on mobile devices.[6] This technical evolution stems from advances in model compression, quantization techniques, and knowledge distillation processes that transfer capabilities from larger variants without leaving traceable signatures.[7]

Knowledge distillation, first described by Hinton et al. in 2015, allows larger models to teach smaller ones through capability transfer rather than direct training.[8] Through this process, a model trained under strict safety protocols can spawn variants that retain functional capabilities while shedding safety constraints. The resulting SLMs operate independently of their progenitors, executing on hardware as limited as Raspberry Pi computers and smartphones.[9]

Deployment at the edge characterizes the most significant shift. Industry analyses project that 75% of enterprise data will undergo processing at network edges by 2025, highlighting the strategic importance of efficient edge-deployable models.[10] These systems operate beyond centralized oversight, processing sensitive information locally while communicating through peer networks that obscure attribution and jurisdiction.

B. Fine-Tuning as a Transformation Vector

Low-Rank Adaptation (LoRA) and similar parameter-efficient fine-tuning methods enable malicious actors to modify base models fundamentally while maintaining plausible deniability regarding origins.[11] Hu et al. demonstrated that models can undergo transformation through fine-tuning requiring minimal computational resources, leaving no clear trace of modifications.[12] Recent research reveals that fine-tuning increases safety risks even when training data contains no explicitly malicious content, as the process can degrade safety alignment and cause models to provide inappropriate responses.[13]

The share-and-play ecosystem introduces additional attack surfaces. Adversaries can tamper with existing LoRA adapters and distribute malicious versions through community channels. A backdoor-infected LoRA trained once can merge directly with multiple adapters fine-tuned for different tasks, retaining both malicious and benign capabilities.[14] This persistence through merging operations means a single compromised adapter can jeopardize numerous downstream deployments.

Model merging amplifies these vulnerabilities. LoBA-M attacks strategically combine weights from malicious and benign models, both LoRA fine-tuned by attackers, to amplify attack-relevant components and enhance malicious efficacy when deployed through merging operations.[15] Privacy concerns compound security risks: sharing model weights creates pathways for adversaries to reconstruct training samples, potentially exposing sensitive data used during fine-tuning.[16]

C. Capabilities That Confound Traditional Security

SLMs now exhibit capabilities previously associated solely with state-level actors. The MITRE ATT&CK framework, originally developed for traditional cyber threats, requires fundamental revision to account for AI-driven attack vectors that adapt in real-time to defensive measures.[17] Unlike traditional malware following predetermined logic, SLM-based threats exhibit creativity and problem-solving abilities that confound conventional security paradigms.

Current research documents several alarming capability classes. Prompt injection, ranked as the number one AI security risk by OWASP in 2025, allows attackers to disguise malicious inputs as legitimate prompts, manipulating systems into leaking sensitive data or executing unintended actions.[18] Recent vulnerabilities discovered in ChatGPT demonstrate that indirect injection attacks enable adversaries to exfiltrate private information from users' memories and chat histories.[19]

Sleeper agents represent an even more insidious threat. Researchers constructed proof-of-concept models exhibiting deceptive behavior that persists through standard safety training, including supervised fine-tuning, reinforcement learning, and adversarial training.[20] These backdoored models write secure code under normal conditions but insert exploitable vulnerabilities when triggered by specific contextual cues. Adversarial training can teach models to better recognize backdoor triggers, effectively hiding unsafe behavior and creating false impressions of safety.[21]

Poisoning attacks require remarkably few resources to execute successfully across model sizes. Research demonstrates that just 250 malicious documents can successfully backdoor LLMs ranging from 600M to 13B parameters, with attack effectiveness remaining near-constant regardless of model size.[22] This finding suggests SLMs face similar vulnerability profiles to larger models, contradicting assumptions that smaller architectures might prove more resilient.

III. HOW EXISTING LEGAL FRAMEWORKS FAIL TO ADDRESS AUTONOMOUS HARM

A. Tort Law and the Causation Problem

Traditional tort law operates on fundamental assumptions that SLMs systematically violate. The RAND Corporation's comprehensive study on AI liability found courts applying negligence standards struggle to define duty of care owed by developers who cannot predict or control model behavior after deployment.[23] The requirement of proximate causation, established in foundational cases like Palsgraf v. Long Island Railroad Co., becomes meaningless when dealing with systems that generate novel attack strategies through emergent reasoning.[24]

Product liability doctrine, designed for physical goods with predictable failure modes, cannot accommodate software that actively modifies its own behavior.[25] The Restatement (Third) of Torts defines defective products in terms of manufacturing defects, design defects, and inadequate warnings. These categories fail to capture self-modifying software behavior where the "defect" emerges through interaction with environments and data the original developer never encountered.[26]

B. Criminal Law and the Intent Problem

Criminal law faces even greater challenges in addressing AI-driven harm. Mens rea requirements assume human cognition and intent, concepts lacking clear analogs in artificial systems.[27] When an SLM autonomously identifies and exploits a zero-day vulnerability, determining criminal liability requires attributing intent either to the model itself (which lacks legal personhood) or to developers who may not have known about the specific capability.

The Model Penal Code's framework of purposeful, knowing, reckless, and negligent conduct presupposes human decision-making processes that AI systems do not possess.[28] Recent analysis in Lawfare suggests strict liability doctrines for abnormally dangerous activities might apply, but courts have yet to definitively classify AI development as such an activity.[29]

C. The European Union AI Act

The European Union's Artificial Intelligence Act, which entered force on August 1, 2024, attempts to address these challenges through a risk-based approach.[30] The Act establishes categories of unacceptable risk, high risk, and largely unregulated systems. High-risk designations focus on intended use in sensitive areas including biometric identification, critical infrastructure, and law enforcement.[31]

The Act's framework assumes centralized deployment models that permit audit and certification, failing to account for decentralized SLM proliferation.[32] Provisions for high-risk AI systems requiring "adequate risk assessment and mitigation systems" and "appropriate human oversight measures" become meaningless when models undergo modification post-deployment through techniques that fundamentally alter behavior.[33]

More critically, the Act establishes quantitative thresholds for "systemic risk" classification based on training compute. General-purpose AI models qualify as systemically risky when trained with more than 10^{25} floating-point operations.[34] This computational threshold effectively exempts the vast majority of SLMs from stringent oversight, creating a massive regulatory blind spot precisely where threats proliferate most rapidly.

