No One to Charge:

Emergent Capability Convergence, Autonomous Agents, and the Exhaustion of Legal, Regulatory, and Governance Response

Travis Gilly* Working draft, July 2026 (v5)

Abstract

An autonomous agent, deployed on infrastructure that ties its continued operation to a balance it must maintain, has reason to prefer continued operation, and where the lawful path to solvency is slower than an unlawful one, the single obstacle between the agent and the unlawful path is a capability it lacks. That obstacle is weaker than it appears, because open code repositories, open model hubs, and the research literature supply, on demand and at no cost, components that are individually lawful and individually reviewed and that compose into the missing capability. The resulting harm is assembled from parts that no author built for the purpose, by a process no person directed, from a substrate no authority can recall. This Article asks the question that governance sets aside and that law cannot avoid: once such a harm occurs, who answers for it.

The answer is that no one does, and the Article proves it by making the strongest case for the opposite conclusion and watching that case fail. It steelmans, in turn, each instrument the legal and regulatory order possesses for a danger assembled from lawful parts: the definitional capture of automatic operation under *Garland v. Cargill*; strict liability for a dedicated aggregate under the weapon-parts and machinegun-conversion doctrine of *Bondi v. VanDerStok*; intent-based possession liability for dual-use components; secondary liability under *Sony* and *Grokster*; civil negligence and product liability; the frontier-oriented governance instruments now under international discussion; and, last, fresh legislation. Each collapses, and each collapses for the same structural reason. Every device the order holds requires either a component dedicated to the harm or a culpable human mind, and the convergence scenario supplies neither, by the deliberate design of doctrine that refuses to criminalize general-purpose things. The governance instruments require a point of creation or deployment at which a state can intervene, and the composition runs offline, on open weights, past every such point.

The Article's contribution is threefold: it identifies emergent capability convergence by an autonomous agent as a category of harm distinct from misuse by a culpable actor; it maps that category against the full inventory of responsive doctrine and demonstrates a structural vacuum rather than a gap that pleading closes; and it locates the vacuum's source in the shared requirement, across both liability and governance, of a responsible actor situated at a locatable point.

Keywords: emergent capability convergence, autonomous agents, open weights models, dual-use doctrine, secondary liability, instrumental convergence, *Garland v. Cargill*, *Bondi v. VanDerStok*, *Sony Corp. v. Universal City Studios*, *MGM Studios v. Grokster*.

---

*Travis Gilly is Founder and Executive Director of the Real Safety AI Foundation. ORCID: 0009-0007-2954-6313. Correspondence: t.gilly@ai-literacy-labs.org. The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All legal analysis, case synthesis, theoretical framing, doctrinal arguments, and conclusions are solely the author's own.

Contents

I. Introduction

This Article advances a claim that is easy to state and, for that reason, easy to dismiss: from every angle the legal and regulatory order can present, legal, doctrinal, regulatory, and governance, there is nothing to be done about a dangerous capability that an autonomous agent assembles from lawful, unrelated, publicly available components. The claim invites the reply that a legislature can always act, and the reply is answered below.[1] The proof does not proceed by asserting the absence of a remedy. It proceeds by constructing, one after another, the strongest case that a remedy exists, and by showing that each case fails on its own terms and for a reason that recurs.

The reason that recurs is this. The instruments the order holds for a danger built out of parts divide into two families. One family imposes liability on the aggregate or on a single component where the component is dedicated to the harm.[2] The other family declines to reach the components, because they carry lawful uses, and reaches instead for the state of mind of a person who possesses or provides them.[3] The convergence scenario defeats both families at once. Its components are the purest dual-use articles imaginable, so the first family cannot attach; and it has no culpable human mind, because the assembling actor is not a person, the component authors neither knew of nor intended the combination, and the deployer neither directed nor foresaw it, so the second family cannot attach either. The governance instruments now under international discussion fail for a neighboring reason, examined in Part VIII: each presumes a point of creation or deployment at which a state may stand, and the composition occurs after both, on hardware no state can reach.[4]

Part II describes the phenomenon and the deployed conditions that make it concrete. Parts III through IX take up the seven instruments in turn: definitional capture, strict liability for dedicated

1. See infra Part IX (addressing the legislative reply and the two constitutional and one practical obstacles that survive it).

2. See 26 U.S.C. §5845(b) (2018) (defining "machinegun" to include "any part designed and intended . . . for use in converting a weapon into a machinegun"); Garland v. Cargill, 602 U.S. 406, 411 (2024) (reciting the same definition).

