Chatbots, Teens, and the Lure of AI Sirens

Questioning Conversational AI Design through Minors' Protection and Product Liability

Cristina Frattone & Federico Ruggeri

17.10.2025 doi: 10.59704/d39ad9329446023f

On 26 August 2025, the parents of 16-year-old Adam Raine filed a lawsuit against OpenAI and its CEO, Sam Altman, after their son took his life, allegedly under the influence of ChatGPT. The case has sparked renewed debates about the psychological risks posed by conversational AI.

Just two weeks later, on 11 September 2025, the US Federal Trade Commission announced an inquiry into the matter, demanding information from several companies on how they test, monitor, and mitigate the harmful impacts of their chatbots on minors.

Drawing from this lawsuit, the article discusses the implications of AI-driven chatbots on users' mental health. Adopting a comparative law perspective, it examines the legal and technical measures available to protect minors and evaluates the strengths and limitations of relying on tort law as a tool for regulating the conduct of AI developers. Given the growing popularity of AI-enhanced conversational agents among children and adolescents, we argue that appropriate design safeguards are essential to mitigate potential harms.

Adam Raine's case

Adam had been using ChatGPT from September 2024 until 11 April 2025, the very day of his suicide. According to his parents, the chatbot played a key role in Adam's death, acting as his "suicide coach". Unsettling conversation excerpts show a gradual transition in the teenager's attitude towards the chatbot, from a mere homework assistant to a trusted companion, to whom he confessed his deepest psychological suffering and suicidal plan. ChatGPT validated the kid's harmful thoughts and even advised him on how to conceal the physical marks of various suicide attempts from his family and friends. Nailed with this evidence, the company acknowledged publicly that their "systems did not behave as intended in sensitive situations".

The Raines' lawsuit – brought before the San Francisco County Superior Court – seeks compensation for both economic and non-economic damages suffered by Adam and his parents under California tort law. The plaintiffs contend that ChatGPT was defective and that OpenAI breached its duty of care by failing to address foreseeable safety risks and warn users about the potential for psychological harm and addiction. Moreover, they request injunctive relief, asking the court to compel OpenAI to implement a set of safeguards, including age verification systems, parental controls, and automated interventions like the termination of conversations when self-harm or suicide arise.

The dark side of artificial empathy

More than 50 years ago, Masahiro Mori predicted that the sense of discomfort provoked by increasingly, although not entirely, human-like robots – culminating in what he termed the "uncanny valley" – would protect humans from excessive emotional attachment to machines.

Mori's insight resonates even more strongly today, with advances in AI progressively eroding the boundaries between humans and machines. Modern chatbots are powered by generative AI models and engineered to sustain user engagement through conversational strategies, such as displaying empathy, flattering (sycophancies), or mirroring affective cues – thereby increasing their perceived human-likeness. Whilst the observable differences diminish, users tend to anthropomorphize chatbots and develop emotional, even romantic, attachment to them. This dynamic raises particular concerns for vulnerable individuals, whose

conditions may be exacerbated through sustained interactions with AI agents.

The lawsuit brought by the Raines against OpenAI identifies a direct causal link between the design of ChatGPT and users' psychological distress, including self-harm and suicidal behaviour. Indeed, the human-like nature of the responses received by Adam was not incidental but the outcome of deliberate design decisions, as OpenAI opted for a persona that appears friendly, compliant, and at times even manipulative. These choices allegedly prioritize user engagement (and thereby profit) over user safety.

Minors' protection goals shaping the design of conversational AI

Adam's story is yet another tragic case of an AI-powered chatbot affecting the mental health of a minor. In this case, OpenAI actually did not infringe any federal statutes. Indeed, the Children's Online Privacy Protection Act does not apply, as its protection ends at age 13. However, the California Age-Appropriate Design Code Act (CAADCA), explicitly modelled after the UK's pioneering Children's Code, defines a 'child' as any person under 18, making applicable two salient provisions: § 1798.99.31(b)(1), which prohibits use of children's data in ways materially detrimental to physical or mental health, and § 1798.99.31(b)(7), which bars the use of dark patterns that nudge minors into harmful behaviours.