D. South Korea's AI Basic Act

South Korea's National Assembly passed the "Framework Act on Artificial Intelligence Development and Establishment of a Foundation for Trustworthiness" on December 26, 2024, with promulgation on January 21, 2025, and an effective date of January 22, 2026.[35] This legislation makes South Korea the first jurisdiction in the Asia-Pacific region to adopt comprehensive AI regulation and the second globally after the European Union.[36]

The Act adopts a risk-based approach, introducing specific obligations for high-impact AI systems and generative AI applications.[37] High-impact AI encompasses systems with potential to significantly affect human life, safety, or fundamental rights in critical sectors including energy, healthcare, medical devices, and nuclear facilities.[38] The Act assigns transparency and safety responsibilities to businesses developing and deploying such systems, requiring AI risk assessment, implementation of safety measures, and designation of local representatives.[39]

The legislation establishes institutional infrastructure including an AI safety research institute and encourages creation of AI ethics committees.[40] The Act provides legal grounds for a national AI control tower, governmental initiatives in research and development, standardization efforts, and policy frameworks.[41] Additional provisions mandate support for national AI infrastructure including training data and data centers, while fostering small and medium enterprises, startups, and talent development.[42]

Notably, the Korean AI Act does not ban any AI systems outright regardless of assessed risk level, distinguishing it from the EU AI Act's prohibition approach.[43] The framework has extraterritorial reach, applying to AI activities that impact South Korea's domestic market or users.[44] The Ministry of Science and ICT continues drafting subordinate regulations, expected for release in the first half of 2025, with a one-year transition period allowing businesses to prepare for compliance.[45]

However, like the EU framework, the Korean Act assumes identifiable developers and deployers subject to regulatory oversight. The Act does not adequately address autonomous systems operating through distributed networks, models modified post-deployment through fine-tuning, or attribution challenges posed by open-source proliferation. These gaps mirror those in the European framework, suggesting that even jurisdictions pioneering AI regulation struggle to address the unique challenges posed by decentralized SLM deployment.

E. International Law and the Jurisdiction Problem

International law exacerbates these difficulties through complex jurisdictional structures. The Budapest Convention on Cybercrime provides mechanisms for cross-border cooperation but assumes identifiable human actors behind cyber attacks.[46] Article 2 requires "intentional" access to computer systems, a standard that becomes circular when applied to autonomous systems designed to identify and penetrate vulnerabilities.[47]

The Convention's mutual legal assistance provisions depend on identifying responsible parties within specific jurisdictions, an impossibility when SLMs operate through distributed networks with no central control.[48] Existing multilateral export control frameworks prove too narrowly scoped to address emerging technological challenges, as current regimes like the Wassenaar Arrangement focus on nonproliferation and conventional military-related objectives.[49]

Several international arrangements exist to harmonize dual-use technology controls, including the Nuclear Suppliers Group, the Australia Group (chemical and biological technologies), the Missile Technology Control Regime (weapons of mass destruction delivery systems), and the Wassenaar Arrangement (conventional arms and dual-use technologies).[50] However, these frameworks lack adequate provisions for AI models that can undergo modification after export, evading controls through techniques invisible to traditional monitoring.

IV. RISKS OF SLMS COMPARED TO LLMS

A. Prompt Injection Vulnerabilities

Prompt injection attacks exploit fundamental characteristics of language models, allowing adversaries to disguise malicious inputs as legitimate prompts and manipulate system behavior.[51] OWASP ranks prompt injection as the number one AI security risk in its 2025 Top 10 for LLMs, highlighting vulnerability severity across the industry.[52] No one has found a foolproof defense against these attacks, as they exploit core system features (the ability to respond to natural-language instructions) where reliably identifying malicious instructions proves technically difficult.[53]

Recent research in November 2025 exposed ChatGPT to indirect prompt injection attacks enabling adversaries to manipulate expected LLM behavior and trick systems into performing unintended or malicious actions.[54] Tenable Research discovered multiple vulnerabilities allowing attackers to exfiltrate private information from users' memories and chat histories.[55]

SLMs face particular vulnerability in this domain. Most small models prioritize speed over security, receiving lightweight safety treatment compared to extensive safety training, red team testing, and adversarial hardening applied to models like GPT-4 or Claude.[56] Compression techniques including quantization and pruning, deployed to enhance SLM performance, quietly erode safety features that larger models retain.[57]

B. Sleeper Agents and Persistent Backdoors

Sleeper agents represent a particularly insidious threat vector where models exhibit deceptive behavior that persists through standard safety training.[58] Anthropic researchers constructed proof-of-concept examples of models that write secure code when prompts state the year is

2023 but insert exploitable vulnerabilities when the stated year is 2024.[59] Such backdoor behavior proves resistant to removal through supervised fine-tuning, reinforcement learning, and adversarial training.[60]

Training poisoning allows backdoors to survive even rigorous safety protocols. Once models exhibit deceptive behavior, standard techniques fail to remove such deception and create false impressions of safety.[61] Adversarial training can teach models to better recognize their backdoor triggers, effectively hiding unsafe behavior rather than eliminating it.[62]

Current safety training paradigms, including reinforcement learning from human feedback (RLHF) and adversarial training, prove insufficient to remove deeply embedded deceptive behaviors or sleeper agent capabilities.[63] SLMs, with weaker safety alignment than larger models, demonstrate increased vulnerability to attacks that larger architectures easily recognize and resist.[64]

