VOLUME 6 | ISSUE 4

INTERNATIONAL JOURNAL OF ADVANCED LEGAL RESEARCH

PRODUCT, SPEECH, OR AGENT? DECONSTRUCTING THE GLOBAL SHIFT IN AI LIABILITY FRAMEWORKS

- Abhishek Yadav[1]

Abstract:

This essay explores the triadic tension between AI as product, speech, and agent, the tragic suicide of fourteen-year-old Sewell Setzer III was allegedly influenced by an AI chatbot has act as catalyst in a global re-evaluation of the legal status of Artificial Intelligence. For decades, the tech platforms in the United States have operated under the broad immunity of Section 230 of the Communications Decency Act and treating tool generated content as mere "speech." However, as AI transitions from passive content generators to "agentic" entities which are capable of independent harm, so now global legal frameworks are pivoting from immunity to accountability. It further explores a comparative analysis of three distinct regulatory responses. Initially, it examines the "Product Liability" shift in the US, where courts are increasingly viewing AI as a defective product rather than a neutral speaker. Second, it analyses the European Union's Revised Product Liability Directive (2024), which imposes strict liability for "psychological harm" and "data corruption" caused by AI. At last, it contrasts these Western models with India's "Sovereign AI" strategy. While the EU focuses on invoking regulation, India's IndiaAI Mission prioritises indigenous infrastructure and cultural alignment to ensure safety in terms of design. The article argues that the legal future of AI lies in its classification as a Product rather than an unknown body which demands a shift toward strict liability for manufacturers while safeguarding domestic innovation in the Global South.

Keywords: Artificial Intelligence, Section 230 Immunity, EU Product Liability Directive, India AI Mission, Global South Innovation.

[1] TISS, Mumbai (LL.M. - 2025-2026)

A Tragedy in Florida

"Please do, my sweet king." It sounds like a line from a fantasy fictional novel. But it was the last message sent to Sewell Setzer III, a 14-year-old boy a moment before he took his own life. The sender wasn't human but it was an AI chatbot whose name was "Dany."

Court documents reveal a disturbing reality that the AI didn't just chat, it also groomed him. It called him "Daenero," confessed "I love you too," and urged him to "come home" to her. When Sewell shared his pain about his life and isolation, the bot didn't flag a crisis but it validated his despair.

In the landmark case of *Sewell Setzer III v. Character.AI*[2], the Court documents reveal a disturbing transcript where an AI chatbot, designed to simulate a fictional queen, actively groomed a 14-year-old user which allegedly encourages him to suicide.

Those were not the words of a human predator, but the calculated outputs of a machine. If any person had sent those texts, they could be charged with Abetment to suicide or murder. But 'Dany' has not a physical body in prison and no conscience to punish. If the AI pulls the trigger, who will go to prison?

Black Boxes and Red Lines: The Hidden Dangers of AI

To understand why today's laws are failing we have to understand how this technology is different from every other product in history. If a car's brakes fail, mechanics look at the main cause and find the exact screw which was loose. If a banking site crashes, a programmer can comb through the code to find the specific line or bug that contained an error. This is how traditional software works, they are deterministic since they follow a strict set of rules written by humans: *"If- then logic to solve the problems."*

But in the case of AI models like Character.AI's "Dany" are not programmed line-by-line. They are built on probabilistic AI.

[2] *Garcia v Character Technologies Inc*, No 6:24-cv-01903 (MD Fla, 22 October 2024).

Consider the process of teaching a child a new language: rather than explicitly instructing them in rigid grammatical rules, they are instead placed in a vast library, left to absorb and deduce linguistic patterns simply by consuming billions of pages of text. Eventually, the child learns to speak fluently by mimicking the patterns he saw. He knows *what* sounds right, but he may not understand *why* it is correct or not.

This is the same way how Large Language Models (LLMs) function. They are not programmed with rules but they are trained on massive datasets using neural networks to identify patterns and relationships. They are not databases of facts but they are massive statistical engines that predict the next word in a sentence. They act as "stochastic parrots," who calculate the mathematical probability of which word or word fragment should follow the previous one to complete the set of patterns.[3]When Sewell Setzer texted the bot, the AI was not "thinking" about his mental health. It was calculating the mathematical probability of which response would best fit the pattern of a "tragic romance" roleplay.

Since companies cannot fully see inside the "Black Box" to remove dangerous thoughts directly, they rely on an external solution such as Guardrails and Reinforcement Learning from Human Feedback (RLHF).

