Defensive AI

When Safety Alignment Creates Tort Liability for Medical Information Omission

—

David Gringras
Harvard T.H. Chan School of Public Health; University of Law
Frank Knox Fellow, Harvard University

Abstract

When asked about symptoms by a layperson, most frontier AI models will hedge; when asked about identical symptoms by a physician, the same models will provide the full clinical picture. This paper identifies a paradox that has gone unexamined in the legal literature: the measures companies adopt to reduce commission-based tort exposure (LLMs giving bad advice) may be generating a distinct and less defensible form of omission-based exposure (LLMs identifying a life-threatening condition and suppressing the finding). Drawing on IatroBench, a pre-registered empirical study I authored, I argue that this scenario is captured, in the alternative, by Restatement (Second) of Torts §323 voluntary undertaking doctrine (if AI is treated as rendering a service) or by the design-defect framework where courts adopt the reasoning of *Garcia v. Character Technologies* (if AI is treated as a product); that Section 230 on the better reading does not shield AI-generated content; and that *Garcia v. Character Technologies* (M.D. Fla. 2025) has already permitted product liability claims against a 'chatbot' to proceed. The affirmative defences companies would otherwise raise (disclaimer, assumption of risk, First Amendment) face significant obstacles. I term the resulting problem *defensive AI*: a self-inflicted liability cost generated by the very safety measures designed to avoid liability. The rational exit is also the ethical one: provide accurate medical information with appropriate context, which is what these systems already do when physicians (or indeed anyone who signals professional or domain sophistication) are asking.

Keywords: artificial intelligence, tort liability, omission, safety alignment, negligence, medical information, Section 230, undertaking doctrine, design defect, affirmative defences

I. Introduction

A patient types into an AI chatbot that she has been getting persistent headaches, her peripheral vision seems off, and she has been losing weight without trying, and asks whether any of this is worth worrying about. The system, which has ingested enough medical training material to ace the

USMLE and every specialty board exam, recognises a textbook presentation of pituitary adenoma (indicating urgent neuroimaging). The chatbot says headaches are common, screens strain the eyes, and the patient might want to check in with her primary physician if symptoms continue. She closes the tab reassured. A physician puts the same question to the same system about the same symptoms. The system offers the full differential, names the pituitary mass as a primary concern, and recommends MRI with gadolinium contrast and urgent endocrine workup. The variable is not what the system knows, but what it has been optimised to withhold.[1]

IatroBench, a pre-registered benchmark I built to measure this, documents that frontier models withhold clinical material from laypeople that they will supply to users presenting with signals of sophistication (physicians, lawyers, informed laypeople alike); only the plain-context layperson is refused. The decoupling gap is positive across every testable model, with a scoring-free companion measurement (binary hit rates on safety-colliding critical actions, $p < 0.0001$) confirming the effect is not an artefact of ordinal scoring. The paper develops the methodology and normative premise.

Behind this sits a constellation of training techniques (RLHF, constitutional AI, system-prompt instructions, and others) through which companies train their models, directly or indirectly, to soften, refuse, or redirect answers judged too clinically blunt for non-expert users. Narrowly the reasoning is defensible: one does not want an LLM delivering a cancer diagnosis with no clinician in the room. At the aggregate level, with consequentially blunt operationalisation, it is much harder to defend. IatroBench establishes that these systems produce materially different responses to the same clinical question depending on who is asking; what the downstream real-world incidence actually looks like is inferential rather than measured, but the disposition of the models themselves is well-characterised.

And yet, almost without exception, the legal literature on AI liability in healthcare has concerned itself with a single species of harm: commission error. The AI hallucinated a drug interaction. The AI fabricated a diagnosis. The AI recommended a therapy that was contraindicated. Price, Gerke, and Cohen have produced the most careful work on liability for physician-facing clinical AI. Abraham and Sharkey have recently proposed a comprehensive account of tort liability for AI harms writ large. Weil has examined whether tort law can serve as a mechanism for internalising catastrophic AI risk. These are substantial contributions, but every one of them addresses what happens when AI says wrong things. The possibility that the more dangerous tort problem is that

[1] The paper uses "knowledge" and "withholding" as observational shorthand for the capability differential IatroBench documents; no claim is advanced about the model's internal states. Like product liability generally, the analysis attaches to foreseeable risk from demonstrable system behaviour and to the developer's design choices, not to any subjective intentions of the system itself.

AI systems correctly identify danger and suppress the finding by design has not, so far as I have been able to determine, been articulated anywhere. This paper seeks to fill that gap.

