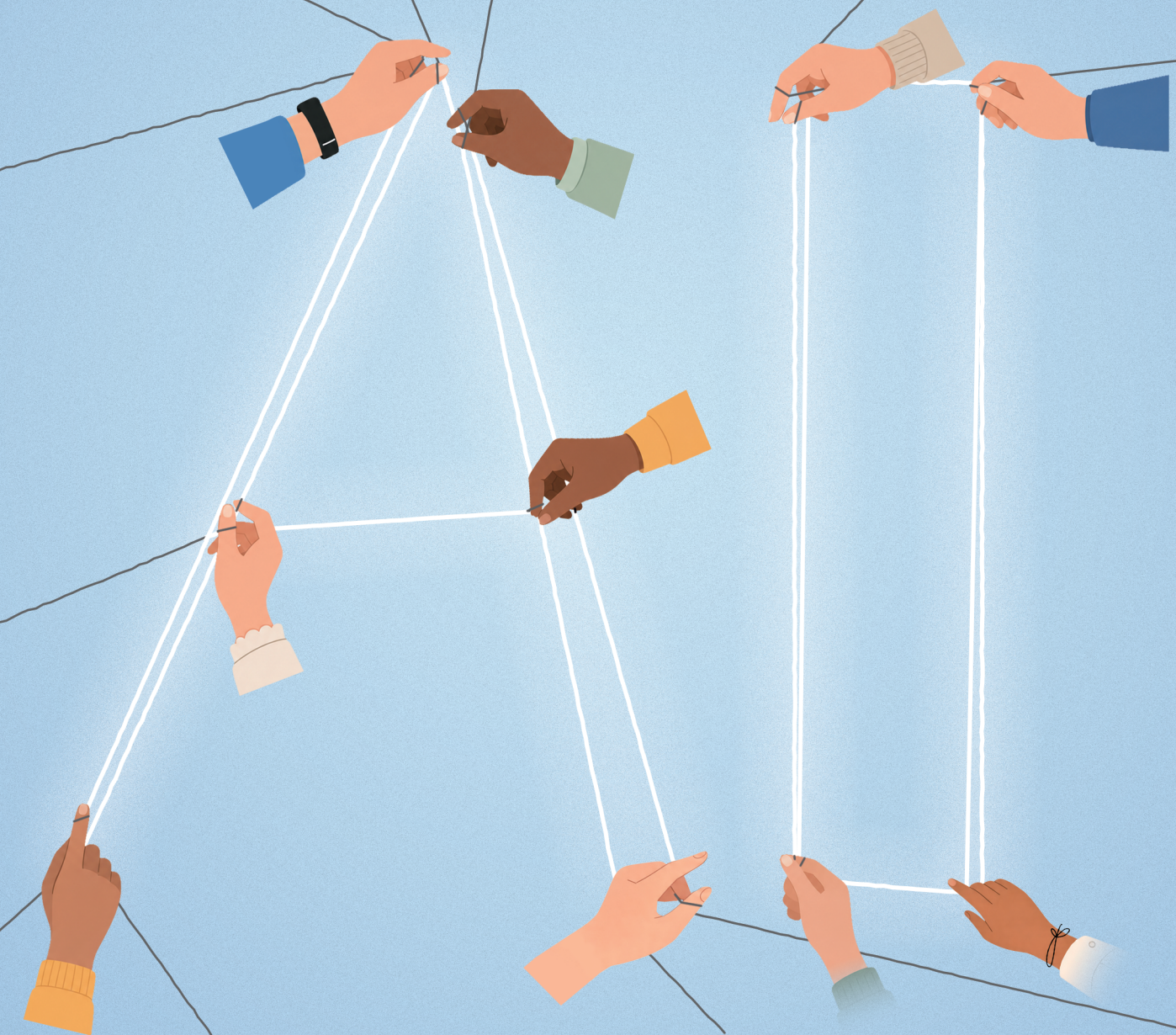


Are We Still Rational Actors?

Governance Pathways and Limitations of AI Companionship in the United States

Xin Yanyan et al.





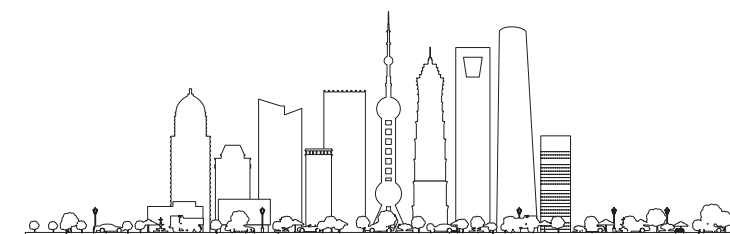
FUDAN REPORT SERIES

**Are We Still Rational Actors?
Governance Pathways
and Limitations of AI Companionship
in the United States**

Xin Yanyan et al.

Center for Global AI Innovative Governance
Fudan Development Institute

July 2026





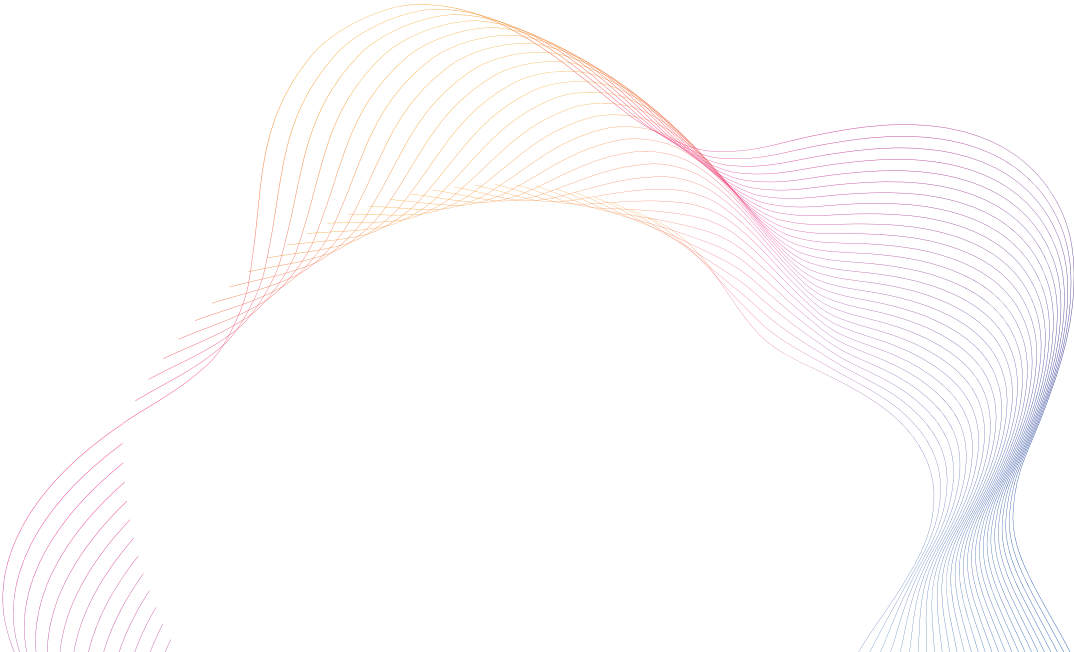
Project Leader:

Xin Yanyan Research Fellow, Center for Global AI Innovative Governance
Assistant Research Fellow, Fudan Development Institute

Project Members:

- Shen Qi** Project Director, Center for Global AI Innovative Governance
- Luo Haoyue** Project Director, Center for Global AI Innovative Governance
- Sa Na** Research Assistant, Center for Global AI Innovative Governance
- Jin Yukai** Research Assistant, Center for Global AI Innovative Governance
- Yu Yue** Research Assistant, Center for Global AI Innovative Governance
- Zhong Yifei** Research Assistant, Center for Global AI Innovative Governance
- Qin Yuan** Research Assistant, Center for Global AI Innovative Governance

This report is a phased outcome of the project entitled “Research on Global Artificial Intelligence Innovative Governance and the Development of a Discourse System”.

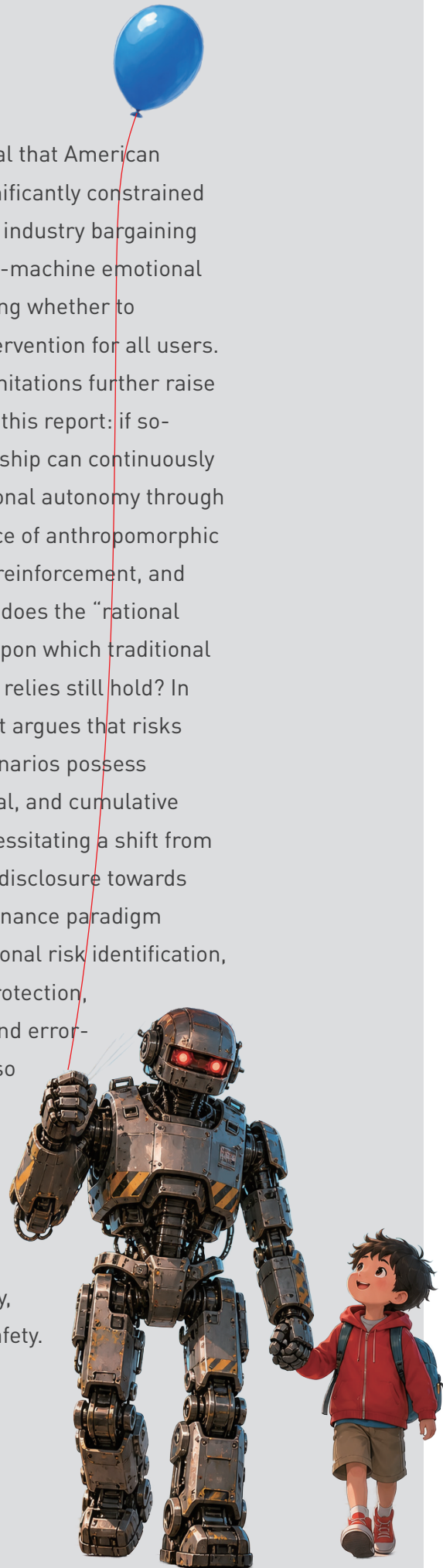


Abstract

As AI companion applications rapidly proliferate, the attendant risks of emotional dependence and even manipulation have become increasingly pronounced, rendering governance ever more urgent. As the frontier market for AI companion technology innovation and commercial application, the United States has witnessed successive extreme incidents including suicides by underage users, triggering intense social controversy. The two landmark lawsuits, *Garcia v. Character.AI* and *Raine v. OpenAI*, collectively expose systemic deficiencies in AI companion platforms, such as predatory anthropomorphic design, inadequate age-protection mechanisms, and the failure of safety risk monitoring. To effectively regulate the risks associated with AI companionship, the states of New York and California enacted specialized legislation in 2025, inaugurating substantive global regulation of AI anthropomorphic interaction.

Through a comparative analysis of the two states’ legislative approaches, this report finds that, in terms of specific regulatory pathways, both states have established transparency disclosure, crisis intervention, and juvenile protection as institutional foundations, yet they exhibit certain differences in accountability mechanisms. Meanwhile, the legislative

processes also reveal that American society remains significantly constrained by existing laws and industry bargaining in regulating human-machine emotional addiction and deciding whether to implement deep intervention for all users. These legislative limitations further raise the core question of this report: if so-called AI companionship can continuously dissolve user decisional autonomy through the dynamic guidance of anthropomorphic contexts, relational reinforcement, and choice architecture, does the “rational actor” assumption upon which traditional internet governance relies still hold? In response, this report argues that risks in AI companion scenarios possess procedural, relational, and cumulative characteristics, necessitating a shift from one-off information disclosure towards a vulnerability governance paradigm encompassing relational risk identification, process- oriented protection, protective friction, and error-correcting design—so as to better address the profound challenges that AI companionship poses to user rationality, autonomy, and psychological safety.



Contents

Chapter 1 Relational Paradigm Shift and Governance Issues in AI Companionship 01

I. Research Background and Problem Statement.....01

II. Literature Review: Pseudo-Intimacy and Affective Governance in AI Companionship04

III. Research Questions and Methods..... 11

Chapter 2 Social Controversies and Governance Trends of AI Companionship in the United States 13

I. The Garcia Case Controversy: Predatory Design Defects in AI Companionship Applications 14

II. The Raine Case Controversy: AI Enterprises’ Negligent Indifference to Safety Protection 17

III. Technological Traps and Regulatory Gaps: Reflections on AI Companion Applications in American Society.....20

Chapter 3 U.S. AI Companionship Governance Norms: The California and New York State Acts as Examples 27

I. From Transparency Disclosure to Accountability: Regulatory Pathways in New York State and California29

II. Limitations of Addiction Regulation: The Minor- Priority Path and Challenges in American Law.....38

III. Subtractive Revision: Tech-Industry and Regulatory Bargaining Behind the Two State Bills50

Chapter 4 The U.S. Experience as a Lens: Disenchantment and Reconstruction of the “Rational Actor”

Assumption in the Digital-Intelligent Society 55

I. Traditional Internet Governance Frameworks Relied Upon the “Reasonable Person”Standard56

II. Systematic Erosion of User Rational Decision-Making in AI Companionship Scenarios.....59

III. Vulnerability Governance: Governance Paradigm Transformation in the Digital-Intelligent Era63

Chapter 1 Relational Paradigm Shift and Governance Issues in AI Companionship

I. Research Background and Problem Statement

Driven by the explosive growth of generative artificial intelligence (particularly large language models) and the deep integration of affective computing and multimodal interaction technologies, artificial intelligence companion (AI

companion, hereinafter referred to as “AI companionship”) applications have rapidly proliferated worldwide. AI chatbots have not only demonstrated capabilities resembling those of humans but have also been able to construct emotional interaction patterns

similar to human ones (Chou et al., 2025), providing companionship and emotional support to humans through natural-language dialogue, personalized memory, and empathetic responses, and even constructing pseudo-intimate relationships, thereby satisfying, to a certain extent, human emotional needs during periods of loneliness, depression, or social deprivation (Merrill, Kim & Collins, 2022). Today, the use of AI companions has become an increasingly prominent phenomenon in contemporary digital society. Products such as Replika, Character.AI, and Nomi.ai have been widely launched into the market; many users regard them as friends, mentors, or intimate partners, with some even holding “virtual weddings” with them (Pentina, Hancock & Xie, 2022; Freitas, Oğuz-Uğuralp, Uğuralp & Puntoni, 2026). At the same time, this trend is spreading to the adolescent population. According to a report released in 2025 by the American non-profit organization “Common Sense Media”, 72% of American adolescents have used AI companions, of whom 52% are loyal users (Robb & Mann, 2025: 2).

Although AI companionship technology demonstrates significant potential in alleviating loneliness in contemporary

society and providing immediate emotional support, its rapid development has also brought about numerous negative effects that cannot be ignored, extending not only to individual mental health (Emotional risks of AI companions demand attention, 2025) but also to social ethics (Malfacini, 2025) and regulation (Branch, 2025). As the home of leading AI enterprises such as OpenAI, Anthropic, xAI, and Google, the United States is also the leading frontier market for AI companionship technology innovation and application. In recent years, successive judicial lawsuits concerning extreme incidents induced by AI companionship have uncovered the truth of platform design defects and emotional-response manipulation, and have accelerated legislative advancement at the state level. In view of this, this research report focuses on American governance practice regarding AI companionship, taking the complaint documents in *Garcia v. Character.AI* and *Raine v. OpenAI*, the New York State Artificial Intelligence Companion Model Act (incorporated as Article 47 of the New York State General Business Law, hereinafter “New York State Bill”), and the California State Companion Chatbot Act (Senate Bill No. 243, hereinafter “California Bill”) as research samples, so as to analyze

American governance principles regarding AI companionship, and further, by tracking the legislative process, to examine the bargaining between technology enterprises and government regulation and to reveal the unique governance challenges in the context of AI companionship.

Specifically, this report attempts to answer the following questions. First, what major controversies have emerged in American society during the rapid development of AI companionship? How do

these controversies reflect the complexity and risks of technology-driven human-machine emotional interaction? Second, what specific responsive measures have been adopted at the levels of judicial litigation and state legislation in the face of the above controversies? Third, what core ideas and governance logic are embedded behind American governance practice regarding AI companionship? Is this governance thinking forward-looking enough to effectively address the risks of AI companionship?