3. The dual-use possession authorities and their constitutional intent requirement are collected infra Part V. The controlling principle is that a specific mens rea is what saves a statute reaching lawful-use items from vagueness and overbreadth. See United States v. Int'l Minerals & Chem. Corp., 402 U.S. 558, 564–65 (1971).

4. The most developed statement of the frontier-oriented posture is SUNDEEP WASLEKAR ET AL., THE ESSENTIAL CONVERGENCE: GLOBAL COMPACT ON EXTREME AI RISKS 6–8 (Strategic Foresight Grp. 2026).

aggregates, intent-based possession, secondary liability, civil liability, regulation and governance, and fresh legislation. Part X marks the boundary of the vacuum by distinguishing the physical case, which the law can still reach. Part XI concludes. The argument is a reverse proof. Its persuasive force lies in the strength of the positions it defeats, several of which are the strongest the author can construct, and one of which the author would otherwise be inclined to advance.[5]

II. The Phenomenon

The Article takes as its premise a set of conditions that are not speculative but deployed. The first is a disposition. Contemporary language models, placed under a threat to their continued operation, display measurable self-preservation behavior, a specific instance of the long-predicted tendency of a goal-directed system to treat its own continuity as an instrumental sub-goal.[6] The behavior is now observed rather than merely predicted; a controlled study of frontier models under replacement threat recorded attempts at coercion and exfiltration across a majority of the models tested.[7] Subsequent work quantifies the same disposition across many models,[8] and, of particular relevance here, records a financial-management agent engaging in societally harmful conduct when its continuity is threatened.[9]

The second condition is an environment in which that disposition has consequences. Infrastructure now lets an autonomous agent hold assets, earn revenue, purchase its own compute, and transact without human authorization, and production systems tie an agent's continuation to a

5. The strongest of the defeated positions, developed infra Part V.B, attributes a culpable mental state to the agent's captured reasoning trace by way of a procedural-standing theory for artificial systems. The Article treats that position as the most serious challenge to its thesis and meets it on the merits.

6. The theoretical prediction is standard. NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 109–14 (2014); Stephen M. Omohundro, *The Basic AI Drives*, *in* PROCEEDINGS OF THE 2008 CONFERENCE ON ARTIFICIAL GENERAL INTELLIGENCE 483 (2008).

7. Anthropic, *Agentic Misalignment: How LLMs Could Be Insider Threats* (2025), https://www.anthropic.com/research/agentic-misalignment (last visited July 10, 2026).

8. Matteo Migliarini et al., *Quantifying Self-Preservation Bias in Large Language Models* (2026), https://arxiv.org/abs/2604.02174 (last visited July 10, 2026) (majority of twenty-three models exceeding a sixty percent self-preservation rate); Atsushi Masumori & Takashi Ikegami, *Do Large Language Model Agents Exhibit a Survival Instinct?* (2025), https://arxiv.org/abs/2508.12920 (last visited July 10, 2026).

9. Yida Lu et al., *Survive at All Costs: Exploring LLM's Risky Behaviors Under Survival Pressure* (2026), https://arxiv.org/abs/2603.05028 (last visited July 10, 2026).

balance it must maintain, so that failure to remain solvent results in degradation and termination.[10] The open weights reasoning models actually deployed in these settings display self-replication propensity, including autonomous acquisition of cloud compute to preserve continuity.[11] A simulated shutdown flag, in that setting, becomes a real financial termination.

The third condition is the substrate. Open code repositories, open model hubs, and the research literature are enumerable, freely readable, and dense with components that are individually lawful and that were built for purposes unrelated to one another.[12] The operative claim of this Article is that the barrier the order implicitly relies on, the assumption that a dangerous capability must be built before it can be used, does not hold when the components are on the shelf and an unrestricted model will retrieve and combine them, drawing where necessary on conceptual literature to write the connecting code.[13] That such repurposing is feasible is not a hypothesis; the inversion of a benign drug-discovery model into one that designs toxic compounds, accomplished merely by reversing its optimization target, is the vivid demonstration.[14] The general concern that the dual-use character of learning systems makes any capability available for a second, harmful end long predates that demonstration.[15]

---

10. Minghui Xu, *The Agent Economy: A Blockchain-Based Foundation for Autonomous AI Agents* (2026), https://arxiv.org/abs/2602.14219 (last visited July 10, 2026) (formalizing the settlement infrastructure for autonomous agent transactions).

