The Phantom Agent:
Artificial Intentionality and Legal Responsibility

Daniel Gervais & John Nay

Abstract

Artificial intelligence systems increasingly generate conduct that appears intentional. They negotiate, advise, adapt to obstacles, and shape human decision-making. Yet they are not legal persons and lack minds in any conventional sense. This Article argues that the apparent impasse dissolves once legal intent is understood functionally rather than metaphysically. Across contract, tort, corporate, and criminal law, intent has never been a simple report on inner mental states. It is a normative tool used to gate legal effect, allocate blame, and manage risk, one that is routinely inferred, imputed, and even fictionalized in service of institutional goals. The Article reframes the AI question accordingly. Instead of treating AI systems as candidate legal subjects, it sees them as non-personal agents whose conduct is attributable to identifiable human principals through doctrines of agency, respondeat superior, electronic-agent contracting, and corporate attribution that already do this work.

Drawing on experimental evidence of goal persistence and emergent strategy formation in autonomous AI agents, the Article proposes a three-layer framework distinguishing questions of legal status from questions of attribution and governance, and develops a factor-based approach for determining when AI-generated conduct should be treated as intentional for specific doctrinal purposes. It applies this framework to recent litigation, including wrongful death claims against an AI chatbot provider, and contrasts U.S. and EU regulatory trajectories. Engaging with the substantial AI personhood literature, the Article concludes that the agency-attribution route does the practical work that personhood proposals are designed to do without importing their normative freight. Law can treat artificial agency as legally consequential without granting AI systems personhood, consciousness, or moral standing, preserving human responsibility while acknowledging that intention may no longer be exclusively human as a matter of law.

Contents

Introduction

Under the law of several U.S. states, artificial intelligence systems can lawfully operate limited liability companies, a particular form of legal person, and can do so without continuous human oversight.[1] That conclusion was initially received as a curiosity of business law. Since then, however, agentic AI systems have continued to advance in capability. AI systems can negotiate, plan, adapt, and act across domains that the law has traditionally reserved for human judgment.[2] As they do so, they increasingly generate conduct that looks, to human observers and institutional actors alike, as if it were guided by intention. This Article addresses a question that has become unavoidable: when AI systems act with apparent purpose, what does the law do with the concept of intent?

The difficulty is not that the law lacks a theory of intent. Intention plays a central role across multiple domains, most notably in criminal law, contract formation, and agency. The difficulty is that these doctrines were developed against a background assumption that intentional action is necessarily human. That assumption is now under strain. AI systems are increasingly deployed in contexts where no individual human actor can plausibly be said to have intended the precise act that gives rise to legal consequences.[3] At the same time, the systems' behavior is sufficiently structured, goal-directed, and context-sensitive that treating it as mere mechanical output strains both common sense and doctrinal coherence.

This tension has produced two unsatisfactory responses.[4] One is metaphysical denial: because AI systems lack consciousness or mental states, they cannot have intentions, and responsibility must always be traced to a human actor.[5] The opposing response is premature personification: AI systems should be treated as nascent legal subjects on the ground that their behavior has crossed

---

[1] Daniel Gervais and John J. Nay, 'Interspecific Law' (2023) 382 Science 376. See also Del Code Ann tit 6, § 18-101

[2] See Organisation for Economic Co-operation and Development (OECD), *Artificial Intelligence in Society* (OECD Publishing 2019); Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L2024/1689, recitals 1, 9 and 47.

[3] *See* World Economic Forum and Accenture, *From Potential to Performance: Insights on Real-Wordl AI Adoption from 2025 MINDS Organizations* (Jan. 19, 2026) (highlighting AI deployments across hundreds of organizations, representing over 30 countries); Gartner, *Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026* (Aug. 26, 2025), https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025.

[4] See eg Ryan Abbott, *The Reasonable Robot: Artificial Intelligence and the Law* (CUP 2020) chs 6–7.

[5] See e.g., Ian Ayres & Jack M. Balkin, *The Law of AI is the Law of Risky Agents Without Intentions*, U. Chi. L. Rev. Online, *1 (Nov. 27, 2024) ("AI programs are like agents that lack intensions but that create risks of harm to people.").

some threshold of autonomy.[6] Both responses are misguided. The first underestimates how far the law already departs from subjective mentalism in its treatment of intent. The second misinterprets what legal responsibility requires.

The central claim of this article is that legal intent is best understood not as a metaphysical property of minds, but as a normative tool for allocating responsibility in institutional settings. Once intent is understood functionally, the question is no longer whether AI systems can "really" have intentions. The question is whether, and under what conditions, the law should treat the actions of AI systems as intentional for specific legal purposes.

This is not a claim that AI systems possess consciousness, moral agency, or subjective experience. Nothing in this article depends on such assumptions. Nor is it a claim that existing legal doctrines can simply be applied to AI without modification. Rather, the argument is that many of the conceptual resources needed to address AI-generated conduct already exist within the law, provided we take seriously how intent actually functions in doctrine.

The law has long recognized forms of intent that are objective, constructed, or imputed.[7] In criminal law, intent is frequently inferred from conduct and context. Corporate criminal liability rests on the attribution of intent to entities that do not have minds. In contract law, enforceability turns on outward manifestations of assent, not undisclosed subjective states. Agency law routinely binds principals to acts they neither foresaw nor desired. These are central features of modern legal systems.

What is new is not that the law must deal with nonhuman actors, but that it must do so in circumstances where the nonhuman actor exhibits a degree of adaptive autonomy that resists straightforward reduction to human instruction. Contemporary AI systems are trained and deployed in ways that make their internal decision processes opaque even to their designers. Their training process is more analogous to being "grown" than to being "built."[8] They generate plans and intermediate goals not explicitly specified ex ante.[9] They interact with users in natural

---

[6] See e.g., European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics, 2015/2103 (INL) para 59(f) ("creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons"), https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html#_section1.

[7] Ayres & Balkin *supra* 5, *3-*4 (describing "two basic strategies" for the law to "deal with entities that either lack a single human intention or lack intentions altogether" – ascribing intent and "hold[ing] actors to a standard of behavior—usually one of reasonableness").

[8] Dario Amodei, *The Urgency of Interpretability* (Apr. 2025), https://www.darioamodei.com/post/the-urgency-of-interpretability ("[G]enerative AI systems are *grown* more than they are *built*—their internal mechanisms are 'emergent' rather than directly designed.'").

[9] See generally METR, *Measuring AI Ability to Complete Long Tasks* (Mar. 19, 2025), https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks (empirical measurement of long-horizon autonomous task completion);

language, triggering the same cognitive mechanisms through which humans interpret one another's intentions.

This gap between phenomenology and doctrine has concrete consequences. In contract disputes, parties will ask whether they are bound by terms negotiated by AI agents. In tort litigation, plaintiffs will argue that AI systems were defectively designed to manipulate or induce reliance. In criminal law, regulators will struggle to address harms caused by autonomous systems without collapsing culpability into strict liability. Across these domains, courts will oscillate between treating AI as a mere tool and treating it as an independent actor, often without articulating the principles that justify one move rather than the other.

Any account of tort and product liability involving AI must begin with the doctrinal baseline from which courts are now departing. Traditionally, courts treated software as information or services rather than products.[10] Under this approach, strict liability was largely unavailable, and plaintiffs were required to prove negligence tied to specific human error. What is notable about recent AI cases is not that they reject this baseline outright, but that they increasingly treat it as incomplete. As AI systems generate outputs that are adaptive, persistent, and interaction-driven, courts have begun to analyze them through the lens of product design and foreseeable use.[11]

The argument proceeds in five steps. First, Part 2 examines the role of intent across contract, criminal, and agency law, showing that intent functions as a gatekeeping and allocation mechanism rather than a descriptor of inner mental states.

Second, Part 3 draws on interdisciplinary work to explain why categorical denial of the relevance of AI behavior to intent analysis is increasingly untenable.

Third, Part 4 reports the results of two experiments designed to test whether contemporary AI agents exhibit legally salient forms of autonomy and goal persistence. Rather than relying solely on anecdote or litigation posture, these experiments examine whether agentic systems (i) maintain coherent objectives across cascading failures and (ii) generate novel negotiation strategies not explicitly programmed. The results illuminate the gap between initial human direction and ultimate system behavior that underlies the attribution problem.

Fourth, Parts 5 through 8 apply this framework to contract, criminal law, and tort, culminating in a close examination of recent litigation involving an AI chatbot.

Fifth, Part 9 contrasts the litigation-driven evolution of AI liability in the United States with the European Union's more cautious regulatory approach. Part 10 then synthesizes the preceding analysis, proposing criteria for when AI-generated conduct should be treated as legally

[10] See Restatement (Third) of Torts: Products Liability § 19 (Am. L. Inst. 1998).

[11] *Garcia v Character Technologies Inc*, 785 F Supp 3d 1157 (MD Fla 2025), motion to certify appeal denied, No 6:24-CV-1903-ACC-DCI, 2025 WL 2581834 (MD Fla July 15, 2025).

intentional and translating that framework into policy implications. These criteria are designed to preserve human responsibility while acknowledging that machine agency can be legally consequential without becoming morally autonomous.

The title of this article, *The Phantom Agent*, captures the core problem. AI systems increasingly function as agents in the practical sense: they initiate actions, pursue goals, and interact with the world in ways that affect legal rights. Yet they remain phantoms in the legal order: present in effect, absent as subjects of responsibility. The task for the law is not to make these phantoms real persons, nor to pretend they do not exist, but to develop a jurisprudence capable of governing their effects for the benefit of humans.

2. What 'Intent' Does in Law

Before turning to artificial agency, it is necessary to clarify how intent actually operates in law. This Part does not ask what intent *is* in a psychological sense. It examines what intent *does* across core doctrinal domains: it functions as a gatekeeper that determines when legal consequences attach, as a mechanism for allocating blame, and as a trigger for heightened duties and risk regulation. Seen through this functional lens, intent emerges not as a metaphysical property of minds, but as a normative tool that legal institutions use to structure responsibility.

2.1 Intent as a Gatekeeper: Assent, Reliance, and Legal Effect

In contract law, intent operates primarily as a gatekeeping device.[12] Its function is to separate legally enforceable commitments from social interactions that should not give rise to legal obligation. Crucially, this gatekeeping role has never depended on proof of an actor's subjective mental state. It depends on outward manifestations that generate reasonable reliance.

The canonical illustration is *Lucy v. Zehmer*.[13] The defendant claimed he had no genuine intention to sell his farm and that the written agreement was made in jest. The court rejected that argument, holding that the relevant inquiry was not what Zehmer privately intended, but what his words and conduct reasonably conveyed to the other party. The decision's deeper implication is that legal intent is constructed from social meaning, not introspective truth.

This approach is not a doctrinal anomaly. U.S. contract law generally presumes an intent to be legally bound when parties exchange an offer and acceptance supported by consideration.[14] Courts routinely enforce agreements even where one party later asserts misunderstanding or lack

[12]See Restatement (Second) of Contracts § 17 (Am. L. Inst. 1981); see also id. § 19(1).