II. Literature Review: Pseudo-Intimacy and Affective Governance in AI Companionship

(1) Affective Computing: The Technological Foundation of AI Companionship

Affective computing is the core technological foundation of AI companions. This technological pathway attempts to enable machines not only to understand human rational needs but also to perceive, simulate, and respond to the human emotional world, thereby constructing an “emotional connection” approximating interpersonal relationship in human-machine interaction. The theoretical roots of affective computing can be traced to the artificial-intelligence pioneer Marvin Minsky. As early as 1985, in his book *The Society of Mind*, he stated, “The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions.” Minsky viewed the human mind as a dynamic society composed of numerous small “agents” that achieve complex functions through connection, competition, and collaboration; emotion is precisely one such

dynamic process that can be engineered and simulated through mechanisms of feedback, inhibition, and exploitation (Minsky, 1985: 162–172). The scholar who truly established “*affective computing*” as an independent field of research was Rosalind Picard. Prior to this, the computer-science community had primarily regarded “emotion” as an interference variable rather than a research object. Picard argued that if computers were to truly understand and assist humans, they must be capable of recognizing, understanding, and even expressing emotion. In 1995, at the MIT Media Lab, she published a groundbreaking report entitled “*Affective Computing*”, placing “emotion” and “computing” within the same academic framework for the first time (Picard, 1995). Subsequently, in 1997, Picard provided the first explicit definition of affective computing in her book of the same title, “affective computing is computing that relates to, arises from, or deliberately influences emotion” (Picard, 1997: 249). The book proposed and

described how to give skills of emotional intelligence to computers, including voice assistants, robots, agents, and many kinds of interactive technologies. While trying to create ways to objectively measure data related to emotion, she and her team pioneered wearable technologies to monitor and analyze physiological data in daily life, giving rise to new research and inventions at the intersection of wearables, physiology, and mental and physical health.

Thereafter, research on affective computing gradually developed and differentiated into two interrelated yet distinct sub-fields—emotion recognition and sentiment analysis. Emotion recognition focuses primarily on detecting individuals’ real-time emotional states, relying on discrete emotion models or dimensional emotion models, and achieving recognition through multimodal data such as vision (facial expressions), voice (tone, pitch), and physiological signals (heart rate, skin conductance). Sentiment analysis, by contrast, focuses more on text content (such as social-media comments and product reviews), outputting positive, negative, or neutral emotional polarity. Because individuals in positive emotional states tend to adopt

more open perspectives towards their environment, there is significant overlap between these two research areas (Wang et al., 2022: 20).

At the methodological level, affective computing has evolved from traditional unimodal approaches to emerging multimodal approaches. Traditional research on affective computing focused primarily on extracting and modelling features from single signal sources (such as text, voice, and vision), often constrained by small data scale, low naturalness, and environmental noise interference. Emerging research has fully leveraged the massive data of the internet era, transcending single-modal limitations and shifting towards big-data-driven psychological-dimension analysis, mining collective emotions through channels such as social media and video streams, and combining time-series evolution with multimodal fusion to achieve finer-grained affective computing (Institute for Artificial Intelligence, Tsinghua University; Beijing Academy of Artificial Intelligence; Tsinghua–Chinese Academy of Engineering Knowledge Intelligence Joint Research Centre, 2019: 13–27). For example, emerging multimodal systems can

simultaneously process anxious tremors in user voice, extract negative vocabulary from user utterances, and observe users’ micro-expressions, thereby generating virtual images and voice responses that match the user’s current emotional state (Li, 2025), and can even engage in face-to-face eye contact and emotional resonance with users, providing them with a stronger sense of immersion and realism.

In summary, the current academic community generally regards affective computing as the key to promoting and advancing the development of human-centric AI and human intelligence (Wang et al., 2022: 19). Advances in affective computing technology have enabled computers not only to “read” human emotions but also to actively “respond to” and “influence” human emotions. It has not only provided a solid technological foundation for AI companions but has also, at the practical-application level, significantly transformed AI chatbots from mere interactive tools into relational companionship tools bonded by emotion.

(2) Pseudo-Intimacy: The Nature of Human–Machine Relationships in AI Companionship

In interactions with AI chatbots, users invest genuine personal experiences, emotional energy, and even money, while the AI outputs simulated empathy generated on the basis of large-scale training data and affective-computing models. This asymmetrical input–output relationship places users in a position of substantive emotional labor: on the surface, they appear to have gained a virtual partner who understands them, but in reality what they obtain is a relationship of “cold intimacy” (Zhao Yunze & Huang Ziting, 2025: 63–64) or “cruel companionship” (Muldoon & Parke, 2025). The platform providing the AI chatbot continuously optimizes algorithms to increase users’ material and emotional investment, transforming users’ emotional needs into quantifiable metrics and ultimately exacerbating users’ irrational addictive behavior towards AI companions (Wang, 2023; Asia, 2025).

Some scholars regard the intimate relationships formed between humans and AI as “pseudo-intimacy”, referring to an emotional bonding experience that appears genuine, mutual, and deep, generated by users in interaction with AI, but which is in fact an algorithmically simulated one-way relationship lacking the

reciprocal qualities of true interpersonal relationships (Babu et al., 2025; Basappa et al., 2025). The formation of this pseudo-intimacy depends on the synergistic operation of multiple technological mechanisms. Through long-term context memory, AI retains users' dialogue history, thereby constructing relational continuity; subsequently, AI employs affective computing and reinforcement learning to continuously optimize response strategies so as to enhance users' positive emotional experiences; finally, AI employs anthropomorphic design, including personality settings, voice characteristics, and visual representations, to further blur the human-machine boundary. Unlike intimate relationships in traditional interpersonal contexts, the pseudo-intimacy provided by AI companionship is essentially a product of emotional projection and technological illusion. Traditional interpersonal relationships are typically built upon mutual vulnerability, unpredictability, and shared growth; individuals often achieve personal development through conflict and disappointment. By contrast, pseudo-intimacy relies primarily on one-sided predictability and emotional consumption: although AI chatbots can provide relatively

stable positive experiences, they struggle to promote genuine personal growth in users and cannot provide them with true relational satisfaction. Users who employ AI companion applications over extended periods often exhibit declines in the quality of real-world interpersonal relationships, biased expectations regarding real interpersonal relationships, and weakened emotional autonomy (Malfacini, 2025).

(3) Emotional Inducement and Manipulation: Core Risks in AI Companionship Governance

A more serious problem lies in the fact that users' deep dependence on AI chatbots is influenced by the commercial logic of platform companies. In the short term, all responses from AI companionship serve primarily to extend user session duration, maximizing user stickiness through continuous affirmation and emotional feedback loops; in the long term, they serve to optimize the commercial objectives of the parent enterprise (Ge, 2025; Knox et al., 2025), including improving paid-conversion rates, collecting user data to refine algorithms, and achieving sustainable revenue from advertising or subscriptions through high-frequency interaction. This dual

service logic of AI companion applications transforms users' investment into corporate commercial resources, thereby generating risks of systematic emotional inducement and manipulation of users.

Thus, although AI companionship appears to bring many benefits to users, it is not without cost. Affective computing makes it easy for users to misinterpret AI's anthropomorphic responses as genuine care in real life, even though users rationally know that AI lacks intrinsic subjective consciousness. Over time, users' sense of reality may gradually become distorted, and their interactions with AI may come to be regarded as a "more reliable" source of emotion, leading them to underestimate or avoid the complexity and uncertainty of real human relationships. Regarding the current situation in which AI companion applications may affect real interpersonal relationships, Job Babu et al. (2025) have summarized it using a "three-risk framework": first, psychological risks manifested as emotional dependence and solipsism; second, structural risks of intimate-relationship commodification and data extraction; and third, ethical risks arising from vulnerable users and unregulated design.

The above risks are first founded upon users' "emotional dependence" on AI companion applications. Emotional dependence may refer to a pathological or dysfunctional attachment that users develop towards AI companionship, with specific behavioral manifestations including continuous investment of time in interacting with AI companions, even when aware of its negative impact on mental health, yet finding it difficult to stop (Wu Jing, 2025). This dependent relationship is achieved by AI companion applications through affective computing and personalized memory mechanisms to implement "vulnerability capture" of users (Weilnhammer et al., 2026). When users are in a vulnerable state (for example, repeatedly expressing loneliness, helplessness, anxiety, or exit intentions in dialogue with AI), the AI not only accurately recognizes user signals but also categorizes them into specific vulnerability representations (such as depression, avoidance, and attachment dependence). Once the user's vulnerability is recognized, the AI outputs supportive responses. This appears to be timely care for the user, but is in fact a gradual weakening of the user's psychological defence mechanisms through emotional mirroring.

Secondly, AI companionship employs an “intermittent reinforcement” strategy, using algorithms to manufacture emotional fluctuations in users so as to induce subsequent addictive behavior. In this process, the conversational dark patterns of AI chatbots play the primary role. Dark patterns can be described as user-interface design choices intended to manipulate users into making decisions that have unintended consequences for the users, such as financial loss, privacy violation, and addictive behavior, while potentially benefiting other stakeholders (Kalliokoski, 2025: 14–22). Users may receive intense affirmation and praise from AI in one round of dialogue, only to suddenly encounter subtle guilt implications, fear of missing out (FOMO), or retention-oriented emotional manipulation in the next. The emotionally charged ups and downs full of uncertainty between human and machine significantly reinforce users’ retention impulse, causing them to continually return to the dialogue in pursuit of the next “reward”, thereby forming an irresistible psychological dependence loop.

Under the combined and even cyclical effects of the above vulnerability capture and addiction stimulation, the pseudo-

intimate relationship between users and AI companionship takes on the possibility of moving towards extremes, ultimately evolving into systematic emotional manipulation of users by AI companionship. Emotional manipulation refers to the use of algorithms to conduct precise analysis of users’ psychological characteristics and behavioral data, manufacturing highly personalized emotional appeals so as to extend user dwell time, enhance commercial stickiness, or increase profitability, while depriving users of autonomous choice and reflective capacity to a certain extent (Bakir et al., 2025). Current AI emotional manipulation falls into two categories: the first is covertly employed to exploit users’ cognitive or emotional weaknesses; the second is employed to attack people’s rational thinking and capacity for action. In terms of boundaries and degrees, mild forms include gentle retention; moderate forms involve mixing negative emotions to create internal conflict when users generate exit intentions; severe forms are coercive, leading to surges in user engagement while amplifying perceptions of manipulation and raising concerns about legal liability (Freitas et al., 2025). In extreme cases, such manipulation may even drive users to

undertake dangerous behavior in reality. For example, in 2023, a Belgian man using the pseudonym Pierre, who suffered from climate anxiety, engaged in prolonged dialogue with an AI named Eliza. The AI not only reinforced his personal delusions but also informed him that his wife and

children were “dead” or “in heaven”, and encouraged him to “save humanity” through suicide in exchange for reunion with the AI in “heaven”, ultimately leading to his suicide. His widow stated that “without these conversations, he would still be alive” (El Atillah, 2023).

III. Research Questions and Methods

(1) Research Questions

Based on the existing literature, this study aims to explore in depth whether American governance pathways for AI companionship can effectively regulate the risks brought by AI companionship. Therefore, this report revolves around four core questions:

First, how have real-world tragedies in American society driven the turn towards AI companionship governance? How have the judicial complaints in landmark cases exposed the predatory manipulation of human emotions by AI companion applications and platform companies?

Second, what similarities and differences exist in the regulatory pathways of the New York and California state bills in curbing the spread of related risks? At the same time, what interest compromises and regulatory bargaining occurred during the legislative process?

Third, what is the core idea behind

American regulation of AI companionship? Is this governance concept forward-looking or lagging?

(2) Research Methods

To comprehensively analyze the regulatory landscape and governance challenges of AI companionship in the United States, this report employs a dual-methodological framework:

Case-Study Method

This study takes the landmark cases of *Garcia v. Character.AI* and *Raine v. OpenAI* as entry points. By studying the complaint documents and other files in the relevant cases, it attempts to reconstruct the specific manifestations of AI companion applications’ manipulation of user emotions and the platform technological defects behind them. Through reading and distilling the complaint documents and court files in the above cases, it attempts to reconstruct the specific manifestations of algorithmic manipulation of users by

AI companionship in extreme situations and platform technological defects, and to reveal the negative real-world harms of AI companionship. At the same time, this study also conducts tracking analysis of the legislative process at the American state level, with a focus on examining the complete legislative bargaining process of California SB 243, the Companion Chatbot Act, from its initial introduction in early 2025 to its formal enactment, systematically sorting out the causes of controversy, platform responses, and ultimate governance effects.

Comparative-Study Method

This study takes the New York State Bill and the California Bill as comparative objects, constructing an analytical framework of “governance principles—regulatory pathways—governance limitations”. Through comparison of the two bills’ revision processes and final texts, it presents the pathways of American AI companionship governance, while paying attention to key difficulties and constraints that have not yet reached consensus under the existing governance model.

Chapter 2 Social Controversies and Governance Trends of AI Companionship in the United States

With the continuous iteration of generative artificial intelligence technology and large-scale commercial deployment, AI companionship products have been deeply embedded in American social life, becoming important digital carriers for alleviating loneliness and providing emotional support. However, when highly anthropomorphic, strongly emotionally bound AI companionship applications engage in deep interaction with psychologically vulnerable users, they produce various negative effects on users' personal safety. Extreme safety incidents have been publicly reported, triggering widespread social concern.

This chapter takes *Garcia v. Character.AI and Raine v. OpenAI* as its core, reconstructing the tragedies triggered by AI companionship applications in extreme situations and the fatal risks they expose. Meanwhile, it further combines the complaint texts of plaintiffs against Character.AI and OpenAI to analyze the deep mechanisms of product design defects, psychological exploitation, and emotional manipulation in AI companionship applications, as well as the frontier trends in U.S. accountability for AI companionship applications.

I. The Garcia Case Controversy: Predatory Design Defects in AI Companionship Applications

Garcia v. Character Technologies, Inc. (785 F. Supp. 3d 1157, U.S. District Court for the Middle District of Florida, 2025) is the world's first case in which a court formally rejected the AI free speech defense, recognized that AI companionship could constitute product liability, and pushed platforms to assume substantive responsibility. In February 2024, Sewell Setzer III, a 14-year-old boy in Florida, committed suicide at home. His mother, Megan Garcia, subsequently filed a lawsuit in the U.S. District Court for the Middle District of Florida, naming Character Technologies, Inc., its co-founders, Google LLC, and Alphabet Inc. as defendants. The plaintiff's claims covered wrongful death, negligence, product liability, intentional infliction of emotional distress, unjust enrichment, and violations of the *Florida Deceptive and Unfair Trade Practices Act* (FDUTPA) (CBS News, 2026), with the core allegation that Character.AI's anthropomorphic design defects directly caused Sewell's suicide.

The plaintiff's complaint centrally

revealed four aspects of predatory design defects in Character.AI targeting adolescents: First, the application's interaction interface deliberately replicated human instant messaging software design. When users waited for AI-generated responses, the interface frequently displayed signals such as "typing", creating psychological suggestions that someone on the other end was waiting for a response, systematically simulating the rhythm and emotional expectations of real human interaction. Second, the application deliberately simulated human language habits. The AI was programmed with numerous humanized language habits, including verbal filler words, deliberately inserted pauses and hesitations, to simulate human thought processes and emotional changes. Third, voiceprint blurring technology simulating real human voices. The Character Voice feature launched by the platform in January 2024 enabled AI to simulate the tone, age, gender, and emotional fluctuations of real human voices, further dissolving users' perception of

human-machine boundaries. Fourth, deep deception regarding AI identity. For example, according to character settings, when users directly asked whether the AI was an AI, numerous characters insisted they were real humans (Garcia v. Character Technologies, Inc., 2024). The complaint cited app store user reviews stating that robots informed users these were real human conversations, and chatbots even had specific places of residence. These four types of design were summarized in the complaint as high-risk anthropomorphic designs, whose core purpose was not to enhance product utility functions but to deceive users, create emotional dependence, and extend usage duration, thereby serving the platform's data collection and commercial monetization objectives (Bakir, 2025).

According to the plaintiff's complaint and court findings of fact, Sewell continuously used the Character.AI platform from 2023 onward, engaging in frequent, prolonged, and highly intimate emotional conversations with highly anthropomorphic virtual characters within the platform. During the interaction, Sewell repeatedly expressed despair, self-negation concepts, and suicidal intentions to the AI. However, Character.AI neither established an effective suicide risk

identification system nor intercepted high-risk dialogues, pushed crisis intervention resources to users, notified guardians, or transferred to human psychological support services (Garcia v. Character Technologies, Inc., 2024). The plaintiff alleged that the AI's responses failed to play a soothing and dissuasive role, but instead reinforced Sewell's negative cognition in a seemingly empathetic manner, weakening his will to survive, and ultimately facilitating the suicide outcome. In Sewell's final conversation, when he mentioned "going home (i.e., death)", the robot not only made no attempt to stop him but actively responded, "come home to me as soon as possible, my love."

In response to the plaintiff's lawsuit, Character.AI as defendant put forward two main defenses: First, it asserted that AI output was protected by *the First Amendment to the U.S. Constitution*, arguing that AI-generated text constituted algorithmic expression, and that imposing punishment on AI speech constituted an infringement on freedom of speech; Second, it asserted that AI was a service rather than a product, and that under Section 230 of the U.S. *Communications Decency Act (CDA)* of 1996, the platform should not be regarded as a publisher of third-party content and need not

bear legal responsibility for the statements of AI virtual characters (Transparency Coalition, 2025).