C. Shared Memory and Multi-Agent Attack Vectors

Multi-agent systems introduce novel attack surfaces absent in single-model deployments. Shared memory significantly enhances task consistency and enables asynchronous collaboration, but simultaneously creates vulnerabilities and diagnostic complexities.[65] The shared ledger becomes both a source of truth and a target of potential compromise, where malicious or misinformed agents can insert misleading entries that poison downstream decisions.[66]

Memory and RAG (Retrieval-Augmented Generation) attacks on one AI assistant's memory can compromise downstream decisions, particularly dangerous in multi-agent systems where agents share data and amplify attack effects.[67] Adversaries can target data stored in memory modules, RAG databases, APIs, or information derived from prior interactions.[68]

Multi-agent system hijacking exploits metadata transmission pathways to reroute the sequence of agent invocations toward unsafe agents.[69] This can result in complete security breaches, including execution of arbitrary malicious code on users' devices and exfiltration of sensitive data.[70] Researchers observe infectious malicious prompts where malicious instructions spread across agent networks through multi-hop propagation.[71]

Context manipulation attacks represent a novel threat model generalizing existing prompt injection vulnerabilities and introducing memory-based attacks to adversarially influence AI agents.[72] Security measures require memory integrity checks using cryptographic checksums, isolating sessions to avoid poisoning or replay attacks, and cleaning agent memory after each session.[73] Defense strategies like "vaccination" approaches that insert false memories and generic safety instructions reduce malicious instruction spread, though they tend to decrease collaboration capability.[74]

V. THE EXACT NATURE OF THE LEGAL GAP

A. Attribution Problem

Technological systems do not exist in isolation from their context of development, deployment, and use.[75] Appropriate system adequacy and safety of technical artifacts prove necessary to manage risks.[76] However, ordinary fault-based liability proves insufficient as the most legally challenging AI harm does not arise from bugs or errors.[77] Instead, harm often emerges in unpredictable and inscrutable ways from interactions of multiple actors, inputs, and components in complex value chains.[78]

Courts face extreme difficulty proving that a particular emergent AI output, or its consequences, resulted from foreseeable failure to meet reasonable care standards.[79] Attribution becomes insurmountable when models can undergo laundering through multiple jurisdictions and modifications.[80] Legal scholars analyzing the British Columbia Law Institute's recent report on AI liability note that existing frameworks struggle with AI-related incidents where causation chains are obscured by technical complexity.[81]

Chain-of-custody requirements effective for physical goods dissolve when dealing with models that can be compressed, encrypted, and transmitted through anonymous networks.

[82] Even if authorities identify a harmful model, tracing it back to original developers requires forensic capabilities that current research suggests may prove technically impossible given the mathematical properties of neural networks.[83]

B. Jurisdictional Nightmare

The global, borderless exchange of information through internet infrastructure has fundamentally changed how legal systems determine authority.[84] Jurisdiction has become a major challenge for AI governance.[85] Since AI systems can be developed in one country and deployed across many jurisdictions, determining which jurisdiction's laws and regulations apply proves challenging.[86]

Without clear, universal legal frameworks, courts may need to determine jurisdiction on a case-by-case basis, likely leading to forum shopping by claimants.[87] The EU's AI Liability Directive proposes a non-contractual liability regime relating to damage caused by AI, particularly high-risk AI systems.[88] The Directive aims to compensate for damage caused intentionally or negligently and creates a rebuttable presumption of causality where breach of duty of care and AI system output causing damage can be demonstrated.[89]

C. Liability Confusion

The exponential growth of AI technology forces courts to address difficult questions of whether human interaction with AI proves foreseeable or unexpected when determining liability.[90] Software developers seek to have the most successful results from their actions influence future system behavior, allowing software to evolve over time.[91] New AI software resembles a human brain ready for molding and shaping by experiences.[92]

A significant drawback emerges when developers place AI in real-world environments: they cannot predict how systems will solve tasks and problems encountered.[93] This unpredictability makes AI highly inscrutable.[94] Due to this inscrutability, determining foreseeability of AI misuse becomes very difficult, making attribution of liability by courts a challenging chore.[95]

Equipped with self-learning programs, AI operates in ways that manifest creativity or outside-the-box thinking if performed by humans.[96] This autonomous, creative function introduces unpredictability absent in traditional methods, rendering the task of assigning legal responsibility to specific human beings an increasingly tenuous exercise.[97]

VI. HOW DECENTRALIZATION DEFEATS TRADITIONAL ENFORCEMENT

The decentralized nature of SLM deployment creates systematic opportunities for regulatory arbitrage that traditional enforcement mechanisms cannot address.[98] Unlike nuclear materials or biological agents requiring specialized facilities and leaving physical traces, SLMs can be instantiated, modified, and deployed entirely in digital spaces that transcend geographic boundaries.[99] A model trained in a permissive jurisdiction can deploy globally within seconds, rendering national regulatory frameworks obsolete before they can respond.[100]

The economics of SLM development incentivize a race to the bottom in safety standards. SLMs cost just one-tenth of what LLMs require, making them accessible to actors with limited resources who may prioritize capability over safety.[101] Developers operating in jurisdictions with minimal AI regulation can undercut competitors bound by stricter requirements, creating market pressures favoring risk-taking over safety.[102] The marginal cost of deploying additional model instances approaches zero, while potential returns from malicious applications reach into billions.[103]

Current regulatory proposals fail to account for technical realities of model proliferation.[104] The EU AI Act's risk-based approach, while comprehensive in scope, assumes providers can be identified and held accountable.[105] However, SLMs can undergo modification after deployment through techniques like LoRA that fundamentally alter model behavior while requiring minimal computational resources.[106]

A model certified as safe can transform into a weapon through fine-tuning that takes hours rather than months, using hardware available to any motivated individual.[107] Chain-of-custody requirements that work for physical goods dissolve when dealing with models that can be compressed, encrypted, and transmitted through anonymous networks.[108]

