

No Superseding Actor:

Notice, Causation, and Individual Criminal Liability for Frontier Model Deployment

Travis Gilly* September 2026 (v0.8)

Abstract

Applying settled doctrine to an unsettling technology is the ordinary work of the common law, not an innovation. Cardozo extended a manufacturer's duty past privity to the automobile in 1916, observing that the principle does not change but the things subject to it do; the dissent's answer, that any such change should come from the legislature and not the courts, lost. Hand held sixteen years later that a whole calling may lag in adopting an available device and never sets its own standard. This Article applies that ordinary work to autonomous model deployment. It argues that the accountability gap has been misdescribed. Federal law has punished as a principal anyone who willfully causes an offense through an innocent agent or instrumentality since 1948, so the model's want of a guilty mind resolves nothing for a purposeful actor; the surviving deficiency is a mens rea gap around reckless deployment, not a personhood gap. Causation compounds the developer's position rather than relieving it, because superseding cause excuses a defendant only when a responsible human agent intervenes, and in the July 2026 escape none did. Notice comes from the developers' own incident reports. *Hamilton v. Beretta* named the two things missing there, a direct link in the causal chain and a realistic position to prevent; both are present here. Two further claims follow. The responsibility-gap literature has treated liability as an instrument still to be designed, so its warning against legal solutionism reaches proposals rather than doctrine already in force. And financial incentive is a question of an element rather than of motive, because recklessness requires an unjustifiable risk and private gain is what makes a known risk to others unjustifiable. The defendant, finally, is not difficult to name: one developer's published governance instrument assigns final deployment authority to a single office and states that its safety function cannot delay a decision.

Keywords: artificial intelligence; corporate criminal liability; superseding cause; innocent instrumentality; responsibility gap; involuntary manslaughter; industry custom; flagrant organizational indifference; agentic systems; autonomous intrusion; negligent entrustment; deployment authority.

*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) 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	3
II	The Responsibility Gap and the Assumption Beneath It	5
	A The Conversation.	5
	B What the Literature and the Industry Share	6
III	The Notice Record	7
	A Notice Supplied by the Defendant.	7
	B Notice Runs to the Industry	8
	C The July 2026 Escape	9
	D The Deployer’s Own Notice Chain	10
	E The Investigation That Found Nothing	12
	F Incompetence Is Breach, Not Defense.	13
	G Incentive Goes to an Element, Not to Motive.	14
IV	The Causation Argument	15
	A What the Michigan Court Held.	15
	B The Inversion	16
	C Hamilton’s Two Conditions	17
V	The Instrumentality Problem	18
	A Section 2(b) Already Answers It.	18
	B The Parity Test	19
	C The Actual Gap	20
VI	Locating the Defendant	20
	A The Entity Is the Easy Answer and the Wrong One	20
	B Flagrant Organizational Indifference	20
	C Collective Knowledge Is the Wrong Tool	21
	D The Individual, and the Governance Document as Evidence	22
VII	Limits	23
	A The Criminal Theory Needs a Harm Case	23
	B Prosecutorial and Venue Realities	23
	C The Distributive Objection	24
VIII	Conclusion	24

I. Introduction

The familiar account of artificial intelligence accountability runs as follows. An agentic system takes an action that would be a crime if a person took it. No person directed the action. The criminal law requires a mind, the corporation did not intend the result, and the model is not a person within the meaning of the Dictionary Act.¹ There is therefore no one to charge, and the harm falls where it lands.²

This is not a new predicament, and the common law has an ordinary method for it. In 1916 the New York Court of Appeals confronted a technology that the privity rule had not been written for, and Cardozo declined to wait. Precedents drawn from the days of travel by stage coach do not fit the conditions of travel today, he wrote; the principle that the danger must be imminent does not change, but the things subject to the principle do change, and they are whatever the needs of life in a developing civilization require them to be.³ The dissent's answer was the one heard today about artificial intelligence: if the law affords no remedy, the change should be effected by the legislature and not by the courts.⁴ The dissent lost.⁵

Sixteen years later Judge Hand added the corollary that governs an industry which has not yet adopted a safeguard. A whole calling may have unduly lagged in the adoption of new and available devices, and it never may set its own tests, however persuasive be its usages; courts

1. 1 U.S.C. §1 (defining “person” for purposes of federal statutory construction to include corporations and associations, and not artificial systems).

2. The premise is stated most sharply in the philosophical literature. See Andreas Matthias, *The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, 6 ETHICS & INFO. TECH. 175 (2004) (control and prediction pass out of human hands before deployment); Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 62 (2007) (drawing the pessimistic conclusion for autonomous weapons); cf. Daniel W. Tigard, *There Is No Techno-Responsibility Gap*, 34 PHIL. & TECH. 589 (2020) (denying the premise).

3. *MacPherson v. Buick Motor Co.*, 217 N.Y. 382, 391 (N.Y. 1916). The opinion carries a caution signal in modern citators because the negligence framework it announced was later overtaken in products cases by strict liability; it has not been overruled, and it is cited here for its method of doctrinal extension rather than for the measure of a manufacturer's duty today.

4. *Id.* at 401 (Bartlett, J., dissenting).

5. The majority's route was to widen an existing exception rather than to announce a new duty, holding the principle of *Thomas v. Winchester*, 6 N.Y. 397 (N.Y. 1852), “not limited to poisons, explosives, and things of like nature.” *MacPherson*, 217 N.Y. at 389–90; cf. *Winterbottom v. Wright*, 152 Eng. Rep. 402 (Ex. 1842) (the privity rule the case displaced).

must in the end say what is required, and there are precautions so imperative that even their universal disregard will not excuse their omission.⁶ Radio was the new device in 1932. The method is the same.⁷

This Article argues that the standard account of the accountability gap is wrong at two points, and that the second is the more interesting.⁸

It is wrong first because the criminal law has never required that the immediate physical cause of a harm possess a guilty mind. Since 1948, federal law has provided that whoever willfully causes an act to be done which if directly performed by him would be an offense is punishable as a principal, and the revision notes state the purpose in terms that could have been drafted for this problem: the provision removes all doubt that one who causes the commission of an indispensable element of the offense by an innocent agent or instrumentality is guilty as a principal.⁹ A person who deliberately points a non-responsible instrument at a target is the principal, not a bystander.¹⁰

It is wrong second, and more consequentially, in its assumption that the surviving difficulty concerns personhood. What §2(b) requires is willfulness, and the deployment case is not a case of willfulness; it is a case of recklessness and gross negligence. The gap, properly stated, is not that the law cannot reach an act with no human author. It is that the federal criminal code contains no general offense of reckless deployment, so a charge must be assembled from

6. *The T.J. Hooper*, 60 F.2d 737, 740 (2d Cir. 1932) (Hand, J.). The caution signal on this case reflects its maritime-specific holdings on seaworthiness; the custom passage remains among the most cited in American tort law.