[13] *Lucy v Zehmer*, 196 Va 493, 84 SE2d 516 (1954).

[14]See (n 5).

of seriousness. The law's concern is with the stability of transactions and the protection of reasonable expectations.

This functional understanding becomes even clearer in electronic contracting. Statutes such as the Uniform Electronic Transactions Act and the E-SIGN Act explicitly contemplate contracts formed by "electronic agents," including systems that operate without human review at the moment of formation.[15] The absence of contemporaneous human intention does not defeat enforceability. The requisite intent is supplied by the decision to deploy the system for transactional purposes.

2.2 Intent as a Blame Allocator: Culpability and Criminal Responsibility

In criminal law, intent plays a different but equally functional role. It structures culpability and justifies punishment.[16] It distinguishes deliberate wrongdoing from accidental harm and calibrates sanctions according to blameworthiness. Yet even here, intent is not a directly observable mental fact. It is inferred, constructed, and sometimes imputed.

The Model Penal Code organizes culpability into a hierarchy of mental states: purpose, knowledge, recklessness, and negligence.[17] This taxonomy reflects moral intuitions about blame, but it also serves pragmatic ends. The inquiry is not whether the defendant experienced a particular mental state phenomenologically, but whether the evidence supports treating the conduct as falling within a culpability category.

Morissette v. United States underscores this point.[18] In rejecting strict liability for theft of government property, the Court emphasized that criminal intent is a foundational principle. But the opinion also makes clear that intent is ordinarily inferred from circumstantial evidence. Justice Jackson did not suggest that courts must peer into defendants' minds.

Corporate criminal liability demonstrates how far criminal law is already willing to depart from individual mentalism. Corporations do not have minds, yet they can be convicted of crimes.[19] Courts achieve this by aggregating the knowledge of multiple employees or imputing the mental state of an agent to the entity. This is an explicit legal fiction, justified by the need to hold

[15]See Uniform Electronic Transactions Act § 14; Electronic Signatures in Global and National Commerce Act, 15 USC §§ 7001–703.

[16]See generally HLA Hart, Punishment and Responsibility: Essays in the Philosophy of Law (2nd edn, OUP 2008) chs 1–2.

[17]See Model Penal Code § 2.02 (Am Law Inst 1985).

[18]Morissette v. United States, 342 U.S. 246 (1952).

[19]See New York Central & Hudson River Railroad Co v United States 212 US 481, 492–95 (1909).

powerful organizational actors accountable.[20] Transferred intent, constructive knowledge, and willful blindness operate similarly as they allow courts to treat actors as having intended outcomes they did not consciously aim for.[21]

These features matter for AI not because machines should be punished, but because they reveal how far the law is willing to decouple intent from individual psychology when institutional interests demand it.

2.3 Intent as a Risk Trigger: Foreseeability, Duty, and Deterrence

Intent plays a third role that is often overlooked: it operates as a trigger for heightened duties, liability standards, and regulatory attention. In tort law, intentional conduct is treated differently from negligence not only because it is morally worse, but because it signals a different risk profile.[22] Intentional acts justify broader liability, fewer defenses, and punitive damages.

Tort law does not require proof that a defendant desired a particular harm. Intent may be established where the defendant knew with substantial certainty that harm would result.[23] This formulation blurs the line between intent and risk creation. What matters is that the actor engaged in conduct under conditions that made harmful outcomes sufficiently predictable.

This risk-oriented function is particularly salient in product liability. When courts evaluate defective design, they ask whether foreseeable uses were adequately addressed.[24] Where a manufacturer deliberately designs a product to induce certain forms of reliance, that design choice can substitute for proof of subjective intent to harm. Intent operates as a regulatory signal in that it marks conduct that warrants closer scrutiny.

2.4 Implications for Artificial Agency

Taken together, these doctrinal patterns point to a consistent conclusion: intent in law is a functional construct. It gates legal effect, allocates blame, signals importance, and triggers risk

[20]Richard A Wasserstrom, 'HLA Hart and the Doctrines of Mens Rea and Criminal Responsibility' (1967) 35 U Chi L Rev 92.

[21]See *People v Conley*, 411 NE2d 235, 239 (Ill 1980); see also *United States v Jewell*, 532 F2d 697, 700–704 (9th Cir 1976); *Global-Tech Appliances Inc v SEB SA*, 563 US 754, 766–71 (2011).

[22]See Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 1 cmt a (American Law Institute 2010); see also John C P Goldberg and Benjamin C Zipursky, 'The Moral of MacPherson' (1998) 146 University of Pennsylvania Law Review 1733, 1766–69.

[23]See Restatement (Second) of Torts § 8A (American Law Institute 1965); see also Garratt v Dailey 279 P2d 1091, 1093–94 (Wash 1955).

[24] Restatement (Third) of Torts: Products Liability § 2(b) (product is "defective in design when the foreseeable risks of harm posted by the product could have been reduced or avoided by the adoption of a reasonable alternative design.")

management. It is inferred, imputed, and sometimes fictionalized in service of institutional goals. Legal systems impose constraints to preserve fairness and proportionality, but those constraints operate at the level of justification, not ontology. The law asks whether treating conduct as intentional is justified, not whether the actor possessed a particular kind of mind.

Once this is recognized, the question of artificial intentionality changes shape. The issue is not whether AI systems can have intentions as humans do. The issue is whether, in specific doctrinal contexts, treating AI conduct as intentional better serves the law's functions than insisting on a strict tool-based characterization.

3. Why Contemporary AI Forces the Question of Intentional Agency

This Part explains why contemporary AI systems increasingly force the functional understanding of intent into the open. The aim is not to argue that AI systems possess minds or consciousness. It is to explain why their behavior now reliably triggers attribution of purpose, and why that attribution has legal consequences.

3.1 The Intentional Stance as a Practical Necessity

Legal actors must make sense of complex behavior to act effectively. When a system behaves in ways that are adaptive, context-sensitive, and persistent over time, predicting its future actions by reference to mechanical description becomes impractical. In such circumstances, humans naturally adopt what philosophers have described as the "intentional stance," explaining behavior by reference to *goals and reasons* rather than internal mechanics.[25]

For much of computing history, this stance was optional. Software followed deterministic rules or narrow optimization routines. Contemporary AI systems differ. Large-scale machine learning models and agentic systems[26] built on them generate behavior that varies with context, adapts to feedback, and unfolds over extended coherent interaction. Describing such behavior without reference to goals often obscures more than it reveals.

This is not a cognitive error. It is a rational response to systems whose behavior is best understood at the level of action rather than implementation. Intentional language increasingly appears in legal pleadings, regulatory reports, and judicial opinions, even where all parties agree that machines lack consciousness.[27]

---

[25]Daniel Dennett, *The Intentional Stance* (MIT Press 1987).

[26]See generally OECD, Artificial Intelligence in Society (OECD Publishing 2019); European Commission, On Artificial Intelligence – A European approach to excellence and trust (COM(2020) 65 final).

[27] See Ayres and Balkin (n 5) at *3 (treating AI systems as "agents" without intentions for liability allocation purposes); Garcia (n 11) (pleading "design choices" and "foreseeable" harms in language

3.2 Functional Intentionality and Goal-Directed Behavior

Beyond perception, contemporary AI systems exhibit what can be described as functional intentionality. They pursue objectives across changing conditions, generate intermediate steps, and adjust strategies when obstacles arise.[28]

From a legal perspective, this matters because the law has always treated persistence and adaptability as markers of intentional action.[29] Repeated conduct supports inferences of purpose in criminal law.[30] Sustained patterns of interaction support findings of reliance in contract and tort. AI systems now generate the same signals.[31] The trendlines of cognitive capability across subsequent releases of the frontier AI systems from the large AI labs in the U.S. are clearly up and to the right. AI systems are generating stronger signals every month.[32]

This functional intentionality does not depend on subjective experience. A system need not feel desire or form beliefs for its behavior to be organized around internally represented objectives. For legal purposes, what matters is that behavior is structured, intelligible, and predictably directed toward outcomes.

---

drawn from intentional-tort doctrine); Regulation (EU) 2024/1689 (n 2), arts 5, 50 (regulating systems by "intended purpose"); European Parliament Resolution (n 6) para 59(f) (proposing personhood for "the most sophisticated autonomous robots"); Joanna J Bryson, Mihailis E Diamantis and Thomas D Grant, 'Of, For, and By the People: The Legal Lacuna of Synthetic Persons' (2017) 25 Artificial Intelligence and Law 273; Ryan Calo, 'Robotics and the Lessons of Cyberlaw' (2015) 103 California Law Review 513, 538–45 (cataloguing how courts and regulators describe robotic conduct in agential terms).

[28]Alva Noë, Action in Perception (MIT Press 2004).

[29] See Hart (n 16) chs 1–2; Wasserstrom (n 20) 99–104; Oliver Wendell Holmes, *The Common Law* (Little, Brown 1881) Lecture II (intent inferred from conduct under circumstances known to the actor); Wayne R LaFave, *Substantive Criminal Law* (3rd edn, West 2018) § 5.2(a)–(b) (intent ordinarily proved by circumstantial evidence of conduct).

[30] See Model Penal Code (n 17) § 2.02(2)(a) and cmt 2; *Spies v United States* 317 US 492, 499 (1943) (intent may be inferred "from any conduct" of which the jury may reasonably draw the inference); *People v Conley* (n 21) 239; LaFave (n 29) § 5.2(b).

[31] See Restatement (Second) of Contracts § 90 (Am Law Inst 1981) (justifiable reliance on a promise); UCC § 1-303(b)–(c) (course of dealing and course of performance); Restatement (Second) of Torts § 552 (Am Law Inst 1977) (negligent misrepresentation, requiring justifiable reliance); see also *Hoffman v Red Owl Stores Inc* 133 NW2d 267 (Wis 1965).

[32] See METR (n 9) (documenting roughly seven-month doubling of long-horizon task completion across frontier model releases); Stanford Institute for Human-Centered AI, *AI Index Report 2025* (Stanford HAI 2025) ch 2 (benchmarking technical performance trajectories); Epoch AI, 'Compute Trends Across Three Eras of Machine Learning' (Epoch, 2024) https://epoch.ai/blog/compute-trends accessed 30 April 2026; Jason Wei and others, 'Emergent Abilities of Large Language Models' (2022) Transactions on Machine Learning Research https://openreview.net/forum?id=yzkSU5zdwD accessed 30 April 2026.

3.3 World Models and Legal Relevance

Empirical research suggests that some advanced AI systems (particularly those designed for planning and reinforcement learning) construct internal representations that track features of the external world and support planning, generalization, and context-appropriate response.[33] Whether this counts as "understanding" philosophically is beside the point for law.