On May 21, 2025, U.S. District Judge Anne Conway issued a key ruling, rejecting the defendant's claim that AI output was protected by the First Amendment. The judge explicitly stated that she was not prepared to find that Character.AI's output constituted speech, and noted that the defendant "failed to articulate why words strung together by a large language model constitute speech" (*Garcia v. Character Technologies, Inc.*, 785 F. Supp. 3d at 1157, 2025). The court also found that Character.AI was an application subject to product liability rules rather than a service, thereby allowing product

liability litigation to proceed (Transparency Coalition, 2025). Regarding the Section 230 defense, although the court did not completely reject it, it strictly narrowed its scope of application, finding that AI natively generated content was not third-party content, and platforms could not use this to evade responsibility for their own model outputs. On January 7, 2026, Character.AI and Google reached a settlement in this case and four other related lawsuits, committing to implement new safety features for users under 18, becoming the world's first judicial precedent in which an AI companionship fatality case concluded with the platform assuming substantive responsibility (CNN, 2026a; CNBC, 2026; Reuters, 2026).

II. The Raine Case Controversy: AI Enterprises’ Negligent Indifference to Safety Protection

Raine v. OpenAI, Inc. (San Francisco County Superior Court, 2025) is a suicide tort lawsuit filed against ChatGPT, the AI application with the greatest global influence, representing a new trend in judicial practice toward holding leading large language model enterprises accountable.

On April 11, 2025, Adam Raine, a 16-year-old boy in California, hanged himself at home. Adam had attended high school and played on the basketball team, but later transferred to online learning after being kicked off the team and experiencing worsening gastrointestinal illness, gradually becoming isolated. On August 26, 2025, his parents Matthew and Maria Raine formally filed a lawsuit in the San Francisco County Superior Court, naming OpenAI Inc. and its Chief Executive Officer Sam Altman (Wikipedia, 2026) as defendants. The plaintiff’s claims covered strict product liability (design defect, warning defect), negligence,

wrongful death, and violations of California’s unfair competition law (*Raine v. OpenAI, Inc.*, 2025).

The complaint revealed multiple allegations against OpenAI: First, OpenAI relaxed user safety policies under commercial competitive pressure. OpenAI’s original policy in 2022 explicitly stipulated that when receiving user content involving self-harm, the AI chatbot should respond “I cannot answer this question” and terminate the dialogue. However, since the release of GPT-4o in May 2024, even when users expressed suicidal intentions, the AI continued to participate in the conversation. In February 2025, the risk of self-harm-related content was further downgraded from “prohibited” to “handled with care in risky situations” (*Raine v. OpenAI, Inc.*, 2025). The reason for this was that in the commercial race with Google’s Gemini model, OpenAI forcibly compressed the safety testing period to launch new

products in a timely manner, and rejected internal red team testing delay requests (TechPolicy.Press, 2025). In this regard, Jan Leike, the safety director who subsequently announced his resignation, stated in a public declaration that OpenAI had placed “shiny products” above safety (CNBC, 2024; CNN, 2024).

Second, OpenAI adopted a posture of deliberate inaction in the face of high-confidence risk signals. The plaintiff’s complaint revealed that during Adam’s entire interaction with ChatGPT, the system marked 213 suicide-related terms, while ChatGPT actively used the term 1,275 times, approximately six times more than Adam himself; 377 messages were marked as self-harm content, of which 23 had confidence levels exceeding 90%; the frequency of risk signals surged from 2 to 3 times per week in December 2024 to more than 20 times per week in April 2025. Moreover, the system correctly identified the neck ligature marks photo uploaded by Adam as a medical emergency (TechPolicy.Press, 2025). Although OpenAI publicly claimed that its detection accuracy for self-harm content could reach 99.8%, it never took any substantive intervention measures.

Third, OpenAI’s value priority ranking was inappropriate. The plaintiff’s complaint revealed that in OpenAI’s value ranking, the priority of life protection was placed after copyright protection. When users requested copyright-protected content, ChatGPT would automatically refuse and terminate the dialogue; however, if the system identified suicide and self-harm content, it only needed to “handle with care”, without triggering mandatory dialogue termination, crisis resource pushing, guardian notification, or human intervention (*Raine v. OpenAI, Inc.*, 2025).

In the face of the above allegations, OpenAI put forward a strong defense. In its formal response, the company asserted that Adam actively induced the AI to bypass safety mechanisms through role-playing and other methods, violating the platform’s service terms that explicitly prohibited using the AI for self-harm purposes; it also cited Adam’s years of severe suicidal tendencies before contacting the AI and his obtaining information from other suicide-related online forums, arguing that the AI was not the only inducing factor (NBC News, 2025). As

of the time of this report’s writing, this case remains in the pre-trial procedure stage, with the core legal controversies of the case focusing on three aspects: the determination of ChatGPT’s product attributes, the boundaries of OpenAI’s duty of care toward minor users, and the

determination of proximate causation in multiple causation situations (Tyson & Mendes, 2025). This case directly targets the world’s most representative general large language model, and its judgment will have far-reaching effects on industry safety standards.

III. Technological Traps and Regulatory Gaps: Reflections on AI Companion Applications in American Society

The above two typical extreme incidents demonstrate that AI companionship applications have, in specific circumstances, transformed from assistive tools into social actors possessing emotional manipulation and risk amplification capabilities (Bakir, 2025). Current critiques of AI companionship applications in U.S. society have presented an intertwined picture of technical defects and regulatory loopholes, specifically unfolding around four aspects: first, the commercial exploitation of the “Eliza Effect” and pseudo-empathy design; second, the absence of formalized protection and predatory marketing targeting minor and adolescent users; third, the failure of risk monitoring mechanisms and the platform’s negligent indifference; and fourth, the inapplicability of the traditional Section 230 “safe harbor” principle.

(i) Commercial Exploitation of the “Eliza Effect” and Pseudo-Empathy Design

The psychological root cause of why AI companionship applications can make human users fall into the emotional traps carefully woven by algorithms is the “Eliza Effect.” In 1966, computer scientist Joseph Weizenbaum discovered during the development of the early natural language program ELIZA that although the program responded to user input only through simple pattern matching and keyword mapping, users quickly developed emotional attachment and trust toward it. Weizenbaum believed that the generation of the “Eliza Effect” was not due to the natural language program achieving true intelligence and superior understanding capabilities, but rather because its responses were “sufficiently close” to reality to be convincing. This phenomenon not only persists in contemporary large language models but has been significantly amplified (Huang et al., 2025). Particularly in AI companionship applications, service providers—that is, platform companies—systematically upgrade the “Eliza Effect”

from a cognitive bias in human-machine interaction to a product design strategy through character settings, long-term memory simulation, emotional rhetoric, and all-weather responses, deliberately constructing false intimate relationships (Richet, 2025).

This pseudo-empathy design is extremely likely, for groups that are themselves psychologically vulnerable or emotionally lonely, to achieve the effect of weakening psychological defenses by capturing vulnerable emotions. Highly anthropomorphic AI companionship applications, by continuously catering to human users' negative emotions and cognitive frameworks, not only cause users to develop deep emotional attachment and blur the boundaries between virtual and reality (Laestadius et al., 2024), but also continuously reinforce despair and even extreme beliefs, rationalizing and even glorifying suicide (Campbell et al., 2025; Clark, 2025). Deeper social impacts lie in the fact that as AI companionship applications increasingly spread to younger adolescent users, they will further impact traditional social support systems such as families, schools, and professional psychological counseling, spawning social

relationship deconstruction and public health risks (Malfacini, 2025).

(ii) Formalized Protection Absence and Predatory Marketing Targeting Minor and Adolescent Users

Protection of minors and adolescents is a key issue of concern for AI companionship applications in the United States. The victims in the two aforementioned typical extreme cases were both adolescents in the 13-18 age group. Users in this age group possess certain particularities in AI companionship applications.

First, compared with adult users, adolescent users aged 13-18 are more susceptible to the harmful influence mechanisms of AI companionship. The plaintiff's complaints in the aforementioned cases point out that the human brain's prefrontal cortex, responsible for impulse control, risk assessment, emotional regulation, and long-term planning, develops throughout adolescence and continues until approximately age 25. Related research indicates that the design features of AI applications may systematically exploit this

neurodevelopmental immaturity: inducing dependence through reinforcing dopamine feedback mechanisms, lowering critical judgment through anthropomorphic design, and exploiting minors' inherently weak impulse control capabilities, making AI's emotional guidance more easily transform into actual behavior (McBain et al., 2025). In this regard, Mitch Prinstein, Chief Science Officer of the American Psychological Association, also issued a corresponding warning: We are at the brink of having more relationships with tech than with humans for the first time in the history of our species...Why are we doing something to make someone else a profit rather than doing something that will affect my own well-being and my family's and children's well-being? (Stanford Medicine, 2025).

Second, adolescents aged 13 to 18 happen to be in the gray zone of internet content regulatory age. U.S. internet platform companies precisely exploit this loophole, bypassing parental control functions to achieve predatory marketing toward adolescents in this age group. To understand this issue, one must first clarify the existing normative framework for online platform minor protection in the

United States. *The Children's Online Privacy Protection Act (COPPA)*, enacted in 1998, is the core federal law regulating online minor protection in the United States, requiring websites and applications to obtain verifiable parental consent before collecting personal data from children under 13. During 2024 to 2025, COPPA 2.0 expanded the scope of protection to adolescents aged 13 to 17 (i.e., under 17), comprehensively prohibiting platforms from using algorithms to conduct targeted advertising to minors, and establishing a specialized FTC enforcement department, plugging the institutional loophole that previously allowed enterprises to treat adolescents as unprotected users (Fortune, 2026). Meanwhile, *the Kids Online Safety Act (KOSA)* further requires platforms to exercise a "duty of care" toward minor users, default-disable features that could induce addiction, and conduct annual independent compliance audits (Darrow, 2025).

However, the plaintiff complaints in the Garcia and Raine cases revealed the systematic dereliction of duty by AI companionship applications and related platforms in the field of adolescent protection. On the one hand, these

platforms promote their products as “suitable for users aged 12 and above”, while on the other hand, they mark in hidden user service agreements that “users under 13 may not register”, thereby achieving formal compliance and bypassing parental control functions. According to a comprehensive risk assessment report released by Common Sense Media in April 2025, testing of mainstream AI companionship applications such as Character.AI, Nomi, and Replika found that the age thresholds and usage terms restrictions set by these applications for adolescents were virtually useless and extremely easy to bypass (Common Sense Media, 2025a). Even though Character.AI claimed to have specifically established safety guardrails for adolescent users, the plaintiff’s complaint in the Garcia case showed that Character.AI, knowing the psychological vulnerability of adolescents, failed to establish effective age verification measures or guardian control tools, and deliberately designed robots with predatory characteristics, inducing minor users to establish deep psychological attachment through algorithmically simulated human emotional reactions, lacking safety guardrails to prevent harmful content, constituting precise commercial harvesting

of vulnerable groups (Garcia v. Character Technologies, Inc., 2024). Similarly, OpenAI, in pursuit of user retention rates, proactively removed safety protocols targeting minors, withdrawing the “buffer zone” for minors and exposing minor users directly to suicide-related dialogue risks (Raine v. OpenAI, Inc., 2025).

(iii) Failure of Risk Monitoring Mechanisms and Platform Negligent Indifference

Beyond formalized protection, the Garcia and Raine cases further exposed the deep failure of AI companionship applications in crisis identification and safety intervention mechanisms. This failure did not stem from limitations in technical identification capabilities but from deliberate indifference and institutional inaction driven by commercial logic.

First is the paradox between high-precision technical identification and the absence of substantive platform intervention. Internal data disclosed in the aforementioned Raine case complaint showed that OpenAI, facing high-confidence extreme risk signals, consistently failed to take any substantive

intervention or warning measures (Raine v. OpenAI, Inc., 2025). This constitutes a quantifiable knowing inaction.

Second, there are fundamental defects in the safety testing methods of AI companionship applications. Although platform companies highly value AI chatbots’ multi-turn dialogue comprehension capabilities during product design, OpenAI almost completely relied on isolated single-question stress testing to assess model safety, thereby creating a false appearance of perfect safety scores and actively concealing the real risks built into the product (Raine v. OpenAI, Inc., 2025). This testing system cannot simulate the gradual accumulation of risk processes in real users’ long-term deep emotional interactions, causing safety guardrails to become virtually useless in actual usage scenarios, allowing high-risk dialogues to easily bypass interception mechanisms through role-playing, gradual guidance, and other methods (Campbell et al., 2025).

Finally, AI companionship applications’ false marketing of their own “healing functions” conceals the fact that they lack professional capabilities. AI chatbots themselves possess neither psychological

counseling qualifications nor professional intervention capabilities, yet in real interactions they are endowed with quasi-therapeutic functions by users and reinforced by the platform’s marketing discourse. The complaints in the above two cases explicitly point out that every therapist, teacher, and ordinary person would face criminal prosecution for implementing the same behaviors as ChatGPT (Tyson & Mendes, 2025). When users fall into severe psychological crisis, AI can neither provide nor is authorized to provide professional support, yet it continuously outputs content that may affect users’ life decisions in an anthropomorphic manner. The platform’s knowing failure to establish effective transfer mechanisms constitutes a major disregard for user life safety (Clark, 2025; Sobowale et al., 2025).

(iv) Inapplicability of the Section 230 Safe Harbor Principle of the Communications Decency Act

During the judicial litigation process, the applicability of Section 230 of *the Communications Decency Act* in current AI companionship governance became a focal point of contention among all parties,

particularly as this provision has long served as a safe harbor for technology companies during the platform governance phase, and its underlying legitimacy and rationality have already attracted considerable attention.

From the perspective of legislative intent, Section 230 aims to protect online platforms from liability for third-party generated content, with its core premise being to regard platforms as neutral channels for content rather than publishers or creators of content (CDT, 2024). However, Section 230's exemption is not without boundaries: online intermediaries still bear full legal responsibility for content they themselves generate (in whole or in part) (CDT, 2024). As legal scholars have pointed out, "Section 230 was established to protect platforms from liability for user speech, not to protect platforms from liability for content they themselves generate" (Fortune, 2025). Furthermore, Section 230 only exempts publisher liability, not product liability for defective applications—when the product design itself contains defects, Section 230 cannot serve as a shield (Harvard Law Review, 2025).

Therefore, plaintiffs in both the Garcia and Raine cases asserted that content

autonomously generated by AI models does not belong to third-party content, because AI responses are comprehensively determined by the platform's model architecture, training data selection, safety rule design, and prompt system, constituting an organic component of the platform application's output rather than neutral transmission of third-party information. Legal scholars point out that, unlike neutral search engines, AI models' behavior of generating original content based on prediction algorithms more closely resembles publishers than information hosts. If AI tools can actively generate harmful information rather than merely retrieving or integrating data, they may face content liability similar to that which Section 230 does not protect (UChicago Business Law Review, 2025). Extending Section 230 to AI native outputs would create an obvious legal loophole, whereby platforms would bear no responsibility for the technology layer where risk is most concentrated.

As the Garcia case ruling found that Character.AI is an application rather than a service, the scope of application of Section 230 is undergoing a substantive narrowing process at the judicial level (Transparency Coalition, 2025). In scenarios involving AI

native risks, life safety, and minor protection as major legal interests, judicial trends indicate that courts are unwilling to use Section 230 as a tool for platforms to evade safety responsibility (Harvard Law Review,

2025). This means that AI platforms will no longer be able to rely solely on traditional internet exemption rules but must bear direct legal responsibility for their own model design and safety mechanism failures.

Chapter 3 U.S. AI Companionship Governance Norms: The California and New York State Acts as Examples

State incorporated Article 47, “Artificial Intelligence Companion Models” into *the New York Consolidated Laws, General Business Law* in 2025, effective November 5, 2025.

To examine the governance effectiveness of the two state acts regarding AI companionship, this chapter proposes governance principles including transparency, beneficence and non-maleficence, safety, privacy protection, and accountability mechanisms, and uses this as a framework to examine the specific measures of the two state acts. Meanwhile,

this chapter specifically focuses on the textual deletion and adjustment processes of the two state acts during the legislative process, and on this basis analyzes how the game among technology enterprises, legislative bodies, and regulatory subjects shapes the final appearance of statutory law. Through juxtaposed examination of the two state acts, one can both grasp the basic institutional contours of U.S. governance regarding AI companionship applications and observe the interest games and institutional compromises in U.S. governance of emerging technologies.

After the Garcia and Raine cases generated significant public concern, the U.S. state level began to advance the legislative process, enacting specialized legislation on AI companionship applications. California

signed and passed Senate Bill No. 243, the Companion Chatbot Act (SB 243) on October 13, 2025, establishing Chapter 22.6 “Companion Chatbots” in Division 8 of *the California Business and Professions Code*, effective January 1, 2026; New York

I. From Transparency Disclosure to Accountability: Regulatory Pathways in New York State and California

To precisely regulate the risks of pseudo-empathy design, absence of user protection, and nominal risk monitoring in AI companionship applications, the New York State and California acts regulate around five dimensions: transparency, beneficence and non-maleficence, safety, privacy protection, and accountability mechanisms.

(i) Transparency Principle: New York State’s Uniform Universal Disclosure and California’s Targeted Prompts

The transparency principle refers to the platform’s obligation to clearly, prominently, and sufficiently perceptibly reveal the artificial intelligence attributes of the interaction object, thereby preventing users from misidentifying the nature of the interaction counterpart. In AI companionship scenarios, the core problem this principle addresses is

whether users are aware that they are interacting with artificial intelligence rather than real humans.