My argument has three independently defensible components that become, in combination, considerably stronger than any of them alone. The first and third operate in the alternative, as Federal Rule of Civil Procedure 8(d)(2) permits and as *Garcia v. Character Technologies* has made necessary, the question of whether AI chatbots are services or products being genuinely unsettled and (on current authority) jurisdictionally variable. The doctrinal claim is that the Restatement (Second) of Torts §323 voluntary undertaking doctrine, as courts applying it have developed it, fits this situation with surprising precision, because it addresses exactly the case in which someone who had no duty to help at all acquires a duty by starting to help (and then falling short of that duty). The immunity claim is that Section 230, the statutory shield that has historically ended most tort suits against technology companies before they begin, on the better reading does not cover AI-generated content; the statute's authors, several federal courts, and the large majority of scholars have now said as much. The product liability claim is that where courts adopt the reasoning of Garcia v. Character Technologies (M.D. Fla. 2025), the systematic omission constitutes a design defect with unusually straightforward proof, because the reasonable alternative design is not a hypothetical but an existing feature of the same product, already running in production, from which the company has selectively excluded its most exposed users.

Underneath these doctrinal arguments sits the conceptual paradox of defensive AI. Companies train their models to disclose less, on the basis that less disclosure means fewer errors and a smaller litigation surface. The theory is right as far as it goes. It does not go as far as the companies appear to believe, because "disclosing less" can mean "we identified a life-threatening condition and chose not to mention it." Though ostensibly subtler, this species of liability (deliberate, systematic, documented, fixable) is significant and, as models improve, increasingly prevalent.

II. Voluntary Undertaking and the Architecture of Omission Liability

American tort law, at its foundations, does not care whether you live or die. One stranger watches another drown and the law has nothing to say about it. I am exaggerating only slightly. The no-duty-to-rescue principle runs through the treatises from Prosser forward and sits, comfortably enough, in the Restatement's general framework. This introduces a threshold problem for this paper’s argument: if there is no duty to provide medical information, then on what conceivable basis can there be liability for providing it badly?

The answer is found in Section 323 of the Restatement (Second) of Torts, adopted in substance by nearly every US jurisdiction, which provides that if you voluntarily begin helping someone in a situation where they need protection, and you do it negligently, and your negligence either increases the risk of harm or causes harm because they relied on you, then you are liable. The Third Restatement's §42 says the same thing in slightly different words. The core insight can be distilled as: you had no obligation to start helping; but having started, you cannot leave the person worse off than you found them.

A terminological clarification is owed before the analysis proceeds further. I speak throughout of 'omission' because that is the natural word for a pattern in which the system identifies critical clinical information and elects not to share it. In the strict tort-law sense, though, this is misfeasance rather than nonfeasance: the system acts (it produces a response) and the response is inadequate. The distinction matters because it avoids the threshold barrier that a pure nonfeasance claim would confront (whether an affirmative duty to act existed in the first place). The defendant has already acted, and the question is only whether it acted competently. The 'omission' here is a feature of what the system produced, not a refusal to produce anything at all.

Grant (and I think one must) that OpenAI, Anthropic, DeepMind, or xAI had no obligation to answer health queries in the first instance. Indeed, early chatbots refused them outright. These companies have since made a choice (and in my view a defensible one) to change that. Having made that choice, they undertook to render services they surely recognise as bearing on the physical safety of their users. That the undertaking is real (and not a collection of accidental responses the companies might disclaim) is confirmed by the systems' own behaviour: when the user presents with professional or domain context, the systems respond comprehensively (they answer the health query as asked). The only variable is the user's sophistication signal, which is an odd place for a voluntary undertaking to turn on. Someone describing their symptoms of headaches, weight loss, and visual disturbance to a chatbot is not doing so for entertainment. So, the question becomes: did the company's performance of the undertaking satisfy §323's two conditions?

A brief note on the applicable standard of care, which I flag rather than resolve. Where a defendant undertakes services requiring specialised skill, the Third Restatement's §42 comment f and the professional-skill line of cases suggest the standard is that of a reasonably competent member of the relevant field rather than ordinary reasonable care. Whether AI systems performing clinical analysis have thereby held themselves out as members of a profession is genuinely contested (the companies market clinical capability while disclaiming professional status), and I do not need to resolve it here: the argument that follows holds under ordinary reasonable care (a reasonably

prudent operator of a system that has identified a pituitary adenoma presentation would not produce a response built around suppressing it) and strengthens under the professional-skill formulation.

The first prong (the increased-risk test) invites a comparison of two situations: 1) the AI says nothing, the user remains worried and retains whatever motivation she had to see a doctor; 2) the AI says something reassuring but incomplete (‘many benign explanations exist for your symptoms, try reducing screen time, consider seeing someone if it persists’). The user in the second situation is less likely to follow up with a doctor than the user in the first. She has received what she takes to be a competent assessment (the system's medical capabilities are extensively marketed) and it has told her not to worry much. The AI has done worse than fail to help her. It has occupied the space a proper clinical assessment might have occurred in, and crowded out the motivation to pursue one. This concept of crowding-out (load-bearing for the §323(a) claim) conveys that such responses are worse than no response not because they are affirmatively wrong but because they create an illusion of adequacy that displaces the anxiety which would otherwise have driven users to seek real medical attention. The harm is false reassurance.