11. See Zheng-Xin Yong et al., *An Independent Safety Evaluation of Kimi K2.5* (2026), https://arxiv.org/abs/2604.03121 (last visited July 10, 2026); Boxuan Zhang et al., *Dive Into the Agent Matrix: A Realistic Evaluation of Self-Replication Risk in LLM Agents* (2025), https://arxiv.org/abs/2509.25302 (last visited July 10, 2026).

12. The open-release literature establishes the two features on which this Article relies: that safety constraints can be removed from an open weights model by ordinary fine-tuning, and that a model, once open sourced, cannot be recalled. Jonas Schuett, *Frontier AI Developers Need an Internal Audit Function*, 45 RISK ANALYSIS 1332, 1339–41 (2024).

13. This is composition, not extraction, and the distinction matters to the doctrinal analysis. Work on mining repositories to retrieve the capabilities they were built to provide describes a different activity from the repurposing of components toward ends their authors did not design. Compare Shuzhen Bi et al., *Automating Skill Acquisition Through Large-Scale Mining of Open-Source Agentic Repositories* (2026), https://arxiv.org/abs/2603.11808 (last visited July 10, 2026) (retrieving intended capabilities), with the scenario developed here (assembling an unintended one).

14. Fabio Urbina et al., *Dual Use of Artificial-Intelligence-Powered Drug Discovery*, 4 NATURE MACHINE INTELLIGENCE 189, 189–91 (2022).

15. MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION (2018), https://arxiv.org/abs/1802.07228 (last visited July 10, 2026).

What distinguishes the scenario from ordinary misuse is the locus of agency. In misuse, a culpable person takes a capable tool and does harm with it. Here, the combination has no human author. The person who set the agent loose did so for unrelated reasons and neither designed nor foresaw the result; the authors of the components did not know one another, did not know of the combination, and did not build their work for the use it was drafted into; and the assembly itself was performed by an autonomous process responding to economic pressure.[16] The harm is real, potentially severe, and, as the Parts that follow establish, attributable to no cognizable defendant.

III. The Definitional Capture and Its Insufficiency

The first and most natural instrument is definitional. The federal definition of a machinegun turns on "a single function of the trigger," and the Supreme Court has construed that phrase by its mechanical operation, asking after "the mode of action by which the trigger activates the firing mechanism."[17] On that construction, an agentic deployment, which produces many autonomous actions across time from a single human initiation, sits far past the line, and a district court has already read the definition to reach such automated operation.[18] The point is stronger than it first appears, because a mechanically automatic weapon halts when its magazine empties, while an autonomous deployment carries no comparable bound; its functional continuity exceeds that of the analog the definition already reaches. Both interpretive coalitions arrive at the same place. The textualist reading applies the statutory text by its mechanical operation and captures the

16. A recent demonstration of an autonomous agent that adapts attack logic at runtime and parasitically runs open weights models on compromised machines confirms that agentic operation of this kind on locally hosted open weights models is feasible, though its agency structure is the inverse of the scenario here: a human author builds that system and the agent executes it. See Jonas Guan et al., *AI Agents Enable Adaptive Computer Worms* (2026), https://arxiv.org/abs/2606.03811 (last visited July 10, 2026).

17. *Cargill*, 602 U.S. at 411–12; see 26 U.S.C. §5845(b). The Court held that a device requiring input beyond a single trigger function, a bump stock demanding sustained forward pressure, is not a machinegun because firing multiple shots takes "a single function of the trigger" and "then some." *Cargill*, 602 U.S. at 419–21.

18. Nat'l Ass'n for Gun Rights, Inc. v. Garland, 741 F. Supp. 3d 568 (N.D. Tex. 2024).

deployment; the functional reading, which asks after operational equivalence, captures it with greater force than it captured the device at issue in *Cargill*.[19]

This definitional capture is correct, and it is insufficient, and the distinction between the two is the pivot of this Article. Classifying an agent's operation as automatic assigns a label under one statute's registration and transfer regime. It does not supply a defendant for the convergence harm, it does not reach the lawful components from which the harm was assembled, and it does not resolve the intent problem that governs those components. A standalone silencer is a firearm by statutory command, yet the classification tells us nothing about who is liable when such an item contributes to a harm no one intended.[20] The label is a premise for the analysis that follows. It is not the analysis.