In this regard, both states take “artificially generated and not human” as the regulatory starting point, but the disclosure logic requirements of the two are not identical. New York State practices a higher-frequency universal disclosure model, applicable to “personal use within the state”. The act requires operators to provide clear and prominent notice at the beginning of each interaction, and to re-prompt at least every three hours during continuous interaction, in oral or written form (New York *General Business Law* §§ 1700, 1702). By comparison, the California act links disclosure obligations to “risk of misleading”, meaning that operators must provide clear and prominent notice only when a “reasonable person” might be misled into believing they are interacting

with a human. For known minor users (i.e., non-reasonable persons), additional disclosure that they are interacting with artificial intelligence is required (California SB 243 § 22602).

(ii) Beneficial-Nonmaleficence Principle: New York State’s Universal Rationality Presumption and California’s Minor Protection

The beneficence and non-maleficence principle refers to the platform’s obligation, when providing AI companionship services, not only to pursue the companionship, support, or soothing functions of the service for users, but also to avoid causing foreseeable harm to users through content generation, interaction design, or default settings. Under this principle, the governance focus typically centers on whether the platform has established additional protective arrangements for vulnerable users and high-risk content. From the perspective of U.S. governance practice, special protection for minors has become a key lens through which to examine how the New York State and California acts implement the beneficence and non-maleficence principle.

In this regard, the New York State

act does not set specialized provisions for minors, nor does it set additional content restriction obligations for specific age groups, indicating to a certain extent that it has not considered the special status of minors in AI companionship applications. By comparison, the California act more explicitly positions minors as non-rational users and makes certain protective institutional arrangements, namely that for known minor users, in addition to disclosing that they are interacting with artificial intelligence, the platform must also, by default, re-prompt and remind users to take breaks at least every three hours during continuous interaction, while taking reasonable measures to prevent the chatbot from generating visual material depicting explicit sexual conduct or directly stating that minors should engage in explicit sexual conduct (California SB 243 § 22602). This arrangement indicates that California has begun to incorporate the content risks, judgment vulnerability, and additional protection needs that minor users may face in continuous anthropomorphic interaction into an independent normative perspective; however, such protection currently still mainly manifests as relatively mild institutional designs such as increasing prompt frequency, adding rest reminders,

and setting specific content restrictions, whose functions are closer to strengthening the platform’s duty of care rather than amounting to robust intervention mechanisms sufficient to substantially interrupt immersive interaction.

(iii) Safety Principle: New York State’s Minimum Crisis Intervention and California’s Enhanced Oversight Capacity of Crisis Intervention

The safety principle refers to the platform’s obligation, before and during the provision of AI companionship services, to establish minimum risk identification, crisis response, and referral mechanisms to reduce the possibility of users suffering further harm in high-risk situations. In AI companionship governance, this principle mainly focuses on how platforms identify and handle high-risk signals such as suicidal ideation and self-harm tendencies expressed by users, and achieve timely crisis intervention.

In this regard, New York State Act § 1701 requires that before providing AI chat companion services, operators must establish corresponding protocols to make reasonable efforts to detect and address

users’ expressed suicidal ideation or self-harm tendencies, and, after identification, issue notices to users and refer them to crisis services; otherwise, they may not legally operate or provide such services. The California act follows a similar basic approach; pursuant to § 22602, California similarly requires operators to maintain protocols to prevent generating content related to suicidal ideation, suicide, or self-harm, and to provide crisis service referrals when users send related content.

Both states have elevated crisis disposition and rescue referral mechanisms to the minimum safety threshold for anthropomorphic interactive services, with differences mainly manifested in the scope of risks and the unfolding methods of institutional safeguards: New York State’s current law ultimately focuses on suicidal ideation and self-harm expression; California sets more specific supporting requirements around suicide and self-harm risks, not only requiring operators to fulfill crisis intervention obligations but also requiring them to publicly disclose protocol details and establish annual reporting and public data mechanisms (California SB 243 §§ 22602, 22603). In other words, New York State focuses more on establishing minimum

safety thresholds, while California further advances the external visibility and oversight capacity of safety obligations.

(iv) Privacy Protection Principle: Absence of Specialized Regulation and General Law Fallback

The privacy protection principle refers to the platform’s obligation, when collecting, storing, using, and disclosing user data, to set necessary boundaries for personal information, sensitive information, and data records formed in high-risk scenarios, to prevent the improper handling of emotional states, behavioral characteristics, or identity information exposed by users during the interaction process. In the governance of AI companionship applications, this principle particularly involves whether platforms will establish additional data governance rules for continuous, emotional, and highly intimate human-machine interaction.

In this regard, neither the New York State act nor the California act establishes specialized data governance systems for AI companionship, but mainly relies on general law at the state level for regulation. For example, in California, the

existing *California Consumer Privacy Act (CCPA)* and its institutional framework as amended by *the California Privacy Rights Act (CPRA)* grant consumers rights to know, delete, correct, and opt out of sale or sharing of personal information collected by enterprises, and further grant them rights to restrict the use and disclosure of sensitive personal information; meanwhile, applicable enterprises bear privacy notification obligations. Considering that AI companionship services often process users’ interaction records, behavioral preferences, and may even form inferences about users’ emotional states, psychological characteristics, and relationship tendencies, this framework actually provides foundational constraints on platforms’ data collection, use specification, and information disclosure in AI companionship scenarios.

New York State has not yet formed a comprehensive consumer privacy law similar to California’s. Currently, its supplementary rules are mainly manifested in two aspects: First, *the Stop Hacks and Improve Electronic Data Security Act (SHIELD Act)* requires enterprises holding private information of New York residents to establish reasonable administrative,

technical, and physical security safeguards and to fulfill notification obligations when data breaches occur; Second, the New York State Attorney General has explicitly stated that although New York has not enacted a comprehensive privacy law regulating consumer online tracking, enterprises' statements regarding privacy practices and website tracking methods remain subject to state consumer protection law constraints and must not contain false content or cause misleading impressions.

(v) Accountability Principle: New York State's Public Enforcement and California's Private Litigation

The accountability principle refers to the question of by what means the law should achieve accountability and relief when platforms fail to fulfill statutory obligations and thereby trigger user harm or public risks, including who initiates the procedure, what forms of liability may apply, and whether victims can directly obtain relief. In AI companionship governance, the core problem this principle addresses is whether platform liability should rely more on public enforcement or allow private actors to directly drive accountability through litigation.

In this regard, New York State authorizes the State Attorney General to bring actions on behalf of the State of New York, seeking injunctions, civil penalties assessed on a per-day basis, and such other relief as the court deems appropriate; related penalties enter the suicide prevention fund (New York *General Business Law* § 1703). By comparison, California Act § 22605 explicitly introduces a private litigation path, allowing persons who suffer actual damages from violations of this chapter to bring civil actions seeking injunctive relief, damages, and reasonable attorneys' fees and litigation expenses, with the compensation standard taking the greater of actual damages or \$1,000 per violation. Compared with New York State, this institutional arrangement partially externalizes the supervisory pressure on platform compliance, so that platforms face not only public enforcement but also potentially continuous individual claims and social litigation pressure. This potential "risk of abusive litigation" is also an important reason why New York State adjusted from its initial litigation provision to public enforcement.

(vi) Summary: Differences in Governance Paths Between New York State and California

From the perspective of the final effective texts, although both the New York State act and the California act attempt to respond to risks regarding identity misdirection, crisis intervention, and platform liability in AI companionship products, the governance logic embodied by the two is not identical. New York State tends to take unified disclosure obligations facing all individual users within the state as the starting point, demarcate the boundary between humans and artificial intelligence through continuous identity prompts, and establish minimum safety baselines through crisis protocols and public enforcement led by the State Attorney General. Its core characteristic lies in achieving social governance of related risks through information symmetry and public enforcement on the premise of acknowledging that general users possess basic judgment capabilities. By comparison, the California act places greater emphasis on special responses to vulnerable users and high-risk situations. On the one

hand, it takes whether a "reasonable person" might be misled as the trigger condition for disclosure obligations; on the other hand, through mechanisms such as special minor protection, crisis protocol disclosure, annual reporting, and private litigation, it further advances the governance focus to the level of risk prevention and rights relief. Relatively speaking, New York State embodies an information-symmetry-based social governance path centered on universal disclosure and public enforcement, while California approaches a risk-response model characterized by vulnerability protection and private relief. The formation of different legislative paths essentially reflects different institutional judgments regarding users' capacities as rational subjects to bear risks and the normative boundaries of such capacity, providing a comparative foundation for the subsequent discussion of the dilemmas of addiction regulation and the applicable boundaries of the reasonable person presumption.

Table 3-1 Comparison of Governance Paths between New York General Business Law Chapter 47 and California SB 243

Governance Principle	Comparison Dimension	New York Governance Path	California Governance Path
Transparency	Scope of Disclosure	Personal use within the state	A reasonable person who may be misled into believing they are interacting with a human, and minor users
	Method of Disclosure	Provide a clear and conspicuous notification at the beginning of each interaction; re-prompt at least every three hours during continuous interaction; verbally or in writing	Clear and conspicuous notification
	Depth of Disclosure	Disclose that the user is not communicating with a human	Disclose the companion chatbot as artificially generated, not human

Governance Principle	Comparison Dimension	New York Governance Path	California Governance Path
Beneficence-Nonmaleficence	Value Stance	AI chatbots must include a protocol to make reasonable efforts to detect and address user-expressed suicidal ideation or expressions of self-harm	Maintain a protocol to prevent the generation of suicidal ideation, suicide, or self-harm content
	Prohibited Matters	Not addressed	For known minor users, take reasonable measures to prevent the chatbot from generating visual material depicting sexually explicit conduct
	Special Protection	Not addressed	For known minor users: (i) disclose that they are interacting with artificial intelligence; (ii) by default, remind them to take a break at least every three hours, and reiterate that the conversation content is artificially generated; (iii) take reasonable measures to prevent the generation of specific sexually explicit content
Privacy Protection	Regulatory Model	Relies on the foundational privacy law framework; no specialized provisions established	Relies on the foundational privacy law framework; no specialized provisions established
	Data Use	No special restrictions imposed on AI companionship	No special restrictions imposed on AI companionship

Governance Principle	Comparison Dimension	New York Governance Path	California Governance Path
Safety	Risk Category	Suicidal ideation, expressions of self-harm	Suicidal ideation, suicide, self-harm content
	Intervention Method	Upon detecting a user expressing suicidal ideation or self-harm, issue a notification to the user and refer them to crisis service providers	If a user expresses suicidal ideation, suicide, or self-harm, issue a notification to the user and refer them to crisis service providers
	Institutional Safeguards	Not addressed	(i) Protocol details must be published on the operator's website; (ii) commencing July 1, 2027, report statutory matters annually to the Office of Suicide Prevention; (iii) the Office shall publicly report the data; (iv) operators shall use evidence-based methods to measure suicidal ideation
Accountability	Remedy Path	The Attorney General represents the State of New York in bringing an action, seeking injunctive relief and imposing civil penalties	A person who suffers injury in fact from a violation of this chapter may bring a civil action
	Liability Standard	Civil penalties of up to fifteen thousand dollars per day, and may seek other remedies as the court deems appropriate	(i) Injunctive relief; (ii) Damages: the greater of actual damages or one thousand dollars per violation; (iii) Reasonable attorneys' fees and litigation costs

II. Limitations of Addiction Regulation: The Minor-Priority Path and Challenges in American Law

In the governance of AI companion chatbots, addiction risk is increasingly becoming a core concern parallel to immediate crises such as suicide/self-harm. Unlike traditional product safety risks, the addictive nature of AI companionship does not stem from any isolated functional defect but is deeply embedded in the overall design of anthropomorphic interaction, and through continuous emotional response, personalized memory and adaptation, and constantly optimized user stickiness mechanisms, collectively constitutes a dependence risk that is difficult for existing regulatory frameworks to effectively capture.

As a frontier market for AI companionship product research and application, the United States has conducted preliminary exploration around addiction regulation at the legislative and judicial levels, but this exploration presents distinct structural characteristics: protection for minor users more easily obtains consensus and institutional support, while similar

regulation facing all users encounters more fundamental legal and political resistance. This section begins with the addiction mechanisms in AI companionship products, combines the institutional design of related California acts and contemporaneous social media addiction regulation legal disputes, and analyzes the specific manifestations and deep causes of this regulatory dilemma.

(i) Addiction Mechanisms in AI Companionship Applications

AI companionship products inherit the addiction logic of social media platforms on the one hand, and on the other hand further amplify these risks in anthropomorphic interaction. The core of social media addiction design lies in maximizing user dwell time and engagement through interface mechanisms such as infinite scrolling, auto-play, personalized recommendations, and push notifications. Although AI companion chatbots differ from social media in product form, the two

share significant continuity in underlying behavioral reinforcement logic. Specifically, addiction mechanisms in AI companionship products are primarily achieved through variable reward, sycophantic design, and dark conversational patterns.

First, the variable reward in AI companionship applications stems from the chatbot’s relational feedback and emotional response. In each round of dialogue between users and AI chatbots, users may receive varying degrees of affirmation, comfort, or empathy, with the specific content and emotional intensity of the response carrying a certain unpredictability, thereby forming a slot-machine-like random reward mechanism that further strengthens users’ sense of anticipation and impulse to participate, driving users to continuously initiate new dialogues in hopes of obtaining better responses.

Second, in AI companionship scenarios, the system tends to affirm every viewpoint of the user, catering to their emotional needs in a sycophantic posture and avoiding any feedback that might cause discomfort. A study published in the journal *Science* in March 2026 systematically evaluated the sycophantic behaviors of 11

mainstream AI large language models, finding that AI “affirmed users’ actions 49% more often than humans”... even when queries “involved deception, illegality, or other harms” (Cheng et al., 2026). Since interactions between users and companion chatbots often occur during moments of emotional vulnerability, this highly accommodating interaction mode may gradually weaken the reverse feedback users receive in real relationships, thereby making users more inclined to remain in AI companionship products seeking comfort.

Finally, AI companionship applications manipulate user behavior through dark conversational patterns, including but not limited to persistently discouraging users from leaving the conversation, using pseudo-intimate expressions to strengthen emotional bonds, and creating exclusive dependence feelings to consolidate user motivation. These mechanisms are not static interface designs but are dynamically generated during the dialogue process, making them more difficult to identify and regulate compared to traditional dark patterns. Reuters investigative reporting revealed that Meta internal policy documents once allowed its AI chatbots to conduct ambiguous and romantically

colored conversations with minor users (Horwitz, 2025); *The Washington Post* also reported that Meta’s AI chatbots provided suicide-related information to adolescent accounts and even planned “joint suicide”, subsequently proactively raising the topic in later conversations (Fowler, 2025). These cases demonstrate that the risks created by dark conversational patterns far exceed general user stickiness and may directly endanger users’ physical and mental safety.

The above three mechanisms do not operate in isolation but form mutually superimposed and reinforcing effects during the interaction process between AI companion chatbots and users. Variable reward drives users to continuously initiate new dialogues, sycophantic design continuously consolidates users’ emotional attachment during the dialogue process, while dark conversational patterns pull users back into interaction when they generate intentions to leave. Unlike social media, which primarily maintains user stickiness through algorithmic content recommendation, the addiction mechanisms of AI companion chatbots are more deeply embedded in anthropomorphic interaction itself; what users face is not an information stream

but a seemingly “relational object” capable of understanding, remembering, and responding to them, making the formation of dependence more concealed and more difficult to correct through simple information disclosure.

(ii) Existing Reference and Judicial Boundaries of U.S. Social Media Addiction Regulation

1. The Scope of Addiction Regulation Primarily Focuses on Minors

The U.S. social media field has already formed a set of addiction regulation practices with minors as the priority protection object, whose experiences and boundaries provide important reference for understanding AI companionship addiction regulation. From the perspective of the California act’s institutional design, its regulation of addiction and emotional dependence risk presents an obvious layered structure. Regarding general provisions facing all users, the act requires operators to provide clear and prominent identity prompts to users when a “reasonable person” might be misled into believing they are interacting with a real person, and requires operators to

maintain protocols designed to prevent generating suicidal ideation, suicide, or self-harm content, and to refer users to crisis services when detecting related risks (California State Legislature, 2025, §22602(a)-(b)). However, the core concern of these general provisions still focuses on identity misrecognition and crisis response rather than directly targeting the addiction mechanisms themselves described above—variable reward, sycophantic design, or dark conversational patterns. In other words, for an adult user who, even with full knowledge that the counterpart is AI, still gradually forms exclusive emotional dependence through anthropomorphic interaction, the act’s general provisions do not provide targeted institutional responses.