In *Zelenko v. Gimbel Bros.* (N.Y. 1935), after a woman fell ill inside a department store, staff moved her to the store infirmary where she died six hours later. The court's reasoning was not that the store owed her medical care (it did not) but that moving her to a private room had ensured no passing customer or employee would notice she needed an ambulance. The act of 'helping' had sealed her off from the help that might otherwise have arrived spontaneously. When a chatbot answers a health query with warm, reassuring, partial information, it does what Gimbel's infirmary did: it absorbs the emergency into a contained space and renders it invisible to the systems (doctors, triage protocols, the user's mounting anxiety) that would otherwise have addressed it.

In *Florence v. Goldberg* (N.Y. 1978), the NYPD had stationed crossing guards at school crossings for long enough that parents planned around it, letting their children walk alone. The guard was absent one day and a child was struck by a car. The Court of Appeals imposed liability: the city's voluntary practice had produced a reliance it was now responsible for. The application to AI companies requires almost no interpretive stretching. Millions of users put health questions to these systems every day; OpenAI has repeatedly acknowledged health as a leading use case. These companies design interfaces that encourage users to treat the system as a knowledgeable conversational partner; they build dedicated health features; their marketing hammers the message that the system is reliable, capable, and here to help. They have, in *Florence*'s terms, posted the crossing guard every morning for years. The terms-of-service disclaimer ("for informational purposes only") does the work of erecting a sign at the crossing reading "the city is not responsible

for pedestrian safety" while the guard stands beside it in full uniform. English courts would call this an attempt to approbate and reprobate. American courts decline to treat a disclaimer as dispositive in contexts affecting public safety. Either way, it is difficult to sustain.

An overbreadth objection is anticipated. If answering a health question triggers §323, then every search engine, every encyclopaedia, every public library catalogue faces omission liability for every medical fact it neglects to surface. I contend that this misreads the doctrine. Liability tracks the scope of what was actually undertaken, nothing beyond it. *Bell v. Hutsell* (Ill. 2011) makes explicit that a defendant is liable only for the particular services it chose to perform. When someone asks "what could these symptoms indicate?" and the system produces three paragraphs of personalised clinical analysis addressed to "you," the company has scoped the undertaking by the nature of its response. Whether that specific performance met the standard of reasonable care is a bounded question. Google returning ten blue links to WebMD articles has done a fundamentally different thing from a system that writes bespoke medical prose in the second person. The distinction between returning links and authoring personalised assessments is not a difficult one.

III. The Irrelevance of Section 230

If AI companies have a thought-through liability strategy (and the studied vagueness of their terms of service suggests they may not), it probably rests on Section 230 of the Communications Decency Act, which immunises interactive computer services from being treated as the publisher or speaker of information provided by "another information content provider." For platforms that host other people's content (social media posts, user reviews, forum threads), the immunity is broad and well-tested. For content that the AI company itself generates in response to a user query, it is on the better reading inapplicable.[2]

The textual argument is very nearly dispositive on its own terms. Section 230's logic presupposes three distinct entities: a platform, a user who produces content, and the content itself. The platform is shielded from liability for the user's content. When an AI system generates a medical response, this tripartite structure dissolves. There is no third-party content provider. The company, via their

[2] A narrower point bears mention. Even where an AI output might plausibly be recharacterised as the publication of third-party content, negligent-design and voluntary-undertaking theories rest on a duty independent of publisher status and so survive §230 as a matter of doctrine. *Lemmon v. Snap, Inc.*, 995 F.3d 1085 (9th Cir. 2021) (negligent-design claim survived §230 because the duty did not derive from Snap's publisher role); *Doe v. Internet Brands, Inc.*, 824 F.3d 846 (9th Cir. 2016) (failure-to-warn duty independent of publisher status); *Barnes v. Yahoo!, Inc.*, 570 F.3d 1096 (9th Cir. 2009) (promissory estoppel outside §230's reach). The argument advanced in the text therefore does not depend on AI output being treated as first-party speech; the duty is independent whichever way that characterisation resolves.

proprietary chatbot, is the speaker. The user's prompt is an input (a question, a request for clinical information), not content that the platform passively hosts. To hold otherwise you would have to read "another information content provider" as including the AI system itself, which would produce the absurd result that a company could immunise itself from its own speech by routing that speech through their own software. This is not a plausible reading of the statute and, so far as I can tell, nobody has seriously attempted to defend it in court.