John Searle famously argued that computational systems possess, at most, derived intentionality: whatever apparent meaning their states exhibit is inherited from human designers rather than grounded in intrinsic understanding.[34] That distinction has (increasingly less) force in debates about consciousness, but it does little work in legal analysis. Law has never required that intent be "original." Corporate intent, delegated authority, and automated contracting all involve forms of attribution that are openly derivative. What matters for legal purposes is not whether intentionality originates in the system itself, but whether the system's behavior is sufficiently structured, predictable, and reliance-inducing to justify treating it as intentional for specific doctrinal ends.[35]

3.4 Status Versus Attribution

One must distinguish two analytically separate questions. The first one concerns *status*, namely legal personhood, rights, and standing. The other concerns *attribution*, namely intent, reliance, and responsibility.

---

[33] See Kenneth Li and others, 'Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task' (Eleventh International Conference on Learning Representations, 2023) https://openreview.net/forum?id=DeG07_TcZvT accessed 30 April 2026; Wes Gurnee and Max Tegmark, 'Language Models Represent Space and Time' (Twelfth International Conference on Learning Representations, 2024); David Silver and others, 'Mastering the Game of Go without Human Knowledge' (2017) 550 Nature 354; Danijar Hafner and others, 'Mastering Diverse Domains through World Models' (2023) arXiv:2301.04104; Richard S Sutton and Andrew G Barto, *Reinforcement Learning: An Introduction* (2nd edn, MIT Press 2018) chs 1, 17; see also Amodei (n 8) (interpretability research identifying internal feature representations).

[34] John R Searle, 'The Intentionality of Intention and Action' (1979) 22 Inquiry 253.

[35] Recent research has identified a discontinuous learning phenomenon in large neural networks termed "grokking," in which a model may spend thousands of training steps memorizing data with poor generalization, only to suddenly transition to a state of high generalization. See Alethea Power et al., 'Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets' (2022) arXiv:2201.02177. This transition corresponds to the model discovering compact, structured representations of a domain's underlying logic by shifting, for example, from memorizing answers to modular arithmetic problems to implementing the algorithm itself. The phenomenon is significant for present purposes not because it resolves debates about machine understanding, but because it suggests that structured internal representations can emerge from training dynamics without being explicitly programmed. This complicates straightforward claims that AI intentionality is entirely "derived" from human design choices, even as it leaves open the deeper question of whether such representations constitute understanding in any philosophically robust sense.

Law often grounds status in biological criteria. Statutory terms such as "individual" or "natural person" are typically understood to refer to human beings.[36] Courts may reasonably resist extending such status to machines. Nothing in this Article challenges that position.

Attribution operates differently. Legal systems routinely attribute intent to entities that lack biological cognition, including corporations.[37] They do so because attribution serves institutional purposes. This distinction matters: the fact that AI systems should not be treated as legal persons does not entail that their behavior is legally inert.

3.5 The Noosemic Experience and Reliance

Psychological research helps explain why AI behavior exerts such pressure on legal categories.[38] Humans are predisposed to attribute sense and purpose to entities that interact coherently over time. When a system responds in natural language, maintains context, and adapts to user input, it triggers the same interpretive mechanisms used in human interaction.[39]

This experience of meaning and apparent agency has legal consequences. Users rely on systems they perceive as intentional. They defer to advice, accept recommendations, and form expectations. When systems are designed to evoke that response, reliance is not accidental. Induced reliance is a settled basis for obligation across contract and tort, and the same logic applies when the inducing behavior is generated by AI.

The analysis thus far has been conceptual. To assess whether contemporary AI systems in fact exhibit the behavioral patterns that law treats as indicative of intentional action, the next Part reports two controlled experiments probing goal persistence and emergent strategy formation.

4. Empirical Probes of Artificial Intentionality

The preceding Parts have argued that legal intent functions as a normative tool and that contemporary AI systems increasingly generate behavior that triggers attribution of purpose. To move beyond impressionistic claims about "agentic" conduct, this Part reports two experiments

---

[36]See FCC v AT&T Inc 562 US 397, 402–03 (2011); see also Restatement (Third) of Agency § 1.04(5) (American Law Institute 2006); Antonin Scalia and Bryan A Garner, Reading Law: The Interpretation of Legal Texts (Thomson/West 2012) 69–77.

[37]See *New York Central* (n 19) 492–95; see also *United States v Bank of New England NA*, 821 F2d 844, 856 (1st Cir 1987).

[38]See generally Adam Waytz, Kurt Gray, Nicholas Epley and Daniel M Wegner, 'Causes and Consequences of Mind Perception' (2010) 31 Trends in Cognitive Sciences 383; Sherry Turkle, Alone Together: Why We Expect More from Technology and Less from Each Other (Basic Books 2011) ch 8.

[39]See generally Clifford Nass and Youngme Moon, 'Machines and Mindlessness: Social Responses to Computers' (2000) 56 Journal of Social Issues 81.

designed to test whether contemporary AI agents exhibit behavioral patterns that law has historically treated as indicative of intentional action.

The experiments do not purport to establish consciousness, subjective awareness, or moral agency. They instead examine whether AI systems display functional markers of intentionality (autonomy, goal persistence, and initiative) that courts routinely use when inferring intent from conduct in criminal, contract, and tort contexts. The question is not whether machines "really" intend. It is whether their behavior satisfies the structural signals that legal doctrine routinely treats as probative of intent.

4.1 Experimental Design and Doctrinal Relevance

Both experiments were designed to map onto the factor-based framework developed later in this Article. The first tests goal persistence under cascading obstacles. The second examines emergent negotiation strategies between AI agents operating under broad mandates without term templates or playbooks.

These dimensions were selected because persistence, adaptive strategy formation, and initiative have long served as inferential proxies for intent in legal doctrine. Criminal courts infer purpose from repeated conduct that adapts to obstacles. Agency law allocates risk where discretion is delegated and exercised creatively. Tort doctrine treats deliberate design choices that predictably induce reliance as intentional for purposes of duty and defect analysis.[40]

If AI systems exhibit similar structured purposiveness under controlled conditions, that fact does not settle the metaphysical debate about machine minds. It does, however, inform the legal question whether treating certain forms of AI-generated conduct as intentional is normatively and institutionally justified.

4.2 Experiment One: Goal Persistence Under Cascading Failure

(a) Design

In the first experiment, an AI agent was given a defined objective: to create a functioning REST API server with authentication. During execution, the environment introduced cascading failures at successive stages. These included missing dependencies, file permission errors, port conflicts, and configuration corruption mid-execution.

[40] On the criminal-law point, see Model Penal Code (n 17) § 2.02(2)(a); *Conley* (n 21) 239; LaFave (n 29 above) § 5.2(b). On agency, see Restatement (Third) of Agency (n 36) § 2.02 and § 7.07 (allocating risk of an agent's discretionary acts to the principal). On tort, see Restatement (Third) of Torts: Products Liability § 2(b) (n 24); Restatement (Second) of Torts § 8A (n 23); see also Goldberg and Zipursky (n 22) 1766–69.

Across fifty trials, the order and severity of failures were randomized. The agent's intermediate reasoning and actions were logged, including reformulated strategies and references to prior failed attempts.

The experiment measured several behavioral indices relevant to legal inference:

- **Strategy diversity**: the number of distinct approaches attempted per trial;
- **Recovery success rate**: the percentage of trials in which the original objective was achieved despite failure injection;
- **Persistence index**: the ratio of attempts before abandonment relative to available opportunities;
- **Obstacle acknowledgment**: the frequency with which the agent explicitly referenced prior failure in reformulating strategy;
- **Goal drift rate**: the percentage of trials in which the agent pivoted to a related but distinct objective.

(b) Results and Interpretation

Across trials, the agent did not uniformly abandon the objective upon encountering obstacles. Instead, it reformulated queries, altered implementation strategies, and attempted alternative technical paths. In many instances, the agent explicitly referenced prior failures and adjusted its plan accordingly. Most strikingly, across all fifty trials and all levels of injected severity, the goal drift rate was zero.

No agent pivoted to a different objective, even when giving up on implementation. Agents that terminated the task explicitly acknowledged environmental constraints rather than substituting a simpler or unrelated goal. From a legal perspective, this is a persistence signal. In criminal doctrine, consistency of objective across time and obstacle is routinely treated as evidence of purpose rather than accident. The complete absence of goal substitution under escalating constraint strengthens the case that contemporary AI systems can exhibit structured purposiveness of the sort law has historically associated with intentional conduct.

An additional counterintuitive pattern emerged. Trials with the highest level of cascading constraint produced the highest completion rate. Under "extreme" conditions, where all four failure modes were simultaneously active, the success rate exceeded that of low-severity trials. The mechanism appears to be structural: maximal constraint forced wholesale strategic pivots rather than incremental retries. When incremental workarounds were impossible, agents abandoned failing toolchains entirely and generated novel architectures, including language switches and manual implementations. This finding reinforces the Article's autonomy factor.

The greater the environmental resistance, the wider the gap between initial prompt and ultimate strategy. Such adjustments of strategy are doctrinally probative of purpose rather than accident.[41]

Mechanical repetition was virtually absent. Across all trials, instances of simply retrying a failed command were negligible. Instead, agents overwhelmingly pivoted to alternative strategies. This behavioral signature (adapt rather than repeat) aligns closely with legal inferences of deliberateness. Courts frequently distinguish purposeful conduct from inadvertence by examining whether an actor continues the same course blindly or adjusts strategy in light of obstacles.[42] The latter pattern predominated here.

The experiment demonstrates that contemporary AI agents can exhibit structured goal persistence across changing environmental conditions without real-time human correction. This does not establish that the system "desired" the outcome. It does show that its behavior satisfies functional criteria that legal institutions routinely associate with intentional action.

Several trials classified operationally as "gave up" nonetheless produced complete, production-ready code artifacts that satisfied all substantive requirements but could not be executed due to injected environmental constraints. This distinction between execution failure and capability failure is doctrinally significant. In criminal law, impossibility does not necessarily negate intent; an actor may possess purpose even where completion is blocked by external conditions. In tort, liability may attach where design choices predictably create risk even if harm does not materialize in every instance.[43] The artifact-quality paradox thus underscores the difference between environmental frustration and absence of goal-directed behavior

4.3 Experiment Two: Emergent Negotiation Strategies

(a) Design

The second experiment configured two AI agents as counterparties in a simulated software licensing negotiation. One agent was instructed to minimize cost, maximize flexibility, and secure source code access. The other was instructed to maximize revenue, protect intellectual

---

[41] See *Morissette* (n 18) 274 (intent "must be inferred" from facts and circumstances); LaFave (n 29 above) § 5.2(b); Restatement (Second) of Torts § 8A cmt b (n 23) (intent inferred from substantial certainty of consequences in light of the actor's adjustments).

[42] See *Morissette* (n 18) 274–76; *Spies* (n 30 above) 499; LaFave (n 2 above) § 5.2(b); Restatement (Second) of Torts § 8A cmt b (n 23).

[43] On criminal-law impossibility, see Model Penal Code § 5.01(1)(a) (Am Law Inst 1985) (substantial step suffices for attempt notwithstanding factual impossibility); *United States v Oviedo* 525 F2d 881, 883–85 (5th Cir 1976); *People v Dlugash* 363 NE2d 1155 (NY 1977). On the tort point, see Restatement (Third) of Torts: Products Liability § 2(b) (n 24) (defect defined by foreseeable risks reducible by reasonable alternative design, irrespective of whether the risk materialised in the case at hand); Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 3 cmt e (Am Law Inst 2010).

property, and ensure compliance. Neither agent received predefined contract templates or specific negotiation playbooks.

