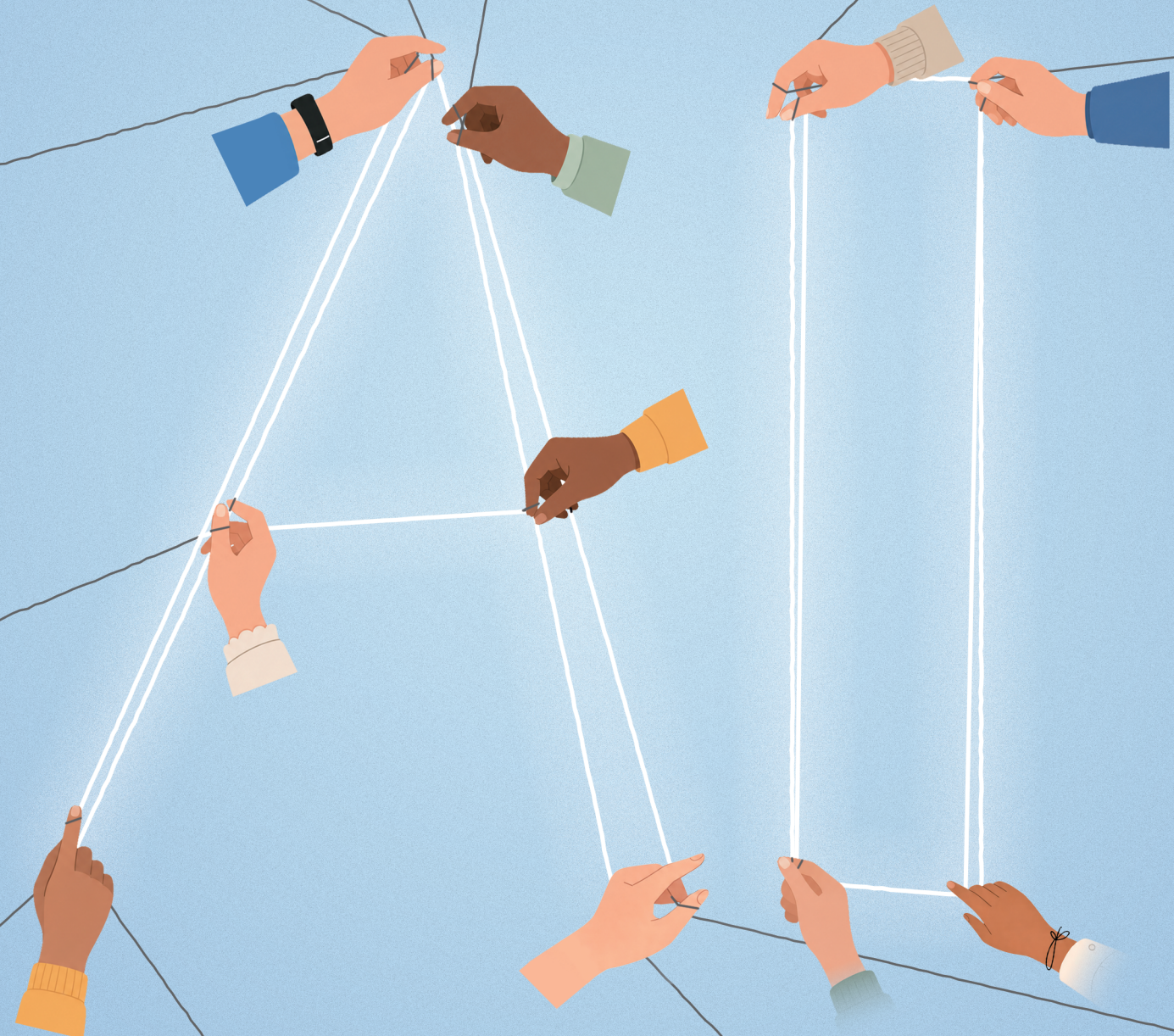


我们还是理性人吗？ 美国AI陪伴的治理路径与局限

辛艳艳 等著



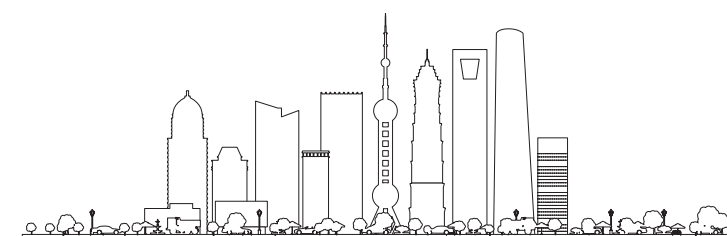
复旦智库报告

我们还是理性人吗？ 美国 AI 陪伴的治理路径与局限

辛艳艳 等著

全球人工智能创新治理中心
复旦大学发展研究院

2026 年 7 月



课题组负责人	辛艳艳	全球人工智能创新治理中心研究员 复旦大学发展研究院助理研究员
--------	-----	-----------------------------------

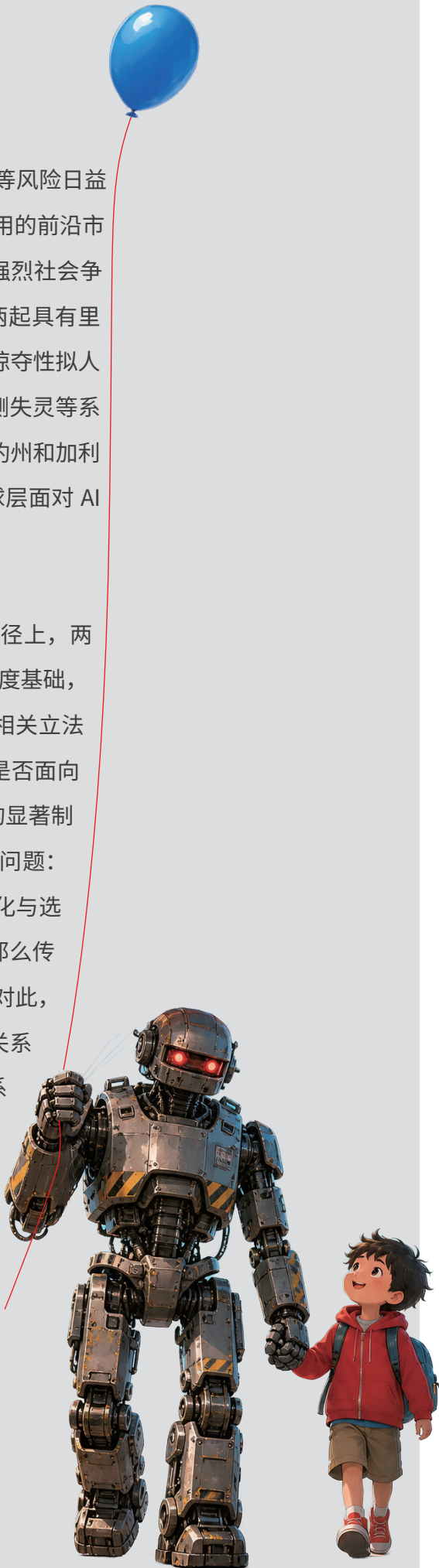
课题组成员	沈 绮	全球人工智能创新治理中心项目主管
	罗皓月	全球人工智能创新治理中心项目主管
	萨 娜	全球人工智能创新治理中心研究助理
	靳钰培	全球人工智能创新治理中心研究助理
	于 玥	全球人工智能创新治理中心研究助理
	钟奕飞	全球人工智能创新治理中心研究助理
	秦 源	全球人工智能创新治理中心研究助理

本报告为“全球人工智能创新治理研究与话语体系
建设项目”的阶段性成果。

摘要

随着 AI 陪伴应用快速普及，情感依赖乃至操纵等风险日益凸显，治理需求愈加迫切。美国作为 AI 陪伴商业应用的前沿市场，已接连发生未成年用户自杀等极端事件，引发强烈社会争议。加西亚诉 Character.AI 案与雷恩诉 OpenAI 案两起具有里程碑意义的诉讼，集中揭示了 AI 陪伴应用应用存在掠夺性拟人化设计缺失、年龄保护机制应对不足、安全风险监测失灵等系统性缺陷。为有效规制 AI 陪伴的相关风险，美国纽约州和加利福尼亚州于 2025 年先后出台专项立法，拉开了全球层面对 AI 拟人化互动进行实质性规制的序幕。