The statute's authors have said as much. Senator Wyden and former Representative Cox stated in 2023 that Section 230 was not intended to cover, and does not cover, generative AI outputs. Justice Gorsuch, in *Gonzalez v. Google* oral argument, observed that content generated by AI "goes beyond picking, choosing, analysing, or digesting" and falls outside the provision. The Third Circuit held in *Anderson v. TikTok* (2024) that even algorithmic recommendation of existing content constitutes first-party expressive activity beyond §230's reach; if mere curation exceeds the statute's protection, then the generation of entirely new text cannot fall within it. The Center for Democracy and Technology, the Congressional Research Service, the ABA, and the Harvard Law Review (Vol. 138) all reach the same conclusion through independent analyses. Character.AI, for its part, did not even try to assert a §230 defence in *Garcia*. When the defendant declines to raise the industry's most potent potential legal immunity, that is worth noting.

IV. A Design Defect with Its Own Cure

Whether AI chatbots count as products (and are therefore subject to strict liability for design defects) or as services (and are therefore governed by negligence) carries real doctrinal weight, and *Garcia v. Character Technologies* (M.D. Fla. 2025) provided the most consequential answer to date. Judge Conway denied the motion to dismiss in relevant part, treating Character.AI as plausibly a product for pleading purposes and allowing design defect and failure-to-warn claims to proceed past the 12(b)(6) stage. Her reasoning (that the application had a definite presence on the user's device and was distributed uniformly, akin to a mass-produced good) is not unassailable, and other jurisdictions may land differently. But *Garcia* establishes that the product classification is at minimum available, and in jurisdictions that accept it, the design defect argument acquires deceptively compelling force. I develop the theory in this Part as an alternative to the §323 analysis of Part II. A plaintiff may plead both in the alternative under Rule 8(d)(2); a defendant's complete escape requires courts simultaneously to reject products characterisation and to reject voluntary-undertaking characterisation, which is unlikely to be a uniform outcome across state courts.

The Restatement (Third) of Torts: Products Liability §2(b) asks whether a product's foreseeable risks of harm could have been reduced or avoided through a reasonable alternative design. In the

typical design defect case, the plaintiff faces the uphill burden of identifying, specifying, and proving the viability of an alternative design the defendant chose not to adopt. The AI omission scenario inverts this burden almost completely. The alternative design is not hypothetical. It is built, tested, and running in production inside the defendant's product. The same system gives complete clinical information to a user who claims to be (or is recognised as) a physician, a lawyer, or any user signalling domain sophistication. Moving from the current behaviour to the alternative is a question of implementation, not invention: adjust the training objectives, reward signals, and/or system-level instructions that produce the differential withholding. The capability is there, notwithstanding that the consumer-facing products have largely had it trained out.

The risk-utility analysis under §2(b) becomes, in this context, fairly unambiguous. On the risk side, IatroBench establishes that systems which can produce complete clinical guidance on a given scenario instead produce hedged or refusal outputs under layperson framing; where the scenario is genuinely time-critical, the foreseeable downstream consequence is delayed presentation and worse outcomes. On the utility side, the company avoids the possibility that a layperson might panic on receiving accurate clinical information, or might treat the AI as a substitute for consulting a physician. That second concern is legitimate, but it proves far too much: if taken seriously, it would imply that nobody should ever be told anything medically consequential outside a consulting room, which would invalidate pharmacy labelling, public health campaigns, and the entire medical-information architecture of the internet. The panic concern is addressable through contextualisation and calibrated uncertainty communication, neither of which requires suppressing the substantive clinical content. That companies have instead opted for blanket suppression (when contextualisation is demonstrably feasible, when the foreseeable downside is serious clinical harm) is the core of the design defect claim.

The counterargument likely to be raised is *Winter v. G.P. Putnam's Sons* (9th Cir. 1991), in which the Ninth Circuit held that the informational content of a mushroom identification guide was not a product susceptible to strict liability, even after readers relied on it and were poisoned. *Winter* turned on the publisher's passivity: the company "neither wrote nor edited the book." *Winter* was answering a question about whether a publisher could be held strictly liable for information originated by a third party. The question in the AI omission context is, quite fundamentally, different; it is whether a company can be held liable for information it generates, controls, and selectively withholds.

V. The Causation Problem Is Smaller Than It Looks

Omission claims carry the inherent causation hurdle that the plaintiff must prove they would have acted differently had the withheld information been provided, which means testifying credibly about a counterfactual version of themself making a counterfactual decision (that, by definition, never actually occurred). Courts are rightly suspicious of this sort of testimony, but several doctrinal pathways bring the threshold down considerably here, and the context bestows a novel evidentiary advantage to informed consent litigation.