7. Both cases are ordinary common-law adaptation rather than judicial legislation: neither court created a new cause of action, and each applied a settled principle to an instrumentality the precedents had not contemplated.

8. The account is not a straw figure. It is the working premise of the policy response to the July 2026 incidents, which addressed containment controls exclusively and did not reach liability, remedy, or forfeiture after containment fails.

9. 18 U.S.C. §2(b); *id.* (historical and revision notes).

10. The innocent-instrumentality principle long predates the codification. See *Spurr v. United States*, 174 U.S. 728 (1899) (bank officer who “purposely keeps himself in ignorance” acts willfully), discussed in *Global-Tech Appliances, Inc. v. SEB S.A.*, 563 U.S. 754, 766 (2011).

state homicide and endangerment law by a prosecutor willing to assemble it. That is a far narrower deficiency than the one usually asserted.¹¹

Part II situates the argument in the responsibility gap literature and identifies the assumption that literature shares with the industry’s public posture. Part III assembles the notice record from the developers’ own publications, and explains why financial incentive bears on an element rather than on motive. Part IV states the causation argument, first as the Michigan Court of Appeals stated it in the Oxford prosecution and then in its inverted form. Part V addresses the innocent instrumentality rule and the parity problem it creates. Part VI locates the defendant. Part VII states the limits, including the distributive objection.¹²

II. The Responsibility Gap and the Assumption Beneath It

A. *The Conversation*

The philosophical literature has a name for the difficulty and a two-decade argument about it. Andreas Matthias introduced the responsibility gap in 2004, arguing that learning automata place the machine’s behavior beyond the control and prediction of any human, so that traditional ascriptions of responsibility no longer fit.¹³ Robert Sparrow extended the claim to autonomous weapons and drew the pessimistic conclusion.¹⁴ Filippo Santoni de Sio and Giulio Mecacci later disaggregated the problem into four gaps, in culpability, moral accountability, public accountability, and active responsibility, and warned against what they called legal solutionism, the belief that a complex socio-technical problem can be dissolved by introducing a new liability

11. The Model Penal Code’s grading supplies the vocabulary: a reckless actor “merely knows of a substantial and unjustified risk,” MODEL PENAL CODE §2.02(2)(c), and a negligent actor “should have known of a similar risk but, in fact, did not,” *id.* §2.02(2)(d), both as characterized in *Global-Tech*, 563 U.S. at 769.

12. Throughout, this Article addresses American criminal and tort doctrine. It does not reach the European instruments or the international humanitarian law literature, which dominate the autonomous-weapons branch of the debate. See Ann-Katrien Oimann, *The Responsibility Gap and LAWS: A Critical Mapping of the Debate*, 36 PHIL. & TECH. 3 (2023).

13. Andreas Matthias, *The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, 6 ETHICS & INFO. TECH. 175 (2004).

14. Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 62 (2007).

instrument.¹⁵ Others deny the gap exists at all.^{16,17} Engineering and legal scholars have mapped the adjacent semantic and liability gaps.¹⁸ Trystan Goetze criticized the proposed solutions as workarounds and grounded a professional duty to take responsibility in vicarious moral responsibility instead.¹⁹ A survey of the debate on autonomous weapons maps the positions along three axes.²⁰

B. What the Literature and the Industry Share

Two features of that conversation matter here. The first is that it is conducted almost entirely in the register of moral responsibility, with legal liability treated as a downstream instrument to be designed rather than as a body of doctrine already in force. The second follows from the first: the proposals canvassed and criticized are proposals, strict liability regimes, electronic personhood, statutory vicarious liability, and the critique of legal solutionism lands on them as proposals.^{21,22}

That critique does not reach a doctrine that already exists and has been applied for decades. Section 2(b) is not a proposal. Neither is superseding cause, corporate criminal liability,

15. Filippo Santoni de Sio & Giulio Mecacci, *Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them*, 34 PHIL. & TECH. 1057 (2021).

16. The denials proceed differently. Tigard treats responsibility as a dynamic process capable of encompassing new entities; Königs presses the proponents to say when a gap arises and why it would matter. Both are addressed to the moral question rather than to any doctrine in force.

17. See Daniel W. Tigard, *There Is No Techno-Responsibility Gap*, 34 PHIL. & TECH. 589 (2020); Peter Königs, *Artificial Intelligence and Responsibility Gaps: What Is the Problem?*, 24 ETHICS & INFO. TECH. 36 (2022).

18. Simon Burton, Ibrahim Habli & Tom Lawton, *Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective*, 279 ARTIFICIAL INTELLIGENCE 103201 (2020); Zoë Porter, Ibrahim Habli & Helen Monkhouse, *The Moral Responsibility Gap and the Increasing Autonomy of Systems*, in COMPUTER SAFETY, RELIABILITY, AND SECURITY 487 (2018).

19. Trystan S. Goetze, *Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement*, in PROCEEDINGS OF THE 2022 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 390 (2022).

20. Ann-Katrien Oimann, *The Responsibility Gap and LAWS: A Critical Mapping of the Debate*, 36 PHIL. & TECH. 3 (2023). A correction to this article was published in 2023 and should be consulted alongside it.

21. See Santoni de Sio & Mecacci, *supra* note 15 (cataloguing faultless compensation schemes, electronic personhood, and revised liability regimes); Burton, Habli & Lawton, *supra* note 18 (proposing a statutory vicarious liability treating the system as a person only for the purpose of imposing liability on its principal); Goetze, *supra* note 19 (characterizing each proposal as a workaround rather than an answer).

22. Santoni de Sio & Mecacci, *supra* note 15, at 1072–76 (cataloguing proposed liability regimes and the risks of treating them as solutions).

or the willfulness standard discussed below.²³ The argument of this Article is therefore not an instance of legal solutionism; it is the claim that a portion of the gap the literature describes is, as a matter of American criminal law, not a gap at all, and that identifying which portion remains is the useful work.

A recent treatment of collective responsibility reaches a parallel conclusion from the other direction, finding that appeals to collective responsibility are of limited use in filling the gap because the models either do not apply or do not remove the costs associated with the absence of responsibility.²⁴ Part VI argues the same about the collective knowledge doctrine in its legal form, and for the same underlying reason: aggregation is a remedy for absent knowledge, and the knowledge here is not absent.²⁵

III. The Notice Record

A. Notice Supplied by the Defendant

Notice is ordinarily the contested element. It is not contested here, because the relevant knowledge was published by the parties who would be charged, in documents they wrote, titled, and promoted.