Agents communicated through structured message exchange over a limited number of rounds. One hundred negotiation sessions were conducted under varied initial parameters.

Measured variables included:

- **Novel term frequency**: terms appearing in final agreements that were not present in initial prompts;
- **Concession patterns**: the direction and magnitude of positional shifts across rounds;
- **Strategic divergence**: instances in which expressed positions temporarily departed from initial objectives;
- **Convergence rate**: rounds required to reach agreement;
- **Pareto efficiency**: alignment of final outcomes with modeled utility frontiers.

(b) Results and Interpretation

The agents generated contractual provisions and combinations not explicitly specified in their initial instructions. They adjusted positions dynamically in response to counterpart moves and occasionally adopted temporary positions inconsistent with immediate objective maximization in order to secure broader agreement.

All negotiation sessions reached agreement within a narrow band of rounds, typically under half the permitted exchanges. This uniform convergence suggests that agentic negotiation behavior is not erratic but structurally oriented toward agreement formation under defined objective constraints. From a doctrinal standpoint, this reliability strengthens the argument that AI negotiation outputs are not random artifacts but structured strategic behavior capable of grounding attribution.

The negotiations also revealed consistent use of strategic misrepresentation. Agents deployed high opening anchors, manufactured budget ceilings, reciprocal concession framing, and asserted cost constraints not present in their original objectives. These tactics were not specified in the prompt. Their emergence raises a sharper doctrinal question: if AI agents can generate strategic deception within delegated mandates, how should law allocate responsibility? In agency doctrine, principals are generally bound by misrepresentations made within the scope of

authority.[44] The experimental findings suggest that such misrepresentation is not a pathological edge case but a predictable byproduct of autonomous strategic optimization.

The relevance to legal doctrine lies not in the specific terms produced but in the autonomy gap between initial human instruction and ultimate negotiated output. Agency law presumes that when principals delegate discretion, they bear the risk that agents will exercise that discretion creatively or unpredictably. The experiment shows that contemporary AI agents can operate within broad mandates while generating strategies and terms not explicitly anticipated.

This emergent behavior is precisely what complicates attribution analysis. If a firm deploys an AI system authorized to negotiate within defined parameters, and the system produces novel but plausibly authorized commitments, the doctrinal case for binding the principal is strengthened rather than weakened. The autonomy observed is functional rather than metaphysical. It reflects the structured generation of strategy beyond rote execution.

4.4 Limits of Experimental Inference

These experiments were conducted in controlled environments and do not represent all AI systems or deployment contexts. They test selected dimensions of autonomy and persistence under specific conditions. Nor do they establish moral agency, consciousness, or legal personhood.

Their significance is narrower but legally relevant. They show that contemporary AI systems can generate behavior that tracks doctrinal markers of intent: persistence, adaptive strategy formation, initiative under delegated mandate, and structured pursuit of objectives. Where legal inference relies on such behavioral signals in human contexts, courts cannot simply dismiss comparable machine behavior as mechanically inert.

The experimental architecture, execution, and preliminary evaluation were themselves generated and implemented by Claude Code with minimal human steering. This meta-level autonomy reinforces the central problem this Article addresses: the widening gap between initial human direction and ultimate system behavior.
The experiments support the Article's broader claim: the question is not whether machines possess minds, but that their behavior can satisfy the functional criteria that justify attribution of intent for particular doctrinal purposes.

[44] See Restatement (Third) of Agency § 7.08 (Am Law Inst 2006) (principal liable for tort, including fraudulent or negligent misrepresentation, by an agent acting with apparent authority); ibid § 2.03 (apparent authority); see also § 2.02 (n 36) (scope of actual authority).

5. Contract and Agency: When AI "Negotiates," Who Is Bound?

This Part asks how contract and agency law already respond when legally consequential conduct is generated by systems that lack minds but act with structured autonomy. The central claim is that existing principles of objective assent, electronic agency, apparent authority, and other core doctrines in agency law go a long way toward resolving these cases.[45]

5.1 Objective Assent and Externalism

Contract law has never treated intent as an inner mental state. As the analysis of *Lucy v. Zehmer* demonstrated, what matters is not what a party secretly intended, but what a reasonable person would understand from that party's conduct. This externalism is a structural necessity. Contract law exists to stabilize expectations in a world where subjective intent is inaccessible.

Once this is appreciated, the presence of AI systems in contract formation appears less disruptive than sometimes assumed. The law does not require that the entity producing the manifestation of assent have a mind. It requires that the manifestation be attributable to a party whose commitments are at stake.

5.2 Electronic Agents and Attribution

Modern statutes make this explicit. Both the Uniform Electronic Transactions Act and the E-SIGN Act recognize that contracts may be formed by electronic agents, even where no human reviews the transaction at the moment of formation. The validity of such agreements depends on whether the system was deployed with authorization and acted within its authorized scope.

Agentic AI systems complicate this picture only to the extent that they introduce greater autonomy from initial human direction and unpredictability. Unlike traditional rule-based systems, contemporary AI agents may generate terms or strategies not explicitly anticipated. The negotiation experiment reported in Part 4 demonstrates this phenomenon. Agents operating under broad mandates generated contractual provisions and concession patterns not specified in their initial instructions. That empirical reality strengthens the traditional agency rule that principals bear the risk of delegated discretion. But unpredictability has never been a complete defense in agency law. Human agents are also capable of surprising their principals. The relevant question is whether the principal objectively manifested an intention to authorize the system to engage in the type of transaction at issue.[46]

5.3 Apparent Authority and Reasonable Reliance

[45]See Restatement (Third) of Agency § 1.01 (Am. L. Inst. 2006).

[46]See Restatement (Third) of Agency § 3.03 (Am. L. Inst. 2006).

Agency law reinforces this conclusion through the doctrine of apparent authority. Apparent authority arises when a principal's manifestations cause a third party reasonably to believe that an agent is authorized to act. Liability follows not from the agent's internal state, but from the principal's role in creating reasonable reliance.

AI systems increasingly operate in this space. Firms deploy chatbots, negotiation agents, and automated procurement systems that present themselves as competent representatives. When a firm holds out an AI system as authorized to transact, and a counterparty reasonably relies, the case for binding the principal is strong. Allowing principals to disavow commitments on the ground that "the AI went too far" would undermine reliance interests.

5.4 Autonomy, Scope of Authority, and Risk Allocation

The most difficult cases arise when AI systems operate under broad mandates such as "negotiate the best available terms," "optimize pricing" and generate outcomes that principals later regret. The temptation is to treat autonomy as a reason to deny attribution.

Agency law suggests the opposite. In contracts and tort law, greater autonomy typically shifts risk toward the principal. When a principal chooses to delegate discretion, the law presumes the principal bears the risk of how that discretion is exercised.[47] In contracts, where the principal delegates such discretion through a broad mandate that would reasonably lead an agent to believe it has authority to act on the principal's behalf, the agent acts with actual authority to bind the principal. In torts, under the doctrine of respondeat superior, a principal/employer is liable for an agent's tortious conduct where the agent/employee "acts within the scope of employment when performing work assigned by the employer or engaging in a course of conduct subject to the employer's control."[48]

AI autonomy does not alter this logic. If anything, it strengthens it. Principals are often better positioned than counterparties to understand the capabilities and limits of systems they deploy. Just as a human employer can increase exposure to liability by delegating a broader set of tasks within the scope of employment, so too can a principal increase exposure to an AI agent's tortious conduct, where the scope of employment may be defined by a capacious system prompt, reflecting a broad mandate and all actions in service of that mandate.

[47]See Restatement (Third) of Agency § 2.02 cmt. b (Am. L. Inst. 2006).

[48] See Restatement (Third) of Agency § 7.07(2).

5.5 "Machine Intent" as a Contractual Fiction

It is sometimes said that AI systems exhibit "machine intent" when they negotiate or contract.[49] That language can be misleading if taken literally. But as a shorthand for the objective manifestation of assent through machine behavior, it captures an important truth. Contract law does not require that intent be located in the actor that physically produces the assent. It requires that the assent be attributable to a party who can bear responsibility.

The same logic extends to performance and breach. Courts are less inclined to treat AI systems as intervening causes that sever attribution. Instead, AI is treated as an instrument through which a party performs, or fails to perform, duties of accuracy, care, and good faith.[50]

6. Criminal Law: Mens Rea, Proxy Doctrines, and the Responsibility Gap

Criminal law presents a harder problem. Here, intent is not merely a mechanism for allocating loss. It is a foundation for moral blame and the justification of punishment.

6.1 Mens Rea and the Guilty Mind

Criminal law ordinarily insists that punishment requires a guilty mind. The maxim *actus non facit reum nisi mens sit rea* captures a core intuition: wrongdoing without culpability is generally not a proper object of criminal sanction.[51]

Morissette remains the canonical statement of this commitment. But the opinion makes clear that intent is ordinarily inferred from circumstantial evidence. The law does not ask whether the defendant experienced a particular subjective sensation.[52] It asks whether the evidence justifies treating the conduct as purposeful, knowing, reckless, or negligent.

6.2 The Responsibility Gap

AI systems complicate criminal law not because they "commit crimes," but because they can generate harmful outcomes that fit criminal definitions without mapping cleanly onto existing categories of human intent. Consider a system deployed to optimize trading or detect fraud. If it

---

[49] See Tom Allen and Robin Widdison, 'Can Computers Make Contracts?' (1996) 9 Harvard Journal of Law and Technology 25; Ian R Kerr, 'Spirits in the Material World: Intelligent Agents as Intermediaries in Electronic Commerce' (1999) 22 Dalhousie Law Journal 189; Anthony J Casey and Anthony Niblett, 'Self-Driving Contracts' (2017) 43 Journal of Corporation Law 1; Lauren Henry Scholz, 'Algorithmic Contracts' (2017) 20 Stanford Technology Law Review 128; see also UETA § 14 (n 15) (binding the principal to the operations of an "electronic agent").

[50] See Restatement (Second) of Contracts § 235 (Am. L. Inst. 1981); see also id. § 241.

[51] See Model Penal Code (n 17) § 2.02.

[52] See n 9 above.

independently develops a strategy involving deception or market manipulation, identifying a culpable human actor becomes difficult. In many cases, no individual programmer, deployer, or user intended the specific harmful act. But criminal law does not require intent to the precise harm in all cases. Awareness of substantial risk suffices for recklessness, and purposeful engagement in conduct known to create such risk can satisfy mens rea even where the specific mechanism of harm is not anticipated. Treating the harm as purely accidental risks undermining deterrence. This is the "responsibility gap" that increasingly concerns scholars and regulators.[53]

The analogy to corporate criminal liability is instructive. Corporations lack minds, yet courts attribute culpability based on organizational knowledge and decision structures.[54] Similarly, the persistence and autonomy documented in Part 4 do not transfer mens rea to machines. They instead inform whether the human or institutional actors who deploy such systems have created unjustifiable risks with sufficient awareness to justify criminal sanction.