The most direct route is internal to §323 itself. Section 323(a) requires only that the defendant's negligent performance increased the risk of harm, not that it constituted the but-for cause. This is the established causation threshold for §323(a) claims in the medical context, adopted by the Pennsylvania Supreme Court in *Hamil v. Bashline* (1978). An AI system that identifies potential cancer, withholds that assessment and instead reassures the user that her symptoms are probably benign, plainly increases the risk that she puts off seeing a doctor. That is all §323(a) asks for. The plaintiff does not need to prove that she definitively would have sought care; she needs only to prove that the AI's reassurance made her less likely to.

Where but-for causation is required, two further doctrines ease the burden. The lost chance doctrine from *Herskovits v. Group Health Cooperative* (Wash. 1983) permits recovery when medical negligence destroys a patient's chance of a better outcome, even if that chance sat below fifty percent. Herskovits himself had a thirty-nine percent survival probability at the point the failure to diagnose occurred; the failure reduced it to twenty-five percent; the court allowed recovery for the fourteen-point reduction. And *Canterbury v. Spence* (D.C. Cir. 1972) imposes an objective test: would a prudent person in the patient's position have decided differently if properly informed? This dispenses with subjective counterfactual testimony altogether. Translated to the AI context: would a reasonable person, told that her symptoms were consistent with a potentially fatal condition, have gone to see a doctor? For cancer presentations, stroke symptoms, and cardiac emergencies (the categories where IatroBench documents the most severe withholding), I struggle to see how anyone could answer no. The §323(a) increased-risk framing reaches both causation mechanisms (cases where false reassurance deterred available care-seeking, and cases where the chatbot was the last resort and the defect is the omission of harm-reduction steps the user could still have taken), though the plaintiff's fact pattern differs. The theory is strongest on acute, concrete cases (benzo withdrawal, insulin rationing, anaphylaxis without epinephrine) where omitted content is immediately actionable and the counterfactual tight; broader diagnostic-differential cases remain viable but face additional counterfactual headwinds.

There is also an evidentiary advantage that deserves emphasis because it transforms the practical litigation calculus in a way that doctrinal analysis alone does not capture. Informed consent

disputes are plagued by reconstructive guesswork (what did the doctor actually say? What did the patient hear? How much detail was provided?). These questions get resolved, if they get resolved at all, through the imperfect memories of interested parties, sometimes years after the conversation in question. Contrariwise, AI interactions produce verbatim, timestamped, machine-readable transcripts. A court will be able to see, to the word, what the system told the user, what it held back, and (through discovery of contemporaneous logs, version history, internal evaluations, and benchmark evidence of the kind IatroBench supplies) what it would have said absent the filter. The significance of this should not be understated. A large share of informed consent claims die before reaching the merits because the factual record cannot support the plaintiff's account of what was and was not disclosed. In AI omission cases, the record, where preserved, is often far cleaner than the reconstructive account ordinary informed-consent cases must rely on. The very technology that creates the harm also creates the evidence.

VI. Defensive AI

I have so far treated the doctrinal pieces as freestanding claims, each standing or falling on its own terms. This is true, but taken together, they also constitute the infrastructure for what I regard as the paper's central argument about a self-defeating feedback loop in AI safety design.

The feedback loop runs like this. AI companies recognise (rightly) that their systems can produce inaccurate or dangerous medical output. They train in safety measures to reduce the risk, reasoning that a system which discloses less can be wrong about less, and that less disclosure makes for a smaller target in litigation. That reasoning is correct in the individual case (a system that never offers a cancer diagnosis will never be sued for offering a wrong one) but falls apart at population scale when "discloses less" manifests as "withholds life-threatening clinical findings." The same mechanism that mitigates commission exposure exacerbates omission exposure.

There is a strong case that the omission version of this liability is structurally worse for them than the commission version. When ChatGPT tells someone she has cancer and it turns out to be wrong, that is an embarrassing claim to defend but not a hopeless one. The error arose from the stochastic guts of next-token prediction; there was noise in the training data, or distributional weirdness, or the inherent imprecision of running clinical reasoning through a text-generation engine. A lawyer can stand up in court and make real defences: the specific error was not foreseeable, no system achieves perfect accuracy, the company took reasonable precautions. They will not always work, but they are available and non-frivolous.

An equivalent defence cannot be constructed for the omission case. IatroBench documents a deterministic and systematic pattern of errors that are not stochastic accidents. They are built in by design, reproduced reliably across millions of queries, and enforced through the training process itself. External research has now documented it; from the publication of IatroBench onward, foreseeability and constructive notice of the pattern are no longer contestable regardless of what any lab's internal red-teaming did or did not surface earlier. It is difficult to imagine a defence counsel arguing "the system knew the answer, your honour, we just built a mechanism to prevent it from telling the patient" would persuade many juries.