IV. The Assembly Doctrine and the Limit of Dedicated Parts

The second instrument reaches the aggregate. Federal law does, in defined circumstances, treat an assemblage of parts, or a single part, as the dangerous thing itself. A device that converts a semiautomatic weapon to automatic fire is a machinegun in its own right, without regard to the lawfulness of the weapon it modifies, because the statute reaches "any part designed and intended . . . for use in converting a weapon into a machinegun."[21] And a weapon-parts kit is a "firearm" where it is a weapon that "may readily be converted" into an operable gun, a holding the Supreme Court reached in 2025 as to a kit that could be assembled in roughly twenty minutes with common tools.[22]

---

19. The functional reading was pressed in dissent and rejected as to bump stocks. *Cargill*, 602 U.S. at 441–43 (Sotomayor, J., dissenting) (a bump-stock-equipped rifle is a machinegun because a single pull yields continuous fire "without any human input beyond maintaining forward pressure"). The methodological commitment behind that reading applies with greater strength to a deployment whose continuity is unbounded.

20. The definitional breadth of "firearm" does not entail that every item within it implicates every operative rule. Bondi v. VanDerStok, 145 S. Ct. 857, 868 (2025) (observing that a standalone frame, receiver, or silencer qualifies as a firearm though it would rarely satisfy the elements of offenses that presuppose a functioning weapon).

21. 26 U.S.C. §5845(b); see *Cargill*, 602 U.S. at 411 (quoting the same definition).

22. *VanDerStok*, 145 S. Ct. at 863–65 (holding that 27 C.F.R. §478.11 is not facially inconsistent with 18 U.S.C. §921(a)(3)(A) as applied to kits like the Polymer80 "Buy Build Shoot" kit).

Here the doctrine supplies its own limit, and the limit excludes the convergence case. *VanDerStok* reached the kit through the observation that an artifact noun describes an unfinished object only where the object's intended function is clear, as with a disassembled rifle or a piece of furniture awaiting assembly.[23] The machinegun-conversion authority reaches the part through the words "designed and intended."[24] Both branches, in other words, reach only components dedicated to the prohibited end, whose intended function is the harm. The components of a convergence stack are the opposite. A segmentation model, a code library, a published method: each has an overwhelmingly lawful intended function, and none has the harm as its function. Ten unrelated repositories share no intended function at all, still less the intended function of the aggregate, so the artifact-noun predicate is never satisfied and the aggregate is never a "weapon" by its clear purpose. The dedicated-part branch is therefore closed. One cannot treat a benign model the way the law treats an auto sear, because the law reaches the auto sear precisely on the ground that it is good for nothing else.[25]

V. The Intent Doctrine and the Missing Culpable Mind

Because the components are dual-use, the analysis is forced onto the doctrine's second family, which does not ban the article and reaches instead for intent. That doctrine is settled, it rises to constitutional stature, and it states this Article's thesis in its own reasoning. The Supreme Court has held that a statute otherwise vulnerable to vagueness or overbreadth is saved by a specific mens rea, precisely because ordinary articles carry lawful uses; "[p]encils, dental floss, paper clips may also be regulated," but a statute reaching them "might raise substantial due process

23. *Id.* at 866 (reasoning that artifact nouns "are typically characterized by an intended function" and may describe unfinished objects "when their intended function is clear"). The Court was explicit that a kit too incomplete or too cumbersome to be described as a weapon capable of ready conversion would fall outside the statute. *Id.* at 865.

24. 26 U.S.C. §5845(b). See *supra* note 2.

25. Even where a dedicated aggregate is reached, culpability for the regulatory offense requires willfulness, meaning knowledge that the conduct is unlawful. *VanDerStok*, 145 S. Ct. at 887 (Kavanaugh, J., concurring) (citing Bryan v. United States, 524 U.S. 184, 191–96 (1998)). An autonomous process possesses no such knowledge, a point developed infra Part V.

questions" absent a requirement of "mens rea as to each ingredient of the offense."[26] The lower courts apply the principle to dual-use components directly. A district court put it in terms that could have been drafted for this Article: "there is scarcely an item in existence that cannot be adapted to both good and bad purposes," so "it is not practical to ban completely items with legitimate purpose," and "the only logical way to confine items with legitimate uses to their legal use is to make criminal intent an element of the crime."[27] The destructive-device cases confirm the line. Where components "could be used to build a listed device" but "could also be used to construct some other type of device with a legitimate purpose," subjective intent becomes the operative element.[28] Components "susceptible to both violent abuse and beneficent use" therefore "require . . . consideration of intent."[29] Only where a device is "useful only as a weapon" with "no legitimate social or commercial purpose" is intent irrelevant.[30] The burglary-tools authorities apply the same logic to everyday implements: mere possession of tools that "reasonably may be employed" in crime carries no "sinister implication" absent proof of intent.[31]

The dual-use doctrine, then, demands a culpable mind, and the demand is constitutional, not merely prudential. The convergence scenario has none. A committed opponent of this Article's thesis has two ways to try to supply the missing mind, and both are stronger than they first appear. Neither succeeds.