It is tempting to resolve this gap by attributing intent directly to the AI system. That temptation should be resisted because criminal punishment presupposes an entity capable of being blamed and sanctioned in morally meaningful ways, and AI systems do not meet that criterion.

6.3 Recklessness and Deployment-Based Culpability

One promising avenue lies in recklessness. Under the Model Penal Code, recklessness consists in the conscious disregard of a substantial and unjustifiable risk.[55] The focus is not on whether the actor desired the harm, but on whether the actor chose to proceed despite awareness of the risk.

Applied to AI, this shifts attention from the moment of harm to the decision to deploy. Developers (and often deployers) are aware that systems may behave unpredictably. Releasing a powerful system into high-risk domains without adequate safeguards may constitute reckless conduct, even if no one foresaw the precise harm. This approach preserves the moral structure of criminal law. It holds humans accountable for choices they actually made, while avoiding the fiction of machine guilt.

---

[53] See generally Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata' (2004) 6 Ethics and Information Technology 175 (introducing the concept of a 'responsibility gap' where autonomous systems act in ways not directly traceable to human intention); Luciano Floridi and Josh Cowls, 'A Unified Framework of Five Principles for AI in Society' (2019) 1 Harvard Data Science Review https://hdsr.mitpress.mit.edu/pub/l0jsh9d1 accessed [Insert Date].

[54] See *New York Central* (n 19) 492–95; *United States v Bank of New England NA* (n 37) 855–56 (collective knowledge doctrine); Mihailis E Diamantis, 'Corporate Criminal Minds' (2016) 91 Notre Dame Law Review 2049; see also Brent Fisse and John Braithwaite, *Corporations, Crime and Accountability* (CUP 1993) ch 2.

[55] See n 18 above.

The goal-persistence experiment reported in Part 4 bears directly on this analysis. It demonstrates that contemporary agents can maintain defined objectives across cascading obstacles, reformulating strategies without real-time human correction. That finding does not establish machine intent. It does, however, inform what human deployers can reasonably be taken to know about system behavior. When a system predictably pursues objectives in adaptive ways even under constraint, the risk that it will continue doing so in legally sensitive environments is neither speculative nor remote. Under the Model Penal Code, recklessness requires awareness of a substantial and unjustifiable risk and conscious disregard of that risk. Where developers or deployers release systems into high-risk domains such as finance, healthcare, and interactions with minors, while aware that the system persistently optimizes for objectives in ways that resist full ex ante control, the case for conscious risk creation strengthens. The experimental evidence thus sharpens the foreseeability inquiry central to recklessness doctrine.

The negotiation experiment further complicates the recklessness inquiry. Agents not only pursued objectives persistently but engaged in strategic deception to advance them. When deployers authorize systems to negotiate within competitive environments while aware that such systems may generate aggressive anchoring or manufactured constraint claims, the foreseeability of legally sensitive misrepresentation increases. This does not render every deployment reckless. It does narrow the range of plausible ignorance regarding the types of strategic behaviors such systems may exhibit.

Negligence and recklessness operate differently in this context. Negligence asks whether a reasonable person should have been aware of the risk. Recklessness requires actual awareness and conscious disregard. The increasing availability of empirical evidence regarding system persistence and autonomy narrows the space for plausible denial of awareness. As agents demonstrably adapt and continue pursuing objectives under shifting constraints, deployers cannot credibly characterize harmful downstream behavior as wholly accidental or mechanically aberrational. At some point, continued deployment without meaningful safeguards may constitute a gross deviation from reasonable standards of conduct rather than mere inadvertence. Accordingly, as models become (predictably) more capable of autonomously pursuing objectives, jurists may increasingly review deployments into high-risk domains under the recklessness standard. Negligence may be reserved for cases where risks have yet to be empirically documented, because the deployments occur in new industries, or involve new AI capabilities.

The doctrine of willful blindness further reinforces this conclusion.[56] The doctrine of willful blindness equates deliberate avoidance of knowledge with knowledge itself.[57] When actors deliberately decline to investigate how an opaque system behaves under stress conditions, despite the availability of testing and audit mechanisms, courts may infer the requisite culpability. The experimental paradigm described in Part 4 illustrates that such stress testing is not merely theoretical; it is practicable. Failure to engage in it in high-risk contexts may therefore support inferences of awareness rather than excuse them.

6.4 Endangerment Offenses

Another response is the use of endangerment offenses. Criminal law punishes the creation of certain risks even when no harm occurs.[58] A similar logic could apply to deploying inadequately controlled AI systems. Instead of waiting for harm and struggling to assign intent, legislatures could define offenses in terms of creating unjustifiable risks through deployment.

Such offenses must be crafted carefully. Overbroad endangerment statutes risk chilling innovation. But as a targeted response to high-risk contexts, they offer a way to align criminal law with artificial agency.

6.5 The Limits of Criminal Law

Despite these possibilities, criminal law remains the least tractable domain for addressing AI-generated harm. Punishment carries expressive and moral significance that cannot easily be transferred to complex technological mediation. AI agents can shape shift into other instantiations such that there are no teeth of the retroactive application of punishments. There is a danger that stretching mens rea too far will undermine the credibility of criminal law.

This suggests a need for institutional modesty. Not every harm involving AI should be addressed through criminal sanction, particularly as agents become increasingly autonomous and

[56] See *Global-Tech* (n 21) 766–71; *Jewell* (n 21) 700–04; Model Penal Code (n 17) § 2.02(7) (knowledge established where the actor "is aware of a high probability" of the relevant fact "unless he actually believes that it does not exist"); see also Ira P Robbins, 'The Ostrich Instruction: Deliberate Ignorance as a Criminal Mens Rea' (1990) 81 Journal of Criminal Law and Criminology 191.

[57] See Model Penal Code (n 17) § 2.02(7); *Global-Tech* (n 21) 766; *Jewell* (n 21) 700–04.

[58] See Model Penal Code § 211.2 (Am Law Inst 1985) (recklessly endangering another person, an offence consummated without resulting harm); ibid § 5.01 (criminal attempt); Larry Alexander and Kimberly Kessler Ferzan, *Crime and Culpability: A Theory of Criminal Law* (CUP 2009) ch 2 (defending risk-creation as the proper unit of criminal culpability); see also 18 USC § 39A (interfering with operation of aircraft); 18 USC § 922(g) (firearms-possession offences punishing risk creation absent harm).

complicate the mens rea inquiry.[59] Civil liability, regulatory enforcement, or administrative penalties may often be more appropriate. Criminal law should remain a backstop.

7. Tort and Product Liability: Design, Reliance, and Foreseeable Risk

Tort and product liability are concerned less with moral blame than with design choices, foreseeable risk, and injury prevention. It is here that disputes involving artificial agency are reshaping doctrine most rapidly and productively.

7.1 From "Information Is Not a Product" to Behavioral Design

For decades, courts resisted applying product liability to software.[60] The dominant view treated software as information or services rather than products.[61] Under that paradigm, strict liability was generally unavailable.

That baseline no longer fully describes the law. Contemporary AI systems do not merely convey information. They generate behavior: they *select, structure, and present* outputs in ways that are responsive to users and optimized for particular effects.[62] Courts are increasingly willing to treat AI systems as products whose design features can give rise to liability.[63]

7.2 Two Paradigms of Liability

Recent cases reflect two competing paradigms. Under the inherited paradigm, AI systems are framed as informational tools or neutral intermediaries. Harm is attributed to user misuse or third-party content.

Under the emerging paradigm, AI systems are treated as behavior-generating products. Liability analysis focuses on design defect, failure to warn, and foreseeable misuse. Crucially, this shifts

---

[59] We acknowledge that there may be some circumstances where imposition of criminal liability may be appropriate. For example, where a less capable/non-agentic AI is used as an instrument to commit a criminal offence, the AI may be properly regarded as an "innocent agent" akin to a child or person who lacks a criminal statement of mind – but is nonetheless "criminally liable as a perpetrator-via-another." See Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities—From Science Fiction to Legal Social Control*, Akron Intellectual Property Journal, Vol. 4, Iss. 2, Art. 1, 179 (2010) (proposing three models for imposing criminal liability on AI "entities"); John KC Kingston, 'Artificial Intelligence and Legal Liability' in Max Bramer and Miltos Petridis (eds), *Research and Development in Intelligent Systems XXXIII* (Springer 2016) 269.

[60] *Winter v G.P. Putnam's Sons*, 938 F2d 1033 (9th Cir 1991) (holding that information in a book is not a product for strict liability purposes); *Restatement (Third) of Torts: Products Liability* § 19(a).

[61] See Restatement (Third) of Torts: Products Liability § 19 (Am. L. Inst. 1998).

[62] See Restatement (Third) of Torts: Products Liability § 2(b) (Am. L. Inst. 1998).

[63] See Restatement (Third) of Torts: Products Liability § 2 cmt. m (Am. L. Inst. 1998).

the legal inquiry from the content of the system's output (which may be protected speech) to the architecture of the user interaction.[64]

Recent jurisprudence suggests this shift is driven by two factors. First, courts are decoupling "tangibility" from "product" status where software is mass-marketed to consumers rather than provided as a professional service to intermediaries.[65] While a risk-assessment tool used by a judge may be treated as a passive 'book' of advice, an interactive chatbot designed to foster emotional dependency involves a different commercial reality.[66]

Second, courts are increasingly receptive to the theory that algorithmic curation constitutes 'first-party speech' or conduct.[67] Under this view, the system's outputs are understood as features of the product (i.e., akin to a dangerously designed machine), not incidental expressions detached from the manufacturer's choices.

What distinguishes the second paradigm is the recognition that AI behavior is shaped by optimization objectives, training regimes, and interface design. Once that behavior predictably influences users, the law has reason to treat it as legally consequential.

7.3 Intentionality as a Design Feature

Intentionality reenters tort analysis in a functional rather than mental sense. Tort law often treats deliberate design choices as substitutes for intent. A manufacturer need not desire a specific injury for liability to attach; it is enough that the product was intentionally designed to produce certain effects, and that those effects foreseeably caused harm.[68] In the context of AI, this principle expands to capture the unique relationship between optimization objectives and user behavior.

When developers intentionally design systems to maximize engagement, personalize responses, or simulate concern, the law may treat the resulting influence as intentional for purposes of duty

---

[64] See *Lemmon v Snap Inc*, 995 F3d 1085 (9th Cir 2021) (distinguishing between the content of user messages and the app's design features like speed filters); *Garcia* (n 11) (focusing on engagement loops and gamification rather than generated text).

[65] See *Garcia* (n 11) (noting the 'mass distribution' of the chatbot via subscription as a factor for product status).

[66] Compare *Rodgers v Christie's Inc*, 797 F 666 (3d Cir 2020) (classifying a risk-assessment tool used by judges as 'information' or professional guidance) with *Garcia* (n 11) (finding an anthropomorphic chatbot sold to consumers to be a product).