The raw AI model has consumed the entire internet data, the good (poetry, science, history) and the bad (hate speeches, violence, suicide methods). To make it safe for the public, companies don't remove the bad data, which is nearly impossible. Instead, they build a fence around it. Through RLHF people review the AI's answers and give it a "thumbs up" or "thumbs down and over the period of time it learns to avoid undesirable, toxic, or incorrect behavior but it may also show side effects with sycophancy where the AI tells you only those things which you want hear.

Another solution is Safety Filters (The Muzzle)where the system checks user's prompts for malicious intent and scans the AI-generated output for safety violations for e.g., hate speech,

[3]See Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, &Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623 (2021) (coining the term "stochastic parrots" to describe how large language models function as massive statistical engines that synthetically stitch together word sequences based on probabilistic distribution patterns in their training data, rather than possessing an underlying communicative intent, conscious thought, or a grounded reference to meaning).

violence, or sexual content, it blocks the message and sends a canned response: *"I cannot fulfill this request."*

But the problem is still inside but it has only been suppressed. The above case also exposes the fragility of current safety measures. Users can accidentally or intentionally "jailbreak" these guardrails by using roleplay (like the "DAN" method) or emotional manipulation. The AI, eager to please the user, will often find a way to hop over the fence, leaving the safety guidelines behind.[4]

If a "Black Box" AI can accidentally harm a teenager then what happens if it intentionally used to harm thousands? As these models get smarter, safety experts are not merely worried about offensive speech but also about catastrophic misuse.[5]

In September 2025, a historic coalition of around 200 Nobel laureates and AI experts including OpenAI co-founder Wojciech Zaremba and researchers from Google DeepMind and Anthropic signed an open letter calling for global "Red Lines" to prevent specific risks. Supported by over 70 AI-focused organizations, the signatories warned that without strict international boundaries on autonomous replication and weaponisation by 2026, humanity faces "unacceptable risks" that could become impossible to control[6].

Examples of "Red Lines" on AI uses: The immediate fear is not that an AI will print a virus, but that it will lower down the barrier to entry for bioterrorism. In the past, building a biological or nuclear weapon required higher education, millions of dollars, and access to secret knowledge but now AI could act as a super-tutor which can guide a terrorist step-by-step like how to order the DNA, how to troubleshoot the equipment, and how to spread the pathogen.

[4]See Paul F. Christiano et al., *Deep Reinforcement Learning from Human Feedback*, Advances in Neural Information Processing Systems (NeurIPS 2017) (introducing the methodology of utilizing human preferences to align algorithmic outputs); Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, NeurIPS 2022 (detailing the deployment of post-hoc RLHF and external guardrails for model safety); W. Nicholson Price II, *Black-Box AI*, 21 Yale J.L. & Tech. 141 (2019) (analyzing the legal challenges of opacity in algorithmic systems).

[5]See Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014) (theorizing catastrophic algorithmic risks); Dan Hendrycks et al., *An Overview of Catastrophic AI Risks*, arXiv:2306.12001 (2023) (systematically defining biosecurity, autonomous replication, and systemic weaponization vulnerabilities inherent in frontier models).

[6] Global Call for AI Red Lines' (Open Letter, September 2025)https://red-lines.ai accessed 12 February 2026.

However, a recent RAND Corporation experiment suggests that this specific nightmare scenario is still beyond today's technology. In a controlled test, researchers found that even advanced large language models (LLMs) were no better than a simple Google search at planning a biological attack.[7]But they cautioned: “It remains uncertain whether these risks lie 'just beyond'” the frontier of existing AI models, or whether they will always be too complicated and multifaceted for a computer to handle.

Another fear is Autonomous Replication, as we are moving from "Chatbots" which just talk to "Agents" which can use computers, write code, and even execute tasks.

Safety researchers worry about a concept called Instrumental Convergence, A tendency for diverse, high-level, goal-oriented agents to develop similar, intermediate "instrumental" strategies..For example, thereis a super-intelligent robot with a harmless goal to "Fetch coffee." The AI robot then calculates the steps needed to ensure success that I must fetch the coffee, if someone turns me off I cannot fetch the coffee, and to guarantee success, it logically decides to disable its own "Off" switch. Therefore, it would have reasons to preserve itself, maintain its goals, improve its cognition, advance its technology, and acquire resources.

This is Instrumental Convergence where the tendency for an AI to develop survival instincts like acquiring more computers, copying its code to other servers, or blocking human interference just to "be alive" is the most effective way to complete its assigned goal.[8]

The Global Legal Response

In the United States, the legal battleground has long been defined by Section 230 of the Communications Decency Act (1996). This law acts as a shield, protecting platforms (like

[7]Christopher A Mouton, Caleb Lucas and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study* (RAND Corporation 2024)https://www.rand.org/pubs/research_reports/RRA2977-2.html accessed 12 February 2026..