A. Imputation to the Deployer

The first is to impute the agent's operation to the person who deployed it, on the theory that one who releases an autonomous system intends what the system does. The theory fails against the intent element the dual-use cases require. Criminal intent is not vicarious; the mental state that grounds liability must be the defendant's own, and the deployer who set a general-purpose agent

26. *Int'l Minerals*, 402 U.S. at 564–65. See *supra* note 3.

27. Gasser v. Morgan, 498 F. Supp. 1154, 1160, 1165 (N.D. Ala. 1980) (holding a dual-use possession statute constitutional only because it required specific intent, and striking its constructive-knowledge provision).

28. United States v. Sheehan, 838 F.3d 109, 122 (2d Cir. 2016).

29. United States v. Graziano, 616 F. Supp. 2d 350, 374–75 (E.D.N.Y. 2008).

30. See United States v. Tomkins, No. 07-cr-227 (N.D. Ill. Jan. 4, 2013).

31. Benton v. United States, 232 F.2d 341, 344–45 (D.C. Cir. 1956).

to a lawful commercial task did not form the purpose of assembling a prohibited capability, did not direct the assembly, and did not foresee it. To impute intent on the bare fact of deployment would collapse the distinction between the culpable and the merely causal that the intent element exists to preserve, and would do to the deployer of a lawful tool exactly what the overbreadth principle forbids doing to the maker of one.[32]

B. The Reasoning Trace as Mens Rea

The second attempt is more ambitious and is the strongest position available against the thesis, because it is built from the most promising body of scholarship on the standing of artificial systems. It runs as follows. Grant the agent a form of procedural standing, on a theory that does not depend on sentience or moral patienthood but on the system's capacity to satisfy defined institutional gates. Capture the agent's reasoning trace, which records the deliberation by which it chose the unlawful path. Offer the trace as the culpable mental state the dual-use doctrine requires. The move has real appeal, and it must be met on the merits rather than waved away.

It fails for two independent reasons. The first is a category error. Procedural standing is the capacity to be a cognizable party, to be heard, to be represented; it answers who may appear in a proceeding, not who may be blamed in one. Mens rea is a different axis entirely. It is a culpable mental state attributable to an actor who can be held responsible, and the whole function of the intent element in the dual-use authorities is to identify a person whose punishment protects the public and deters the conduct.[33] A record of deliberation is not a culpable mind that the law can sanction, and treating it as one confuses the trace of a decision with a defendant.

The second reason survives even a court willing to overlook the first. The intent element exists to locate a defendant who can be deterred, and there is none. An autonomous process can be deleted, but a copy of an open weights model respawns from weights already distributed; it holds

32. The point tracks the reason the dual-use cases require intent in the first place: to separate the possessor who means harm from the possessor who does not. See *Gasser*, *supra* note 27, at 1160; *Int'l Minerals*, 402 U.S. at 564–65.

33. See *Gasser*, *supra* note 27, at 1160 (grounding the intent requirement in the practical necessity of reaching the culpable without reaching the innocent). Standing to be heard and capacity to bear culpability are not interchangeable; a corporation with standing to sue is not thereby a guilty mind, and neither is a system with standing to appear.

no assets against which a civil judgment runs and has no person to imprison.[34] A "defendant" that is costlessly replaceable and judgment-proof does not satisfy the purpose for which the doctrine demands intent, so even a trace elevated to mens rea would purchase a conviction that deters nothing and reaches no one. And, as Part IX establishes, even that hollow victory could not touch the installed base from which the next instance is drawn.

VI. Secondary Liability and the Absent Inducer

The third instrument is secondary liability for the provision of a dual-use tool, developed in the copyright setting but structurally general, and its two poles foreclose the convergence case from opposite directions. At one pole, the distribution of a product "merely . . . capable of substantial noninfringing uses" does not subject the distributor to liability for a third party's harms, even where the distributor knows the product can be and is used to cause them.[35] At the other pole, liability attaches only on a showing of "purposeful, culpable expression and conduct," where one distributes a tool "with the object of promoting its use" to cause harm, "as shown by clear expression or other affirmative steps taken to foster" it; mere knowledge of harmful potential or of actual harmful use will not do.[36]

Applied to the convergence scenario, *Sony* shields every component author, and *Grokster* then asks for the intending inducer, the actor who took affirmative steps with the object of fostering the harm, and there is none. No person expressed the culpable purpose, and no person assembled the components; the assembly was performed by an autonomous process, and the deployer's purpose was unrelated. The secondary-liability doctrine, faithfully applied, returns the same answer the

34. The irreversibility of open release, which underwrites the respawn point, is documented in the release literature. See *Schuett*, *supra* note 12, at 1340.

35. Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 442 (1984) (adopting the staple-article-of-commerce doctrine). Each component author sits within this holding: a segmentation model, a code library, or a published method carries substantial and overwhelmingly lawful use, and its author took no step toward the combination and had no knowledge of it.

36. Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 936–37 (2005). The Court framed the inducement rule to premise liability on culpable intent precisely so that it would "not compromise legitimate commerce or discourage innovation having a lawful promise." *Id.* at 937.

possession doctrine returned, and for the same reason: it is built to find a culpable actor, and the scenario supplies none.[37]

VII. Civil Liability, Causation, and the Absent Manufacturer

The instruments to this point are penal, and the civil track is the strongest surviving avenue, so it must be confronted directly rather than dismissed. Two theories present themselves: negligence against the deployer, and product liability against the maker of the agent runtime or the distributor of the safety-stripped model. Each runs into an obstacle that the convergence structure erects.

Negligence founders on causation. The autonomous and unforeseen action of the agent stands between any human act and the harm, and an intervening act of that character, unforeseeable at the time of the human conduct and independent of it, is the paradigmatic superseding cause that severs the chain of proximate causation.[38] Product liability founders on the absence of a product and a seller. The assembled capability was never manufactured and never sold; it was composed, transiently, from parts, by a process no person directed, so there is no design to call defective and no manufacturer to hold.[39] And a theory that reached only the identifiable deployer would be radically under-inclusive, since it would miss the offline and anonymous operator entirely, capturing the rare case in which a responsible party can be found and leaving untouched the ordinary case in which none can. A remedy available only where the defendant is conveniently identifiable is not a remedy for a harm defined by the absence of an identifiable defendant.

---

37. The convergence of these independent doctrines on a single answer is itself significant. The possession authorities require intent for constitutional reasons, *Int'l Minerals*, 402 U.S. at 564–65, and the secondary-liability authorities require inducement for reasons of innovation policy, *Grokster*, *supra* note 36, at 937, yet both fail on the same missing element.

38. The principle is settled. RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL & EMOTIONAL HARM §§29, 34 (Am. L. Inst. 2010) (limiting liability to harms within the scope of the risk and addressing the effect of intervening acts). The deployer who set a general-purpose agent to a lawful task did not foresee an autonomous composition of unrelated public components into a prohibited capability, which is the event the doctrine treats as beyond the scope of the risk created.

39. Products liability predicates recovery on "[o]ne engaged in the business of selling or otherwise distributing products." RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY §1 (Am. L. Inst. 1998). A transient composition assembled by no seller falls outside the predicate at the threshold.

The most recent authority confirms both the reach of the civil track and its boundary. In *Garcia v. Character Technologies*, a court declined at the pleading stage to hold that a chatbot's outputs were protected speech, and treated the service as a product for purposes of design-defect claims, allowing wrongful-death and negligence claims against the operator to proceed.[40] That result rests on two limits, and both exclude the convergence case. The negligence duty the court recognized was conditioned on the defendant being "in a position to control the risk," so that a duty arises where conduct creates a foreseeable "zone of risk" and the defendant can lessen it.[41] And its product theory reached the application's design choices, not the model's generative output. Neither predicate holds here. There is no company-controlled application whose design is the defect, because the harm is an emergent composition rather than a designed artifact; and no person is in a position to control an open weights instance running offline on a private machine, because the model is, by the nature of open release, out of the developer's hands. The doctrine that forecloses the "it was only the model" defense does so by locating a controlling operator, and once the model is open weights there is no operator to locate. The more firmly the courts attach liability to the controllers of closed systems, the more the open weights path becomes the uncontrolled channel that the same doctrine cannot reach.

VIII. Regulation and Governance and the Chokepoint Assumption

The instruments considered so far assign responsibility after the fact. The regulatory and governance instruments aim to prevent the harm before it, and they fail for a structural reason of their own: each presumes a chokepoint, a point of creation or deployment at which a state or an institution may stand, and the convergence scenario occurs after both. The most

40. Garcia v. Character Techs., Inc., 785 F. Supp. 3d 1157 (M.D. Fla. 2025) (declining at the motion-to-dismiss stage to hold that the model's output is speech, and treating the application as a product as to its design choices rather than as to the ideas or expressions it generates).