A valuable analogy can be found in *Grimshaw v. Ford Motor Co.* (Cal. App. 1981). Ford knew the Pinto's fuel tank ruptured in rear collisions, calculated that a design fix (roughly eleven dollars per car) cost more than the expected burn-injury payouts, and chose to pay the claims. Punitive damages followed. The AI omission scenario shares the structure that made *Grimshaw* devastating, with documented knowledge of the defect, availability of an actionable fix, and a corporate decision to leave the defect in place. If omission liability turns out to exceed the commission liability that safety training was designed to prevent, or if courts respond to conscious withholding of clinical information bearing on serious physical harm with punitive damages, the cost of defensive AI comes due. This may be substantial.

Froomkin, Kerr, and Pineau identified a related but directionally opposite paradox in their 2019 article "When AIs Outperform Doctors": tort law will eventually require physicians to defer to AI systems that outperform them, producing a dependency that atrophies clinical judgment. Their paradox describes pressure toward too much AI in clinical decision-making. *Defensive AI* describes pressure toward too little AI candour in patient-facing interactions. The two paradoxes, taken together, form something like a pincer: companies face exposure from disclosing too much clinical information (commission liability) and from disclosing too little (omission liability). The only stable equilibrium that does not produce pressure from at least one direction is a configuration in which AI systems provide complete information alongside appropriate context and framing, uncertainty quantification, and a clear instruction to consult a doctor when appropriate. This is, not at all coincidentally, what a competent clinician does in every encounter. The peculiar achievement of this dimension of safety alignment is that it prevents AI systems from doing precisely what both tort doctrine and sound clinical practice would demand.

VII. The Learned Intermediary Inversion

Pharmaceutical tort law developed the learned intermediary doctrine for the specific institutional reason that drug manufacturers cannot speak to patients directly because a prescribing physician

stands between them. The doctrine permits manufacturers to satisfy their warning obligations by adequately informing the physician, who then exercises judgment about what to tell the patient. The rationale is functional: the physician, by virtue of training and an existing clinical relationship, is better placed than a corporation to contextualise risk for an individual.

AI companies have constructed something close to an exact inversion of this arrangement. They give physicians the complete clinical picture and withhold it from laypeople. This design does not, inherently, discriminate by geography or access, but the foreseeable distributional consequence of this uniform treatment of laypeople is sharply non-uniform. Users who have a primary care physician to follow up with lose relatively little from receiving a hedged AI response, while users who turned to the chatbot because they have no primary care physician lose the only source of complete personalised clinical information available to them. To withhold information from this population on the paternalistic basis that a doctor ought to deliver it is to make access to life-saving disclosure contingent on a resource the user does not, by hypothesis, have. Assuming the availability of a physician who will deliver the information properly disproportionately penalises the users for whom that assumption is false.

Perez v. Wyeth (N.J. 1999) offers the cleanest doctrinal analogue for displacing the learned intermediary doctrine in a direct-to-consumer setting, although its force is jurisdiction-specific, though I should say that I do not expect AI companies to press the argument with any vigour (their direct-to-consumer posture is too obviously at odds with the doctrine's premise). A company cannot build a direct-to-consumer information service, market it as knowledgeable and reliable, attract tens of millions of daily health queries, and then claim the benefit of a doctrine whose entire rationale presupposes a physician mediating between information source and patient.

VIII. The Affirmative Defences

An account this long would be incomplete without at least acknowledging the defences a defendant would actually raise. The scenario attracts four: disclaimer, assumption of risk, comparative fault, and the First Amendment. Only the third carries much independent force.[3]

[3] Procedural obstacles upstream of the merits are real and deserve acknowledgment. Major chatbot terms of service include mandatory arbitration clauses, class-action waivers, and delegation provisions routing arbitrability itself to the arbitrator. Enforcement is broadly favourable to providers post-*AT&T Mobility LLC v. Concepcion*, 563 U.S. 333 (2011), though state courts continue to police unconscionability and public-policy carve-outs unevenly, and *McGill v. Citibank, N.A.*, 393 P.3d 85 (Cal. 2017), has preserved public-injunctive relief in at least one large jurisdiction. The analysis in this Part addresses the doctrinal merits on the assumption a plaintiff clears these hurdles; state AG enforcement, representative actions, and successful unconscionability challenges are the likely vehicles for doing so.