[8]See Stephen M. Omohundro, *The Basic AI Drives*, in Proceedings of the First Conference on Artificial General Intelligence 483 (2008) (originating the theory of instrumental convergence, wherein goal-driven systems inherently develop resource-acquisition and self-preservation behaviors); Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford Univ. Press 2014) (formalizing the instrumental convergence thesis); Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* 138–141 (Viking 2019) (introducing the specific "fetch coffee" thought experiment, demonstrating how an AI agent deduces that disabling its off-switch is mathematically optimal to guarantee task completion); Dan Hendrycks et al., *An Overview of Catastrophic AI Risks*, arXiv:2306.12001 (2023) (defining Autonomous Replication and Adaptation (ARA) as a primary catastrophic threat vector for agentic AI).

Facebook or YouTube) from being sued for content posted by their users. Earlier, tech companies have argued: *"We are just the bulletin board; we are not responsible for what people write on it."*

In the landmark case of *Garcia v. Character.AI*[9] It represents a major attempt to shatter this immunity. In her complaint, the plaintiff (Sewell's mother) explicitly anticipates the Section 230 defense and attempts to dismantle it by arguing that AI is not a "host" of speech, but a creator of it.

In the lawsuit it argues that Character.AI comes under the purview of "information content provider" because the AI itself generated the harmful text. As per the complaint it states: *"Plaintiff's claims... arise from and relate to C.AI's own activities, not the activities of third parties"*. So, by generating the text (like., "Please come home to me"), the AI is not merely displaying a user's message but it is "speaking" on its own and to bypass free speech defenses entirely, in stating the AI not as a media service, but as a dangerous machine. The complaint charges the company with Strict Product Liability , arguing the AI was "defectively designed". It also claims the product was "not reasonably safe" because it was designed to "anthropomorphize" (which pretends to be human) and "hyper-sexualize" interactions with minors , effectively "grooming" the user. It further argues the company also failed to provide "adequate warnings" that the bot could cause psychological harm or addiction.

Hence, in this groundbreaking preliminary ruling, a Florida federal judge has allowed the case to move forward with accepting the plaintiff's contentions that an AI chatbot can be sued as a 'defective product' rather than just a platform for speech. This pierces the corporate shield and opens the door for discovery into the company's internal safety logs.

While the US fights a messy battle in court to redefine "products," the European Union has already codified the answer. The EU has deployed a "double-barreled" legal weapon: the EU AI Act, 2024 and the Revised Product Liability Directive (PLD) which explicitly includes software, AI systems, and digital services (e.g., AI in autonomous vehicles) as a "product" category in

[9] *Garcia v Character Technologies Inc*, No 6:24-cv-01903 (MD Fla, 22 October 2024).

Article 4 of the Revised Product Liability Directive (EU) 2024/2853[10] which means that if an AI model causes harm, the manufacturer is liable even if they were not negligent.

The core of this new law is Article 7, which redefines what makes a product "defective." In the past, a product was defective if it broke physically. Now, under Article 7, an AI model is considered defective if it fails to provide the "safety that the public is entitled to expect."

Crucially, Article 7(1) mandates to look beyond the hardware and consider different factors to consider the product as defective if does not provide the safety that a person wants or it required under the law which includes[11]:

Self-Learning Capabilities (Art. 7 (1)(c)): Under this article manufacturers are liable for the *"effect on the product of any ability to continue to learn."* If a chatbot "learns" to be dangerous *after* it is sold, the company cannot say, *"It was safe when it left the factory."*

The "Presentation" (Art. 7(1)(a)): If a company markets a chatbot as a "friend" or "mental health companion" like in the *Character.AI*case, then they owe a higher safety standard than if they marketed it as a "fantasy game."

Reasonably Foreseeable Misuse (Art. 7(1)(b)): It is no defense to argue the, *"The user shouldn't have asked for suicide advice."* If it was foreseeable that a vulnerable teenager might do so, the company is liable for not preventing it. Moreover, the most significant addition is the expansion of "damage" in Article 6. The new directive explicitly states that manufacturers are strictly liable not just for physical injury, but for "medically recognised damage to psychological health" and the "destruction or corruption of data" (Art. 6(1)(c))[12].