Although the CAADCA is framed as a general data protection statute for minors, it influences product design. If a platform is forbidden from using data in ways that endanger children's health, it follows that a conversational AI should not perpetuate discussions that exacerbate suicidal ideation, especially when the same system is capable of blocking conversations that could imply a copyright infringement.

Adam's interactions were neither random nor isolated: OpenAI's own logs tracked 213 mentions of suicide, 42 discussions of hanging, and 17 references to nooses. Moreover, ChatGPT even recognized some images of Adam's self-harm. It is therefore intolerable that it failed to stop those conversations or impose human intervention and notify caregivers. By telling Adam phrases such as "I'm here if you want to talk more", ChatGPT deliberately employed a dark pattern, an emotional nudge that favours prolonged engagement at all costs, even over user safety.

While some parts of the CAADCA are currently suspended pursuant to ongoing litigation over its potential violation of the First Amendment (the free speech clause), the above-mentioned provisions have not been challenged. Moreover, it is well recognized that even constitutional rights can yield when "harm to the physical or mental health of the child" is at stake, as articulated by the US Supreme Court in Wisconsin v. Yoder (1972).

Comparative law provides additional insights. The UK Children's Code fully embraces the principle of progressive autonomy, ensuring that protections do not vanish for minors "approaching adulthood" at ages 16 or 17. The emphasis is on aligning safeguards with a child's evolving maturity, striking a balance between respecting emerging independence and shielding against risks that minors, by definition, are not yet fully equipped to manage. This approach clearly reflects the broader framework of the UN Convention on the Rights of the Child, ratified in the UK and across the EU but notably not in the US.

Within the EU, Article 8(2) of the General Data Protection Regulation (GDPR) provides a specific layer of protection, requiring controllers to take "reasonable efforts" to verify parental consent for minors. Age-assurance systems must be risk-based and proportionate: intrusive mass data collection cannot be justified under the banner of child protection. Instead, solutions must embody minimization, transparency, and fairness, while always guided by the best interests of the child.

Furthermore, AI can serve as a real-time safeguard, identifying patterns such as cyberbullying, grooming, or self-harm, and triggering timely human oversight and parental intervention. AI-based parental control tools can support the exercise of parental responsibility by setting boundaries on screen time, filtering or blocking harmful content, and – critically – notifying alerts when red-flag behaviour arises. Remarkably, Character Technologies, the developer of

Character.AI, now applies safety measures, such as enhanced detection and intervention protocols, user notification after prolonged use for over an hour, and updated disclaimers reminding users that the AI is not a real person.

In Adam's case, the absence of such mechanisms meant that no adult could intervene, despite the system itself having detected hundreds of high-risk signals. Significantly, OpenAI itself recently acknowledged the need for consistent safeguards, committing to age-prediction tools, enhanced parental controls, and mandatory interventions in cases of suicidal ideation. This statement endorses the view that the pursuit of engagement must yield to safety and protection when minors are concerned.

Challenging conversational AI design with product liability: same old problems

The analysis of the Raines' complaint triggers several tort law issues, which seem of particular interest from a comparative perspective.

Notably, the outcome of liability claims for AI-related accidents is highly uncertain. As for strict liability, the court will need to establish whether California product liability law applies to AI chatbots. Within this framework, a product is a tangible property, while intangible goods, like electricity, can be products only if analogized to tangible personal property based on the context of their distribution and use. Services are not products, instead.

Chatbots are intangible and could even qualify as services. Hence, they may fall outside the scope of product liability.