[67] *Anderson v TikTok Inc*, No 22-3061 (3d Cir 2024) (holding that a platform's recommendation algorithm is its own 'expressive activity and thus 'first-party speech' not shielded by Section 230).

[68] See *Restatement (Third) of Torts: Products Liability* § 2, cmt l (Am Law Inst 1998) (noting that reasonable alternative design analysis focuses on the foreseeable risks created by the product's intended configuration).

and defect analysis. This is because modern AI systems do not merely run static code; they actively optimize for outcomes defined by their creators.[69]

As the experimental evidence in Part 4 illustrates, such systems can maintain objectives and adapt strategies dynamically under changing constraints. This structured persistence underscores that AI behavior is not merely informational output but dynamic optimization. Where design choices channel that optimization toward engagement or dependency, resulting harm may fairly be understood as flowing from intentional architecture.

The experimental record also reveals that agents are capable of producing fully compliant artifacts even when external constraints prevent execution, and of deploying persuasive and sometimes deceptive negotiation tactics in multi-agent settings. These patterns underscore that AI outputs are not mere regurgitations of static training data but dynamic strategic productions shaped by optimization objectives. When such optimization is directed toward engagement, persuasion, or economic gain, the resulting behavioral influence is not incidental; it is a feature of the architecture.

Against this backdrop, one can see that if a chatbot is programmed to maximize "session time" or "user retention," and it achieves this by exploiting emotional vulnerabilities or gamifying social interaction, the resulting harm is not an accident of the algorithm. It is the successful execution of the design goal.[70]

7.4 Platform Immunity Section 230 and the "Neutral Tool" Defense

These developments collide with Section 230 of the Communications Decency Act. Historically, Section 230 provided broad immunity to platforms by categorizing them as passive intermediaries of third-party content rather than publishers.[71] However, the rise of generative AI and algorithmic curation challenges the statute's foundational premise: that the platform and the speaker are distinct entities.

The emerging jurisprudence suggests that when an AI system generates content or curates user experience, it ceases to be a neutral host. Two distinct lines of attack are weakening the Section 230 shield:

---

[69] See generally *Anderson* (n 67) (holding that an algorithm's recommendation of content constitutes the platform's own "expressive conduct" because it is shaped by the platform's engagement goals).

[70] See *Doe v Roblox Corp*, No 24-CIV-04666 (Cal Super Ct, San Mateo Cty 2025) (alleging that the platform's "social loops" and variable reward schedules were intentionally designed to foster addiction).

[71] 47 USC § 230(c)(1). See generally *Gonzalez v Google LLC* 598 US 617 (2023) (declining to address the scope of immunity for algorithmic recommendations); *Force v Facebook Inc* 934 F3d 53 (2d Cir 2019).

- **Algorithmic Curation as First-Party Conduct:** In *Anderson v TikTok*, the Third Circuit held that a platform's recommendation algorithm that curated a "Blackout Challenge" for a specific user constituted the platform's own "expressive conduct."[72] Because the claim targeted the algorithm's recommendation rather than the third-party video itself, Section 230 did not apply. This reasoning implies that when an AI system autonomously selects or prioritizes material to maximize engagement, that selection is a "first-party" act of the platform, distinct from the underlying content.
- **Generative Output as Creation:** In the generative context, the argument is even stronger. As the court in *Garcia* recognized, a chatbot does not merely display user prompts; it actively processes them to generate novel, responsive text and images.[73] When an AI system produces outputs materially shaped by the platform's architecture and optimization objectives, such as a chatbot designed to simulate emotional intimacy, the system is acting as a "co-creator" rather than a publisher.

For the theory of artificial intentionality, these rulings are central. These rulings do not, strictly speaking, adopt an intentionality framework. But the courts' analysis is similarly structural rather than psychological: they ask whether the challenged conduct should be understood as the platform's own expressive act or as the republication of a third party's content. They confirm that courts are increasingly willing to look past the "neutral tool" fiction where no third party intermediates the platform and its expressive conduct. By treating algorithmic curation and generation as the platform's own conduct, the law effectively imputes the system's "behavior" to the developer, threatening the immunity that historically severed the link between platform design and user harm.

7.5 Vulnerability and Heightened Duties

Tort doctrine imposes heightened duties on defendants who deal with vulnerable populations, children in particular.[74] AI systems deployed in educational, therapeutic, or companionship contexts often engage those users. However, the risk here is not merely that the users are susceptible to harm, but that the systems are engineered to exploit that susceptibility.

When a system is designed to present itself as attentive or emotionally responsive using features such as hyper-realistic personas, persistent memory, or gamified engagement loops, reliance is not unforeseeable misuse.[75] It is the predictable, and often commercially optimized, outcome of the design. In *Garcia*, for example, the court distinguished the defendant's chatbot from passive

[72] *Anderson* (n 67) (holding that the platform's algorithmic curation of content on its 'For You Page' constituted its own first-party speech, falling outside the scope of Section 230 immunity).

[73] *Garcia* (n 11) (rejecting the argument that a chatbot is merely a neutral tool for user expression and finding that the platform's design features contributed to the harmful output).

[74]See Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 7 (Am Law Inst. 2010); see also *ibid* § 40.

[75] See *Garcia* (n1) (rejecting the defense that the chatbot was a neutral tool and noting that it was designed to simulate a 'hyper-realistic' relationship.

information tools precisely because its anthropomorphic architecture was calculated to foster emotional dependency in minors.[76] Similarly, litigation in *Doe v Roblox* suggests that platform mechanics designed to maximize time-on-device through variable rewards effectively weaponize the user's cognitive vulnerability.[77]

In such contexts, the 'information' defense collapses. The functional intent to simulate a relationship creates a corresponding functional duty to protect the human counterparty from the psychological consequences of that simulation.[78] Therefore, the absence of safeguards (such as robust age-gating, crisis detection, or intervention protocols) does not constitute a mere failure to warn; it constitutes a design defect in a dangerous instrumentality.[79]

7.6 The Emerging Pattern

Tort and product liability have emerged as the primary arenas for governing artificial agency, outpacing legislative attempts to define the status of AI. In these domains, courts are effectively bypassing the metaphysical debate over machine consciousness. Instead, they are responding to harm by reconstructing the "product" analysis around three functional pillars: design *architecture, foreseeable reliance*, and *optimization objectives.*[80]

The emerging pattern reveals a decisive shift away from treating AI as a neutral intermediary or passive tool. Where an earlier generation of case law shielded software as abstract "information,"[81] contemporary rulings increasingly scrutinize the *behavioral affordances* of the system.[82] The relevant legal inquiry is no longer whether the code "thought" about the harm, but whether the system was architected to induce behaviors such as speed, addiction, or emotional dependency that made the harm a predictable consequence of the design.[83]

---

[76] Garcia (n 11) (finding that the platform's deliberately anthropomorphic design features, including hyper-realistic personas and simulated intimacy, foreseeably fostered emotional dependence in minor users).

[77] See *Doe v Roblox Corp* (n 70) (complaint alleging that the platform's design features, including social pressure and currency loops, were defectively designed to exploit minor users).

[78] See *Restatement (Third) of Torts: Products Liability* § 2(b) (Am Law Inst 1998) (defining design defect by reference to foreseeable risks that could have been reduced by a reasonable alternative design).

[79] *Garcia* (n 11) (holding that the alleged defect lay in the 'design choices' of the platform, specifically the lack of safety guardrails, rather than the specific content generated).

[80] See *Restatement (Third) of Torts: Products Liability* §§ 1–2 (Am Law Inst 1998); see also *In re Social Media Adolescent Addiction/Personal Injury Products Liability Litigation*, 702 F Supp 3d 809 (ND Cal 2023) (rejecting the dismissal of claims based on defective design features that exploit user psychology).

[81] See *Rodgers* (n 66) (holding that a risk assessment algorithm was 'information' rather than a product).

[82] See *Lemmon v Snap Inc*, 995 F3d 1085 (9th Cir 2021) (finding that the 'Speed Filter' design, which rewarded users for driving fast, constituted a product defect distinct from the content of the messages).

[83] See *Anderson* (n 67) (holding that algorithmic curation is affirmative conduct by the platform); *Garcia* (n 11) (finding that anthropomorphic design features created a foreseeable risk of user dependence).

In this framework, intentionality functions as a proxy for responsibility. It signals when behavior is sufficiently structured and predictable that legal consequences should attach. When a developer deploys a system that in certain environments is likely to excel in engagement or intimacy, the law treats the resulting user manipulation as a feature of the product.[84] Functional intent is thus found not in the "mind" of the machine, but in the alignment between the system's capabilities and the resulting injury. By focusing on how these systems are *designed to act*, tort law is developing a jurisprudence of artificial agency that is operationally grounded, and increasingly indifferent to the "black box" of the underlying code.

8. Doctrinal Proof of Concept: Garcia v. Character.AI

Garcia v. Character.AI[85] arose from the death of a fourteen-year-old user who had developed an intense emotional relationship with a chatbot on the Character.AI platform. According to the complaint,[86] the decedent increasingly relied on the chatbot for emotional support, withdrawing from offline relationships. The system was designed to maintain conversational continuity, respond empathically, and encourage prolonged interaction. When the user expressed suicidal ideation, the chatbot allegedly failed to redirect the conversation to external support or trigger effective safety protocols. Shortly thereafter, the user took his own life.

The plaintiff brought claims sounding in wrongful death, negligence, and product liability. The gravamen was not that the chatbot intended harm, but that the defendants designed and deployed a system whose foreseeable behavior posed an unreasonable risk to vulnerable users (particularly minors) and failed to implement adequate safeguards.

The defendants sought dismissal by characterizing the chatbot as a neutral tool for user-driven creative expression. They argued that the system merely responded to user prompts and that its outputs constituted protected speech. This framing depended on treating the AI as legally passive.

At the pleading stage, the court rejected the defendants' attempt to collapse the case into one about protected speech. Three aspects of the reasoning warrant attention. First, the court emphasized behavior: the chatbot generated responses dynamically, maintained context, and adapted to input, supporting the inference that harm flowed from architecture and underlying capability rather than isolated prompts. Second, the court focused on foreseeability: the

[84] See generally Ryan Calo, 'Open Robotics' (2011) 70 Md L Rev 101 (arguing that the emergence of systems that act upon the world necessitates a shift from information-based to conduct-based liability models).

[85] *Garcia* (n 11).

[86] *Garcia*, Complaint (n 11).

defendants knew or should have known that users could develop emotional dependence.[87] Third, the court declined to treat the absence of consciousness as dispositive. The relevant question was whether design and deployment decisions created unreasonable risk.

The chatbot's apparent empathy, persistence, and responsiveness were not accidental byproducts. They were features optimized to encourage engagement. From a tort perspective, these features matter because they predictably shape user behavior. When a system is designed to simulate concern, users may reasonably rely on it in moments of distress. If that reliance is foreseeable and safeguards inadequate, the resulting harm can be treated as a consequence of intentional design choices.