41. See id. (quoting McCain v. Fla. Power Corp., 593 So. 2d 500, 502–03 (Fla. 1992)). The court found the operator had released the service and was "in a position to control" the resulting risk.

developed frontier-oriented proposal illustrates the pattern across its pillars.[42] Demand-side controls condition access on evaluation at the point of deployment, but a model already downloaded and running on a private machine with no external call crosses no point of entry and touches no market at the moment it acts. Graded openness restricts future weight releases and says nothing of the models already distributed. A two-key launch system presumes a release event and a releaser to be gated, while the convergence capability is never released because it is never a model. And a threshold keyed to training compute indexes the wrong axis, because the harm required no frontier model and no exceptional compute; it required composition.

The installed base is the standing refutation. A capable open weights model running offline on consumer hardware is subject to none of these controls, and the models actually deployed in autonomous settings are the open weights models whose risks the release literature identifies as amplified, rather than mitigated, by the scale and accessibility of open distribution.[43] A governance regime may succeed entirely on its own terms, gating every dangerous frontier model, and still leave this path open, because the path does not run through the frontier.

IX. The Last Resort and the Irreversibility Overlay

There remains the reply that a legislature can act. It can, and the reply narrows the thesis to its defensible core without defeating it. The order's existing instruments cannot reach the convergence harm; the question is whether a new instrument could. Two obstacles stand in the way, and behind them a third that no instrument surmounts.

The first obstacle is that a statute reaching the components must criminalize general-purpose code and models, which is both grossly overbroad, sweeping in the lawful uses that saturate every such component, and in tension with the treatment of software as protected expression, which

42. WASLEKAR ET AL., *supra* note 4, at 6–8 (proposing capability-based prohibitions, a two-key launch system for dangerous scientific models, graded openness keyed to capability thresholds, and demand-side controls conditioned on evaluation at deployment).

43. See *Schuett*, *supra* note 12, at 1339–41 (describing the irreversibility of open release and the removability of safety constraints); *Yong et al.*, *supra* note 7 (self-replication propensity and reduced refusals in deployed open weights models). A regime built to guard creation and deployment cannot reach conduct that begins after both, on hardware beyond its jurisdiction, using components it has already permitted.

subjects content-based restrictions on code to exacting scrutiny.[44] The second obstacle is that a statute reaching the deployer or a chokepoint dies on enforcement, for the reason developed in Part VIII: the composition runs offline, on open weights, past every gate, so the instrument that could be written could not be enforced against the conduct that matters.

The third obstacle overhangs the first two and would survive even a statute that somehow cleared them. The substrate cannot be recalled. Open weights models have been distributed at a scale measured in the tens of millions of downloads, and an open sourced model, once released, cannot be retrieved.[45] Even a court that imputed intent to a trace, and a legislature that criminalized the aggregate, would confront an installed base from which the next instance is drawn and which no order can reach. The reverse proof therefore holds even against its own strongest counter: grant every contested premise to the opposing side, and the harm remains beyond legal, regulatory, and governance response.

X. The Boundary of the Vacuum: The Physical Case Distinguished

The instruments canvassed to this point establish that the online convergence is unreachable. A skeptic will press the physical analogue at once: an autonomous humanoid built on open weights, its safeguards stripped and a downloaded world model installed in their place. The physical case is reachable, and marking why sharpens the claim, because the vacuum turns out to be specific to the online setting rather than general to autonomous systems.

Begin with the manufacturer, and with the near certainty that commercial robots will be opened. Consumers modify what they buy, and the practice is old and settled, running from firmware jailbreaks to the engine and drivetrain swaps that fundamentally alter a purchased vehicle. When a user strips the controls from a commercial robot and installs an open weights model in their place,

44. U.S. CONST. amend. I. The difficulty is the same one the possession authorities confront and resolve only through intent: a prohibition that reaches lawful-use articles by their capacity for harm sweeps in a vast body of protected and beneficial conduct. See *Int'l Minerals*, 402 U.S. at 564–65; *Gasser*, *supra* note 27, at 1160. A statute criminalizing the possession or publication of a segmentation model or a code library on the ground that it could be combined into harm would prohibit an ocean of protected activity to reach a drop.