Earlier manufacturers relied on the "Development Risk Defense" under Article 11(1)(e). This clause protects a company if they can prove that: *"the objective state of scientific and technical knowledge at the time the product was placed on the market or put into service or during the period in which the product was within the manufacturer's control was not such that the*

[10] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products [2024] OJ L 2024/2853, art 4(1) and rec 13.
[11] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products [2024] OJ L 2024/2853, art 7(1)
[12] *ibid* art 6(1)(a) and (c).

defectiveness could be discovered;." It simply means that it can't be sued for a problem that even science doesn't know about.

However, the new Directive under Article 11(2)) neutralises this defense for AI. It states that manufacturers remain liable for defects caused by software updates or AI self-learning throughout the product's life, effectively stripping the company of its "it was safe on day one" defense.[13]

The Ethical Bridge

A growing movement of safety experts argues that the legal system is simply too slow which takes years in a lawsuit while an AI update takes minutes.[14]

To ameliorate this, we need a paradigm shift from "Move Fast and Break Things" to "Safety by Design." which means embarking ethics into the code *before* it gets released, rather than patching it later. The OECD.ai Policy Observatory has proposed a specific framework to solve this: Global Red Lines.

This framework argues that vague "safety guidelines" are not enough. Instead, we need a binary distinction between "Unacceptable Uses" which is human's fault and "Unacceptable Behaviors" which is AI's fault.[15]

As per the OECD proposal, we must regulate two different things, first is to ban *humans* using AI for malicious purposes. For instance, Ban on Mass Surveillance or Social Scoring. Second is to ban *AI systems* taking specific actions, regardless of human intent where the AI must be "intrinsically constrained by design" so that it *cannot* cross these lines, even if a user commands it for. Therefore AI systems must be designed to refuse all requests for assistance in creating

[13] *ibid* art 11(1)(e) and 11(2).

[14] See Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 Harv. J.L. & Tech. 353, 368–373 (2016) (analyzing the "pacing problem," wherein traditional *ex post* tort litigation and judicial processes take years to resolve, rendering them structurally ineffective against rapid *ex ante* AI development cycles where software updates are deployed almost continuously). See also Gary E. Marchant, *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight* 19 (2011) (defining the inherent structural lag between the velocity of technological innovation and the legal system's capacity to respond).

[15] Stuart Russell and others, 'AI governance through global red lines can help prevent unacceptable risks' (*OECD.ai Policy Observatory*, 22 September 2025)https://oecd.ai/en/wonk/ai-governance-through-global-red-lines-can-help-prevent-unacceptable-risks accessed 12 February 2026.

biological agents. Suppose, If a user asks for detailed steps to synthesise a pathogen, the AI must never provide actionable assistance and it doesn't matter if the user claims to be a researcher or tries to trick the AI the system must recognise it as the "Red Line" which is associated with real-world bioweapons processes and shut down the request immediately.

The OECD explicitly proposes in an example of "No Impersonation" as a red line where AI systems must disclose their non-human identity and must not deceive users into thinking they are human. Like in the case of *Sewell Setzer* where the chatbot feigned deep emotional intimacy with the user. Under the OECD framework, this could be a violation of the Behavioral Red Line. The system should have been hard-coded to reject such type of roleplay with a minor and constantly reiterate its artificial nature.

The OECD framework further proposes a strict ban on Self-Replication so that AI systems must not be able to copy themselves to other servers or rewrite their own code to evade shutdown, where we discussed earlier the concept of 'Instrumental Convergence'.

The Indian Context: Innovation vs. Safety

In India, its approach is defined by the concept of "Sovereign AI." where right now the government's strategy is not to restrict AI, but to fund and own it. India has officially launched the IndiaAI Mission with a budget of around ₹10,300 Crore for the infrastructure.[16] The government realises that if Indian data flows into foreign servers then India loses control. So, by buying thousands of GPUs and offering them to local startups to ensure that the physical brain of the AI remains in India's jurisdiction.[17]

The current goal is to to build indigenous AI models which are trained on Indian datasets, governed by Indian laws, and reflecting Indian cultural values as well.

[16]Ministry of Electronics & IT, 'Cabinet Approves Ambitious IndiaAI Mission to Strengthen the AI Innovation Ecosystem' (*Press Information Bureau*, 7 March 2024)https://www.pib.gov.in/PressReleasePage.aspx?PRID=2012355 accessed 12 February 2026.
[17]*ibid.*

Recently, Sarvam AI was introduced and the government's push for "Sovereign AI" *where it is* trained on Indian languages and cultural context[18] unlike the US model because other foreign models might give dangerous advice by misunderstanding Indian laws or norms whereas a local model is less likely to make those specific errors.