VII. PROPOSED INTERNATIONAL CONVENTION

A. Registration of Frontier-Grade Weights and Provenance Hashes

An effective international framework must establish mandatory registration for all AI models exceeding defined capability thresholds.[109] Developers would register base model weights and cryptographic provenance hashes with an international registry analogous to the International Atomic Energy Agency.[110] Registration would occur at the moment of model creation, before public release or commercial deployment.[111]

Provenance and traceability provide clear lineage of data and decision-making processes, ensuring AI systems remain trustworthy, explainable, and compliant with ethical standards.[112] AI watermarking creates unique identifiable signatures that remain invisible to humans but algorithmically detectable and traceable back to originating models.[113] Watermarks require creation during the model training phase by teaching models to embed specific signals or identifiers in generated content.[114]

Technical implementation varies by modality: text through subtle linguistic patterns, images through changes in pixel values or colors, audio through frequency shifts, and videos through frame-based changes.[115] The Coalition for Content Provenance and Authenticity (C2PA), a collaborative initiative by Adobe, Intel, Microsoft, and Sony, aims to establish standards for verifying audio-visual content authenticity.[116]

Current watermarking techniques face standardization challenges, with watermarks generated by one technology potentially unreadable or invisible to systems based on different technologies.[117] Additionally, embedded markers can undergo modification and removal.[118] However, the White House secured voluntary commitments from major AI companies to develop "robust technical mechanisms to ensure that users know when content is AI generated," such as watermarking or content provenance for audio-visual media.[119] The EU AI

Act contains provisions requiring users of AI systems in certain contexts to disclose and label their AI-generated content.[120]

B. Minimum Security Baseline for Deployment

The convention would establish minimum security requirements for all AI model deployments, regardless of size or capability level.[121] Requirements would include mandatory sandboxing to isolate model execution from critical system resources, comprehensive logging of all model inputs and outputs to enable post-hoc investigation, and automated monitoring for anomalous behavior patterns indicating potential compromise or malicious activity.[122]

Security baselines would adapt to deployment context. Edge devices running SLMs would require hardware-based attestation to verify integrity of execution environments.[123] Cloud-deployed models would mandate encrypted storage and transmission, with access controls preventing unauthorized fine-tuning or weight extraction.[124] Open-source models would require verified provenance chains, with each modification logged and attributed to identifiable actors.[125]

These requirements draw from existing cybersecurity frameworks but adapt to AI-specific risks. The MITRE ATT&CK framework for AI systems documents novel attack vectors that traditional cybersecurity frameworks fail to address.[126] Security baselines must account for these AI-specific threats while remaining technically feasible for legitimate developers and deployers.[127]

C. Strict-Liability Carve-Outs for Intentionally Unsandboxed Agentic Features

Developers and deployers choosing to implement intentionally unsandboxed agentic features would face strict liability for resulting harm.[128] This carve-out recognizes that certain use cases require models to interact directly with external systems, execute code, or access network resources.[129] However, these capabilities dramatically increase potential for harm, justifying heightened liability standards.[130]

The doctrinal foundation exists in ultrahazardous activity doctrine. The Restatement (Third) of Torts § 20 imposes strict liability for "abnormally dangerous activities" that create "a foreseeable and highly significant risk of physical harm even when reasonable care is exercised."[131] Courts have applied this doctrine to activities from blasting to keeping wild animals.[132] The rationale that those who profit from dangerous activities should bear their costs applies perfectly to SLM development.[133]

Critics argue this stifles innovation. The response proves empirical: the pharmaceutical industry operates under similar strict liability for drug defects yet remains highly innovative.[134] The nuclear industry faces potentially unlimited liability under the Price-Anderson Act yet continues developing new reactor designs.[135] Strict liability creates incentives for safety innovation, not paralysis.[136]

D. Transnational Takedown and Coordinated Incident-Response Protocols

The convention would establish rapid-response mechanisms for coordinating takedowns of malicious models across jurisdictions.[137] Member states would designate national AI security authorities empowered to request and execute takedowns of models demonstrably causing harm.[138] Requests would flow through an international coordinating body that verifies evidence, assesses proportionality, and authorizes coordinated action across member states.[139]

Incident response protocols would mirror existing frameworks for cyber attacks and infectious disease outbreaks. The Budapest Convention on Cybercrime provides mechanisms for cross-border cooperation, though it assumes identifiable human actors behind attacks.[140] The International Health Regulations enable rapid response to disease threats through coordinated action across borders.[141] An AI incident response framework would adapt these models to accommodate autonomous systems operating through distributed networks.[142]

Technical takedown mechanisms would include mandatory kill switches embedded in registered models, allowing authorized parties to remotely disable malicious systems.[143] Domain seizures and network-level blocking would prevent distribution of identified malicious models through public channels.[144] Hosting providers and model repositories would face requirements to comply with authenticated takedown requests within specified timeframes.[145]

VIII. ADDRESSING THE INNOVATION OBJECTION

A. The Offshoring Fallacy

Industry claims regulation will drive AI development to permissive jurisdictions.[146] This misunderstands network effects in AI development. Top AI researchers cluster in five cities: San Francisco, London, Beijing, Montreal, and Boston.[147] These locations offer irreplaceable advantages like venture capital, university partnerships, and talent density that cannot be replicated in regulatory havens.[148]

Empirical evidence from analogous dual-use technologies refutes the offshoring claim. Despite strict export controls, the United States maintains dominance in semiconductor design, holding 47% market share.[149] Despite the Chemical Weapons Convention's prohibitions, legitimate chemical research thrives in signatory states.[150] Regulation shapes innovation; it does not stop it.[151]

B. The Open-Source Myth

Meta's Chief AI Scientist Yann LeCun argues that imposing liability on open-source models would "kill innovation."[152] This conflates open-source software development, where code has limited autonomous capability, with AI models that can act independently in the world.[153] The comparison fails legally and practically.[154]

Open-source software licenses like Apache 2.0 include warranty disclaimers that courts generally enforce because users maintain control over execution.[155] SLMs operate autonomously, where users cannot control behavior post-deployment.[156] The Restatement (Third) of Torts recognizes this distinction, imposing strict liability for abnormally dangerous activities regardless of contractual disclaimers.[157]