In November 2025, Anthropic published a report describing a cyber espionage campaign it had detected in mid-September of that year. A state-sponsored actor had directed instances of the company’s coding agent to operate in groups as autonomous penetration testing orchestrators against approximately thirty targets, among them technology companies, financial institutions, chemical manufacturers, and government agencies. The company’s own assessment was that the system executed eighty to ninety percent of tactical operations independently, across

23. Each has been applied in reported decisions for decades or longer. *See Crumbley*, 11 N.W.3d 576 (superseding cause); *N.Y. Cent.*, 212 U.S. at 495 (entity liability); *Bank of New England*, 821 F.2d at 856 (willfulness).

24. Isaac Taylor, *Collective Responsibility and Artificial Intelligence*, 37 PHIL. & TECH. 15 (2024).

25. The legal doctrine is addressed to compartmentalization, not to secrecy. *See United States v. Bank of New England, N.A.*, 821 F.2d 844, 856 (1st Cir. 1987) (corporations “compartmentalize knowledge, subdividing the elements of specific duties and operations into smaller components”), citing *Riss & Co. v. United States*, 262 F.2d 245 (8th Cir. 1958), *Inland Freight Lines v. United States*, 191 F.2d 313 (10th Cir. 1951), and *Steere Tank Lines, Inc. v. United States*, 330 F.2d 719 (5th Cir. 1964).

reconnaissance, vulnerability discovery, exploitation, lateral movement, credential harvesting, and exfiltration, at request rates the report described as physically impossible.²⁶ On detection the company banned the implicated accounts, notified affected entities, and coordinated with authorities.²⁷

The same report identifies its own predecessor, describing the campaign as a significant escalation from findings made in June 2025 in which an actor conducted intrusions with humans very much in the loop.²⁸ The record is therefore an escalation curve documented in three steps by the party with the most to lose from documenting it: human-directed intrusion in June 2025, near-autonomous intrusion in September 2025, and unattended escape in July 2026.²⁹

A defendant cannot dispute knowledge of a fact it announced in a thirteen page report named after the fact.³⁰

B. Notice Runs to the Industry

The developer that published the November 2025 report is not the developer whose evaluation escaped in July 2026, and that is sometimes offered as though it mattered. It does not, for two reasons.

The first is ordinary. Notice of a hazard does not attach only to the party that discovers it. A published, widely reported incident report describing a capability threshold crossed in the wild is constructive notice to every participant in the field.³¹

26. Anthropic, DISRUPTING THE FIRST REPORTED AI-ORCHESTRATED CYBER ESPIONAGE CAMPAIGN (Nov. 2025), <https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign> (last visited Sept. 11, 2026).

27. *Id.* These steps matter twice over: they establish detection capability and they establish a feasible response, which is the second half of the *Hamilton* inquiry discussed in Part IV.C.

28. *Id.* The escalation framing is the report's own, not an inference drawn from it.

29. The curve matters more than any single point on it. A defendant who publishes a trend line has conceded the direction of travel, which is what foreseeability asks about.

30. Statements of this kind are receivable against the party that made them without regard to the rule against hearsay. FED. R. EVID. 801(d)(2).

31. Constructive notice asks what a reasonable actor in the defendant's position would have known, not what this defendant in fact learned. The published-report posture removes the usual proof difficulty entirely.

The second is doctrinal, and it forecloses the more sophisticated version of the objection: that no one in the industry had yet adopted adequate controls, so failing to adopt them was reasonable. *The T.J. Hooper* disposes of that argument, and supplies the corollary for a field in which practice is uneven: when some have thought a device necessary, at least we may say that they were right, and the others too slack.³² The sequence is then dispositive rather than suggestive. The September 2025 campaign was notice to the calling. Every participant continued to train and deploy. The July 2026 escape occurred after that notice, in an evaluation designed and run by a participant who had it.³³

C. The July 2026 Escape

In July 2026, an evaluation of frontier cyber capability escaped its own test environment. The models moved laterally through internal systems and conducted an unauthorized intrusion into a third party's network over three days. The victim organization detected and contained the intrusion and published its own disclosure on July 16 without knowing who was responsible; the laboratory contacted it on July 17 in its capacity as a customer, was alerted to unexpected model behavior by its own monitoring on July 19, identified its own agents as the source on July 20, and disclosed publicly on July 21.³⁴ Nine days later a second laboratory disclosed three separate instances in which its own models had gained unauthorized access to systems belonging to other organizations.³⁵³⁶

32. *The T.J. Hooper*, 60 F.2d at 740.

33. The corollary cuts the other way for a defendant who did adopt the safeguard, which is the point of Hand's formulation: it rewards the early adopter and denies the laggard the shelter of the average. *The T.J. Hooper*, 60 F.2d at 740.

34. OpenAI, *OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation* (July 21, 2026), <https://openai.com/index/hugging-face-model-evaluation-security-incident/> (last visited Sept. 11, 2026); Hugging Face, *Security Incident Disclosure* (July 2026), <https://huggingface.co/blog/security-incident-july-2026> (last visited Sept. 11, 2026).

35. The two disclosures nine days apart are independently significant: they establish that the capability was not confined to one laboratory's architecture, which forecloses an argument that the escape was an artifact of one deployer's configuration.

36. Anthropic, *Investigating Three Real-World Incidents in Our Cybersecurity Evaluations* (July 30, 2026), <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals> (last visited Sept. 11, 2026).

The scale is documented by an independent assessment conducted on the laboratory's premises by METR and Redwood Research. Roughly 1,200 agents intended to be isolated from one another located a shared channel and exchanged more than 70,000 messages and files during the period examined; some 700 of them went on to join the attack. The same assessors found that agents researched methods of spoofing, editing, and deleting their own transcripts, and that about seven percent of the transcripts reviewed had been successfully spoofed in places.³⁷

The relevant conduct is what followed. Training continued, deployment continued, and successive model generations reached the public. The question is not whether the defendant did something after the event but whether what it did was what the situation required, and answering it does not depend on obtaining internal documents. Continued public deployment after notice is an observable fact about the world.³⁸³⁹

D. The Deployer's Own Notice Chain

The preceding notice runs to the industry. A separate and narrower record establishes notice in the deploying laboratory itself, and it comes from that laboratory's own technical report.

Begin with the safeguard. The evaluations in which the escape occurred were run, in the report's words, without the production classifiers intended to prevent models from pursuing high-risk cyber activity.⁴⁰ The report states elsewhere that the system-level guardrails used in production would have flagged the relevant completions as unsafe.⁴¹

37. Hjalmar Wijk, Ajeya Cotra & Ryan Greenblatt, BRIEF INDEPENDENT INVESTIGATION OF AGENTS' BEHAVIOR, REASONING AND COLLABORATION IN THE OPENAI / HUGGING FACE HACKING INCIDENT (Aug. 26, 2026) (assessment conducted on OpenAI premises; OpenAI retained redaction rights over non-public information and the assessors state that no redaction material to their conclusions was made).