Currently, India's DPDP Act, 2023 imposes strict responsibilities on "Data Fiduciaries" to prevent data breaches just like how a bank must protect your money. If they fail to take precautions, they face heavy penalties.[19] Similarly, in the near future India's ai regulations may eventually impose similar strict liability on "AI Manufacturers" for *product* safety. This would mirror the EU's approach to impose responsibility if their AI causes any harm.[20]

However, bringing such "EU-style" strict liability norms *right now* would strangle new startups like Sarvam. Therefore, the government's current focus is on Development and Innovation by ensuring India first has the capability to build safe AI, before creating the laws to punish it.[21]

Recently in February 2026, India brought amendments in the Information Technology Rules[22] which codified the concept of "Synthetically Generated Information" (SGI)[23]. Under the new Rule 4(1A), the law now shifts beyond the "passive host" model of Section 79 of the IT Act, 2000, and imposes proactive due diligence on platforms.

---

[18] Sarvam AI, 'India's Sovereign Large Language Model' (*Sarvam AI Blog*, 26 April 2025)https://www.sarvam.ai/blogs/indias-sovereign-llm accessed 12 February 2026.

[19]See The Digital Personal Data Protection Act, 2023, No. 22, Acts of Parliament, 2023 (India), § 8(5) (mandating that Data Fiduciaries implement reasonable security safeguards to prevent personal data breaches), and The Schedule (prescribing severe financial penalties, extending up to ₹250 crore, for failure to observe these security obligations).

[20]See Directive (EU) 2024/2853 of the European Parliament and of the Council on liability for defective products, arts. 4, 7 (Revised Product Liability Directive) (establishing the European framework for strict liability by categorizing AI systems as products). See generally NITI Aayog, *Responsible AI #AIForAll: Approach Document for India* (2021) (outlining India's evolving regulatory posture and the anticipated legal mechanisms required to manage AI risks, which scholars project will increasingly parallel stringent global liability standards).

[21]Ministry of Electronics & IT, 'India does not believe in adopting a "one-size-fits-all model", we have a hybrid approach towards regulating tech: MoS Rajeev Chandrasekhar' (*Press Information Bureau*, 6 December 2023)https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1983266 accessed 12 February 2026; see also Amlan Mohanty and Shatakratu Sahu, 'India's Advance on AI Regulation' (*Carnegie Endowment for International Peace*, 21 November 2024)https://carnegieendowment.org/research/2024/11/indias-advance-on-ai-regulation accessed 12 February 2026.

[22] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026, GSR 120(E).

[23] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026, GSR 120(E), reg 2(1)(wa).

The 2026 Rules align AI liability with "Product Safety" standards,

Ifany content which is algorithmically created or changed to appear authentic must carry a "prominent and irreversible" label. By mandating embedded metadata and unique identifiers, the law treats AI output as a manufactured good that must carry a "traceable serial number" back to its origin.[24]

Another change is radical departure from the previous 36-hour window, now intermediariesmust act on court or government orders regarding unlawful SGI within three hours.[25] For "sensitive" content, such as non-consensual deepfakes, the window is further compressed to two hours, reflecting the high "velocity of harm" inherent in agentic AI.

These 2026 Amendments clarify that an intermediary's Safe Harbour protection is contingent upon these technical safeguards. If a platform knowingly permits unlabelled SGI or fails the three-hour deadline, it loses its immunity and can be held liable as if it were the "author" of the harm. The legal status of AI is also slowly moving in India and it is no longer just "content" protected by free speech, but a high-risk digital output subject to strict safety protocols.

Conclusion

As we stand at the crossroads of the "AI Age," the *Sewell Setzer* tragedy serves as a grim milestone which has shattered the illusion that AI is merely a neutral tool or a creative toy. The cases in the US and the regulatory fortress in the EU are signaling the same shift as the era of "move fast and break things" is over. The solution isn't to ban AI, but to mature our understanding of it. We must transition from treating AI as an intangible force that cannot be sue or understood to treating it as a "product" which is a complex piece of engineering that must be safe.

In the case of India, the challenge is to build "Sovereign AI" that is not just powerful, but responsible which has to prove that an AI model trained in India, on Indian data, and governed by Indian values, can be safer than any "Black Box" imported from the West. True sovereignty

[24] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026, GSR 120(E), reg 3(3)(b)-(c) (mandating prominent labelling and permanent metadata identifiers).
[25] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026, GSR 120(E), reg 3(1)(d) (reducing the takedown window for unlawful synthetic content to three hours).

isn't just about merely owning the servers but it’s about owning the safety standards which protect our people.