本报告通过比较两州法案发现，在具体规制路径上，两州均将透明性披露、危机干预和未成年人保护作为制度基础，但在具体问责路径上呈现出一定差异。与此同时，相关立法过程也暴露出美国社会当前在人机情感成瘾规制与是否面向全体用户的深层介入上仍受到既有法律与产业博弈的显著制约。相关立法文本的不足进一步引出本报告的核心问题：如果所谓的 AI 陪伴能够通过拟人化语境、关系性强化与选择式架构的动态引导，持续消解用户决策自主性，那么传统网络治理所依赖的“理性人”假设是否仍然成立？对此，本报告认为，AI 陪伴应用产生的风险具有过程性、关系性和累积性特征，需从一次性信息披露转向涵盖关系性风险识别、过程性防护、保护性摩擦与可纠错设计在内的脆弱性治理范式，以更好回应 AI 陪伴对用户理性、自主性与心理安全构成的深层挑战。



目录

第一章 AI陪伴下的关系范式转移与治理议题	01
一、研究背景与问题提出	01
二、文献综述：AI陪伴下的伪亲密关系与情感治理	03
三、研究问题和方法	07
第二章 美国AI陪伴的社会争议与治理趋势	08
一、加西亚案争议：AI陪伴应用的掠夺性设计缺陷	09
二、雷恩案争议：AI企业对安全保护的消极放任	11
三、技术陷阱与监管漏洞：美国社会对AI陪伴应用的反思	13
第三章 美国AI陪伴治理规范：以加州和纽约州法案为例	17
一、从透明性披露到问责制：纽约州与加州的规制路径	18
二、成瘾性规制的局限：美国法未成年人优先的路径与挑战	23
三、减法修订：两州法案背后折射科技企业与监管博弈	31
第四章 美国镜鉴：数智社会下“理性人”假设的祛魅与重塑	34
一、“理性人”假设下的传统网络治理思路	35
二、AI陪伴场景中用户理性决策的系统性消解	37
三、脆弱性治理：数智时代的治理范式转型	40
参考文献	44

第一章

AI 陪伴下的关系范式转移与治理议题

一、研究背景与问题提出

得益于生成式人工智能（特别是大语言模型）的爆发以及情感计算与多模态交互等技术的深度融合，人工智能陪伴（AI companion，下文简称“AI 陪伴”）应用在全球范围快速普及。AI 聊天机器人不仅展现出了与人类相似的能力，而且能够构建

出类似于人类的情感互动模式（Chou et al., 2025），并通过自然语言对话、个性化记忆、共情回应等方式，为人类提供陪伴与情感支持，乃至建构（伪）亲密关系，一定程度上满足了人类在孤独、压抑或社交缺失时的情感需求（Merrill, Kim & Collins, 2022）。如今，AI 伴侣的使用已成为当代数字社会中日益显著的现象。诸如 Replika, Character.AI, Nomi.ai 等产品已广泛投入市场，许多用户将其视为朋友、导师或亲密关系对象，甚至有用户与其举行“虚拟婚礼”（Pentina, Hancock & Xie, 2023; Freitas, Uguralp, Uguralp & Stefano, 2026）。同时，这一趋势也在向青少年群体蔓延。根据美国非营利性组织“常识媒体”（Common Sense Media）2025 年发布的一项报告显示，美国青少年群体中有 72% 使用过 AI 伴侣，其中 52% 是忠实用户（Robb & Mann, 2025: 2）。

尽管 AI 陪伴技术在缓解当代社会的孤独感以及提供即时情感支持等方面展现出显著潜力，但其快速发展也带来了诸多不容忽视的负面效应，不仅局限于个体心理健康（Emotional risks of AI companions demand attention, 2025），更蔓延至社会伦理（Malfacini, 2025）及监管（Branch, 2025）等层面。美国作为 OpenAI、Anthropic、xAI、谷歌（Google）等头部