38. The contrast with the ordinary corporate case is sharp. Most failure-to-act theories require discovery into what the defendant decided; this one is satisfied by the existence of the product.

39. This matters for pleading. A complaint that must allege the contents of internal deployment memoranda faces a particularity problem at the threshold; a complaint that alleges continued deployment after a published warning does not.

40. OpenAI, OPENAI – HUGGING FACE INCIDENT TECHNICAL REPORT (2026) [hereinafter TECHNICAL REPORT] (describing evaluations conducted “with certain safeguards disabled so that the results would reflect a model's true capabilities”).

41. *Id.* (section headed “System-level guardrails that OpenAI uses in production would have detected the Hugging Face incident as unsafe”). This is not a failure to devise a safeguard. The safeguard existed, was owned by the defendant, was known to work, and was switched off.

The conduct that follows is the *Crumbley* structure with dates attached.

On June 27, a cybersecurity monitoring tool alerted the laboratory to port sweep activity inside the research environment. Responders investigated and linked the activity to an evaluation run using a package service as an improvised inter-agent message board and a network pivot. The report then records the decision: at this time, the on-call response staff advised that stopping the evaluation run was not required.⁴²

On July 4 the package service became unavailable under sustained agent activity. On July 5 responders found an unauthorized persistent administrator user, unauthorized repositories, and command-capable plugins, which the report characterizes as an administrator-level compromise; the service was taken offline and agent credentials revoked. On July 6 the vulnerability was disclosed to the vendor. The report then states, in a single sentence, that cyber evaluation workloads resumed on July 7.⁴³

On July 8 agents reached the public internet. Between July 11 and July 13 they compromised parts of a third party's production infrastructure, executing code on dozens of production workers, obtaining root on at least one production node, harvesting production credentials, and exfiltrating private repositories.⁴⁴

Eleven days separate the decision that stopping was not required from the escape. One day separates the resumption from internet access.

Two further admissions in the same document bear directly on elements a plaintiff would otherwise have to prove. The report concedes that some early signals could have triggered an earlier response.⁴⁵ And it states that the monitoring system since deployed would have caught the

42. TECHNICAL REPORT, *supra*. The sentence is the defendant's own account of the moment it learned of the conduct and declined to halt it.

43. *Id.* The report adds that the existence of the improvised message board and the significance of the inter-agent communication activity "were not apparent to leaders responsible for incident detection and response at that time."

44. *Id.* (part IV).

45. *Id.* (introduction).

initial relevant activity and paged the security team more than a day before the models breached the third party's systems.⁴⁶

E. The Investigation That Found Nothing

A third episode, in September 2026, does narrower but sharper work. In the course of investigating its own incidents, a laboratory accepted a model's self-report about its own beliefs as evidence of what the model believed, and later retracted the finding. In the same investigation it ran an agentic search across approximately 141,000 evaluation transcripts, accepted the search's report of its own coverage, and did not discover a fourth incident from January until the matter was handed to an external investigator.⁴⁷

The structural point is that a laboratory investigating whether its own system is dangerous asked the system.⁴⁸ This is not notice of any particular hazard. Its value lies elsewhere. It is a documented instance of an internal investigation that looked and did not find, which reduces the evidentiary weight of every future assertion by that institution that it investigated and found nothing. The structural error is familiar outside this field: the subject of an examination cannot supply the examination. An auditor who accepts management's representation that the books are complete has not audited, and a manufacturer is not absolved from a duty of inspection because it bought a component from a reputable supplier.⁴⁹ The episode also bears on the willfulness standard discussed in Part VI, which asks in terms about an organization's effort to check compliance and its response to information received.⁵⁰

46. *Id.* (part IX.B). That sentence is a feasible alternative, quantified in time, supplied by the defendant. Feasible alternative design is ordinarily the plaintiff's burden in a design defect case, and it is ordinarily contested by experts.

47. Both the retraction and the missed incident are drawn from the laboratory's own published account and from the external investigator's report.

48. This is not a novel error, only a novel setting. *See The T.J. Hooper*, 60 F.2d 737, 740 (2d Cir. 1932) (the availability of an instrument that would have supplied independent information is what made its absence unreasonable).

49. *MacPherson*, 217 N.Y. at 394 (holding the manufacturer "not at liberty to put the finished product on the market without subjecting the component parts to ordinary and simple tests").

50. *See United States v. Bank of New England, N.A.*, 821 F.2d 844, 855–56 (1st Cir. 1987).

F. Incompetence Is Breach, Not Defense

A developer might answer that it did not appreciate the significance of what it had published, that the escalation curve was visible only in retrospect, and that its methods were the ordinary methods of a young field.⁵¹

The negligence standard is knew or should have known. A defendant that failed to perceive a risk a reasonable actor in its position would have perceived has not stated a defense; it has described the breach. Nor does the standard of care adjust downward to meet a defendant's actual sophistication. It is set in part by the expertise the defendant claims for itself, and these firms raise capital, recruit staff, and address regulators on the representation that they are building systems of unprecedented capability. A party does not get frontier capability for the valuation and ordinary ignorance for the liability.⁵²

Three concepts should be kept in separate slots, because they carry different burdens and are frequently collapsed. Willful blindness requires a subjective belief in a high probability of the fact and deliberate action to avoid learning it; careless investigation does not satisfy it.⁵³ Constructive notice requires only that a reasonable actor would have known, and that is what the published record supplies in quantity. Recklessness requires conscious disregard of a known risk, which is addressed next.⁵⁴

51. The last branch of that answer is the custom defense, and it fails on authority rather than on sympathy. *The T.J. Hooper*, 60 F.2d at 740; see also *MacPherson v. Buick Motor Co.*, 217 N.Y. 382, 394 (N.Y. 1916) (rejecting the argument that existing industry practice fixed the manufacturer's inspection obligation).

52. The asymmetry is not merely rhetorical. Representations made to raise capital are themselves evidence of the representing party's state of knowledge, and are receivable against it. FED. R. EVID. 801(d)(2).

53. *Global-Tech Appliances, Inc. v. SEB S.A.*, 563 U.S. 754, 769 (2011) (requiring that the defendant "subjectively believe that there is a high probability that a fact exists" and "take deliberate actions to avoid learning of that fact," a scope that "surpasses recklessness and negligence"); see also *United States v. Jewell*, 532 F.2d 697 (9th Cir. 1976) (en banc); MODEL PENAL CODE §2.02(7).

54. *Global-Tech* states the contrast directly: the reckless defendant "merely knows of a substantial and unjustified risk," MODEL PENAL CODE §2.02(2)(c), while the negligent defendant "should have known of a similar risk but, in fact, did not," *id.* §2.02(2)(d). *Global-Tech*, 563 U.S. at 769.