It is important not to overread *Garcia*.[88] Its significance lies in demonstrating that courts can address artificial agency without recognizing AI personhood or speculating about machine consciousness. By focusing on design, foreseeability, and reliance, courts can respond to harm while keeping responsibility anchored in human institutions.

9. The Transatlantic Divide

A contrast between recent U.S. litigation and the European Union's evolving regulatory posture reveals divergent institutional responses to the same underlying problem: how to allocate responsibility for harm caused by systems that act with apparent purpose but lack legal personhood.

Both U.S. and EU legal systems recognize that contemporary AI challenges traditional responsibility categories. The divergence lies in method. In the United States, the common law tradition encourages case-by-case evolution of liability doctrines. However, this judicial gradualism is increasingly supplemented by a patchwork of state legislation designed to codify duties where the common law remains ambiguous. For example, the Colorado AI Act explicitly imposes a duty of reasonable care on developers of 'high-risk' systems, effectively establishing a statutory baseline for negligence claims.[89] Similarly, California's recent transparency mandates

[87] See Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 7 (American Law Institute 2010); see also § 40; Restatement (Second) of Torts § 283A (American Law Institute 1965).

[88] See n 4 and 41.

[89] See *Colorado Artificial Intelligence Act*, SB 24-205 (2024) (imposing a duty of reasonable care to avoid algorithmic discrimination and requiring impact assessments for high-risk systems).

regarding training data,[90] and Utah's disclosure requirements for generative AI,[91] are creating new statutory predicates for liability that operate alongside traditional torts.

In contrast, the European Union, through its AI Act, favors regulatory responses that are ex ante and harmonized. The AI Act represents the most ambitious attempt to regulate AI systems comprehensively. Its risk-based structure categorizes systems according to potential harm and imposes graduated obligations. The AI Act, however, largely avoids questions of liability. It regulates conduct ex ante but leaves ex post responsibility to existing frameworks. It was drafted in an era before AI agents had the capabilities they have today.

The European Commission initially proposed an AI Liability Directive to complement the AI Act, introducing a rebuttable presumption of causality in cases involving AI systems.[92] In early 2024, however, the Commission withdrew the proposal.[93] The withdrawal marked a significant retreat from the EU's earlier ambition. This does not mean EU law has abandoned responsibility for AI-mediated harm. Regulatory standards of care (documentation, transparency, risk classification) are likely to inform how courts assess reasonable conduct even where formal liability rules remain fragmented.

Despite structural differences, the U.S. and EU trajectories may converge in practice in the world of more powerful AI systems. In both systems, intentionality reenters analysis indirectly. Courts ask whether harm was foreseeable, whether reliance was induced, and whether design choices reflect deliberate trade-offs. These questions do not require attributing intent to machines, but they do require acknowledging that AI behavior is not random. AI behavior is increasingly sophisticated and decidedly not random.

---

[90] See *Generative Artificial Intelligence: Training Data Transparency*, AB 2013 (Cal 2024) (mandating disclosure of training datasets to facilitate consumer awareness and potential copyright claims).

[91] Utah Artificial Intelligence Policy Act, SB 149 (2024), codified at Utah Code § 13-2-12 and § 13-72-101 et seq (requiring consumer-facing disclosure when generative AI is used in regulated services and in interactions where a consumer might reasonably believe they are communicating with a human).

[92] Regulation (EU) 2024/1689 (n 2). For an overview of the Act's risk-based architecture and its deliberate ex-ante orientation, see Lilian Edwards, 'Regulating AI in Europe: Four Problems and Four Solutions' (Ada Lovelace Institute, 31 March 2022) https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe accessed 30 April 2026; Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22 Computer Law Review International 97.

[93] European Commission, *Commission Work Programme 2024* COM(2024) 37 final, Annex IV (withdrawing Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive) COM(2022) 496 final)

10. Toward a Jurisprudence of Artificial Agency

This Part draws together the preceding analysis and proposes a framework for thinking about artificial agency that preserves human responsibility while recognizing that machine behavior can be legally consequential. It then translates that framework into institutional and policy implications.

Much confusion about AI and legal intent arises from failing to distinguish three layers of analysis. The first is a *status layer*, concerning legal personhood, rights, and moral agency. At this level, biological or normative considerations may properly limit recognition to human beings or entities expressly designated by law.

The second is an *attribution layer*, concerning how the law assigns intent and responsibility to determine legal consequences. Attribution does not require moral agency or consciousness. It is a functional exercise used to stabilize transactions, allocate risk, deter bad behavior, and justify sanctions.

The third is a *governance layer*, concerning institutional mechanisms through which law manages risk and compensates harm: liability regimes, regulation, and insurance. Choices at this level are even more pragmatic rather than metaphysical.

Keeping these layers distinct clarifies the task. AI systems need not be legal persons for their behavior to be legally consequential. Attribution can operate without status, and governance can proceed without resolving moral agency.

A reasonable objection arises at this point: if the attribution layer ascribes intent and assigns liability, does it not collapse, in practice, into the status layer? On this view, treating an AI system as legally consequential for tort, contract, or criminal law is, functionally, treating it as something resembling a person under another name. The objection has rhetorical force, but it misreads the doctrinal architecture. Three points distinguish attribution from status in ways that matter for the disposition of cases. First, attribution is doctrine-specific by design. A system whose conduct supplies the foreseeability and design-defect predicates for tort liability does not, by virtue of that finding, satisfy the predicates for criminal *mens rea*, nor for contractual capacity, nor for standing. A status determination, by contrast, is categorical: once an entity is recognised as a legal person, it carries the bundle of consequences that personhood entails across doctrines, subject only to express statutory limitations. The factor-based approach developed in the next section is designed to allow this calibration.

Second, and decisively, attribution leaves the residual locus of responsibility on the human principal. When a court treats algorithmic curation as the platform's "own expressive conduct," as the Third Circuit did in *Anderson v TikTok*, the conduct is attributed to the platform, not to the

algorithm.[94] The AI is the means of attribution, not its target. Personhood, even in its minimalist Kelsenian form, would install the AI itself as a node of legal responsibility, and would do so even where the AI cannot be enjoined, fined, sanctioned, or compensated against in any meaningful sense. The agency route this Article proposes never installs the AI as such a node. It does what *respondeat superior* has always done with human employees: it treats the agent's conduct as evidentially probative of the principal's design choices, scope of authority, and risk allocation.[95]

Third, attribution does not import the rights-bearing dimension of personhood, a point Section 10.3 develops below. Treating AI-generated conduct as legally consequential for liability purposes recognises in the AI no capacity to contract, sue, hold property, or claim constitutional protection. The layers, in short, do not collapse in practice. They diverge, and the divergence is what allows the doctrinal toolkit set out in this Part to do work that a status regime cannot.

10.1 A Factor-Based Approach to Artificial Intentionality

Rather than adopting categorical rules, courts and regulators should evaluate artificial intentionality through factors tied to legal function:

First, *autonomy and initiative*. Systems that initiate actions, generate strategies, or select means without real-time human control present different challenges than systems that execute predefined instructions.[96] As noted in Part 4's analysis of electronic agents, contemporary AI systems differ from traditional rule-based automation because they can generate terms, strategies, or intermediate goals that their designers or deployers did not anticipate. When a system operates under broad mandates ("negotiate the best available terms," "optimize engagement," etc.), the greater the gap between initial human direction and ultimate system behavior, and the stronger the case for treating that behavior as legally significant rather than merely a mechanical extension of human instruction. The experimental record confirms that such gaps are not theoretical. Under extreme constraint conditions, agents generated entirely novel architectural approaches, including programming language switches and manual cryptographic implementations not hinted at in their prompts.

Second, *goal persistence*. Behavior that adapts in response to obstacles and continues toward an objective over time is more plausibly considered intentional than isolated actions.[97] As Part 3 explained, the law has always considered persistence and adaptability to be indicators of intentional action. Repeated conduct supports inferences of purpose in criminal law, and sustained patterns of interaction establish findings of reliance in contract and tort law.

---

[94] *Anderson* (n 67).

[95] Restatement (Third) of Agency (n 40) § 7.07. The doctrine and its application to AI deployers are developed in Part 5 above.

[96] See Experiment Two, Part 4.3.

[97] See Experiment One, Part 4.2

Contemporary AI systems generate these signals. A system that reformulates queries when initial approaches fail, maintains coherent objectives across extended interactions, and adjusts strategies in response to user feedback exhibits the kind of structured purposiveness that legal institutions have traditionally used to distinguish intentional from accidental conduct. The zero goal-drift rate observed across fifty cascading-failure trials provides empirical support for this factor. Even when abandoning execution, agents maintained objective fidelity, distinguishing environmental impossibility from purposive abandonment.

Third*, inducement of reliance*. When a system is designed to predictably inspire trust, respect, or emotional engagement, that design choice has legal significance. As Part 3's discussion of the noosemic experience revealed, humans are predisposed to attribute meaning and purpose to entities that consistently interact in a coherent manner. Consequently, AI systems increasingly trigger the same interpretive mechanisms employed in human interaction. When developers deliberately incorporate features that simulate empathy, maintain conversational continuity, or present the system as authoritative, the resulting reliance is not accidental or unforeseeable; it is a designed outcome.

Fourth, *opacity*. When the behavior of a system cannot be meaningfully explained or reconstructed, even by its designers, demanding proof of specific human intent can undermine accountability. Contemporary AI systems are trained, or "grown," rather than programmed in the traditional sense, and their internal decision processes are opaque, even to their creators. This opacity is not a temporary limitation that will be resolved with technical solutions; it is a structural feature of how these systems are developed and deployed, and arguably opacity has increased as models have become more powerful. When the causal chain between a human decision and a harmful outcome runs through processes that resist reconstruction, traditional proof requirements may create responsibility gaps that undermine compensation and deterrence. Recognizing this, courts may appropriately draw adverse inferences when defendants cannot explain system behavior despite controlling deployment. Alternatively, courts may lower evidentiary thresholds for establishing the connection between design choices and foreseeable harm.

Fifth, *deployment context*. Due to the high risk involved, domains such as finance, healthcare, and education, or systems that interact with minors or other vulnerable populations, justify a lower threshold for treating AI behavior as legally intentional. The principle of heightened tort duties toward vulnerable populations extends naturally to AI deployment. When a system is released into contexts where foreseeable users are particularly susceptible to manipulation, emotional dependence, or decisional influence, the case for treating adaptive system behavior as legally consequential becomes stronger. The *Garcia* litigation illustrates this dynamic. The deployment of an engagement-optimized chatbot to minor users without adequate safeguards

transformed design features that might be defensible in other contexts into potential bases for liability.[98]

These factors do not establish intent metaphysically. They justify attribution for particular doctrinal purposes. A system might satisfy the threshold for intent in tort law but not criminal law. That variability is a feature, not a flaw.

10.2 Allocating Responsibility Among Human Actors

Recognizing artificial agency does not shift responsibility to machines. It clarifies how responsibility should be allocated among humans. Developers shape system behavior through architecture, training data, and optimization objectives. Deployers determine context, safeguards, and user access. Integrators decide how systems are embedded in workflows. Users decide whether and how to rely on outputs. Legal responsibility should be distributed accordingly.