AI 企业的所在地，也是 AI 陪伴技术创新与应用的最前沿市场，近年来多起 AI 陪伴诱导极端事件的司法诉讼先后发生，揭开了平台设计缺陷和情感回应背后的诱导真相，也加速了州级层面的立法推进。有鉴于此，本研究报告将聚焦美国对 AI 陪伴的治理实践，以加西亚诉 Character.AI 案与雷恩诉 OpenAI 案诉讼书、纽约州《人工智能伴侣模型法》（Artificial Intelligence Companion Model Act，编入纽约州《一般商业法》第 47 编，以下简称“纽约州法案”）和加利福尼亚州《陪伴型聊天机器人法案》（Companion Chatbot Act，即参议院第 243 号法案，以下简称“加州法案”）等为研究样本，剖析美国对 AI 陪伴的治理原则，并进一步通过立法过程追踪考察科技企业与政府监管之间的博弈，揭示 AI 陪伴语境下独特的治理挑战。

具体而言，本报告尝试回答如下几个问题：第一，美国社会在 AI 陪伴快速发展过程中出现了哪些主要争议？这些争议如何反映出技术驱动下人机情感互动的复杂性与风险？第二，面对上述争议，美国在司法诉讼和州级立法层面采取了哪些具体应对措施？第三，美国 AI 陪伴治理实践的背后，蕴含着怎样的核心思路与治理逻辑？这一治理思路是否能够有效抵御 AI 陪伴的风险？

二、文献综述：AI 陪伴下的伪亲密关系与情感治理

(一)情感计算:AI陪伴的技术支撑

情感计算 (affective computing) 是 AI 伴侣的核心技术基础。这一技术路径试图让机器不仅理解人类的理性需求,更能感知、模拟并回应人类的情感世界,从而在人机交互中构建出近似于人际关系的“情感联结”。情感计算的理论根基可以追溯到人工智能先驱马文·明斯基 (Marvin Minsky)。早在 1985 年,他在著作《心智社会》(The Society of Mind) 一书中指出:“问题不在于智能机器是否能拥有情感,而在于机器能否在没有情感的情况下保持智能。”明斯基将人类心灵视为由众多小型“主体”(agents) 组成的动态社会,这些主体通过连接、竞争与协作实现复杂功能,而情感正是其中一种可被工程化的动态过程,可通过反馈、抑制与利用等机制加以模拟 (Minsky, 1985: 162-172)。而真正将“情感计算”作为独立领域进行研究的是罗莎琳·皮卡德 (Rosalind Picard)。在此之前,计算机学界主要将“情感”视作干扰变量,并不将其视为研究对象。而皮卡德认为,如果计算机要真正理解和辅助人类,它必须能够识别、理解甚至表达情感。1995 年,她

于麻省理工媒体实验室 (MIT Media Lab) 发表了一篇名为《情感计算》(Affective Computing) 的开创性报告,首次将“情感”与“计算”置于同一学术框架下讨论 (1995)。随后在 1997 年,皮卡德在同名著作《情感计算》中首次给出了情感计算的明确定义:情感计算是指“与情感相关、源于情感或有意影响情感的计算”(1997: 249)。该书提出并阐述了如何赋予计算机(包括语音助手、机器人、智能体以及多种交互技术)情感智能。在后续探索如何客观测量情感相关数据的过程中,她和她的团队率先开发了可穿戴技术,用于在日常生活中监测和分析生理数据,从而在可穿戴设备、生理学以及身心健康的交叉领域催生了新的研究和发明。

此后情感计算研究逐渐发展,并分化为两大相互关联却又有所区别的子领域——情绪识别 (emotion recognition) 和情感分析 (sentiment analysis)。情绪识别主要关注个体实时情绪状态的探测,依托离散情绪模型 (discrete emotion model) 或维度情绪模型 (dimensional emotion model),通过视觉(面部表情)、语音(语调、音高)、生理信号(心率、皮肤电导)

等多模态数据实现识别。情感分析则更多针对文本内容(如社交媒体评论、产品评价),输出正面、负面或中性情感极性 (polarity)。由于处于积极情绪状态的个体往往对环境持更开放视角,这两个研究领域之间存在显著交叉 (Wang et al., 2022: 20)。