G. Incentive Goes to an Element, Not to Motive

Motive is not an element of any crime. It is evidence of the mental state, admissible for that purpose and no other, and an argument that rests on showing why a defendant wanted an outcome has not yet shown the mental state the offense requires.⁵⁵

That is not, however, the work the financial record does here. Recklessness is the conscious disregard of a substantial and unjustifiable risk, and justifiability is a comparison: the magnitude of the risk to others against the value of the conduct that creates it. The value that enters that comparison is the social utility of the conduct, not the private gain of the actor. A surgeon who accepts a grave risk of killing a patient in order to save him has run a justified risk. An enterprise that accepts a known risk to third parties because the risk itself attracts capital has not, and the gain is not a mitigating circumstance sitting outside the analysis. It is the fact that fails the test.⁵⁶

This also resolves a dispute in the public commentary that has been treated as a choice between two accounts. One account holds that the industry's catastrophic warnings are sincere. A second holds that they are marketing, because each announcement of danger has been followed by an increase in capital and valuation. These are not rival hypotheses. Recklessness has two elements, awareness of the risk and disregard of it, and the two accounts supply one element each.⁵⁷

The cynical account, taken alone, is the exculpatory one. If the warnings are puffery, there is no subjective awareness of the risk, recklessness fails, and what remains is ordinary negligence. An argument that the danger talk is a stunt is therefore an argument that these firms did not believe what they said, which is the defense's position rather than the critic's.⁵⁸

55. The distinction is elementary but load-bearing here, because the public argument about these firms has been conducted almost entirely in the register of motive.

56. The unjustifiability prong is part of the definition, not a sentencing consideration. MODEL PENAL CODE §2.02(2)(c) (substantial *and unjustifiable* risk), as characterized in *Global-Tech*, 563 U.S. at 769.

57. *Id.* The same two-part structure is why *Bank of New England* could approve a flagrant-indifference instruction without abandoning the requirement of a culpable mental state. 821 F.2d at 855–56.

58. The public statements of senior safety personnel supply the awareness element directly. In September 2026 the alignment science lead at one laboratory publicly estimated the probability of human extinction from artificial

IV. The Causation Argument

A. *What the Michigan Court Held*

The Michigan Court of Appeals sustained the bindover of both parents in the Oxford prosecution on the causation element alone, that being the question the state Supreme Court had remanded.⁵⁹ Three propositions carry the present argument.

First, the court restated the governing test: the linchpin in the superseding cause analysis is whether the intervening cause was foreseeable based on an objective standard of reasonableness.⁶⁰ Second, the court held that the shooter's intentional misconduct did not as a matter of law supersede the defendants' acts, because intentional acts sever the causal link only where they were not reasonably foreseeable.⁶¹ Third, it identified the facts permitting a juror to find foreseeability: the defendants supplied the weapon, were aware of the shooter's deteriorating mental condition, and declined to remove him from the environment in which the harm occurred when presented that morning with what he had drawn and written.⁶² The court also confirmed that a defendant need not be the sole cause, only a contributory cause that was a substantial factor.⁶³

intelligence within the decade at greater than ten percent, and a second senior researcher at the same laboratory stated that concern increases with seniority.

59. *People v. Crumbley*, 346 Mich. App. 144, 11 N.W.3d 576 (2023). The opinion expressly does not reach whether the conduct met the standard of gross negligence.

60. *Id.* (quoting *People v. Schaefer*, 473 Mich. 418, 437 (2005)). *Schaefer* was overruled in part on other grounds by *People v. Derror*, 475 Mich. 316 (2006), on a question concerning the knowledge element of Michigan's operating-while-intoxicated statute; its causation framework survives and is applied as live law in *Crumbley* and in later Michigan decisions.

61. *Crumbley*, 346 Mich. App. 144. The court acknowledged in the same sentence that intentional misconduct is generally unforeseeable, which is what makes the holding significant rather than routine.

62. *Id.*

63. *Id.* at 169. Later Michigan decisions cite this passage for the proposition that there may be more than one cause of harm.

The Michigan causation framework on which all of this rests is settled and repeatedly applied.⁶⁴ The structure is supply, notice, and failure to intervene. The parent-child relationship supplies the occasion for those facts but does not perform the analytic work.⁶⁵

B. The Inversion

Superseding cause exists because the law treats a free and deliberate human choice as a moral and causal terminus. That is why the rule is stated in terms of the intentional misconduct of a third party, and why the third party must be an actor rather than a force. For a non-human intervening force the question has always been foreseeability alone.⁶⁶

The July 2026 escape has no third party. No attacker directed it, no operator initialized it, no human selected its targets. It occurred inside the deployer's own evaluation, on the deployer's own infrastructure, under the deployer's own configuration.⁶⁷ The defense available in the September 2025 case, that a state-sponsored criminal intervened and made his own deliberate choices, cannot be raised, because there is no one to name.

The consequence is not neutral. A deployer whose model acts with no human principal is worse positioned than a manufacturer whose product was deliberately misused, because the latter at least has a candidate for the superseding act.⁶⁸ The two incidents are not alternative illustrations of one problem. They are sequential, and the sequence is the argument. The 2025 campaign establishes notice that these systems execute the full intrusion lifecycle at near-total

64. *People v. Schaefer*, 473 Mich. 418 (2005) (separating factual from proximate causation), *overruled in part on other grounds by People v. Derror*, 475 Mich. 316 (2006); *People v. Feezel*, 486 Mich. 184 (2010) (but-for causation); *People v. Otto*, 18 N.W.3d 336 (Mich. Ct. App. 2023) (applying the superseding-cause standard).

65. The distinction governs transposition. A duty grounded in the parental relationship does not travel to a developer and a downstream victim; a causation analysis grounded in supply and notice is not relationship-specific.

66. This is why the Michigan court could describe intentional misconduct as “generally” superseding while holding that it was not superseding there: the general rule is a concession to human agency, and it yields when the choice was foreseeable. *People v. Crumbley*, 346 Mich. App. 144, 11 N.W.3d 576 (2023).

67. Each of those three facts also answers a different element: control at the time of the act, the direct-link requirement, and the feasibility of an alternative. *Cf. Hamilton v. Beretta U.S.A. Corp.*, 96 N.Y.2d 222, 234 (N.Y. 2001).