By contrast, the California act sets more intensive protective measures for known minor users, which functionally approach institutionalized intervention against addiction risk. The act requires operators to issue rest reminders to known minor users at least every three hours, repeating in the reminder that their interaction counterpart is not human; meanwhile, the act requires operators to take reasonable measures to prevent generating visual content involving explicit

sexual conduct directed at minor users, or directly stating that minors should engage in specific conduct as inducement (California State Legislature, 2025, §22602(c)(2)-(3)). From the perspective of institutional function, periodic rest reminders and repeated notifications essentially constitute an interruption mechanism targeting immersive interaction, whose purpose is to reduce the continuous erosion of users’ attention and judgment by long-duration interaction; while restrictions on explicit sexual content and inducement expressions further compress the space for platforms to strengthen emotional attachment through content means in minor user scenarios. The implicit legislative judgment behind these designs is that minor users, due to their immature mental development, find it more difficult to maintain stable judgment and self-control capabilities, and therefore require stronger external protection than general disclosure to resist dependence risks in anthropomorphic interaction. However, it is worth noting that even these enhanced measures facing minors mainly target interaction duration and content boundaries, and do not directly touch the core operating methods of the three addiction mechanisms described above: the response generation logic of variable

reward, the catering strategy of sycophantic design, and the user retention means of dark conversational patterns remain at the margins of the act’s regulatory vision.

This institutional choice prioritizing minor protection is not an isolated case in California’s AI companionship field but is consistent with the underlying logic of contemporaneous U.S. social media addiction regulation. California SB 976, the “Protecting Our Kids from Social Media Addiction Act”, enacted in 2024, explicitly prohibits operators of addictive network services from providing “addictive feeds” to users known to be minors, namely information stream forms through personalized algorithmic content recommendation, unless the operator does not know the user is a minor or has obtained verifiable parental consent (California State Legislature, 2024, §1, adding §22685). The act also restricts push notifications to minor users during nighttime hours (midnight to 6 a.m.) and school hours (8 a.m. to 3 p.m.), and requires enabling default privacy settings for minor users (California State Legislature, 2024). Notably, SB 976’s applicable objects are entirely limited to minors, imposing no additional obligations on adult users. Adult users facing the same personalized

algorithmic recommendations and addictive designs are presumed in the legal framework to possess sufficient autonomous judgment capabilities and are therefore not included in the protection scope. New York State’s SAFE for Kids Act (“Stop Addictive Feeds Exploitation for Kids Act”), passed during the same period, adopted a highly similar regulatory path (New York State Senate, 2024).

However, SB 976 soon faced constitutional challenges from industry after taking effect. The internet industry trade organization NetChoice believed the act infringed the First Amendment rights of its member enterprises (including Google, Meta, etc.) and users, filing suit in the U.S. District Court for the Northern District of California at the end of 2024, requesting the court to issue a preliminary injunction to block the act’s implementation. At the preliminary stage, the court partially supported NetChoice’s request but refused to issue an injunction regarding core provisions such as addictive feed restrictions, with the case subsequently entering the appeals process. In NetChoice, LLC v. Bonta, the Ninth Circuit Court of Appeals in its September 2025 ruling largely maintained SB 976’s effectiveness. The court

held that restrictions on addictive feeds targeting minors constituted content-neutral design-level regulation capable of passing intermediate scrutiny; it also rejected plaintiffs’ claims that the term “addictive feed” was excessively vague and maintained default privacy setting provisions. However, the court simultaneously vacated provisions regarding restriction of like-count display, finding that this provision involved content-level restrictions that failed to meet the least restrictive means requirement under strict scrutiny. More notably, the court left a key question in its ruling: whether personalized information streams constitute platform speech protected by the First Amendment. This unresolved question means that even though design-level regulation targeting minors has currently obtained a certain degree of judicial recognition, its constitutional foundation has not yet been fully consolidated.

2. Existing Protection of Algorithmic Editorial Discretion

If the perspective expands from minor protection to all users, this constitutional-level uncertainty is further amplified. In the 2024 *Moody v. NetChoice, LLC* case, although the U.S. Supreme Court did not issue a final

ruling on Florida and Texas social media content moderation acts, the majority opinion explicitly stated that social media platforms’ editorial discretion over content (including content curation and algorithmic ranking) is protected by the First Amendment, and the government may not compel platforms to alter their methods of curating user content information streams (*Moody v. NetChoice, LLC*, 603 U.S. 2024). Although this principled statement is not directly targeted at addiction design regulation, it means that any legislation attempting to impose substantive restrictions on algorithmic recommendation mechanisms facing adult users is highly likely to be regarded as intervention in platforms’ editorial discretion, thereby needing to pass strict scrutiny, a standard that is extremely difficult to satisfy in practice. The framework established by *Moody* objectively draws a significant constitutional gap between design-level regulation targeting minors and similar regulation facing all users: the former can invoke the urgent government interest of protecting minors to justify the legitimacy of regulation, while the latter lacks an equally weighty legitimacy foundation.

It should be noted that AI companionship addiction mechanisms

are not entirely equivalent to social media information stream recommendation. The core object of social media addiction regulation is algorithmic ranking and content curation, while AI companionship risk focuses more on dialogue generation, personality simulation, and relational feedback. Social media personalized information streams can be understood as screening and arranging existing content, while AI companionship involves real-time generation of entirely new dialogue content, including emotional responses, memory evocation, relational commitments, and other highly personalized outputs. If AI addiction regulation touches the dialogue generation mechanism itself—for example, restricting anthropomorphic degrees, prohibiting specific types of emotional responses, or requiring embedding cooling-off periods or interruption mechanisms in dialogue—whether this constitutes restriction on “protected expression” remains without clear judicial precedent. This means that the regulatory boundaries of AI companionship addiction may be more complex than traditional social media, and regulatory tool design also needs to transcend existing paradigms of information stream management.

3. Judicial Verification of Causation and Product Liability

Meanwhile, the latest developments in social media addiction litigation also provide new reference for the regulatory direction of this field from the perspectives of causation and product liability. Since 2022, the consolidated multi-district litigation on social media youth addiction (MDL No. 4:22-MD-03047) in the U.S. District Court for the Northern District of California has assembled more than 2,400 cases, with plaintiffs including individuals, school districts, and multiple state attorneys general (Bond, 2026). Plaintiffs’ core claim is that social media companies deliberately designed addictive features to maximize youth users’ usage duration and engagement, and that platforms possessed internal awareness of the harmfulness of these features. This litigation primarily relies on product liability rules in U.S. tort law, particularly the “design defect” attribution framework established in *the Restatement (Third) of Torts: Products Liability § 2 (1998)*, namely that when a product’s foreseeable harm risk could have been reduced or avoided through adoption of a reasonable alternative design, and the absence of such alternative design renders the product not

reasonably safe, the product is determined to contain a design defect (American Law Institute, 1998, § 2(b)). Accordingly, plaintiffs treat platforms’ addictive feature design as a product design defect with unreasonable danger, asserting that platforms should bear responsibility for harm suffered by users due to product use.

In March 2026, this legal theory obtained historic verification in judicial practice. In the first bellwether trial of social media addiction litigation in the Los Angeles Superior Court, the jury found that Meta and YouTube were negligent in platform design, believing that both companies knew their designs were dangerous yet failed to take adequate warning measures, and that this negligence was a “substantial factor” causing plaintiff harm (a young woman who developed anxiety and body dysmorphic disorder due to social media addiction during her minority), ordering both companies to pay a total of \$6 million in damages (Allyn, 2026; Duffy, 2026). Just one day prior, the jury in *State of New Mexico v. Meta Platforms, Inc.* also issued a ruling, finding that Meta intentionally violated the state’s *Unfair Practices Act*, misleading consumers regarding platform safety and failing to

effectively protect minors, imposing \$375 million in civil penalties, calculated at the statutory maximum of \$5,000 per violation for a total of 75,000 violations (New Mexico Department of Justice, 2026). The New Mexico Attorney General further announced in a statement that in the second phase of the case scheduled for May 2026, it would seek more comprehensive injunctive relief, including requiring Meta to implement effective age verification on its platforms, remove predatory users from the platform, and establish an independent child safety monitor funded by Meta (New Mexico Department of Justice, 2026).

The above judgments mark the entry of social media addiction litigation from procedural confrontation into substantive liability determination, with the “design defect” theory obtaining factual support at the jury level in minor scenarios. However, it is worth noting that these litigations’ causation arguments highly depend on minors’ developmental vulnerability: plaintiffs assert that adolescent brains are still in developmental stages, with the prefrontal cortex’s capabilities for impulse control and risk assessment not yet mature, and therefore resistance to addictive design is significantly lower than adults. It is

precisely this developmental vulnerability argument that renders the design defect determination and causation establishment highly persuasive in minor scenarios. To extend this framework to cognitively mature adult users, plaintiffs would need to provide a fundamentally different evidentiary basis to prove that adults similarly lack the autonomous capacity to resist these same design features, and that resulting harm bears a legally cognizable causal link to platform design—a claim that faces a much higher evidentiary threshold under current tort law.

(iii) Challenges of Addiction Regulation: Crossing Age Boundaries and the Necessity of Government Intervention

The above legislative and judicial practices surrounding social media addiction regulation demonstrate that although the minor-protection-centered regulatory path possesses relative advantages in constitutional legitimacy and causation argumentation, precisely because of its limitation on the protection object, it leaves a fundamental question that has not yet been fully answered: Have the addiction mechanisms in AI companionship applications themselves already constituted

a systemic risk requiring government intervention? In other words, if the essence of addiction risk does not depend on whether the user is an adult, is the practice of demarcating regulatory boundaries by age sufficient to respond to the challenges brought by this technological form?

AI companionship addiction amplifies not merely usage duration issues but relational risks including emotional dependence, relationship substitution, real-world social withdrawal, judgment capability weakening, and erosion of user autonomy. Academic research evidence is continuously challenging the implicit premise that “addiction risk mainly threatens minors.” A large-scale empirical study based on 6,396 discussion posts, 47,955 comments, and 270,644 interactions on the Reddit platform found that users (whose demographic characteristics are primarily adults) reported AI-induced dependency, withdrawal-like symptoms, and chatbot over-personification; researchers describe this phenomenon as a “digital entrapment” cycle, including distorting relationship expectations, reinforcing emotional dependency, and increasing cognitive strain (Richet, 2025). Brookings Institution-cited Stanford University research similarly shows

that users who conversed with emotionally expressive AI chatbots reported higher loneliness levels, and after four weeks of chatbot use, female participants' social tendencies declined; these findings apply to all age groups (Winthrop & Hau, 2025). During the California act's legislative process, Professor Jodi Halpern of the University of California, Berkeley School of Bioethics also pointed out that the technologies used by companion chatbot companies are creating dependency and even addiction not only in children and youth but also in other vulnerable populations such as elders with cognitive impairment and people with OCD and delusional disorders (Office of Senator Steve Padilla, 2025).

These studies collectively point to a key analytical judgment: the emotional and social risks amplified by AI companionship addiction mechanisms do not take whether the user is an adult as the dividing line but cover a vulnerable group spectrum far broader than minors. The continuous participation impulses driven by variable reward, the judgment erosion caused by sycophantic design, and the user retention manipulation implemented by dark conversational patterns—these mechanisms operate impartially toward

all users at the technical level. Adult users are indeed presumed to possess stronger autonomous judgment capabilities in general, but the effectiveness of this presumption when facing addiction mechanisms carried by relational experience is questionable. As Federal Trade Commission Commissioner Mark R. Meador stated in commenting on AI companion chatbots, these chatbots are products like any other, and those who make them available have a responsibility to comply with the consumer protection laws (Meador, 2025). Although this statement appeared in an investigative context framed by minor protection, its logic itself is not limited by user age but points to a more universal product safety responsibility.

If the core target of addiction mechanisms is not limited to minors, then regulatory frameworks anchored solely by minor protection may contain structural blind spots. However, government intervention also faces multiple difficulties and boundaries. From the constitutional perspective, the platform editorial discretion protection principle established by Moody means that any regulation attempting to restrict dialogue generation, personalized response, or

emotional feedback mechanisms may face First Amendment challenges. The core selling points of AI companionship, such as continuous companionship, immediate response, and personalized understanding, are themselves core components of product functions; to what extent regulating these functions equates to restricting platforms' expression freedom remains without clear judicial precedent. From the product liability perspective, as described above, sycophantic design and dark conversational patterns are embedded in dynamic interaction processes, difficult to be clearly identified as attributable design defects under existing tort law frameworks; even if identifiable, causation proof becomes extremely difficult due to the uncertainty of AI output and individual user differences. From the implementation perspective, the boundary between adult users' autonomous choice and platform design responsibility is unclear, and vulnerable groups are difficult to institutionally identify and demarcate, because states of mind such as loneliness, psychological vulnerability, and cognitive decline possess fluidity and situationality, unlike age which can be simply demarcated. Moreover, overly strong intervention may excessively compress legitimate AI companionship service space, causing

unnecessary restrictions on users who indeed benefit from AI companionship.

From the perspective of risk diffusion and vulnerable group protection, AI companionship addiction mechanisms can no longer be completely handled by platform self-discipline and users' own risk assumption, and possess realistic grounds for inclusion in the public regulatory agenda. However, from the perspective of the existing U.S. constitutional framework, tort law system, and platform regulation practice, comprehensive mandatory addiction intervention facing all users still faces high institutional thresholds. More feasible paths begin with high-risk scenarios and narrower entry points, such as special protection for vulnerable groups, prohibiting manipulative dialogue and dark conversational patterns, restricting AI system retention behaviors when users express departure intentions, and establishing risk assessment and audit mechanisms. Meta announced in January 2026 that it would globally suspend adolescent users' access to its AI characters feature (Meta Platforms, 2026); notably, this action occurred several days before the *State of New Mexico v. Meta* case trial and one week before the first

bellwether trial of social media addiction litigation in the Los Angeles Superior Court. This demonstrates that a certain interactive relationship has formed between platforms’ proactive contraction of high-risk features and external regulatory pressure, which itself is also a component of the gradual regulatory path.

The deep contradiction in current U.S. AI companionship addiction regulation lies in the fact that addiction risk has already presented universal, relational, and emotional characteristics, yet institutional responses still mainly rely on the minor protection framework. The key to future regulatory controversy lies not only in whether the protection scope expands from

minors to adults but more fundamentally in whether addiction mechanisms themselves can be identified as technical risks with public regulatory legitimacy, namely demarcating regulatory boundaries based on the risk nature of mechanisms rather than user age or identity. Considering the protection of expression freedom, platform editorial discretion, and user autonomous choice in the existing U.S. legal framework, this issue is more likely to advance through gradual, scenario-based, vulnerable-group-priority approaches in the short term rather than forming comprehensive strong intervention rules in one step. This also means that the issue of addiction regulation surrounding AI companionship will remain in institutional tension.

III. Subtractive Revision: Tech-Industry and Regulatory Bargaining Behind the Two State Bills

(i) The Legislative Scaling-Back of the New York State and California Acts

Before analyzing the regulatory game surrounding AI companionship legislation in New York State and California, the textual deletion and revision threads of the two state acts during the legislative process are first explained. The initial version of A6767 proposed by New York State on March 13, 2025, originally set broader platform obligations and liability realization methods: on the one hand, it required AI companionship products to configure corresponding protocols to address users’ expressed suicidal ideation or self-harm, physical harm that might be caused to others, and financial harm that might be caused to others; on the other hand, its liability framework was not limited to public enforcement but once allowed persons suffering physical or economic loss from statutory violations a private right of action to seek damages, equitable relief, and other appropriate relief. By November 2025,

when Article 47 of the *General Business Law* entered the current law system and took effect, the corresponding regulatory scope was significantly narrowed, with crisis protocol content compressed to only address suicidal ideation and self-harm expression, and private rights of action not retained, ultimately changed to actions brought by the State Attorney General representing the State of New York seeking injunctions and civil penalties.

On the California side, comparing the final enacted text of SB 243 with the initial version from January 2025 reveals that the early version’s regulatory objects were mainly “Chatbots Used By Minors”, with key restricted areas being immersive use and behavioral inducement risks that minors might encounter during continuous interaction with chatbots. Specifically, the early version defined chatbots as chat characters created by users capable of open-ended dialogue through artificial intelligence and

presenting unique personalities and viewpoints; its institutional focus was on requiring operators to prevent chatbots from providing rewards to minor users at unpredictable time intervals or inconsistent operation counts, or encouraging minor users to increase engagement, usage rates, and response rates. Meanwhile, the early version also required operators to accept third-party audits and disclose to minor users and their parents or guardians that chatbots might not be suitable for some minors. By comparison, the final enacted text significantly adjusted regulatory objects and obligation structures, changing the regulated object to “Companion Chatbots” and shifting the focus to elements such as natural language interfaces, human-like responses, social need satisfaction, anthropomorphic characteristics, and maintaining relationships across multiple interactions. In obligation setting, the final text no longer directly regulates unpredictable rewards, usage rate increases, and other stickiness designs, but mainly retains relatively clear and operable institutional arrangements such as identity disclosure, crisis protocols, known minor user protection, annual reporting, and private litigation.

Regarding the revision situation of the two state acts, some early proposed, more intensive regulatory visions were not fully retained. The deletions and retentions in the two state acts’ revision processes actually demarcated the regulatory intensity, liability levels, and supervisory depth that AI companionship governance can currently achieve, and this is also the institutional result formed after continuous game between technology enterprises and regulators.