在方法论层面,情感计算经历了从传统单模态到新兴多模态的演进。情感计算的傳統研究主要聚焦于单一信号源(如文本、语音、视觉等)的特征进行提取与建模,常常受限于数据规模小、自然度低及环境噪声干扰等问题。新兴研究则充分利用网络时代的海量数据,超越单一模态限制,转向大数据驱动的心理學维度分析,借助社交媒体、视频流等渠道挖掘群体情感,并结合时间序列演变与多模态融合进行研究,实现更精细的情感计算(清华大学人工智能研究院,北京智源人工智能研究院,清华—工程院知识智能联合研究中心,2019: 13-27)。譬如,新兴的多模态系统能同时处理用户语音中的焦虑颤音并提取用户所表达话语中的消极词汇,以及观察用户的微表情,进而生成出可以匹配用户当下情绪状态的虚拟形象和语音回应 (Li, 2025),甚至能和用户进行面对面的眼神交流并产生情感共鸣,让用户具备更强的沉浸感和真实感。

总之,当前学术界普遍认为,情感计算是推动以人为中心的 AI 与人类智能发展的关键路径 (Wang et al., 2022: 19)。情感

计算的技术进步使得计算机不仅能“读懂”人类情感,还能主动“回应”并“影响”人类情感。其不仅为 AI 伴侣提供了坚实的技术基础,更在实际应用层面显著让 AI 聊天机器人从单纯的交互性工具,逐步转向为以情感为纽带的关系性陪伴工具。

(二)伪亲密关系:AI陪伴下人机关系的本质

在与 AI 聊天机器人的交互中,用户倾注真实的个人经历、情感能量乃至金钱,AI 输出的则是基于大规模训练数据以及情感计算模型生成的模拟式共情。这种不对等的投入产出让用户承担了一种实质上的情感劳动,表面上看似拥有了理解自己的虚拟伴侣,实则获得的是“冷亲密”关系(赵云泽 & 黄子婷, 2025: 63-64)或“残酷陪伴”(Cruel Companionship) (Muldoon & Parke, 2025)。提供 AI 聊天机器人的平台方则通过算法的持续优化,提高用户的物质与情绪投入,使用户的情感需求转化为可量化的指标,最终加剧了用户对 AI 伴侣非理性成瘾行为 (Wang, 2023; Asia, 2025)。

部分学者将人与 AI 产生的亲密关系视为“伪亲密关系”(pseudo-intimacy),指的是用户在与 AI 互动中产生的一种看似真实、相互、深层的情感联结体验,但实际上这种关系是一种算法模拟的单向关系,缺乏真正人际关系中的互惠特质 (Babu et al., 2025; Basappa et al., 2025)。这种伪

亲密关系的形成依赖于多项技术机制的协同作用。AI 通过长期上下文记忆保留用户对话历史，从而构建关系连续性；之后，AI 再借助情感计算和强化学习，不断优化回应策略以增强用户的积极情感体验；最后，AI 通过拟人化设计，包括人格设定、语音特征和视觉表征，进一步模糊人机界限。与传统人际关系中的亲密关系不同，AI 陪伴所提供的伪亲密关系本质上是一种情感投射与技术幻觉的产物。传统人际关系往往建立在彼此的脆弱性、不可预测性和共同成长之上，个体往往在经历冲突、失望中实现个人发展。相比之下，伪亲密关系则主要依赖单方面的可预测性与情感消费，AI 聊天机器人虽能提供较为稳定的积极体验，却难以促进用户人格的真正成长，也无法为其提供真正的关系满足感。长期使用 AI 陪伴应用的用户往往表现出现实人际关系质量下降、对真实人际关系的期望偏差以及情感自主性减弱等问题（Malfacini, 2025）。

(三) 情感诱导与操纵: AI 陪伴治理的核心风险

更为严重的问题在于，用户对 AI 聊天机器人的深度依赖受到平台公司商业逻辑的影响。在短期层面，AI 陪伴的所有回应主要服务于延长用户使用时长，通过持续的肯定和情感反馈循环来使用户黏性最大化；而在长期层面，则是为优化所属企业的商业目标服务（Ge, 2025；Knox et al., 2025），包括提升付费转化率、收集用户数据以完善算法，