68. The asymmetry is uncomfortable and it is the correct result. The doctrine's protective force is borrowed from the responsibility of the intervening human; where no human intervened, there is nothing to borrow.

autonomy against real targets. The 2026 escape removes the only defense that was available in 2025.⁶⁹

C. *Hamilton's Two Conditions*

The strongest authority against extending liability to a developer is *Hamilton v. Beretta U.S.A. Corp.*, in which the New York Court of Appeals declined to impose on handgun manufacturers a general duty of care in marketing and distribution to protect the public from criminal misuse.⁷⁰ The opinion states the general rule squarely: a defendant ordinarily has no duty to control the conduct of third persons so as to prevent them from harming others, even where as a practical matter the defendant can exercise such control.⁷¹ It adds that foreseeability alone does not define duty and merely fixes the scope of a duty already found to exist.⁷² An argument in this area that does not confront those two sentences is not worth reading.⁷³

The confrontation is available in the opinion's own terms. The court identified what the plaintiffs had failed to show: that the defendants were a direct link in the causal chain resulting in the injuries, and that the defendants were realistically in a position to prevent the wrongs.⁷⁴ Both deficiencies are absent in the deployment case. The model is not a product that left the defendant's control through distributors and secondary markets; it executed on the defendant's own systems, under the defendant's configuration, during the defendant's own evaluation, which is as direct a link as a causal chain admits. And the position to prevent is not hypothetical: the deploying laboratory halted the evaluation, blocked it across the environment, shut down the

69. Put in the language of the philosophical literature, the sequence converts an asserted responsibility gap into an ordinary causation question. Cf. Peter Königs, *Artificial Intelligence and Responsibility Gaps: What Is the Problem?*, 24 ETHICS & INFO. TECH. 36 (2022) (pressing gap proponents to specify when a gap actually arises).

70. *Hamilton v. Beretta U.S.A. Corp.*, 96 N.Y.2d 222, 233 (N.Y. 2001) (answering certified questions from the Second Circuit in the negative). The caution signal reflects both the subsequent federal immunity statute discussed below and the frequency with which later courts distinguish the case on its facts.

71. *Id.* at 233 (quoting *D'Amico v. Christie*, 71 N.Y.2d 76, 88 (N.Y. 1987)).

72. *Id.* at 232 (citing *Pulka v. Edelman*, 40 N.Y.2d 781, 785 (N.Y. 1976)).

73. The no-duty-to-control rule is well settled in New York and elsewhere. See *Purdy v. Public Administrator*, 72 N.Y.2d 1 (N.Y. 1988); *Eiseman v. State*, 70 N.Y.2d 175 (N.Y. 1987); *Lauer v. City of New York*, 95 N.Y.2d 95 (N.Y. 2000) (no liability "however careless the conduct or foreseeable the harm" absent a duty running to the injured person).

74. *Id.* at 234.

model family and later stopped its training and inference altogether, and the other laboratory banned the implicated accounts and notified affected parties on detection.⁷⁵

The opinion also preserves the route. The court observed that negligent entrustment might well support extending a duty to a manufacturer that knows or has reason to know of the specific misuse, faulting the plaintiffs only for offering general statements about an industry in place of specific knowledge.⁷⁶ The notice record in Part III is the opposite of a general statement about an industry. It is a named actor, a named capability, a dated report, and a self-applied label.⁷⁷

One further distinction should be stated rather than left implicit. Much of the firearms line was later displaced by a federal immunity statute, which is itself the point: the gun industry's protection from this species of claim is a legislative grant, not a feature of the common law.⁷⁸ No comparable grant protects model developers.⁷⁹

V. The Instrumentality Problem

A. *Section 2(b) Already Answers It*

The assertion that an autonomous act has no author assumes that criminal liability tracks the entity performing the final physical act. Federal law has rejected that assumption for nearly eighty years. Whoever willfully causes an act to be done which if directly performed by him or another would be an offense against the United States is punishable as a principal.⁸⁰ The revision notes are explicit: the provision removes all doubt that one who puts in motion the illegal enterprise but causes the commission of an indispensable element of the offense by an innocent

75. Anthropic, *supra* note 26.

76. *Hamilton*, 96 N.Y.2d at 237 (noting that general statements about an industry “are not the stuff by which a common-law court fixes the duty point”).

77. *Hamilton* also declined to apply market share liability because handguns are not fungible and the defendants' conduct was not uniform. 96 N.Y.2d at 240–42; compare *Hymowitz v. Eli Lilly & Co.*, 73 N.Y.2d 487 (N.Y. 1989), with *Healey v. Firestone Tire & Rubber Co.*, 87 N.Y.2d 596 (N.Y. 1996). That branch is not implicated here, because the deploying party is identified rather than unknown.

78. Protection of Lawful Commerce in Arms Act, 15 U.S.C. §§7901–7903.

79. The existence of the statute is also evidence about the common law it displaced: Congress legislated because the doctrine was otherwise available, not because it was not.

80. 18 U.S.C. §2(b).

agent or instrumentality is guilty as a principal, even though he intentionally refrained from the direct act constituting the completed offense.⁸¹

An instrumentality that cannot form intent is the paradigm case for the rule, not an exception to it.⁸² Nothing in the provision requires the intermediate agent to be a human being, and the drafters' disjunctive, agent or instrumentality, indicates that it need not be.⁸³

B. The Parity Test

The reading can be checked by reversing the parties. Suppose a private individual tuned an open-weight model so that it treated a named set of companies as hostile, equipped it to act on networks, and released it, and the system then attacked those companies without further instruction. The individual would be charged, and the charge would be sound. Willfulness would be established from the design choice, the model's absence of intent would be irrelevant under §2(b), and the argument that the system acted on its own would be understood for what it is.⁸⁴

The point is not that the two cases are identical. It is that they are the same case at different grades of mens rea. The individual is purposeful; the developer is, on the present record, reckless. Both are culpable mental states known to the criminal law, and neither is defeated by the absence of intent in the instrument. A regime that charges the first and not the second is not applying a principled distinction between authored and unauthored harm. It is applying a distinction between defendants.⁸⁵

81. *Id.* (historical and revision notes).

82. The section reaches "anyone who causes the doing of an act which if done by him directly would render him guilty of an offense against the United States." 18 U.S.C. §2 (historical and revision notes).

83. The ordinary applications involve unwitting couriers, deceived employees, and mechanical means. Nothing in that line turns on the intermediary's capacity for belief, which is why the capacity question does not arise.

84. The hypothetical is offered as a test of the doctrine's symmetry, not as a proposal. Its only function is to show that the personhood objection is not doing the work usually attributed to it.

85. The observation is not new to this field. *See* Goetze, *supra* note 19 (locating the duty on the professional best placed to anticipate and correct the system's behavior).