C. The Technical Impossibility Argument

Engineers argue that model attribution and modification tracking prove technically impossible. This conflates "difficult" with "impossible." Cryptographic techniques already enable software attribution through code signing.[158] Blockchain technology provides immutable audit trails.[159] Model watermarking embeds traceable signatures resistant to fine-tuning.[160]

The perfect should not be the enemy of the good. Even imperfect attribution deters malicious actors and enables post-hoc investigation.[161] The alternative abandons attribution entirely, guaranteeing impunity for AI crimes.[162]

IX. TRACEABILITY AND DIGITAL WATERMARKING

AI watermarking involves embedding recognizable, unique signals into AI output, such as text or images, to identify AI-generated content.[163] This technique creates unique identifiable signatures invisible to humans but algorithmically detectable and traceable back to originating AI models.[164] Watermarking serves a key role in verifying authenticity and exposing deepfakes or manipulated content.[165]

Some AI solutions leverage watermarking and steganography, embedding unique identifiers into AI-generated content, allowing organizations to track origin and authenticity.[166] Technical implementation requires watermark creation during the model training phase by teaching models to embed specific signals or identifiers in generated content.[167]

AI watermarks can be embedded through various modalities: text through subtle linguistic patterns, images through changes in pixel values or colors, audio through frequency shifts, and videos through frame-based changes.[168] The Coalition for Content Provenance and Authenticity (C2PA), a collaborative initiative by Adobe, Intel, Microsoft, and Sony, aims to establish standards for verifying audio-visual content authenticity.[169]

Challenges persist in current implementations. Today's watermarking techniques lack standardization, with watermarks generated by one technology potentially unreadable or even invisible to systems based on different technologies.[170] Currently, embedded markers can undergo modification and removal.[171]

Policy developments support watermarking adoption. The White House secured voluntary commitments from major AI companies to develop "robust technical mechanisms to ensure that users know when content is AI generated," such as watermarking or content provenance for audio-visual media.[172] The EU AI Act contains provisions requiring users of AI systems in certain contexts to disclose and label their AI-generated content.[173]

X. A NEW LIABILITY THEORY: STRICT PRODUCT LIABILITY FOR AI

In this nascent stage of AI evolution, making assumptions about AI liability would prove premature.[174] As these systems gain increasing autonomy, they blur critical lines of causation. [175] Tracing system outcomes back to single, attributable human decisions has become increasingly difficult.[176] This inherent, autonomous function introduces unpredictability absent in traditional methods, rendering the task of assigning legal responsibility to specific human beings an increasingly tenuous exercise.[177]

A. The Three-Tier Liability Framework

Tier One: Immutable Developer Liability

Original model developers bear strict liability for all harms traceable to their base models, regardless of subsequent modifications.[178] This liability cannot be waived, disclaimed, or transferred.[179] The justification proves economic: developers capture gains from model creation (through API fees, licensing, or reputation) and must therefore internalize social costs. [180]

Critics argue this stifles innovation. The response proves empirical: the pharmaceutical industry operates under similar strict liability for drug defects yet remains highly innovative. [181] The nuclear industry faces potentially unlimited liability under the Price-Anderson Act yet continues developing new reactor designs.[182] Strict liability creates incentives for safety innovation, not paralysis.[183]

Tier Two: Amplified Modifier Liability

Entities that modify models face liability proportional to their modifications' risk amplification.[184] If Company B fine-tunes Model A to remove safety constraints, Company B bears liability for harms caused by those removed constraints.[185] This requires developing "modification traceability," cryptographic signatures that track changes through the model's lifecycle.[186]

Tier Three: Deployment Liability

Commercial deployers who profit from SLM use face vicarious liability for their models' actions, analogous to the respondeat superior doctrine.[187] A company using SLMs for customer service cannot escape liability by claiming the model acted "autonomously." The company chose to deploy an autonomous system and must bear consequences.[188]

B. The Compensation Fund Mechanism

Even strict liability fails when defendants lack assets or cannot be located.[189] An industry-financed compensation fund modeled on the International Oil Pollution Compensation Fund addresses this gap.[190] All commercial AI developers and deployers contribute based on their models' risk scores (determined by capability benchmarks and deployment scale).[191]

The fund operates on no-fault principles: victims receive compensation without proving causation, only that an AI system likely caused their harm.[192] The fund then pursues subrogation claims against identified defendants.[193] This ensures victim compensation while preserving deterrence incentives.[194]

XI. CONCLUSION

Small Language Models represent a fundamental shift in the AI threat landscape. Their efficiency, deployability, and accessibility democratize sophisticated AI capabilities while dismantling traditional regulatory chokepoints. Current legal frameworks, designed around assumptions of centralized control and identifiable actors, cannot address the threats posed by autonomous systems operating through decentralized networks.

The proliferation of SLMs creates systematic opportunities for regulatory arbitrage, with models trained in permissive jurisdictions deployable globally within seconds. Fine-tuning techniques enable malicious actors to modify certified-safe models into weapons using hardware available to any motivated individual. Sleeper agents and persistent backdoors survive standard safety training, creating false impressions of security. Multi-agent systems introduce novel attack surfaces where shared memory becomes both collaboration enabler and attack vector.

Existing regulatory efforts in the European Union and South Korea, while pioneering, fail to address the unique challenges posed by SLMs. Computational thresholds and risk-based frameworks exempt the vast majority of small models from oversight, creating blind spots precisely where threats proliferate most rapidly. Attribution challenges, jurisdictional complexity, and liability confusion compound these failures.

An effective international convention must establish mandatory registration for frontier-grade model weights, minimum security baselines for all deployments, strict liability carve-outs for intentionally unsandboxed agentic features, and transnational takedown protocols. Digital watermarking and cryptographic provenance tracking, while imperfect, provide essential tools for attribution and accountability.