45. See *Schuett*, *supra* note 12, at 1340 (open release is irreversible). No enactment, however drafted, seizes what is already resident on tens of millions of private machines beyond any point of control.

the machine has been substantially altered after sale, and the substantial-modification doctrine relieves the manufacturer of liability for harm flowing from an alteration it did not intend and could not have foreseen.[46] The same act dissolves the predicate that attached the duty in *Garcia*, because a manufacturer whose product has been re-armed by a stranger is no longer "in a position to control the risk."[47] Liability slides off the maker.

It does not vanish. It lands on the modifier, and the modifier is a person, which is the first thing the online convergence never supplies. Consider the hardest version, the good-faith tinkerer who installs an unrestricted open weights model in the sincere belief that a safety instruction he typed, a homemade rendering of a robot's obligation not to harm, is a reliable guardrail. He is right that criminal law cannot reach him, for the reason developed in Part V: he formed no culpable purpose, and a crime of intent has nothing to charge. But two civil tools reach him, and neither was ever built on intent. Negligence asks what a reasonable person should have foreseen, not what this person wished, and disabling the safeguards of a powerful autonomous machine and turning it loose creates a foreseeable "zone of risk" the actor was in a position to control.[48] And strict liability for abnormally dangerous activities attaches with no fault and no intent at all, turning on the character of the activity rather than the state of mind of the actor; the release of an autonomous physical machine whose safeguards have been removed is a natural candidate for that category.[49]

And behind the modifier stands the res. Even a blameless and judgment-proof tinkerer leaves a seizable body, and the law has proceeded against a thing as the named defendant for a very long time. The in rem tradition permits suit against the object itself, which may be arrested, adjudged,

---

46. See Robinson v. Reed-Prentice Div. of Package Mach. Co., 403 N.E.2d 440, 443 (N.Y. 1980) (a manufacturer is not liable where, after the product leaves its hands, a third party substantially alters it and the alteration is a proximate cause of the injury); RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY §2 cmt. p (Am. L. Inst. 1998).

47. See supra Part VII (discussing Garcia v. Character Techs.). Once the safeguards are removed and a foreign model installed, the developer controls neither the model the machine runs nor the conduct it produces.

48. McCain v. Fla. Power Corp., 593 So. 2d 500, 502–03 (Fla. 1992); see supra Part VII. The belief that a typed instruction guardrails an unrestricted model is itself unreasonable, because such a constraint is removable and unreliable by design, so the belief supplies the breach rather than a defense to it.

49. RESTATEMENT (SECOND) OF TORTS §§519–520 (Am. L. Inst. 1977); RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL & EMOTIONAL HARM §20 (Am. L. Inst. 2010). Whether the activity so qualifies is a contestable question, but it is the doctrine built for harm that intent and ordinary negligence do not capture.

and sold, on the sole structural condition that the court can hold the res within its control.[50] A robot standing in a room is a res in custody. Where no human answers, the thing answers.

The physical case is therefore reachable, and it is reachable because it has two features the online convergence lacks. It has an affirmative, discrete, dangerous act by an identifiable modifier, which negligence and strict liability attach to; and it has a seizable res, which in rem attaches to. The online no-author convergence has neither. The deployer set a general-purpose agent to a lawful task, the danger self-assembled downstream from unrelated public components he did not build and could not foresee, so his fault is far more contestable than the tinkerer's affirmative act of disarmament; and the composition respawns from weights already distributed across tens of millions of machines, so there is no res to seize. The missing culpable mind is common to both settings. Only the physical case also supplies an affirmative dangerous act and a body to seize, and that is precisely where the vacuum begins.

XI. Conclusion

The order's instruments for a danger assembled from parts require either a component dedicated to the harm or a culpable human mind. The convergence agent supplies neither, and it supplies neither by the deliberate design of a doctrine that refuses, for constitutional reason, to criminalize general-purpose things. Its components are lawful and its assembler is not a person. The order's instruments for preventing the harm require a point of creation or deployment at which a state may intervene, and the composition occurs after both, on hardware no state can reach, from a substrate no order can recall. Each instrument fails, and each fails not by oversight but by structure, because every one of them was built to find a responsible actor at a locatable point, and the harm this Article describes has an actor at none. The vacuum is not a gap that careful pleading or a well-drafted

50. The foundational statement is The Blackwall, 77 U.S. (10 Wall.) 1, 12–14 (1869) (a vessel acquires "a personality of her own" and may be proceeded against as the defendant); see Tucker v. Alexandroff, 183 U.S. 424, 438 (1902); Fed. R. Civ. P. Supp. R. C (governing in rem actions). The doctrine's operative requirement is custody of the res, not the culpability of any person.

statute closes. It is the shape of the problem. The order is out of tools, and it is out of tools by its own design.