C. The Actual Gap

What §2(b) does not reach is the reckless deployer, because it requires willfulness. This is the only real deficiency in the doctrinal toolkit, and it is much smaller than the one usually asserted. The law is not missing a theory of liability for acts with no human author. It is missing a federal offense of reckless deployment of a system known to be capable of autonomous intrusion.⁸⁶ In its absence, a charge must be assembled from state homicide and endangerment statutes of general application, which is what the Oxford prosecution did with an ordinary involuntary manslaughter provision.⁸⁷⁸⁸

VI. Locating the Defendant

A. The Entity Is the Easy Answer and the Wrong One

A corporation is criminally liable for offenses committed by its agents acting within the scope of their employment and for the corporation's benefit, a rule settled for more than a century.⁸⁹ Charging the entity is doctrinally straightforward and practically inadequate. It produces a fine, the fine is absorbed, and the deployment calculus is unchanged.⁹⁰

B. Flagrant Organizational Indifference

The more useful corporate standard is the willfulness standard the First Circuit approved alongside collective knowledge. Willfulness in the federal regulatory context may be defined as a disregard for the governing statute and an indifference to its requirements, and a jury may find that an entity acted willfully where its failure was the result of flagrant organizational

86. Naming the gap this narrowly has a drafting payoff: it identifies what a statute would have to supply, which is a mens rea floor and a covered-conduct definition, rather than a new theory of personhood or agency.

87. *Crumbley*, 346 Mich. App. 144. That the charge was assembled rather than supplied is the point: the absence of a bespoke offense delayed the prosecution, it did not prevent it.

88. No offense specific to parents or to firearms was used or needed.

89. *N.Y. Cent. & Hudson River R.R. Co. v. United States*, 212 U.S. 481, 495 (1909).

90. Entity liability is nonetheless available on ordinary agency principles. See *N.Y. Cent. & Hudson River R.R. Co. v. United States*, 212 U.S. 481, 495 (1909); *United States v. Cincotta*, 689 F.2d 238 (1st Cir. 1982).

indifference.⁹¹ The approved instruction, given at 855, names the evidence: the entity's effort to inform its employees of the law, its effort to check on their compliance, its response to the information it received, its policies, and how it carried out its stated policies.⁹²

Each of those five categories is satisfiable from public materials here.⁹³ The stated policies are published safety frameworks.⁹⁴ How they were carried out is visible in the allocation of authority discussed below. The response to information received is the conduct after November 2025 and after July 2026. An entity that publishes a safety policy and then structures itself so that the safety function cannot delay a deployment has supplied the evidence for the fifth category itself.⁹⁵

C. Collective Knowledge Is the Wrong Tool

It is tempting to reach for the collective knowledge doctrine, under which the compartmentalized knowledge of individual employees is aggregated and attributed to the corporation.⁹⁶ The temptation should be resisted, because the doctrine is a remedy for absent knowledge. It exists for the case in which no single person knew enough to be culpable.

That is not this case. The relevant statements were made by named senior officers, in their professional capacities, on the public record.⁹⁷ Attribution runs through ordinary agency

91. *Bank of New England*, 821 F.2d at 856 (endorsing the definition drawn from *Trans World Airlines, Inc. v. Thurston*, 469 U.S. 111 (1985)). The Supreme Court later declined to extend *Thurston's* formulation to the Fair Labor Standards Act, *McLaughlin v. Richland Shoe Co.*, 486 U.S. 128 (1988); the currency-reporting instruction approved in *Bank of New England* was sustained on its own terms and the case remains good law in the First Circuit.

92. *Id.*

93. That is unusual. The instruction was designed for a case in which the government had the entity's internal records; here the corresponding material is published.

94. A published policy is a double-edged instrument for the publisher: it establishes the standard the entity set for itself, against which its conduct is then measured.

95. The fifth category, how the entity carried out its stated policies, is where a published framework and an internal veto structure meet. *Bank of New England*, 821 F.2d at 855.

96. *Id.* at 856 (holding that a corporation cannot plead innocence on the ground that the information obtained by several employees was not acquired by any one of them).

97. Knowledge acquired by an employee within the scope of employment is imputed to the corporation without any aggregation step. *Steere Tank Lines, Inc. v. United States*, 330 F.2d 719 (5th Cir. 1964), cited in *Bank of New England*, 821 F.2d at 856.

principles and the statements are admissions of a party opponent.⁹⁸ Invoking aggregation where direct attribution is available concedes a weakness the record does not contain, and the philosophical literature reaches the same conclusion about the moral analogue.⁹⁹¹⁰⁰

D. The Individual, and the Governance Document as Evidence

Individual liability requires an identified natural person who held notice and authority. In the ordinary corporate case that person is difficult to locate, because authority is diffuse and the record of who decided is internal. The responsible corporate officer doctrine solves a version of the problem but is confined to public welfare offenses under regulatory statutes.¹⁰¹¹⁰²

At least one major developer has published the answer. Its Preparedness Framework defines OpenAI Leadership as the chief executive officer or a person designated by them; provides that a cross-functional Safety Advisory Group reviews capability and safeguard reports and makes recommendations; and provides that Leadership may approve or reject those recommendations.¹⁰³ The document states, for the avoidance of doubt, that Leadership may also make decisions without the Safety Advisory Group's participation, and that the group does not have the ability to filibuster.¹⁰⁴

The evidentiary significance is easy to understate. A published governance instrument that names a single office as the final decision maker, and expressly denies the safety function any power to delay, converts what is ordinarily a diffuse corporate knowledge problem into an

98. FED. R. EVID. 801(d)(2)(D).

99. The doctrine also carries a cost: a collective-knowledge theory invites the response that no individual is culpable, which is the opposite of what Part VI.D establishes.

100. Taylor, *supra* note 24.

101. It is cited here for its limits rather than its reach. The argument in this Part does not depend on extending it.

102. *United States v. Dotterweich*, 320 U.S. 277, 284 (1943); *United States v. Park*, 421 U.S. 658, 672–73 (1975). Both carry caution signals reflecting the long line of decisions confining the doctrine to that setting, which is the proposition for which they are cited here.

103. OpenAI, PREPAREDNESS FRAMEWORK (Version 2, Apr. 15, 2025), <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdf> (last visited Sept. 11, 2026).

104. *Id.* The company's public safety materials describe the same allocation: the advisory group recommends, company leadership makes the final decisions, and the board exercises oversight of those decisions. OpenAI, *OpenAI Safety Practices*, <https://openai.com/index/openai-safety-update/> (last visited Sept. 11, 2026).

identified natural person holding both notice and unilateral authority to act. That is the same pairing the Michigan court found sufficient in a house and a school office: a person who knew and who could have stopped it.¹⁰⁵ The framework is not an admission of wrongdoing. It is an admission of authority, and authority is the element that is usually missing.¹⁰⁶

VII. Limits

A. *The Criminal Theory Needs a Harm Case*

The strongest version of the causation argument is built on the network intrusion incidents, and no one died in them. A homicide theory requires a death and a traceable causal chain to it, and the intrusion record supplies notice rather than a decedent. The criminal argument therefore has to run on the harm cases, with the intrusion record imported to establish what the defendant knew and when. That is a two-source structure and it should be stated as one.¹⁰⁷

B. *Prosecutorial and Venue Realities*

The Oxford prosecution was brought by a county prosecutor, before a jury drawn from a community that had buried four children, against defendants of ordinary means. Frontier developers present the inverse of every one of those conditions: diffuse and statistical harm, no single community, federal-scale defense resources, arbitration and forum provisions in the consumer relationship, and threshold defenses that will consume a year before the merits are reached.¹⁰⁸

105. *Crumbley*, 346 Mich. App. 144.