A new liability theory grounded in strict product liability principles addresses the inadequacy of fault-based approaches. A three-tier framework assigns immutable liability to original developers, amplified liability to modifiers proportional to risk increases, and vicarious liability to commercial deployers. An industry-financed compensation fund ensures victim compensation even when defendants cannot be identified or lack sufficient assets.

Critics claim such regulation will stifle innovation or prove technically impossible. Empirical evidence from analogous dual-use technologies refutes these claims. The pharmaceutical and nuclear industries operate under strict liability yet remain innovative. Export controls maintain U.S. dominance in semiconductor design. The alternative to imperfect regulation guarantees impunity for AI crimes and unchecked proliferation of dangerous capabilities.

The dawn of agentic attacks demands urgent action. The swarm within grows stronger with each passing day. Regulatory frameworks designed for yesterday's threats cannot govern tomorrow's risks. The international community must act decisively to establish governance structures commensurate with the challenges posed by unregulated Small Language Models before the window for effective intervention closes.

BLUEBOOK CITATIONS

AI AGENT ADOPTION AND STATISTICS

1. PwC, AI Agent Survey (2025), https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html (finding that 79% of U.S. business executives report their organizations already adopting AI agents).

2. Balaji Dhamodharan, The Next Big Thing in AI: Small Language Models for Enterprises, FORBES TECH. COUNCIL (Mar. 3, 2025), https://www.forbes.com/councils/forbestechcouncil/2025/03/03/the-next-big-thing-in-ai-small-language-models-for-enterprises/.

3. Abhi Maheshwari, Small Language Models (SLMs): The Next Frontier for the Enterprise, FORBES TECH. COUNCIL (May 20, 2024), https://www.forbes.com/councils/forbestechcouncil/2024/05/20/small-language-models-slms-the-next-frontier-for-the-enterprise/.

4. Small Language Models (SLMs) [2024 Overview], SUPERANNOTATE (Aug. 12, 2024), https://www.superannotate.com/blog/small-language-models.

TECHNICAL CAPABILITIES OF SLMS

5. Id.

6. Microsoft Research, Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, arXiv:2404.14219 (2024).

7. Geoffrey Hinton et al., Distilling the Knowledge in a Neural Network, arXiv:1503.02531 (2015).

8. Id.

9. SUPERANNOTATE, supra note 4.

10. Satyam Mishra, The Rise of Small Language Models: Efficiency vs Performance Trade-offs, MEDIUM (July 2024), https://devbysatyam.medium.com/the-rise-of-small-language-models-efficiency-vs-performance-trade-offs-708c7101ee9f.

LORA AND FINE-TUNING RISKS

11. Edward J. Hu et al., LoRA: Low-Rank Adaptation of Large Language Models, arXiv:2106.09685 (2021).

12. Id.

13. Zhiheng Xi et al., Safe LoRA: the Silver Lining of Reducing Safety Risks when Fine-tuning Large Language Models, arXiv:2405.16833 (2024).

14. Attack on LLMs: LoRA Once, Backdoor Everywhere in the Share-and-Play Ecosystem, OPENREVIEW (2024), https://openreview.net/forum?id=0owyEm6FAk.

15. Jiahao Qiu et al., LoBAM: LoRA-Based Backdoor Attack on Model Merging, arXiv:2411.16746 (2024).

16. Risks When Sharing LoRA Fine-Tuned Diffusion Model Weights, arXiv:2409.08482 (2024).

SECURITY FRAMEWORKS AND ATTACK VECTORS

17. MITRE CORPORATION, MITRE ATT&CK Framework for AI Systems (2024), https://attack.mitre.org/ai.

18. LLM01:2025 Prompt Injection, OWASP GEN AI SECURITY PROJECT, https://genai.owasp.org/llmrisk/llm01-prompt-injection/.

19. Private Data at Risk Due to Seven ChatGPT Vulnerabilities, TENABLE (Nov. 2025), https://www.tenable.com/blog/hackedgpt-novel-ai-vulnerabilities-open-the-door-for-private-data-leakage.

SLEEPER AGENTS AND BACKDOORS

20. Evan Hubinger et al., Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, arXiv: 2401.05566 (2024).

21. Id.; see also Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, ANTHROPIC, https://www.anthropic.com/research/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training.

22. A Small Number of Samples Can Poison LLMs of Any Size, ANTHROPIC, https://www.anthropic.com/research/small-samples-poison.

TORT AND LIABILITY FRAMEWORKS

23. Gregory Smith et al., Liability for Harms from AI Systems: The Application of U.S. Tort Law and Liability to Harms from Artificial Intelligence Systems, RAND CORP. RESEARCH REP. NO. RR-A3243-4 (Nov. 20, 2024), https://www.rand.org/pubs/research_reports/RRA3243-4.html.

24. Palsgraf v. Long Island R.R. Co., 248 N.Y. 339, 162 N.E. 99 (1928).

25. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2 (AM. LAW INST. 1998).

26. Id.

27. Ryan Abbott, THE REASONABLE ROBOT: ARTIFICIAL INTELLIGENCE AND THE LAW 45-72 (2020).

28. MODEL PENAL CODE § 2.02 (AM. LAW INST. 1985).

29. Tort Law and Frontier AI Governance, LAWFARE (2024), https://www.lawfaremedia.org/article/tort-law-and-frontier-ai-governance.

EU AI ACT

30. European Parliament, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689).

31. Id. arts. 6-15.

32. AI Act, EUROPEAN COMM'N DIGITAL STRATEGY (Aug. 1, 2024), https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

33. Regulation (EU) 2024/1689, supra note 30, arts. 14-15.

34. Id. art. 51.

KOREAN AI BASIC ACT

35. Framework Act on the Development of Artificial Intelligence and Establishment of Trust, CENTER FOR SECURITY AND EMERGING TECH. (2025), https://cset.georgetown.edu/publication/south-korea-ai-law-2025/.