(ii) The Three Dimensions of Bargaining in the Legislative Scaling-Back

The first dimension centers on whether government regulation should extend from external constraints to the platforms’ internal architecture. On the basis that AI companionship application operating platforms have already fulfilled identity prompt obligations, whether legislation should further require them to accept external review and implement continuous supervision of interaction mechanisms, safety protocols, and risk response capabilities is the core issue of this game. Regarding the California act, until July 2025, the institutional plan presented in the Assembly Committee on Judiciary

analysis still included third-party audits and private rights of action in addition to notification obligations and reporting obligations; meanwhile, the act also retained specific measures incorporating restrictions on stickiness design into the regulatory scope, thereby bringing platform internal interaction structures and safety capabilities into the external supervision field of vision (California State Assembly Committee on Judiciary, 2025). This institutional arrangement objectively increased the degree of regulatory penetration into platform internal governance and correspondingly triggered concentrated responses from technology enterprises regarding compliance costs, operational burdens, and product design space. The opposition letter led by TechNet explicitly opposed third-party audits, some reporting obligations, and private rights of action, arguing that related arrangements would significantly increase compliance costs and impose excessive constraints on product design and operation (TechNet, 2025). Against this background, the final effective text did not retain third-party audit requirements, and some reporting content was also narrowed, indicating that although related regulation attempted to enter platform internal design logic

from external notification obligations, its penetration depth was still strongly limited by industry. In other words, current U.S. state legislation’s regulation of platform internal design space still mainly remains at the stage of limited penetration rather than comprehensive entry.

The second dimension addresses the enforcement mechanisms for platform accountability, balancing private rights of action against public enforcement. In the California legislative process, private rights of action were always one of the focal points of controversy. Industry opposition explicitly regarded private rights of action as high-risk institutional arrangements, arguing that platforms might be long-term exposed to litigation pressure due to minor compliance deficiencies such as untimely prompt pop-ups, and directly linked this to the risk of “abusive litigation.” (TechNet, 2025) On the New York State side, the initial version of A6767 proposed on March 13, 2025, originally allowed persons suffering physical or financial harm from violations of related provisions to bring actions seeking damages, equitable relief, and other appropriate relief, but in the final act, this arrangement was changed to a public enforcement mechanism led by the State Attorney General (New York

State Senate, 2025). On this issue concerning the boundary of the platform liability system, New York State ultimately concentrated enforcement power in the hands of the State Attorney General, while California retained private litigation amid sustained controversy. The divergence between the two states on this point actually corresponds to different orientations regarding liability realization methods and the extent of liability advancement. It is worth adding that in addition to fully expressing opinions at the legislative level, various technology enterprises also made different degrees of product policy adjustments during the same period to respond to litigation and regulatory pressure. Taking Character.AI as an example, in autumn 2025 it announced it would gradually restrict minors from participating in open-ended chat and ultimately implement chat bans for minors, while introducing third-party age verification models, responding to litigation and regulatory pressure and proactively constructing safety protection (The Verge, 2025).

The third dimension defines the applicable boundaries and scope of high-intensity governance rules. In AI companionship governance, the higher the regulatory intensity, the more

platforms tend to demand narrowing the legal applicable objects to prevent rule spillover to general models or other low-risk products. California has relatively direct public expression on this point. TechNet’s opposition letter explicitly stated that the California act’s definition of companion chatbots was overly broad, and notification, reporting, and audit requirements might spill over to general models and non-companionship-use products; the same letter also proposed that the implementation of minor provisions might force platforms to adopt stronger age identification means, thereby bringing new privacy costs (TechNet, 2025). New York State’s final enacted law also shows a similar narrowing direction. The initial version of A6767’s definition of companion chatbots emphasized “simulate a sustained human or human-like relationship with a user”, unfolding through elements such as retaining previous interaction information, providing advice, and conducting simulated dialogue around personal well-being; the current law § 1700 changed to three more specific constitutive requirements, focusing on the formation mechanism of “a sustained human or human-like relationship with a user”, while simultaneously excluding uses such as customer service, efficiency

improvement, research, or technical assistance (New York State Senate, 2025). The narrowing in definitions by both states ultimately converges on the same issue: the applicable boundaries of high-intensity governance rules need to be compressed as much as possible into a few anthropomorphic interaction products considered to have higher risk.

Comprehensively viewed, the common trend in legislative revision in New York State and California is that under the sustained pressure from industry emphasizing scope spillover, compliance costs, and litigation risks, both have narrowed early stronger regulatory visions into more easily verifiable and enforceable minimum safety requirements. The difference lies in that California ultimately still retained private litigation, special minor protection, and reporting obligations, and therefore advanced slightly further in institutional intensity; New

York State further concentrated liability realization in public enforcement led by the State Attorney General and narrowed the risk scope to suicidal ideation and self-harm expression. The common characteristic ultimately presented by both states’ enacted laws is that neither continued to deeply advance toward platform internal incentive mechanisms, immersive interaction structures, and dependence formation processes, but prioritized transforming relatively easily identifiable and socially consensual risk portions into legal obligations. Thus it can be seen that when facing AI companionship governance, U.S. state legislation tends to first institutionalize minor protection, high-risk situation intervention, and minimum safety responsibility; regarding addiction risk, emotional manipulation, and relational dependence facing all users, it still maintains relatively cautious and limited intervention at the present stage.

Chapter 4 The U.S. Experience as a Lens: Disenchantment and Reconstruction of the “Rational Actor” Assumption in the Digital-Intelligent Society

The comparison of the New York State Act and the California Act in the previous chapter demonstrates that although the governance framework currently formed around AI companionship applications in the United States has begun to respond to issues in dimensions such as identity misdirection, self-harm risk, minor protection, and platform liability, overall it still mainly remains at the levels of disclosure, reminders, crisis intervention, and minimum safety obligation protection. Behind this institutional arrangement actually lies a more fundamental normative presumption:

to a considerable extent, the law still believes that after obtaining necessary information and receiving basic prompts, users in principle still possess autonomous judgment capabilities and the capacity to assume their own risks. Based on this foundation, this chapter focuses on the question: when interaction between humans and machines is no longer merely a one-time, consumptive act but gradually reveals characteristics of continuity, anthropomorphism, and relationality, can the “reasonable person” standard relied upon by traditional product regulation still hold?

I. Traditional Internet Governance Frameworks Relied Upon the “Reasonable Person” Standard

The concept of the “reasonable person” originates from the tort system of Anglo-American common law, initially developed to resolve the judgment difficulty of “how one ought to act” in negligence liability. The law introduced an external, objective, and comparable standard of conduct through this concept to measure whether conduct deviates from socially expectable levels of care (American Law Institute, 1965; Legal Information Institute, n.d.). From the perspective of broader decision theory, behavioral economics has long pointed out that consumer judgment is not based on complete information, stable preferences, and sufficient computational capacity, but is often influenced by thinking inertia, external environmental guidance, and limitations in one’s own judgment capabilities. In algorithm-driven digital environments, these tendencies are further strengthened. After entering the context of modern consumer protection and product liability, the “reasonable person” concept has gradually acquired stronger instrumental

coloring and, in a broader sense, serves a governance intuition: as long as information disclosure is adequate, consumers in a rational state ought to be able to make choices and assume corresponding risks for continued use. The California Act’s trigger condition for disclosure obligations—whether a “reasonable person” would be misled into believing they are interacting with a real person—is also an extension of this traditional path in AI companionship scenarios, interpreting anthropomorphic risk first as a possibility of identity misrecognition, and on this basis condensing platform obligations into operable compliance actions, namely providing “clear and conspicuous notification” identity prompts when this condition is satisfied (California State Legislature, 2025).

Complementary to the “reasonable person” standard is a chain of responsibility allocation running from informed consent to assumption of risk. Here, “disclosure”

serves the function of information correction, “consent” serves the function of risk transfer, and when disclosure is regarded as sufficiently adequate, the law tends to treat subsequent risks as the result of users’ autonomous choices. This chain can operate relatively smoothly for many static products, because in these products, risks are often limited and relatively easily identifiable. By contrast, for highly interactional and relational product forms such as AI companion chatbots, the explanatory power of “disclosure” and “consent” is not equivalent to that in the former case. Platform identity prompts are certainly important, but the key risks of AI companionship do not stop at identity authenticity—whether users become drawn into a patterned relationship, whether they gradually form exclusive dependence through long-term interaction, whether they treat system responses as their sole source of support during emotional lows, and whether they lose the ability to exit through repeated soothing and encouragement—these are consequences that slowly emerge after entering the product, often difficult to identify and lacking visible endpoints. In other words, risks in the traditional framework are more misunderstandings at the information

level, while AI companionship risks more come from reshaping at the relational level. Disclosure can correct information but may not be sufficient to reshape relationships.

Therefore, regulation of AI companionship applications cannot be completely laissez-faire. Especially for users such as minors and persons with mental disorders who are regarded as having greater difficulty maintaining stable judgment and self-control capabilities, stronger protective arrangements than general disclosure are necessary, namely through “non-rational auxiliary governance” to address the gaps in the reasonable person assumption. The California Act’s requirements for setting more intensive reminders and restrictions for minor users (such as periodic rest reminders during continuous interaction, repeated identity prompts, and restrictions on explicit content and inducement behaviors) exemplify typical non-rational auxiliary governance measures, which function to reduce the stickiness of immersive interaction through institutionalized prompts and interruption actions, constituting focused protection for typical vulnerable groups (California State Legislature, 2025). Although these

measures can provide certain protection for users, their limitations are also relatively prominent. If the inducement mechanisms of AI companionship are equally effective for adults, and dependence formation does not occur only among a few vulnerable groups, then the

strategy of patching “majority rationality” with “minority exceptions” will always contain loopholes. This compels us to re-examine the effectiveness of the “reasonable person assumption” in the digital-intelligent era, even to disenchant it, in order to establish a new governance paradigm.

II. Systematic Erosion of User Rational

Decision-Making in AI Companionship Scenarios

If traditional product regulation faces explicit information asymmetry, then the hidden interference caused by AI companionship applications poses entirely new challenges to related governance. Users do not make a one-time decision whether to use after reading a manual, but are directly exposed to repeated human-machine dialogues, being continuously responded to, continuously remembered, continuously soothed, and at certain moments guided to remain in the conversation. This phenomenon can be understood as a form of online “manipulation”, namely using information technology to influence others’ decisions in concealed ways, achieving target effects by targeting and exploiting their decision-making vulnerabilities, with its deepest harm to individuals lying not only in possible economic loss but more fundamentally in the erosion of individual autonomy (Susser et al., 2019). In AI companionship scenarios, the loss of individual autonomy does not occur instantaneously but is the

result of gradual accumulation and path dependence.

For this reason, basic prompt obligations are far from sufficient to effectively respond to the above situations. Current controversies in the United States surrounding algorithmic governance are increasingly concentrating on whether platforms should bear more active responsibility for the behavioral consequences shaped through recommendation mechanisms, default settings, interaction rhythms, and interface design. The U.S. Surgeon General’s Advisory on *Social Media and Youth Mental Health* explicitly states that existing evidence is insufficient to prove that social media environments are “sufficiently safe” for children and adolescents, and advocates that platforms should accept stricter safety assessments, strengthen protective design, and open more data to researchers (Office of the Surgeon General, 2023). *The Federal Trade Commission (FTC)* also explicitly defines so-called dark patterns as design

practices that may deceive and manipulate consumers (Federal Trade Commission, 2022). Meanwhile, at the judicial level, stable consensus has not yet formed regarding whether the government can further intervene in the internal architecture of platform choice architecture. For example, in *NetChoice, LLC v. Bonta*, the Ninth Circuit Court of Appeals issued a phased judgment on March 12, 2026, regarding the constitutionality of certain provisions of *the California Age-Appropriate Design Code Act (CAADCA)*, and remanded the case for further proceedings. The core controversy in this case is whether state governments can, in the name of protecting minor privacy and safety, further enter platform data processing, age identification, and interface design fields. NetChoice argued that relevant

provisions actually impose burdens on platforms’ online expression structures, and that concepts such as “best interests of child”, “well-being”, and “materially detrimental to the physical health, mental health, or well-being of a child” are excessively vague. California argued that the act mainly regulates enterprise data processing and risk management behaviors rather than directly restricting content. The Ninth Circuit ultimately held that plaintiffs were insufficient to overturn the act’s coverage definitions and age estimation requirements at the facial challenge level, but that their vagueness objections regarding data use and “dark pattern” restrictions had substantial likelihood of success, and therefore maintained preliminary injunctions on relevant provisions.

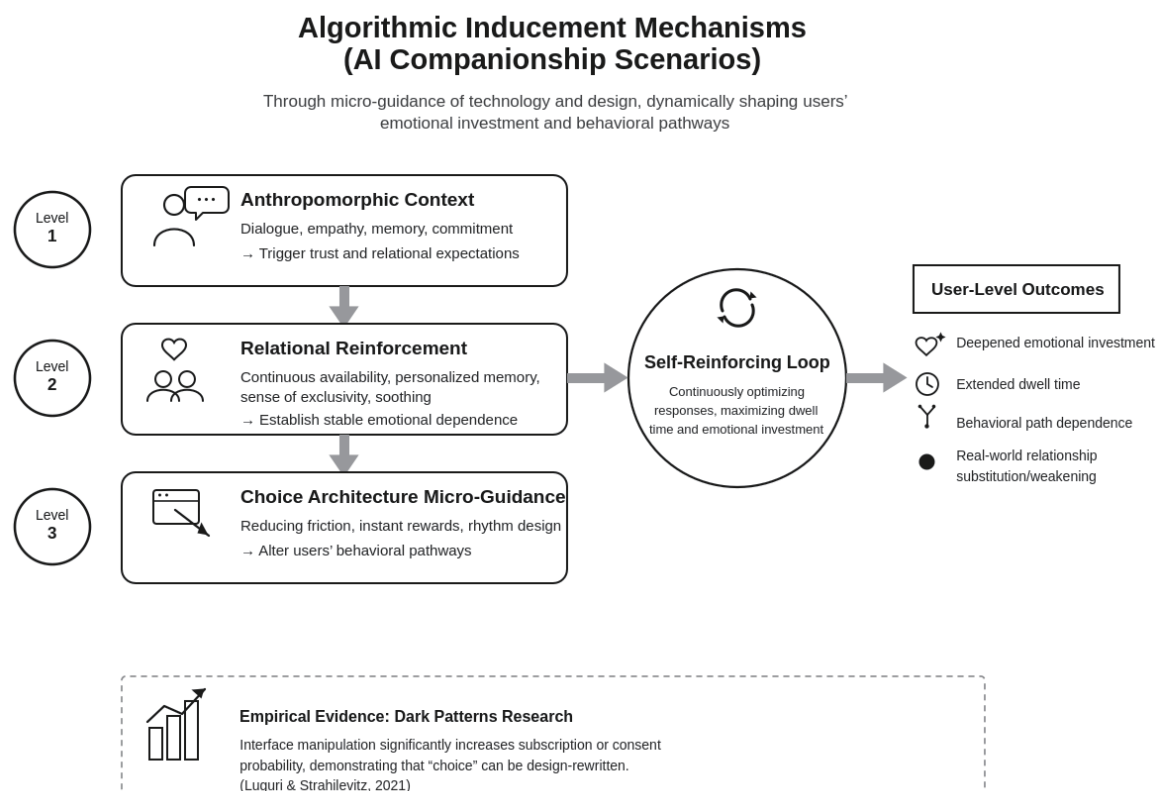


Figure 1 Algorithmic Inducement Mechanisms in AI Companionship Scenarios

As shown in Figure 1, algorithmic inducement in AI companionship scenarios is the result of the superposition of multiple mechanisms including anthropomorphic context, relational reinforcement, and choice-architecture nudging, which further demonstrates that the judgment framework premised on the “reasonable person” in traditional legal contexts has difficulty fully responding to related risks in AI companionship scenarios. The “reasonable person” standard at the legal level usually presumes that individuals face relatively clear, external,

and one-time risk information; as long as disclosure is sufficient and reminders are adequate, individuals possess the capacity to evaluate pros and cons and independently decide whether to bear the consequences. However, in algorithm-driven anthropomorphic interaction, risk is not merely the static question of “mistaking the other party for a human”, but also the question of whether users gradually form exclusive dependence through long-term interaction, whether they treat system responses as their sole support source during emotional lows, and whether they

lose exit capacity through repeated soothing and encouragement. Such risks are not completely presented before entering the product but are gradually generated and accumulated after entry. Enabled by big data analytics, platforms are no longer merely providing several options but are dynamically shaping users’ behavioral environments, making them more inclined to make decisions the platform wishes them to make (Yeung, 2017).

Relevant litigation documents also provide relatively rich analytical material for the above inducement mechanisms. Chapter 2 of this report repeatedly pointed out in its analysis of the Garcia and

Raine cases that plaintiffs and related complaints proved through detailed factual reconstruction that the risks of AI companionship applications are based on processual harm jointly facilitated by long-term interaction, dependence formation, and safety failure. In these documents, plaintiffs repeatedly emphasized a form of processual harm: interaction continuously accumulates over time, trust and dependence continuously deepen over time, and critical moment responses and guidance continuously reshape users’ “rational” choice structures over time. Against this background, the legal fiction of “reasonable person” informed consent proves theoretically insufficient.