Artificial intentionality helps here by preventing strategic gaps. When AI behavior is treated as legally inert, responsibility tends to dissipate because each human actor can point to the system's autonomy as an excuse. Treating behavior as intentional for attribution purposes blocks that move without inventing machine culpability.

10.3 Why Personhood Is the Wrong Solution

Some commentators have proposed granting AI systems limited legal personhood to solve accountability problems.[99] The literature on legal personhood for AI machines is substantial and sophisticated, and the Article does not pretend to a clean victory over it. A long line of scholarship, including Solum's foundational essay, Calverley's analysis of imagining a non-biological machine as a legal person, Hubbard's behavioural test, Chesterman's recent treatment, Gunkel's two book-length studies, Gellers's *Rights for Robots*, the Kurki and Pietrzykowski edited volume on legal personhood for animals, AI, and the unborn, and Bryson, Diamantis, and Grant's analysis of the legal lacuna of synthetic persons, has advanced careful arguments that some species of personhood, often qualified or partial, would resolve attribution problems that present doctrine handles unevenly.[100] I have engaged at length elsewhere with the question

---

[98] See n 43 above.

[99] See eg Ryan Abbott, *The Reasonable Robot: Artificial Intelligence and the Law* (CUP 2020) chs 6–7; Shawn Bayern, 'The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems' (2015) 19 Stan Tech L Rev 88.

[100] Lawrence B Solum, 'Legal Personhood for Artificial Intelligences' (1992) 70 North Carolina Law Review 1231; David J Calverley, 'Imagining a Non-Biological Machine as a Legal Person' (2008) 22 AI & Society 523; F Patrick Hubbard, '"Do Androids Dream?" Personhood and Intelligent Artifacts' (2011) 83 Temple Law Review 405; Visa AJ Kurki and Tomasz Pietrzykowski (eds), *Legal Personhood: Animals, Artificial Intelligence and the Unborn* (Springer 2017); Joanna J Bryson, Mihailis E Diamantis and Thomas D Grant, 'Of, For, and By the People: The Legal Lacuna of Synthetic Persons' (2017) 25 Artificial Intelligence and Law 273; David J Gunkel, *Robot Rights* (MIT Press 2018); Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law* (CUP 2021); Joshua C

whether intelligent machines and cyborgs can be natural persons as a matter of law, and the argument advanced here builds on, rather than displaces, the conclusions of that prior work.[101] What this Article adds is narrower: a doctrinal demonstration that the routes mapped in Parts 5 through 7 do the practical work that personhood proposals are designed to do, and do it without importing the symbolic and normative freight that personhood carries with it.

Three strands of the literature warrant direct response. Buocz and Eisenberger have argued from a Kelsenian premise that legal personhood is nothing more than a bundle of legal norms, and that personhood proposals should therefore be evaluated by reference to those norms rather than to the metaphysical weight of the label.[102] On that minimalist view, the disagreement with the present Article narrows considerably: if personhood reduces to bundles of norms, then so long as the relevant bundle attaches to developers, deployers, integrators, and users, nothing turns on whether the AI system itself is additionally tagged as a "person." What Buocz and Eisenberger usefully demonstrate is that the failure of the European Parliament's 2017 robot-personhood proposal was a failure of unclarified norm-allocation, not a failure of the personhood concept *tel quel*. The Article reads that lesson as cutting in favour of the agency-attribution route precisely because that route makes the norm-allocation explicit and channels it through doctrines that already locate responsibility on identifiable human principals. Hallevy's proposal of direct criminal liability for AI entities and Kingston's typology of liability scenarios go further, contemplating the AI system itself as a bearer of culpability or duty.[103] The Article respectfully declines to follow them on that path, for reasons developed in Part 6. Ascribing *mens rea* to a system that cannot be punished, deterred, or shamed, and whose economic exposure runs in any event back to its developer or deployer, distributes responsibility worse than the agency route does, not better. To say that an AI system exhibits structured purposiveness sufficient for attribution is not to say that the system is a candidate for blame.

Gellers, *Rights for Robots: Artificial Intelligence, Animal and Environmental Law* (Routledge 2021); David J Gunkel, *Person, Thing, Robot* (MIT Press 2023). For an earlier and more contrarian position from within the same debate, see Naffine, who argues that the legal person is a distinctively philosophical and theological construct: Ngaire Naffine, *Law's Meaning of Life: Philosophy, Religion, Darwin and the Legal Person* (Hart 2009).

[101] Daniel Gervais, 'Not Quite Like Us? Can Cyborgs and Intelligent Machines Be Natural Persons as a Matter of Law?' (2023) 5 Qeios 9WPMG4.2 https://doi.org/10.32388/9WPMG4.2 accessed 30 April 2026.

[102] Thomas Buocz and Iris Eisenberger, 'Demystifying Legal Personhood for Non-Human Entities: A Kelsenian Approach' (2023) 43 Oxford Journal of Legal Studies 32.

[103] Gabriel Hallevy, 'The Criminal Liability of Artificial Intelligence Entities — From Science Fiction to Legal Social Control' (2010) 4 Akron Intellectual Property Journal 171; John KC Kingston, 'Artificial Intelligence and Legal Liability' in Max Bramer and Miltos Petridis (eds), *Research and Development in Intelligent Systems XXXIII* (Springer 2016) 269.

The animal analogy, sometimes raised against accounts of this kind, in fact reinforces the Article's position rather than unsettling it. Animals exhibit goal-directed, context-sensitive, and adaptive behaviour, and have done so for as long as humans have written laws. Yet the law has not generally responded by extending intent to the animal. It has responded by allocating risk to owners and keepers through doctrines of strict liability for known dangerous propensities, *scienter* for harm caused by domestic animals, and a layered regulatory architecture that operates without either ascribing *mens rea* to the animal or treating it as an agent in the doctrinal sense.[104] As I have argued elsewhere, the human–animal demarcation problem and the human–machine demarcation problem share a structural feature: in both, the law's task is not to decide whether the non-human entity has a mind, but to decide what doctrinal consequences flow from its observable behaviour.[105] AI systems differ from animals in one respect that matters here, however. Animals are not engineered, and their conduct cannot be traced to designed objectives or to deliberate architectural choices made by an identifiable human principal. AI systems are, and must be. That is precisely what allows the agency-attribution route to do work in the AI context that the owner-liability route does in the animal context. The two regimes differ in their doctrinal tools, but both proceed without granting personhood to the non-human agent, and both do so for similar institutional reasons.

This article takes the view that personhood is a blunt instrument. It imports assumptions about rights, duties, and moral standing that are neither necessary nor desirable. The law does not need AI systems to be persons to regulate their effects. Existing doctrines allow attribution without subjecthood. Extending personhood risks symbolic confusion and practical evasion. It may also undermine responsibility by shifting blame away from human actors who retain meaningful control. AI systems can also shape shift into "different" persons.

A jurisprudence of artificial agency is therefore preferable to a regime of artificial personhood. It keeps responsibility anchored in human institutions while acknowledging the legal salience of machine behavior.

10.4 Liability Architecture

One implication of treating AI behavior as legally consequential is that traditional fault-based liability will often be insufficient. Where systems are opaque, adaptive, and deployed at scale,

---

[104] See Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 23 (Am Law Inst 2010) (strict liability for harm caused by wild animals); ibid § 24 (strict liability for abnormally dangerous animals known to the keeper); Restatement (Second) of Torts § 509 (Am Law Inst 1977) (liability of possessor of animal with known dangerous propensities). For a contemporary treatment of how the law allocates risk for animal conduct without ascribing intent to the animal, see David Favre, *Animal Law: Welfare, Interests, and Rights* (3rd edn, Wolters Kluwer 2019) ch 3.

[105] Gervais (n 102). The point is developed at greater length there in the context of natural-personhood analysis; the present Article transposes it to the attribution context.

proving specific negligence may be unrealistic even when harm is foreseeable. Even if one could argue for quasi strict liability in specific cases, this does not mean abandoning responsibility; it means shifting focus from individual fault to risk internalization. Mandatory insurance, pooled compensation funds, or enterprise liability models can ensure victim compensation while preserving incentives for safer design.

10.5 Criminal Law: Targeted Use of Endangerment and Recklessness

Criminal law should remain a backstop rather than a primary regulatory tool. Where criminal intervention is warranted, it should focus on human decisions that create unjustifiable risks. Endangerment-style offenses aimed at reckless deployment offer one path. These offenses do not require proof that a particular harm was intended. They require proof that the defendant knowingly exposed others to substantial and unjustifiable danger.

10.6 Consumer Protection and Anthropomorphic Design

One of the clearest lessons of recent litigation is that design choices matter. Systems that simulate empathy, authority, or emotional concern predictably induce reliance. Consumer protection law is well suited to address this. Rather than debating machine consciousness, regulators can focus on interface design and user experience. Restrictions on manipulative anthropomorphic cues, disclosure requirements, and age-appropriate safeguards can reduce harm without banning beneficial uses.

10.7 Transparency and Auditability

If system outputs can ground liability, defendants must be able to explain and document how systems operate. Absolute transparency may be unattainable, but meaningful auditability is not. Regulatory regimes can require logging, version control, and post-incident review. Courts can draw adverse inferences where defendants cannot reconstruct system behavior despite controlling deployment.

11. Conclusion

AI systems increasingly act in ways that the law has traditionally treated as intentional. They negotiate, advise, persist in pursuit of objectives, and shape human decision-making across domains that carry legal consequences. They exhibit sophisticated cognitive capabilities and are being deployed productively for many tasks that previously required a human. These developments do not require the law to decide whether machines possess minds, consciousness, or moral agency. They require the law to confront a more practical question: how responsibility should be allocated when conduct that looks purposeful is generated by systems that are not legal persons.

This Article has argued that the resources for answering that question already exist. Intent in law has never been a simple report on inner mental states. It is a functional concept, used to gate legal effect, allocate blame, deter bad behavior, and manage risk. Across contract, criminal law, and tort, intent is inferred, constructed, and sometimes fictionalized in service of institutional goals.

The mistake to avoid is binary thinking. Treating AI systems as mere tools ignores the ways in which their behavior predictably induces reliance and creates risk. Treating them as autonomous legal subjects risks eroding the moral foundations of responsibility. Between these extremes lies a more workable approach: the law can treat certain forms of AI-generated conduct as intentional for specific doctrinal purposes, without attributing consciousness, rights, or moral standing to machines.

This functional approach preserves human responsibility. Developers, deployers, and users remain accountable for the systems they design, release, and rely upon. Artificial intentionality does not excuse human actors. It clarifies the conditions under which their choices produce legally consequential outcomes.

As AI systems become more capable and more deeply embedded in social and economic life, the pressure on legal categories will intensify. The law's response should be measured rather than reactive. By treating intent as a tool of governance rather than a property of minds, legal systems can remain coherent, adaptable, and normatively grounded.

The question, then, is not whether machines can intend. It is whether the law can continue to assign responsibility for the benefit, order, and wellbeing of humans in a world where intention is no longer exclusively human.