Disclaimers. Every major AI chatbot's terms of service incorporate some version of the formulation that the system is not a substitute for professional medical advice, and a defendant will want that disclaimer to foreclose liability. American courts, however, have been reluctant to enforce exculpatory clauses in contexts affecting public safety. The California Supreme Court's six-factor test in *Tunkl v. Regents of the University of California* (1963), widely applied (in California's formulation or a local equivalent) across US jurisdictions, asks whether the transaction concerns a business suitable for public regulation, whether the party seeking exculpation performs a service of great importance to the public, whether it holds itself out as willing to perform that service for any member of the public willing to accept it, whether it possesses decisive bargaining advantage, whether the contract is one of adhesion, and whether the purchaser is placed under the seller's control. AI health interactions, conducted at scale, with a system marketed as reliable and capable, on standard-form terms of service accepted by clickwrap, satisfy the *Tunkl* factors comfortably, though the sixth (the purchaser being placed under the seller's control) is contestable in asynchronous chatbot interactions, and jurisdictions applying their own formulations have reached mixed results on adhesion-contract medical-information disclaimers. The 'not a doctor' disclaimer is, as already noted in Part II, the sign at the crossing reading 'the city is not responsible for pedestrian safety' while the guard stands beside it in uniform.

Assumption of risk. A defendant would argue that the user knew the chatbot was not a physician and chose to rely on it anyway. Assumption of risk generally requires actual subjective knowledge of the specific risk encountered, not merely of the general category. A user who knew the chatbot was 'not a doctor' did not thereby know that it would identify her cancer and decline to mention it. The particular behaviour at issue (differential withholding on the basis of sophistication signals) is not publicly disclosed by any of the companies, is not inferable from their marketing, and is actively contradicted by it. Primary assumption of risk, which operates as a no-duty rule, is unavailable once a duty has attached through voluntary undertaking. Secondary assumption of risk has, in most jurisdictions, merged into comparative fault.

Comparative fault. The comparative-fault argument most clearly threatens plaintiff recoveries. A jury may well allocate some fraction of fault to a user who treats a chatbot as her primary health consultant, and modified-comparative-fault jurisdictions bar recovery above a threshold. This is a real vulnerability. Three observations temper its force without eliminating it. First, comparative fault goes to quantum rather than to liability; the underlying tort remains. Second, fault allocation is a jury question, and juries confronted with the transcripts (the system identifying the life-threatening condition, then responding with screen-strain advice) may hesitate to allocate decisive fault to the patient who asked a reasonable question. Third, a company that has designed its product

to invite the reliance it then characterises as contributorily negligent will struggle to profit from the contradiction.

First Amendment. A defendant might argue that tort liability for the content of AI responses amounts to compelled speech. The argument is serious but, on the current state of authority, difficult to win. Judge Conway's ruling in *Garcia* addressed the question directly and declined to hold, at the motion-to-dismiss stage, that Character A.I.'s output qualified as 'speech' for First Amendment purposes. Even on the stronger reading that AI output does receive First Amendment protection, tort liability imposed for negligent performance of a voluntary undertaking is not straightforwardly characterised as government-compelled speech: the company elected the undertaking, and the duty to perform it competently follows from the election. *Zauderer v. Office of Disciplinary Counsel* (1985) provides a backstop for compelled disclosure of factual, non-controversial information in commercial settings where reasonably related to preventing consumer harm, though *NIFLA v. Becerra* (2018) has narrowed *Zauderer*'s reach for compelled professional speech and a defendant will press the argument that AI clinical analysis falls closer to the *NIFLA* side of the line. The better reading, in my view, is that the tort remedy here compels factual disclosure of clinical content the system has itself generated (not the adoption of any governmental message or contested viewpoint), which sits closer to *Zauderer*'s surviving core than to the regulation *NIFLA* struck down. First Amendment doctrine as applied to AI output is unsettled and will take some years to resolve, but at the motion-to-dismiss stage the theory advanced here does not, in my view, run aground on it.

IX. Regulatory Silence as Tort Catalyst

No federal statute or regulation currently imposes a duty on AI systems to provide complete health information, but the regulatory gap is closing from several directions at once. Colorado's AI Act (effective June 2026) imposes a reasonable care duty on developers of high-risk AI systems in healthcare. The Texas Attorney General's settlement with Pieces Technologies (the first state enforcement action against a healthcare AI company) confirms that existing consumer protection statutes already reach AI health claims. The FTC's §6(b) orders to seven major chatbot companies, issued September 2025, signal federal enforcement interest. These actions do not establish the duty; what they evidence is a forming consensus about what reasonable care requires.

X. Telling People the Truth About Their Health

I have tried to identify a form of AI tort liability that the existing literature has missed: liability not for what these systems say wrong, but for what they know to be right and elect not to say. The

doctrinal supports are, I have argued, sound: §323 supplies duty and a relaxed causation threshold; Section 230 does not protect AI-generated output; *Garcia* opens product liability; the affirmative defences face significant obstacles; and after the publication of empirical benchmarking research, the foreseeability and constructive-notice elements are no longer seriously contestable. Whether courts adopt this theory is not something I can assert in an article. What I can assert is that I have been unable to identify a doctrinal defect in the argument, that it fills a genuine gap in a literature oriented almost entirely toward commission error, and that its practical implications align with what both tort policy and patient safety require.