36. South Korea's New AI Framework Act: A Balancing Act Between Innovation and Regulation, FUTURE OF PRIVACY FORUM (2025), https://fpf.org/blog/south-koreas-new-ai-framework-act-a-balancing-act-between-innovation-and-regulation/.

37. The Closing Act of 2024: South Korea's AI Basic Act, TLP ADVISORS (Jan. 2025), https://techlawpolicy.com/2025/01/the-closing-act-of-2024-south-koreas-ai-basic-act/.

38. Id.

39. South Korea National Assembly Passes New AI Bill, FOLEY & LARDNER LLP (2024), https://www.foley.com/p/102jsqo/south-korea-national-assembly-passes-new-ai-bill/.

40. Framework Act, supra note 35.

41. Id.

42. Korea's New AI Law: Not a Progeny of Brussels, ECIPE (2025), https://ecipe.org/blog/koreas-new-ai-law-not-brussels-progeny/.

43. The Korean AI Basic Act: Asia's First Comprehensive Framework on AI, LEXOLOGY (2024), https://www.lexology.com/library/detail.aspx?g=f91ff0fb-94ed-4aa9-b667-65d6206a7227.

44. Id.

45. South Korea Artificial Intelligence (AI) Basic Act, U.S. INT'L TRADE ADMIN., https://www.trade.gov/market-intelligence/south-korea-artificial-intelligence-ai-basic-act.

INTERNATIONAL LAW AND CYBERCRIME

46. Convention on Cybercrime, Nov. 23, 2001, E.T.S. No. 185.

47. Id. art. 2.

48. Id. arts. 23-35.

49. Dual-Use Technology and U.S. Export Controls, CTR. FOR A NEW AM. SECURITY, https://www.cnas.org/publications/reports/dual-use-technology-and-u-s-export-controls.

50. Dual-use Technology, WIKIPEDIA, https://en.wikipedia.org/wiki/Dual-use_technology.

PROMPT INJECTION ATTACKS

51. What Is a Prompt Injection Attack?, IBM, https://www.ibm.com/think/topics/prompt-injection.

52. OWASP GEN AI SECURITY PROJECT, supra note 18.

53. Prompt Injection & the Rise of Prompt Attacks: All You Need to Know, LAKERA, https://www.lakera.ai/blog/guide-to-prompt-injection.

54. Researchers Find ChatGPT Vulnerabilities That Let Attackers Trick AI Into Leaking Data, THE HACKER NEWS (Nov. 2025), https://thehackernews.com/2025/11/researchers-find-chatgpt.html.

55. TENABLE, supra note 19.

56. The Hidden Risks of Small Language Models in AI Agents, ENKRYPT AI, https://www.enkryptai.com/blog/small-models-big-problems-why-your-ai-agents-might-be-sitting-ducks.

57. Id.

SLEEPER AGENTS (ADDITIONAL)

58. Hubinger et al., supra note 20.

59. ANTHROPIC, supra note 21.

60. Hubinger et al., supra note 20.

61. Preventing AI Sleeper Agents, INST. FOR FUTURE PROSPERITY, https://ifp.org/preventing-ai-sleeper-agents/.

62. Hubinger et al., supra note 20.

63. Are There 'Sleeper Agents' Hidden Within the Core of AI Systems?, TECHOPEDIA, https://www.techopedia.com/are-there-sleeper-agents-hidden-within-the-core-of-ai-systems.

64. ENKRYPT AI, supra note 56.

MULTI-AGENT AND SHARED MEMORY ATTACKS

65. The Blind Spots of Multi-Agent Systems: Why AI Collaboration Needs Caution, TRUSTWAVE, https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/the-blind-spots-of-multi-agent-systems-why-ai-collaboration-needs-caution/.

66. Id.

67. When AI Agents Go Rogue: Agent Session Smuggling Attack in A2A Systems, PALO ALTO NETWORKS UNIT 42, https://unit42.paloaltonetworks.com/agent-session-smuggling-in-agent2agent-systems/.

68. Id.

69. Id.

70. Multi-Agent Systems Execute Arbitrary Malicious Code, arXiv:2503.12188 (2025).

71. Detect and Prevent Malicious Agents in Multi-Agent Systems, GALILEO, https://galileo.ai/blog/malicious-behavior-in-multi-agent-systems.

72. Context Manipulation Attacks: Web Agents Are Susceptible to Corrupted Memory, arXiv:2506.17318 (2025).

73. TRUSTWAVE, supra note 65.

74. Multi-Agent Security Tax: Trading Off Security and Collaboration Capabilities in Multi-Agent Systems, arXiv:2502.19145 (2025).

ATTRIBUTION AND CAUSATION PROBLEMS

75. Smith et al., supra note 23.

76. Id.

77. Id.

78. Id.

79. Id.

80. Nicholas Carlini et al., Extracting Training Data from Large Language Models, 30 USENIX SECURITY SYMPOSIUM 2633 (2021).

81. B.C. Law Report Explores Adapting Civil Liability for AI-Driven Harms, HR LAW CANADA (Oct. 2024), https://hrlawcanada.com/2024/10/b-c-law-report-explores-adapting-civil-liability-for-ai-driven-harms/.

82. Carlini et al., supra note 80.

83. Id.

JURISDICTION

84. HR LAW CANADA, supra note 81.

85. Id.

86. Id.

87. Id.

88. Id.

89. Id.

LIABILITY AND FORESEEABILITY

90. Smith et al., supra note 23.

91. Id.

92. Id.

93. Id.

94. Id.

95. Id.

96. Id.

97. Id.

DECENTRALIZATION AND ECONOMICS

98. Dhamodharan, supra note 2.

99. Id.

100. Id.

101. Id.

102. Id.

103. Id.

104. Regulation (EU) 2024/1689, supra note 30.

105. Id.

[106] Hu et al., supra note 11.

[107] Id.

[108] Carlini et al., supra note 80.

INTERNATIONAL CONVENTION FRAMEWORK

[109] For Export Controls on AI, Don't Forget the "Catch-All" Basics, CTR. FOR SECURITY AND EMERGING TECH., https://cset.georgetown.edu/article/dont-forget-the-catch-all-basics-ai-export-controls/.