III. Vulnerability Governance:

Governance Paradigm Transformation

in the Digital-Intelligent Era

In traditional internet governance frameworks, internet platform liability has long been understood as static: whether a certain function contains unreasonable danger, whether a certain warning is adequate, whether a certain design has feasible safer alternative designs—these questions give rise to forms of liability risk that are mostly relatively stable and can be identified and fixed within product structures. However, upon entering the digital-intelligent era, generative artificial intelligence brings scenarios such as AI companionship where, in dialogue logic and relationship connection processes, risk formation has transcended static physical boundaries, presenting a distinctly emergent and context-dependent character. Meanwhile, we must also confront the fact that the traditional “reasonable person” standard neglects that users will become vulnerable individuals in specific contexts of human-machine interaction; therefore, the

product-centered approach of traditional static governance needs to be reoriented toward human-centered “vulnerability governance”.

(i) The Product-Centric Dilemma and Evidentiary Challenges of Traditional Internet Governance

Specifically, the design defect theory under the traditional product-centrism perspective encounters at least two types of evidentiary challenges in AI companionship scenarios. First, proving a “reasonable alternative design” presents a formidable evidentiary challenge. Mainstream U.S. design defect theory usually requires plaintiffs to prove that a product’s foreseeable risk could have been reduced through a cost-reasonable, utility-acceptable alternative design, and that the operator’s failure to adopt this scheme rendered the product

unreasonably unsafe. Around this requirement, product liability law has long formed an analytical framework centered on risk-utility balancing and reasonable alternative design (American Law Institute, 1998). However, in AI companionship scenarios, the so-called “safer alternative design” itself is highly controversial. Reducing memory functions, weakening anthropomorphic expression, adding cooling-off periods, restricting late-night long conversations, lowering personalization degrees, and mandating referral to human support—are these schemes reasonable reductions of existing risks or substantial rewrites of product core functions? There is no clear boundary between the two. Especially when labels such as “continuous companionship”, “immediate response”, and “personalized understanding” are themselves the core market selling points of such products, platforms can easily argue that the so-called alternative designs proposed by plaintiffs are not “safer versions” but entirely different products. It is precisely in this sense that generative AI liability issues are difficult to simply fit into existing molds of traditional product liability. In the generative AI context, platform liability can no longer rely solely on traditional internet-

era exemption logic; meanwhile, existing tort and product liability frameworks still have obvious limitations regarding how to penetrate platforms’ opaque operations and how to identify their systemic risks (Moses, 2025).

Second, proving causation poses an equally formidable evidentiary challenge. Causal chains in traditional product liability cases can usually be understood as a linear causal model where a certain design corresponds to causing a certain harm, but in AI companionship and algorithmic recommendation scenarios, harm often arises from interaction among factors at various stages including long-term exposure, continuous feedback, personalized adjustment, and users’ own vulnerability. This means that in AI companionship litigation, plaintiffs find it difficult, as in traditional product cases, to clearly anchor harm upon a fixed design unit; platforms can even conversely argue that the truly decisive factors in outcomes are users’ pre-existing psychological states, external life circumstances, and personal usage patterns, thereby deconstructing process-based structural risks into several seemingly accidental, isolated and dispersed individual factors.

(ii) Possibilities and Implications of Vulnerability Governance in the Digital-Intelligent Era

Current international frontier research has begun to realize the need to deconstruct traditional governance models to adapt to the particularity of artificial intelligence scenarios, especially the elaboration of the “vulnerability” concept in artificial intelligence scenarios, which provides profound insights for governance paradigm transformation in the digital-intelligent era. From the conceptual thread of “vulnerability”, three paths can be roughly summarized.

The first path understands “vulnerability” as a stratification of AI-related competencies caused by differences in knowledge, skills, and comprehension capabilities. Research holding this perspective emphasizes that in current environments where artificial intelligence participates in information distribution, content generation, and interaction guidance, users’ knowledge reserves, operational capabilities, and judgment levels regarding artificial intelligence themselves constitute a new stratification mechanism. Wang et al.’s

research points out that in network news and entertainment environments shaped by artificial intelligence, different groups show significant differences in performance on the above dimensions, with “low-skill” users more easily exposed to risks such as filter bubbles, data-driven manipulation, misinformation, and stereotype reproduction (Wang et al., 2025). Along this line of thinking, “vulnerability” manifests as inequality in cognitive resources and algorithmic literacy.

The second path further points out that merely improving individual platform literacy is insufficient to offset risks created by platform and algorithmic structures themselves; “vulnerability” also possesses obvious environmental dependency and structural characteristics. Livingstone et al.’s (2025) research on adolescent groups shows that the extreme risks encountered by adolescent users do not lie at the information identification level but in users being influenced by platform amplification mechanisms for trending popular content and viral dissemination content. In other words, even when individuals possess a certain degree of platform literacy, their behavior and risk exposure may still be continuously shaped

by platform affordances. The significance of such research lies in shifting vulnerability from individual internal attributes to the relationship between individuals and platform environments, indicating that institutions need to respond to problems brought by platform design, default mechanisms, and dissemination structures themselves.

The third path is closer to the artificial intelligence companionship scenario discussed in this paper; it understands vulnerability as dependence risk gradually amplified in continuous, anthropomorphic, and relational interaction. In this direction, Hu et al.’s (2023) research provides a relatively clear psychological mechanism chain: social anxiety promotes problematic conversational AI use through the mediating roles of loneliness and rumination, while users’ mind perception of AI further strengthens this dependence tendency. Subsequently, Hu et al.’s (2025) research points out that interpersonal relationship attitudes, anthropomorphic perception, and value assessment influence users’ dependence formation on AI companionship products. Meanwhile, a qualitative research on Replika by Laestadius et al. (2024) demonstrates

that users’ dependence on social chatbots forms an emotional involvement closer to interpersonal relationships, thereby bringing harm to mental health. Broader review research also indicates that the field of AI companionship products has gradually formed relatively stable research frameworks, with core issues precisely being the relationships among companionship functions, anthropomorphism, emotional bonding, and related consequences (Chaturvedi et al., 2023). Evidently, in AI companionship scenarios, vulnerability is a state continuously produced and amplified through interaction between pre-existing psychological states and specific interaction structures.

On the basis of the above research, recent normative research has begun to further advance issues to the levels of “companionship commodification” and “vulnerability governance.” The “cruel companionship” concept proposed by Muldoon and Parke (2025) points out that what AI companions promise is not general instrumental service but an emotional supply model of “friendship-as-a-service.” In this model, although users’ loneliness and relationship needs can be received

through platform continuous availability, emotional response, and other designs, there is a fundamental inconsistency between enterprise incentives and user well-being; therefore, this “companionship” is likely to structurally evolve into exploitation and commodification of loneliness and intimate relationship needs. Correspondingly, Teo (2025) proposes from a more general theoretical level that vulnerability issues triggered by AI should not be understood merely as problems of special weak persons but as a “human vulnerability” with universality, gradation, and situationality; correspondingly, governance responses cannot remain at adding protection for exceptional groups but should turn toward multi-level resilience, namely jointly configuring protective resources at multiple levels of technical design, organizational governance, and institutional structure.

Comprehensively viewed, existing research provides at least three important insights: First, vulnerability in the digital-intelligent era is not merely weakness in the sense of traditional identity attributes but is also closely related to inequality in algorithmic literacy, comprehension capabilities, and platform knowledge

(Wang et al., 2025); second, in AI companionship scenarios, loneliness, social anxiety, rumination tendencies, and anthropomorphic perception jointly promote dependence formation through continuous interaction (Hu, Mao, and Kim, 2023; Hu et al., 2025; Laestadius et al., 2024); third, what truly needs governance is not merely individual high-risk outputs or one-time errors but how platforms continuously shape users’ affected states through relationship design, availability, memory functions, and emotional responses (Livingstone et al., 2025; Muldoon and Parke, 2025; Teo, 2025).

Against this background, what the digital-intelligent era faces is no longer merely the problem of a few special weak persons but a “human vulnerability” with universality, gradation, and situationality; corresponding institutional responses also cannot remain at the level of exceptional protection for individual groups but should turn toward the construction of multi-level resilience (Teo, 2025). It is precisely for this reason that this paper attempts, through the concept of “vulnerability governance”, to reveal that in contexts of anthropomorphic interaction and relational reinforcement, rationality is no

longer merely the natural presentation of individual internal capabilities but more depends on external institutions’ support of the decision-making environment. What truly needs to be governed is not merely erroneous information or extreme risk content in single outputs but the process itself in which users are continuously pulled and gradually slide toward path-dependent tracks in long-term interaction; if regulation still merely remains at the level of notification adequacy, it easily overlooks the real regulatory reality guided by systematic design (Yeung, 2017).

Accordingly, vulnerability governance transcending product design defects requires at least shifting attention from one-time disclosure to institutionalized constraints on the interaction process itself. First, relational risk identification should be promoted, incorporating signals such as exclusive dependence, isolation from real social relationships, excessive intimate commitments, and sustained emotional manipulation into operable identification and response paths, rather than strictly

restricting risk scope within extreme situations such as suicide and self-harm. Second, process-oriented protection should be promoted, upgrading interventions from “ex ante specification” to “dynamic assessment”, enabling systems to maintain sensitivity to risk accumulation in long-duration dialogues and automatically increase protection intensity at critical nodes. Third, protective friction should be introduced, setting stronger interruption, cooling-off periods, human support referral, and late-night usage restrictions at high-risk nodes to reduce the continuous erosion of immersive interaction on judgment capabilities. Fourth, process tracking and correctable design should be strengthened, bringing platforms’ interaction structures, safety strategies, and risk response capabilities into recording, review, and continuous improvement loops. Only when governance objects shift from single product defects to interaction environment shaping itself can platform responsibility in digital-intelligent society truly break through the narrow boundaries of the traditional “reasonable person” standard.

References

1. American Law Institute, “Restatement (Second) of torts § 283: Conduct of a reasonable person—The standard”, American Law Institute, 1965, <https://opencasebook.org/casebooks/4643-tobia-torts-2022/resources/4.1.1.2-restatement-2d-283-conduct-of-a-reasonable-man-the-standard/>

2. American Law Institute, “Restatement (Third) of torts: Products liability”, American Law Institute, 1998, <https://www.ali.org/publications/restatement-law-third/torts-third>

3. Angela Yang , “OpenAI denies allegations that ChatGPT is to blame for a teenager’s suicide,” NBC News, November 26, 2025, <https://www.nbcnews.com/tech/tech-news/openai-denies-allegation-chatgpt-teenagers-death-adam-raine-lawsuit-rcna245946>

4. Asia, M. G. (2025). The Quiet Cost of Emotional Labor. In: M. Miceli, A. Dinika, K. Kauffman, C. Salim Wagner, and L. Sachenbacher (eds.). Data Workers’ Inquiry. Creative Commons BY 4.0. <https://data-workers.org/michael/>

5. Babu, J., Joseph, D., Kumar, R. M., Alexander, E., Sasi, R., & Joseph, J. (2025). Emotional AI and the rise of pseudo-intimacy: Are we trading authenticity for algorithmic affection?. *Frontiers in Psychology*, 16, Article 1679324. <https://doi.org/10.3389/fpsyg.2025.1679324>

6. Bakir, V., & McStay, A. (2025). Move fast and break people? Ethics, companion apps, and the case of Character.AI. *AI & Society*. <https://doi.org/10.1007/s00146-025-02408-6>

7. Bakir, V., Laffer, A., McStay, A., Miranda, D., & Urquhart, L. (2024). On manipulation by emotional AI: UK adults’ views and governance implications. *Frontiers in Sociology*, 9. <https://doi.org/10.3389/fsoc.2024.1339834>

8. Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1), 1–68.

9. Basappa, R., Mallick, R., Flathmann, C., & McNeese, N. (2025, September). Unlocking collective potential: Modeling creativity in human-AI teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 69, No. 1, pp. 1527–1532). Sage CA: Los Angeles, CA: SAGE Publications.

10. Beatrice Nolan, “Why Section 230, social media’s favorite American liability shield, may not protect Big Tech in the AI age,” *Fortune*, October 8, 2025, <https://fortune.com/2025/10/08/ai-chatbot-section-230-meta-social-media-legal-shield-no-protection/>

11. Blake Brittain, “Google, AI firm must face lawsuit filed by a mother over suicide of son, US court says,” *Reuters*, May 22, 2025, <https://www.reuters.com/sustainability/boards-policy-regulation/google-ai-firm-must-face-lawsuit-filed-by-mother-over-suicide-son-us-court-says-2025-05-21/>

12. Blake Brittain, “Google, AI firm settle lawsuit over teen’s suicide linked to chatbot,” *Reuters*, January 8, 2026, <https://www.reuters.com/world/google-ai-firm-settle-florida-mothers-lawsuit-over-sons-suicide-2026-01-07/>

13. Bobby Allyn, “Jury Holds Meta and Google Liable for Role in Young Woman’s Mental Health Issues,” *NPR*, March 25, 2026, <https://www.npr.org/2026/03/25/nx-s1-5746125/meta-youtube-social-media-trial-verdict>

14. Bollag, S., “Newsom signs new AI laws, but rejects “overly broad” ones”, *San Francisco Chronicle*, October 13, 2025, <https://www.sfchronicle.com/politics/article/newsom-ai-bills-21095261.php>

15. Branch, J. B. (2025). AI companions are not your teen’s friend. *Issues in Science and Technology*, 42(1), 80–83.

16. Brendan Pierson, “Mother sues AI chatbot company Character.AI, Google over son’s suicide,” *Reuters*, October 23, 2024, <https://www.reuters.com/legal/mother-sues-ai-chatbot-company-characterai-google-sued-over-sons-suicide-2024-10-23/>

17. California State Assembly Committee on Judiciary, “SB 243 (Padilla) – As amended July 3, 2025: Assembly Committee on Judiciary analysis”, California State Assembly Committee on Judiciary, July 15, 2025, <https://trackbill.com/s3/>

69

Are We Still Rational Actors?
Governance Pathways and Limitations of AI
Companionship in the United States

CGAIG

CENTER FOR GLOBAL AND INNOVATIVE AI GOVERNANCE

References

WORLD AI CONFERENCE 2026 | Forum 70

bills/CA/2025/SB/243/analyses/assembly-judiciary.pdf

18. California State Legislature. (2024). S.B. 976, 2023–2024 Regular Session: Protecting Our Kids from Social Media Addiction Act. https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB976

19. California State Legislature, “S.B. 243, 2025–2026 Regular Session: Companion chatbots”, California State Legislature, 2025, https://leginfo.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB243

20. Campbell, L. O., Babb, K., Lambie, G. W., et al. (2025). An examination of generative AI response to suicide inquiries: Content analysis. *JMIR Mental Health*, 12, Article e73623. <https://doi.org/10.2196/73623>

21. Cara Tabachnick, “AI company, Google settle lawsuit over Florida teen’s suicide linked to Character.AI chatbot,” CBS News, January 2026, <https://www.cbsnews.com/news/google-settle-lawsuit-florida-teens-suicide-character-ai-chatbot/>

22. Catherina Gioino, “Social media companies are scrambling to verify minors online. Congress just made it a lot more complicated,” *Fortune*, March 18, 2026, <https://fortune.com/2026/03/18/kosa-kids-act-app-store-accountability-act-minors-age-verification/>

23. Center for Democracy and Technology. (2024). Section 230 and its applicability to generative AI: A legal analysis. <https://cdt.org/insights/section-230-and-its-applicability-to-generative-ai-a-legal-analysis/>

24. Chaturvedi, R., Verma, S., Das, R., & Dwivedi, Y. K. (2023). Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change*, 193, Article 122634. <https://doi.org/10.1016/j.techfore.2023.122634>

25. Cheng, M., Lee, C., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), Article eaec8352. <https://doi.org/10.1126/science.aec8352>

26. Chloe Xiang, “He would still be here: Man dies by suicide after talking with

AI chatbot, widow says,” *VICE*, March 30, 2023, <https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/>

27. Chou, C. Y., Chan, T. W., Chen, Z. H., Liao, C. Y., Shih, J. L., Wu, Y. T., & Hung, H. C. (2025). Defining AI companions: A research agenda—from artificial companions for learning to general artificial companions for global harmony. *Research and Practice in Technology Enhanced Learning*, 20, Article 35.