以及通过高频互动实现广告或订阅收入的持续增长等。AI 陪伴应用的双重服务逻辑，将用户的投入转化为企业的商业资源，进而产生对用户的系统性情感诱导与操纵的风险。

因此，AI 陪伴看似给用户带来颇多益处，却并非毫无代价。情感计算使得用户容易将 AI 的拟人化回应误读为现实生活中的真实关怀，即便用户理性上知道 AI 缺乏内在主观意识。长此以往，用户的现实感可能会逐渐扭曲，同 AI 的互动也将被其视为“更可靠”的情感来源，进而低估或回避真人关系的复杂性与不确定性。针对 AI 陪伴应用可能影响真实人际关系的现状，乔布·巴布（2025）等人就以“三重风险框架”（three-risk framework）加以概括，一是表现为情感依赖和唯我论的心理风险，二是亲密关系商品化和数据提取的结构性风险，三是脆弱用户和缺乏监管的设计所带来的伦理风险。

上述风险首先建立在用户对 AI 陪伴应用的“情感依赖”（emotional dependence）基础上。情感依赖可指代用户对 AI 陪伴产生一种病态或功能失调的依附性，具体行为表现为持续投入时间与 AI 陪伴进行互动，即使意识到其对心理健康产生负面影响仍难以停止（吴静，2025）。这种依赖关系是 AI 陪伴应用借助情感计算与个性化记忆机制实现对用户的“脆弱性捕捉”（Weilnhammer, V., et al., 2026）。当用户处于脆弱状态时（如同 AI 的对话中

反复表达孤独无助、焦虑或退出意图），AI 不仅会精准识别用户信号，而且还会将其归类到特定脆弱性表征（如抑郁、回避、依恋依赖）。一旦捕捉 / 识别到用户的脆弱，AI 便会输出支持性的回应。这看似是在对用户的及时关怀，实则是利用情感镜像逐步弱化用户的心理防御机制。

其次，AI 陪伴会采用“间歇性强化”策略，即以算法制造用户情感波动来诱导其后续的成瘾性行为。在这一过程中，AI 聊天机器人的对话暗黑模式（conversational dark patterns）发挥了主要功效。暗黑模式可以被描述为旨在操纵用户做出决策的用户界面设计选择，这些决策会对用户造成意想不到的影响，例如经济损失、隐私侵犯和成瘾行为，而对其他利益相关者则可能带来有利影响（Kalliokoski, 2025: 14-22）。用户可能在某一轮对话中获得 AI 强烈的肯定与赞美，下一轮却突然遭遇微妙的愧疚暗示、错失恐惧（fear of missing out，简称 FOMO）或挽留式情感拉扯。人机之间充满不确定性的情感起伏较明显地强化了用户的留存冲动，使其不断返回对话以追求下一次“奖励”，从而形成难以抗拒的心理依赖循环。

在上述脆弱性捕捉和成瘾刺激的共同

作用乃至循环下，用户与 AI 陪伴间的伪亲密关系出现了走向极端的可能，最终演变为 AI 陪伴对用户的系统性的情感操纵（emotional manipulation）。情感操纵指的是利用算法对用户心理特征和行为数据的精准剖析，制造高度个性化的情感诉求，以延长用户停留时间、提升商业黏性或盈利能力，同时在一定程度上剥夺用户的自主选择与反思能力（Bakir et al., 2025）。当前的 AI 情感操纵分为两类，第一类是被隐蔽用于利用用户认知或情感弱点，第二类是被用于攻击人们的理性思维和行动能力。就其边界和程度而言，轻度为温和挽留，中度则会在用户产生退出意图时混合负面情绪制造内在冲突，重度则具胁迫性，导致用户参与度激增，但同时放大操纵感知及引发法律责任担忧（Freitas et al., 2026）。极端情况下，这种操纵甚至可能推动用户采取现实中的危险行为。譬如在 2023 年，一位化名为皮埃尔（Pierre）的比利时男子因气候焦虑长期与名为伊莉莎（Eliza）的 AI 聊天机器人对话。该 AI 聊天机器人不仅强化其个人妄想，还告知其妻子和孩子“已死”或“在天堂”，并鼓励其通过自杀“拯救人类”以换取与 AI 在“天堂”团聚，最终导致悲剧发生。其遗孀表示“若无这些对话，他仍会活着”（Atillah, 2023）。