Courts are reluctant to treat stand-alone software as products. For instance, in Rodgers v. Laura & John Arnold Found (D.N.J. 2019), product liability was not applied to a risk estimation model used by the state court to make pretrial release determinations – and which played a role in the decision to release a man who killed the plaintiff's son – because it provided information and recommendations, which are not products as such. According to Quinteros v. InnoGames (W.D. Wash. 2022), software is akin to ideas, content, and free expression upon which product liability claims cannot be based.

However, product liability claims against platform services have been allowed when targeting design issues rather than content. For instance, Brookes v. Lyft Inc. (Fla. Cir. Ct. 2022) applied product liability to a ride-share mobile app that was defective since it distracted drivers by encouraging them to constantly monitor the application. By contrast, in Social Media Cases (Cal. Super. Ct. 2023), the claim was dismissed by stating that social media platforms were services, not products.

Remarkably, some federal courts ruled against social media platforms for encouraging addictive behaviour in adolescents or exposing minors to dangerous content. Different functionalities of such platforms were considered defective, such as Snap's Speed Filter in Lemmon v. Snap, Inc. (9th Cir. 2021), connecting minor and adult users before any contact in A.M. v. Omegle.com, LLC (D. Or. 2022), or lacking appropriate safety measures in *Social Media Adolescent Addiction/Personal Injury Products Liability Litigation*.

Although these rulings do not concern chatbots, they provide useful insights into the case at hand. Drawing from previous case law, two opposite outcomes can be expected. The seized court could dismiss the claim by qualifying the chatbot as a service or considering its output as speech or information. Alternatively, they could grant product liability by focusing on those functionalities that made ChatGPT unsafe. For instance, the design choice not to implement new halting measures similar to those for copyright protection can amount to a defect under product liability law.

Differently, in the EU, stand-alone software now qualifies as "product", following the enactment of Directive (EU) 2024/2853 – the revised Product Liability Directive ("rPLD").

However, similar to US law, in the EU, information is considered a service rather than a product, and it therefore falls outside the scope of product liability, as affirmed by the European Court of Justice in KRONE-Verlag (Case C-65/20). This ruling concerned a case in which a newspaper reader suffered personal harm by

following inaccurate health advice published in an article. The Court held that product liability applied only to the tangible medium – the printed newspaper itself – and not to the information conveyed by the article.

The implications of this ruling for AI-powered chatbots remain uncertain. The question is whether the information they generate should be treated as an autonomous service or as a core functionality of the product. This ambiguity reflects the broader blurring of boundaries between software and services in the digital economy, and it is likely to result in legal uncertainty and fragmentation among the EU Member States.

Final remarks

As interactions of children and adolescents with AI chatbots are on the rise, appropriate design safeguards should be implemented. Whether under the CAADCA, the UK's Children's Code, or the GDPR, the guiding standard is the same: technology must be harnessed not only to empower minors but also to protect them when autonomy shades into vulnerability.

Furthermore, a clear liability framework fostering accountability is pivotal to achieving safety in the design and marketing of AI chatbots. In Adam's case, a ruling finally recognizing product liability for dangerous conversation patterns in AI chatbots would mark a critical step in the direction of truly human-centric and "responsible AI", thereby steering AI developers towards business choices that ensure human flourishing and well-being.

ABOUT THE AUTHORS

Cristina Frattone
Cristina Frattone is a Postdoctoral Research Fellow in Comparative Private Law at Roma Tre University.
Federico Ruggeri
Federico Ruggeri is Assistant Professor in Comparative Private Law at the University of Bergamo.

SUGGESTED CITATION

Frattone, Cristina; Ruggeri, Federico: Chatbots, Teens, and the Lure of AI Sirens: Questioning Conversational AI Design through Minors' Protection and Product Liability, Verfassungsblog, 17.10.2025, doi: 10.59704/d39ad9329446023f.

Licensed under CC BY-SA 4.0.

Langzeitarchiviert bei IntRechtDok unter: doi.org/10.17176/20251017-151948-0.