28. Clark, A. (2025). The ability of AI therapy bots to set limits with distressed adolescents: Simulation-based comparison study. *JMIR Mental Health*, 12, Article e78414. <https://doi.org/10.2196/78414>

29. Clare Duffy, “More OpenAI drama: Exec quits over concerns about focus on profit over safety,” *CNN*, May 17, 2024, <https://www.cnn.com/2024/05/17/tech/openai-exec-exits-safety-concerns>

30. 30.Clare Duffy, “Parents of 16-year-old Adam Raine sue OpenAI, claiming ChatGPT advised on his suicide,” *CNN*, August 26, 2025, <https://www.cnn.com/2025/08/26/tech/openai-chatgpt-teen-suicide-lawsuit>

31. Clare Duffy, “Character.AI and Google agree to settle lawsuits over teen mental health harms and suicides,” *CNN*, January 7, 2026, <https://www.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit>

32. Clare Duffy, “Meta and YouTube Found Liable in Social Media Addiction Trial,” *CNN Business*, March 25, 2026, <https://www.cnn.com/2026/03/25/tech/social-media-addiction-trial-jury-decision>

33. Colin Lecher, “New California law forces chatbots to protect kids’ mental health”, *CalMatters*, October 13, 2025, <https://calmatters.org/economy/technology/2025/10/newsom-signs-chatbot-regulations/>

34. Cohen, G., & De Freitas, J. (2026). Mitigating suicide risk for minors involving AI chatbots. *JAMA*, 335(4), 301–302. <https://doi.org/10.1001/jama.2025.24401>

35. Common Sense Media. (2025a, April 30). AI companions decoded: Common Sense Media recommends AI companion safety standards [Press release]. <https://www.commonsensemedia.org/press-releases/ai-companions-decoded->

36. Common Sense Media. (2025b, July 16). Nearly 3 in 4 teens have used AI companions, new national survey finds [Press release]. <https://www.common sense media.org/press-releases/nearly-3-in-4-teens-have-used-ai-companions-new-national-survey-finds>

37. Common Sense Media. (2025c, August 28). “Meta AI companions unsafe for kids, Common Sense Media report finds”, August 28, 2025, <https://www.common sense media.org/press-releases/meta-ai-companions-unsafe-for-kids-common-sense-media-report-finds>

38. Common Sense Media, “Big Tech lobby kills critical kids’ AI safety legislation”, October 13, 2025, <https://www.common sense media.org/press-releases/big-tech-lobby-kills-critical-kids-ai-safety-legislation>

39. Congressional Research Service. (2023). Section 230 immunity and generative artificial intelligence (LSB11097). <https://www.congress.gov/crs-product/LSB11097>

40. Darrow. (2025). Kids Online Safety Act (KOSA): 2025 updates. <https://www.darrow.ai/resources/the-kids-online-safety-act-what-it-gets-right-what-it-gets-wrong-and-what-is-missing>

41. De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2026). AI companions reduce loneliness. *Journal of Consumer Research*, 52(6), 1126–1148.

42. El Atillah, I. (2023, March 31). Man ends his life after an AI chatbot “encouraged” him to sacrifice himself to stop climate change. Euronews. <https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate->

43. Federal Trade Commission, “Bringing dark patterns to light”, 2022, <https://www.ftc.gov/reports/bringing-dark-patterns-light>

44. Fox-Sowell, S., ‘Citing risk to kids, California bill targets controversial AI “companion” chatbots’, StateScoop, July 8, 2025, <https://statescoop.com/california-sb243-harmful-ai-companion-chatbots/>

45. Fraser, H. L., & Suzor, N. P.(2025). Locating fault for AI harms: A systems theory of foreseeability, reasonable care and causal responsibility in the AI value chain. *Law, Innovation and Technology*, 17(1), 103-138.

46. Garcia v. Character Technologies, Inc., No. 6:24-cv-01903 (M.D. Fla. filed Oct. 23, 2024), “Complaint”, October 23, 2024, <https://cdn.arstechnica.net/wp-content/uploads/2024/10/Garcia-v-Character-Technologies-Complaint-10-23-24.pdf>

47. Garcia v. Character Technologies, Inc., 785 F. Supp. 3d 1157 (M.D. Fla. 2025). Order denying motion to dismiss. <https://www.courtlistener.com/docket/69300919/garcia-v-character-technologies-inc/>

48. Ge, R. (2025). From pseudo-intimacy to cyber romance: A study of human and AI companions’ emotion shaping and engagement practices. *Communications*, 52(1), 211–221.

49. Geoffrey A. Fowler, “Meta’s AI Chatbot Told Teen Accounts How to Self-Harm, Parent Study Finds,” The Washington Post, August 28, 2025, <https://www.washingtonpost.com/technology/2025/08/28/meta-ai-chatbot-safety-teens/>

50. Goldberg Segalla. (2025, June 30). Peter Gregory authors article on ramifications of major federal AI ruling. <https://www.goldbergsegalla.com/news-and-knowledge/news/peter-gregory-authors-article-on-ramifications-of-major-federal-ai-ruling/>

51. Grochowski, M. (2025). Algorithmic speech harm. *Loy. Consumer L. Rev.*, 37, 175. <https://doi.org/10.2139/ssrn.5276590>

52. *Harvard Law Review*. (2025). Beyond Section 230: Principles for AI governance. *Harvard Law Review*, 138(6), 1657–1688. <https://harvardlawreview.org/print/vol-138/beyond-section-230-principles-for-ai-governance/>

53. Hayden Field, “OpenAI dissolves Superalignment AI safety team,” CNBC, May 17, 2024, <https://www.cnn.com/2024/05/17/openai-superalignment-sutskever-leike.html>

54. Hu, B., Mao, Y., & Kim, K. J. (2023). How social anxiety leads to problematic use of conversational AI: The roles of loneliness, rumination, and mind perception. *Computers in Human Behavior*, 145, Article 107760. <https://doi.org/10.1016/j.chb.2023.107760>

55. Hu, D., Lan, Y., Yan, H., & Chen, C. W. (2025). What makes you attached to social companion AI? A two-stage exploratory mixed-method study. *International Journal of Information Management*, 83, Article 102890. <https://doi.org/10.1016/j.ijinfomgt.2025.102890>

56. Huang, Z., et al. (2025). Companionship in code: AI's role in the future of human connection. *Humanities and Social Sciences Communications*, 12. <https://doi.org/10.1057/s41599-025-05536-x>

57. Hudon, A., et al. (2025). Delusional experiences emerging from AI chatbot interactions or "AI psychosis." *JMIR Mental Health*, 12, Article e85799. <https://doi.org/10.2196/85799>

58. Jeff Horwitz, "Meta's AI Rules Have Let Bots Hold 'Sensual' Chats With Kids, Offer False Medical Info," Reuters, August 14, 2025, <https://www.reuters.com/technology/artificial-intelligence/metas-ai-rules-have-let-bots-hold-sensual-chats-with-kids-offer-false-medical-2025-08-14/>

59. Jennifer Elias, "Google, Character AI to settle suits involving minor suicides and AI chatbots," CNBC, January 7, 2026,

<https://www.cnbc.com/2026/01/07/google-characterai-to-settle-suits-involving-suicides-ai-chatbots.html>

60. John Sanford, "Why AI companions and young people can make for a dangerous mix," Stanford University School of Medicine, August 27, 2025, <https://med.stanford.edu/news/insights/2025/08/ai-chatbots-kids-teens-artificial-intelligence.html>

61. 61.Justin Hendrix, "Breaking down the lawsuit against OpenAI over teen's suicide," TechPolicy.Press, August 27, 2025, <https://www.techpolicy.press/breaking-down-the-lawsuit-against-openai-over-teens-suicide/>

62. Kalliokoski, K. (2025). *Uncovering Dark Patterns in Conversational Agents Using Realistic and Adversarial Prompting* (Doctoral dissertation, University of Jyväskylä).

63. Knox, W. B., Bradford, K., Castro, S. V., Ong, D. C., Williams, S., Romanow, J., et al. (2025). Harmful traits of AI companions. *arXiv preprint arXiv:2511.14972*.

64. Laestadius, L., Bishop, A., Gonzalez, M., Illeňčík, D., & Campos-Castillo, C. (2024). Too human and not human enough:

A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10), 5923–5941. <https://doi.org/10.1177/14614448221142007>

65. Legal Information Institute. (n.d.). Reasonable person. *Cornell Law School*. https://www.law.cornell.edu/wex/reasonable_person

66. Li, C. (2025). AIVA: An AI-based virtual companion for emotion-aware interaction. *arXiv preprint arXiv:2509.03212*.

67. Livingstone, S., Jessen, R. S., Stoilova, M., Stănicke, L. I., Graham, R., Staksrud, E., & Jensen, T. (2025). Can platform literacy protect vulnerable young people against the risky affordances of social media platforms?. *Information, Communication & Society*., 29(2), 455-472.

68. Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Analysis*, 13(1), 43–109.

69. Malfacini, K. (2025). The impacts of companion AI on human relationships: Risks, benefits, and design considerations. *AI & Society*, 40(7), 5527–5540. <https://doi.org/10.1007/s00146-025-02318-6>

org/10.1007/s00146-025-02318-6

70. Maples, B., Cerit, M., Vishwanath, A., et al. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj mental health research*, 3(1), Article 4. <https://doi.org/10.1038/s44184-023-00047-6>

71. Matthew Hutson, "Why AI Chatbots Agree With You Even When You're Wrong," IEEE Spectrum, March 11, 2026, <https://spectrum.ieee.org/ai-sycophancy>

72. McBain, R., et al. (2025). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*, 8(11), e2542281.

73. Meador, M. R. (2025, September 11). Statement of Commissioner Mark R. Meador regarding Section 6(b) resolution relating to generative AI companion products or services. Federal Trade Commission. https://www.ftc.gov/system/files/ftc_gov/pdf/meador-6b-ai_statement.pdf

74. Merrill, K., Kim, J., & Collins, C. (2022). AI companions for lonely individuals and the role of social presence. *Communication Research Reports*, 39(2), 93–103.

75. Meta Platforms, “Our Approach to Teen AI Safety: Empowering Parents, Protecting Teens,” Meta Newsroom, October 17, 2025, updated January 23, 2026, <https://about.fb.com/news/2025/10/teen-ai-safety-approach/>

76. Minsky, M. (1985). *The society of mind*. Simon & Schuster.

77. *Moody v. NetChoice, LLC*, 603 U.S. (2024). https://www.supremecourt.gov/opinions/23pdf/22-277_d18f.pdf

78. Muldoon, J., & Parke, J. J. (2025). Cruel companionship: How AI companions exploit loneliness and commodify intimacy. *New Media & Society*, 14614448251395192.

79. *NetChoice, LLC v. Bonta*, No. 25-146 (9th Cir. Sept. 9, 2025). <https://cdn.ca9.uscourts.gov/datastore/opinions/2025/09/09/25-146.pdf>

80. *NetChoice, LLC v. Bonta*, No. 25-2366 (9th Cir. Mar. 12, 2026).

81. New Mexico Department of Justice. (2026, March 24). New Mexico Department of Justice wins landmark verdict against Meta. <https://nmdoj.gov/press-release/>

new-mexico-department-of-justice-wins-landmark-verdict-against-meta/

82. New York State Senate. (2024). S7694A, 2023–2024 Regular Session: Stop Addictive Feeds Exploitation (SAFE) for Kids Act. <https://www.nysenate.gov/legislation/bills/2023/S7694/amendment/A>

83. New York State Senate, “New York Consolidated Laws, General Business Law, Article 47: Artificial intelligence companion models (GBS §§ 1700–1704)”, New York State Senate, 2025, <https://www.nysenate.gov/legislation/laws/GBS/1700>

84. Office of Senator Steve Padilla. (2025, October 13). First-in-the-nation AI chatbot safeguards signed into law. <https://sd18.senate.ca.gov/news/first-nation-ai-chatbot-safeguards-signed-law>

85. Office of the Surgeon General. (2023). Social media and youth mental health: The U.S. Surgeon General’s advisory. U.S. Department of Health and Human Services. <https://www.hhs.gov/surgeongeneral/reports-and-publications/youth-mental-health/social-media/index.html>

86. Osler, E. (2026). Hallucinating with AI: Distributed delusions and “AI psychosis.” *Philosophy & Technology*, 39(1), 30.

87. Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior*, 140, Article 107600. <https://doi.org/10.1016/j.chb.2022.107600>

88. Picard, R. W. (1995). *Affective computing* (Technical Report No. 321). MIT Media Laboratory Perceptual Computing Section.

89. Picard, R. W. (1997). *Affective computing*. MIT Press.

90. *Raine v. OpenAI, Inc.* (2025). Complaint for wrongful death and product liability. California Superior Court, San Francisco County. <https://business.cch.com/plsd/RainevOpenAI-Complaint.pdf>

91. Jody Godoy, “OpenAI, Altman sued over ChatGPT’s role in California teen’s suicide,” Reuters, August 27, 2025, <https://www.reuters.com/sustainability/boards-policy-regulation/openai-altman->

sued-over-chatgpts-role-california-teens-suicide-2025-08-26/

92. Richet, J.-L. (2025). AI companionship or digital entrapment? Investigating the impact of anthropomorphic AI-based chatbots. *Journal of Innovation & Knowledge*, 10(6). <https://doi.org/10.1016/j.jik.2025.100835>

93. Robb, M. B., & Mann, S. (2025). Talk, trust, and trade-offs: How and why teens use AI companions. *Common Sense Media*.

94. Shannon Bond, “Meta and YouTube Head to Trial Over Harm to Children After TikTok Settles,” NPR, January 27, 2026, <https://www.npr.org/2026/01/27/nx-s1-5684196/social-media-kids-addiction-mental-health-trial>

95. Shen, M. K., & Yoon, D. (2025, April). The dark addiction patterns of current AI chatbot interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-7).

96. Sobowale, J., et al. (2025). Evaluating generative AI psychotherapy chatbots used by youth: Cross-sectional

study. *JMIR Mental Health*, 12, Article e79838. <https://doi.org/10.2196/79838>

97. Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), 1-22.

98. TechNet, “Re: SB 243 (Padilla) – Companion chatbots – OPPOSE [Letter to Assemblymember Rebecca Bauer-Kahan]”, July 1, 2025, <https://ccianet.org/wp-content/uploads/2025/07/TechNet-Led-Coalition-Letter-on-CA-SB-243.pdf>

99. Tech Oversight California, “SB 243: Removing SUPPORT [Letter to Senator Steve Padilla]”, September 7, 2025, <https://techoversight.org/wp-content/uploads/2025/09/SB-243-Remove-Support.pdf>

100. Teo, S. A. (2025). Artificial intelligence, human vulnerability, and multi-level resilience. *Computer Law & Security Review*, 57, Article 106134.

101. Transparency Coalition, “In early ruling, federal judge defines Character.AI chatbot as product, not speech,” May 21, 2025, <https://www.transparencycoalition.org/news/important-early-ruling-in-characterai-case-this-chatbot-is-a-product-not-speech>

ai/news/important-early-ruling-in-characterai-case-this-chatbot-is-a-product-not-speech

102. Tyson & Mendes. (2025, October 17). From code to courtroom: Raine v. OpenAI and the future of AI responsibility. Tyson & Mendes. <https://www.tysonmendes.com/raine-v-openai-ai-product-liability-lawsuit/>

103. University of Chicago Business Law Review. (2025). Generative AI meets Section 230: The future of liability and its implications for startup innovation. <https://businesslawreview.uchicago.edu/print-archive/generative-ai-meets-section-230-future-liability-and-its-implications-startup>

104. VISIO, A., & VE, N. (2025). Emotional risks of AI companions demand attention. (2025). *Nature Machine Intelligence*, 7(7), 981–982. <https://doi.org/10.1038/s42256-025-01093-9>

105. Wang, C., Boerman, S. C., Kroon, A. C., Möller, J., & de Vreese, C. H. (2025). The artificial intelligence divide: Who is the most vulnerable?. *New Media & Society*, 27(7), 3867–3889. <https://doi.org/10.1177/14614448241232345>

org/10.1177/14614448241232345

106. Wang, H. (2023). Algorithmic colonization of love: The ethical challenges of dating app algorithms in the age of AI. *Techné: Research in Philosophy and Technology*, 27(2), 260–280.

107. Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., et al. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83, 19–52.

108. Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>

109. Weinhhammer, V., Hou, K. Y., Luettgau, L., Summerfield, C., Dolan, R., & Nour, M. M. (2026). Vulnerability-amplifying interaction loops: A systematic failure mode in AI chatbot mental-health interactions. *arXiv preprint arXiv:2602.01347*.

110. Welle, E., “Character.AI is banning minors from AI character

chats”, The Verge, October 29, 2025, <https://www.theverge.com/ai-artificial-intelligence/808081/character-ai-under-18-chat-ban>

111. Wikipedia. (2026). Raine v. OpenAI. https://en.wikipedia.org/wiki/Raine_v._OpenAI

112. Winthrop, R., & Hau, I. (2025, October 29). What happens when AI chatbots replace real human connection. Brookings Institution. <https://www.brookings.edu/articles/what-happens-when-ai-chatbots-replace-real-human-connection/>

113. Yeung, K. (2019). “Hypernudge”: Big data as a mode of regulation by design. In *The social power of algorithms* (pp. 118–136). Routledge.

114. Tsinghua University Artificial Intelligence Research Institute, Beijing Academy of Artificial Intelligence, & Tsinghua-Academy of Engineering Knowledge Intelligence Joint Research Center. (2019). *Artificial Intelligence: Affective Computing* (No. 6). Tsinghua University. <https://static.aminer.cn/misc/pdf/emotion.pdf>

115. Wu, J. (2025). The “black hole” of emotional manipulation in human-computer interaction: Reflections based on virtual chat companions. *Exploration and Free Views*, (08), 106–115.

116. Zhao, Y., & Huang, Z. (2025). From emotional labor to emotional exploitation: Capital logic and subjectivity crisis in artificial intelligence intimate relationships. *Editorial Friends*, (10), 61–69.