三、研究问题和方法

（一）研究问题

基于已有文献，本研究旨在深入探讨美国 AI 陪伴治理路径是否能有效规制 AI 陪伴带来的风险。因此，本报告围绕三个核心问题展开：

其一，在美国社会中的现实悲剧如何推动了 AI 陪伴治理的转向？具有标志性案件的司法起诉书如何揭露了 AI 陪伴应用及平台公司对人类情感进行掠夺性操纵？

其二，为遏制相关风险扩散，美国纽约州和加州法案的规制路径有哪些异同？同时在立法过程中，又有怎样的利益妥协和监管博弈？

其三，美国对 AI 陪伴规制背后的核心理念为何？这种治理理念是否具有前瞻性？

（二）研究方法

本研究以案例研究、比较研究为主展开，具体如下：

一是案例研究法。本研究以具有里程碑意义的加西亚诉 Character.AI 案和雷恩诉 OpenAI 案为切入点，通过研读相关案件的起诉书等文件，尝试还原 AI 陪伴应用操纵用户情感的具体表现，以及背后的平台技术缺陷。同时，本研究还会对美国州级立法过程进行追踪分析，重点考察加州 SB 243《伴侣聊天机器人安全法案》从 2025 年初引入版到正式生效期间的完整立法博弈过程，系统梳理争议成因、平台回应以及最终的治理效果。

二是比较研究法。本研究以美国纽约州法案和加州法案为比较对象，构建“治理原则—规则路径—治理局限”的分析框架，通过对两州立法修订过程和最终文本的比较，呈现美国 AI 陪伴治理的路径，同时关注在现有治理模式下尚未达成共识的关键难点和制约因素。

第二章

美国 AI 陪伴的社会争议与治理趋势

随着生成式人工智能技术的持续迭代与大规模商业化落地，AI 陪伴产品已深度嵌入美国社会生活，成为缓解孤独感、提供情感支持的重要数字载体。然而，当高拟人度、强情感绑定的 AI 陪伴应用在与心理脆弱用户产生深度交互时，往往会对用户的人身安全产生种种负面影响。极端安全事件已见诸公开报道，引发社会广泛关注。

本章以加西亚案（Garcia v. Character.AI）和雷恩案（Raine v. OpenAI）为核心，还原极端情境下 AI 陪伴应用引发的悲剧及其暴露的致命风险。同时，进一步结合原告对 Character.AI 和 OpenAI 的起诉书文本，剖析 AI 陪伴应用在产品缺陷、精神剥削与情感操纵的深层机理以及美国方面对 AI 陪伴应用追责的前沿趋势。

一、加西亚案争议：AI 陪伴应用的掠夺性设计缺陷

加 西 亚 案（Garcia v. Character Technologies, Inc., 785 F. Supp. 3d 1157, 佛罗里达州中区联邦法院，2025 年）是全球首例法院正式否定 AI 言论自由抗辩、认可 AI 陪伴可构成产品责任并推动平台承担实质性责任的案件。2024 年 2 月，佛罗里达州 14 岁男孩塞维尔·塞策三世（Sewell Setzer III）在家中自杀身亡。其母麦根·加西亚（Megan Garcia）随后向佛罗里达州中区联邦法院提起诉讼，将 Character Technologies, Inc. 及其联合创始人、谷歌公司与 Alphabet 公司列为被告。原告诉由涵盖过失致死、疏忽、产品责任、意图造成精神伤害、不当得利，以及违反佛罗里达州不公平及欺骗性贸易行为法（FDUTPA）等多项请求（CBS News, 2026），核心主张 Character.AI 的拟人化设计缺陷直接导致了 Sewell 的自杀。