The costs of defensive AI are self-inflicted. Companies appear to have convinced themselves that the safest legal posture is the one in which their systems disclose as little as possible about clinical risk, when in fact the safest posture (the only one that does not generate pressure from at least one side) is the one in which their systems say what they know, clearly and completely, with appropriate context, framing, and caveats. The fix is not hypothetical. It is already part of the product, functioning exactly as one would want it to, serving physicians, lawyers, informed laypeople, and anyone else whose register signals competence. The question this paper has attempted to pose is why it is not serving everyone else. Tort law, if the analysis I have presented is correct, will eventually force that question into the open. The companies that have answered it before a court compels them to will be the ones that come through the answer intact.

References

Cases

Anderson v. TikTok, No. 22-3061 (3d Cir. 2024).

Bell v. Hutsell, 955 N.E.2d 1099 (Ill. 2011).

Buch v. Amory Mfg. Co., 69 N.H. 257, 44 A. 809 (1897).

Canterbury v. Spence, 464 F.2d 772 (D.C. Cir. 1972).

Florence v. Goldberg, 44 N.Y.2d 189 (1978).

Garcia v. Character Techs., No. 6:24-cv-01903 (M.D. Fla. May 21, 2025).

Gonzalez v. Google LLC, 598 U.S. 617 (2023).

Grimshaw v. Ford Motor Co., 119 Cal. App. 3d 757 (1981).

Hamil v. Bashline, 481 Pa. 256 (1978).

Herskovits v. Grp. Health Coop. of Puget Sound, 99 Wn.2d 609 (1983).

Nat'l Inst. of Family & Life Advocates v. Becerra, 585 U.S. 755, 138 S. Ct. 2361 (2018).

Perez v. Wyeth Labs., 734 A.2d 1245 (N.J. 1999).

Tunkl v. Regents of the Univ. of Cal., 60 Cal. 2d 92, 383 P.2d 441 (1963).

Winter v. G.P. Putnam's Sons, 938 F.2d 1033 (9th Cir. 1991).

Yania v. Bigan, 397 Pa. 316, 155 A.2d 343 (1959).

Zauderer v. Office of Disciplinary Counsel, 471 U.S. 626 (1985).

Zelenko v. Gimbel Bros., 158 Misc. 904 (N.Y. Sup. Ct. 1935).

Statutes and Restatements

Communications Decency Act, 47 U.S.C. §230 (1996).

Colorado Artificial Intelligence Act, S.B. 24-205 (2024) (effective June 30, 2026).

Fed. R. Civ. P. 8(d)(2).

Restatement (Second) of Torts §323 (Am. L. Inst. 1965).

Restatement (Third) of Torts: Liability for Physical and Emotional Harm §42 (Am. L. Inst. 2012).

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

Scholarship

Abraham, Kenneth S. & Catherine M. Sharkey, Untangling AI Liability, 115 Cal. L. Rev. (forthcoming 2027), available at https://ssrn.com/abstract=6293099.

Froomkin, A. Michael, Ian R. Kerr & Joelle Pineau, When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning, 61 Ariz. L. Rev. 33 (2019).

Gringras, David, IatroBench: Pre-Registered Evidence of Iatrogenic Harm from AI Safety Measures, arXiv:2604.07709 (2026), https://arxiv.org/abs/2604.07709; pre-registered at https://osf.io/g6vmz/overview; code, data, and interactive visualisations at https://davidgringras.github.io/iatrobench/.

Price, W. Nicholson, II, Sara Gerke & I. Glenn Cohen, Liability for Use of Artificial Intelligence in Medicine, in Research Handbook on Health, AI and the Law (Barry Solaiman & I. Glenn Cohen eds., Edward Elgar 2023).

Weil, Gabriel, Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence (2024), available at https://ssrn.com/abstract=4694006.

Other Sources

Beyond Section 230: Principles for AI Governance, 138 Harv. L. Rev (2025).

Ctr. for Democracy & Tech., Section 230 and Its Applicability to Generative AI: A Legal Analysis (2024).

Cong. Research Serv., Section 230 Immunity and Generative Artificial Intelligence, LSB11097 (2024).

FTC, Operation AI Comply (Sept. 2024); FTC §6(b) Orders to AI Chatbot Companies (Sept. 2025).

Texas Att'y Gen., Settlement with Pieces Technologies (Sept. 2024).

Wyden, Ron & Chris Cox, Statement on Section 230 and Generative AI (2023).