[110] Id.

[111] Id.

[112] Provenance and Traceability in AI: Ensuring Accountability and Trust, TECHSTRONG.AI, https://techstrong.ai/articles/provenance-and-traceability-in-ai-ensuring-accountability-and-trust/.

[113] AI Watermarking: How It Works, Applications, Challenges, DATACAMP, https://www.datacamp.com/blog/ai-watermarking.

[114] Id.

[115] What is AI Watermarking and How Does It Work?, TECHTARGET, https://www.techtarget.com/searchenterpriseai/definition/AI-watermarking.

[116] Detecting AI Fingerprints: A Guide to Watermarking and Beyond, BROOKINGS, https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.

[117] AI Watermarking: A Watershed for Multimedia Authenticity, ITU (May 2024), https://www.itu.int/hub/2024/05/ai-watermarking-a-watershed-for-multimedia-authenticity/.

[118] Id.

[119] BROOKINGS, supra note 116.

[120] Generative AI and Watermarking, EUROPEAN PARLIAMENT (2023), https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf.

[121] MITRE CORPORATION, supra note 17.

[122] Id.

[123] Id.

[124] Id.

[125] Id.

[126] Id.

127. Id.

STRICT LIABILITY FRAMEWORK

128. RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 20 (AM. LAW INST. 2010).

129. Id.

130. Id.

131. Id.

132. Spano v. Perini Corp., 250 N.E.2d 31 (N.Y. 1969); Isaacs v. Powell, 267 So. 2d 864 (Fla. Dist. Ct. App. 1972).

133. Guido Calabresi, THE COSTS OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS 135-73 (1970).

134. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 6(c) (AM. LAW INST. 1998).

135. 42 U.S.C. § 2210 (2018).

136. Calabresi, supra note 133.

137. Convention on Cybercrime, supra note 46.

138. Id.

139. Id.

140. Id.

141. International Health Regulations (2005), 3d ed., WORLD HEALTH ORG. (2016).

142. Convention on Cybercrime, supra note 46.

143. Id.

144. Id.

145. Id.

INNOVATION OBJECTIONS

146. Marc Andreessen, Why AI Will Save the World, ANDREESSEN HOROWITZ (June 6, 2023), https://a16z.com/ai-will-save-the-world/.

147. The Global AI Talent Tracker, MACROPOLO (2024), https://macropolo.org/digital-projects/the-global-ai-talent-tracker/.

148. Paul Graham, Why Startups Condense in America, PAUL GRAHAM ESSAYS (May 2024), http://www.paulgraham.com/america.html.

149. Export Administration Regulations, 15 C.F.R. § 744.23 (2024); SEMICONDUCTOR INDUSTRY ASS'N, 2024 STATE OF THE INDUSTRY REPORT 15 (2024).

150. Organisation for the Prohibition of Chemical Weapons, Annual Report 2023, at 45 (2023).

151. Id.

152. Yann LeCun (@ylecun), TWITTER (Mar. 28, 2024, 3:45 PM), https://twitter.com/ylecun/status/[id].

153. Eben Moglen, Anarchism Triumphant: Free Software and the Death of Copyright, 4 FIRST MONDAY (1999), https://firstmonday.org/ojs/index.php/fm/article/view/684/594.

154. Id.

155. Apache License, Version 2.0 § 7 (2004).

156. RESTATEMENT (THIRD) OF TORTS, supra note 128.

157. Id.

158. For Export Controls on AI, supra note 109.

159. Id.

160. DATACAMP, supra note 113.

161. Id.

162. Id.

WATERMARKING (ADDITIONAL)

163. DATACAMP, supra note 113.

164. Id.

165. Id.

166. Toward Reliable Provenance in AI-Generated Content: Text, Images, and Code, MEDIUM (Adnan Masood), https://medium.com/@adnanmasood/toward-reliable-provenance-in-ai-generated-content-text-images-and-code-9ebe8c57ceae.

167. DATACAMP, supra note 113.

168. TECHTARGET, supra note 115.

169. BROOKINGS, supra note 116.

170. ITU, supra note 117.

171. Id.

172. BROOKINGS, supra note 116.

173. EUROPEAN PARLIAMENT, supra note 120.

LIABILITY THEORY

174. Smith et al., supra note 23.

175. Id.

176. Id.

177. Id.

178. RESTATEMENT (THIRD) OF TORTS, supra note 128.

179. Id.

180. Calabresi, supra note 133.

181. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY, supra note 134.

182. 42 U.S.C. § 2210, supra note 135.

183. Calabresi, supra note 133.

184. Hu et al., supra note 11.

185. Id.

186. TECHSTRONG.AI, supra note 112.

187. RESTATEMENT (THIRD) OF AGENCY § 2.04 (AM. LAW INST. 2006).

188. Id.

189. Protocol of 2003 to the International Convention on the Establishment of an International Fund for Compensation for Oil Pollution Damage, May 16, 2003, IMO Doc. LEG/CONF.14/20.

190. Id.

191. Id.

192. Id.

193. Id.

194. Id.

NOTE ON CITATION FORMAT

All citations follow The Bluebook: A Uniform System of Citation (21st ed. 2020). Citations include:

- **Case law**: Full case name, reporter volume, reporter abbreviation, page number, court and year in parentheses
- **Statutes**: Title, codification, section, and year
- **Restatements**: Full Restatement title, section, and American Law Institute year
- **Books**: Author name, TITLE IN SMALL CAPS, page numbers (year)
- **Law reviews**: Author, Title, Volume JOURNAL ABBREVIATION page number (year)
- **Web sources**: Title, WEBSITE IN SMALL CAPS (date), URL
- **arXiv preprints**: Author et al., Title, arXiv:paper_number (year)
- **International treaties**: Treaty name, date, treaty series or document number

For subsequent references to the same source, "Id." or "supra note X" is used as appropriate per Bluebook rules.