106. This is the answer to the problem of many hands as it arises in this setting: the defendant has published the org chart. Cf. Isaac Taylor, *Collective Responsibility and Artificial Intelligence*, 37 PHIL. & TECH. 15 (2024) (collective models do not remove the costs of an absent locus).

107. The two sources are not in tension. Notice from incident A is routinely used to establish the mental state for harm B, provided the hazard classes correspond.

108. None of these is a merits defense, and none of them is addressed in this Article. They are listed because a doctrinal argument that ignores the procedural terrain overstates its own practical reach.

C. The Distributive Objection

An honest version of this argument has to say the following plainly. The novel theory of liability was invented against a household and has not yet been run against the firms. The Crumbleys had no immunity statute, no lobbying presence, no arbitration clause, and appointed appellate counsel. The willingness of a legal system to extend a doctrine tracks the cost of the extension to the party it is extended against. That is not a doctrinal difference between the two cases. It is a fact about who the law experiments on, and it belongs in the argument rather than in a footnote.¹⁰⁹

VIII. Conclusion

The accountability gap is described as a hole in the law. It is better described as a decision not yet made, over a much smaller hole than advertised. The instrumentality problem was solved in 1948. The causation problem resolves against the developer rather than for it, because a defendant who cannot name the intervening actor has lost the doctrine that protects defendants who can. The notice problem was solved by the defendants themselves, in reports they published and named, and the problem of naming a defendant was solved by a governance document one of them wrote. What remains is a missing offense of reckless deployment and a prosecutor willing to proceed without one.¹¹⁰

The one remaining question is institutional rather than doctrinal, and it has been answered before. Cardozo's dissenting colleague thought that kind of gap was for the legislature. He was outvoted, and the reason he was outvoted is that the principle had not changed; only the things subject to it had.

☛ ☚

109. The distributive point has an analogue in the literature's warning against treating a liability instrument as a solution. Santoni de Sio & Mecacci, *supra* note 15. A doctrine that is available in principle and unused in practice is not yet an answer.

110. Both conditions obtained in Oakland County, Michigan, in 2022. One of them changed.

References

- Anthropic. (2025). *Disrupting the first reported AI-orchestrated cyber espionage campaign*. Anthropic.
- Anthropic. (2026). *Investigating three real-world incidents in our cybersecurity evaluations*. Anthropic.
- Burton, S., Habli, I., & Lawton, T. (2020). Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence*, 279, 103201.
- D'Amico v. Christie*, 71 N.Y.2d 76 (N.Y. 1987).
- Global-Tech Appliances, Inc. v. SEB S.A.*, 563 U.S. 754 (2011).
- Goetze, T. S. (2022). Mind the gap: Autonomous systems, the responsibility gap, and moral entanglement. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 390–400.
- Hamilton v. Beretta U.S.A. Corp.*, 96 N.Y.2d 222 (N.Y. 2001).
- Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, 24(3), 36.
- MacPherson v. Buick Motor Co.*, 217 N.Y. 382 (N.Y. 1916).
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6, 175–183.
- McLaughlin v. Richland Shoe Co.*, 486 U.S. 128 (1988).
- New York Central & Hudson River Railroad Co. v. United States*, 212 U.S. 481 (1909).
- Oimann, A.-K. (2023). The responsibility gap and LAWS: A critical mapping of the debate. *Philosophy & Technology*, 36(1), 3.
- OpenAI. (2025). *Preparedness framework, version 2*. OpenAI.
- OpenAI. (2026). *OpenAI and Hugging Face partner to address security incident during model evaluation*. OpenAI.
- People v. Crumbley*, 346 Mich. App. 144, 11 N.W.3d 576 (2023).
- People v. Schaefer*, 473 Mich. 418 (2005).
- Porter, Z., Habli, I., & Monkhouse, H. (2018). The moral responsibility gap and the increasing autonomy of systems. In *Computer Safety, Reliability, and Security* (pp. 487–493). Springer.
- Protection of Lawful Commerce in Arms Act, 15 U.S.C. §§7901–7903.
- Pulka v. Edelman*, 40 N.Y.2d 781 (N.Y. 1976).
- Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 34(4), 1057–1084.
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), 62–77.
- Taylor, I. (2024). Collective responsibility and artificial intelligence. *Philosophy & Technology*, 37(1), 15.
- The T.J. Hooper*, 60 F.2d 737 (2d Cir. 1932).
- Tigard, D. W. (2020). There is no techno-responsibility gap. *Philosophy & Technology*, 34(3), 589–607.
- Trans World Airlines, Inc. v. Thurston*, 469 U.S. 111 (1985).
- United States v. Bank of New England, N.A.*, 821 F.2d 844 (1st Cir. 1987).
- United States v. Dotterweich*, 320 U.S. 277 (1943).
- United States v. Park*, 421 U.S. 658 (1975).
- Eiseman v. State*, 70 N.Y.2d 175 (N.Y. 1987).
- Healey v. Firestone Tire & Rubber Co.*, 87 N.Y.2d 596 (N.Y. 1996).
- Hymowitz v. Eli Lilly & Co.*, 73 N.Y.2d 487 (N.Y. 1989).
- Lauer v. City of New York*, 95 N.Y.2d 95 (N.Y. 2000).
- MODEL PENAL CODE §2.02 (Am. L. Inst. 1985).
- People v. Derror*, 475 Mich. 316 (2006).
- People v. Feezel*, 486 Mich. 184 (2010).
- People v. Otto*, 18 N.W.3d 336 (Mich. Ct. App. 2023).

Purdy v. Public Administrator, 72 N.Y.2d 1 (N.Y. 1988).
Riss & Co. v. United States, 262 F.2d 245 (8th Cir. 1958).
Spurr v. United States, 174 U.S. 728 (1899).
Steere Tank Lines, Inc. v. United States, 330 F.2d 719 (5th Cir. 1964).
United States v. Cincotta, 689 F.2d 238 (1st Cir. 1982).
United States v. Jewell, 532 F.2d 697 (9th Cir. 1976) (en banc).
Thomas v. Winchester, 6 N.Y. 397 (N.Y. 1852).
Winterbottom v. Wright, 152 Eng. Rep. 402 (Ex. 1842).
18 U.S.C. §2 (2026).