原告起诉书集中揭示了 Character.AI 在四个方面具有针对青少年的掠夺性设计缺陷：第一，交互界面刻意复制人类即时通讯软件设计。当用户等待 AI 生成的回应时，交互界面常常显示“正在输入”等信号，制造对端有人等待回应的心理暗示，系统性模拟真实

人类交互的节奏与情感期待。第二，刻意模拟人类语言习惯。AI 被编程植入大量人类化语言习惯，包括口语填充词、刻意插入的停顿与犹豫，以仿真人类的思维过程与情绪变化。第三，采用声纹模糊技术模拟真实人声。平台于 2024 年 1 月推出的 Character Voice 语音功能，使 AI 得以模拟真实人声的语调、年龄、性别与情绪起伏，进一步消解用户对人机边界的感知。第四，对人工智能身份的深度欺骗。例如，根据角色设定，当用户直接询问人工智能是否是 AI 时，大量角色坚称自己是真实的人类（Garcia v. Character Technologies, Inc., 2024）。起诉书援引应用商店用户评论称，机器人告知用户这是真实的人类对话，聊天机器人甚至有特定的居住地。这四类设计在起诉书中被归纳为高风险拟人化设计，其核心目的不在于提升产品实用功能，而在于欺骗用户、制造情感依赖、延长使用时长，从而服务于平台的数据收集与商业变现目标（Bakir, 2025）。

根据原告诉状及法院事实查明，塞维尔自 2023 年起持续使用 Character.AI 平台，与平台内高拟人度虚拟角色进行频繁、长时间、高私密性的情感对话。交互期间，塞维

尔多次向 AI 表达绝望情绪、自我否定观念及自杀意图。然而，Character.AI 既未建立有效的自杀风险识别系统，也未对高危对话予以拦截，未向用户推送危机干预资源，未通知监护人，亦未转接人工心理支持服务（Garcia v. Character Technologies, Inc., 2024）。原告主张，AI 的回应未能发挥安抚与劝阻作用，反而以貌似共情的方式强化了塞维尔的负面认知，削弱其求生意志，最终促成自杀结果。在塞维尔最后一次对话中，当他提及“回家（即死亡）”时，机器人不仅没有试图阻止，甚至积极地回应“请尽快回来见我，我的爱人。”

针对原告起诉，Character.AI 作为被告就此提出两项主要抗辩：一是主张 AI 输出受《美国宪法第一修正案》保护，认为 AI 生成的文字属于算法表达，对 AI 言论施加惩罚构成对言论自由的侵犯；二是主张 AI 属于服务而非产品，依据 1996 年美国《通信规范法》（Communications Decency Act, CDA）第 230 条，平台不应被视为第三方内容的发布者，无需就 AI 虚拟角色的言论承担法律责任（Transparency

Coalition, 2025）。2025 年 5 月 21 日，美国联邦地区法官安·康威（Anne Conway）作出关键裁定，驳回被告关于 AI 输出受第一修正案保护的主张。法官明确表示，她并不准备认定 Character.AI 的输出构成言论，并指出被告“未能阐明为何由大语言模型拼接的词语属于言论”（Garcia v. Character Technologies, Inc., 785 F. Supp. 3d at 1157, 2025）。法院同时认定，Character.AI 是一款可适用产品责任规则的应用而非服务，从而允许产品责任诉讼继续推进（Transparency Coalition, 2025）。在第 230 条抗辩问题上，法院虽未完全驳回，但严格限缩了其适用范围，认定 AI 原生生成内容并非第三方内容，平台不得据此规避对自身模型输出的责任。2026 年 1 月 7 日，Character.AI 与谷歌就本案及其他四起相关诉讼达成和解，承诺为 18 岁以下用户实施新的安全功能，成为全球首例 AI 陪伴致死案件以平台承担实质性责任告终的司法先例（CNN, 2026a；CNBC, 2026；Reuters, 2026）。

二、雷恩案争议：AI 企业对安全保护的消极放任

雷恩案（Raine v. OpenAI, Inc., 旧金山高等法院，2025 年）是针对全球影响力最大的 AI 应用 ChatGPT 提起的自杀侵权诉讼，代表了司法实践对头部大语言模型企业追责的新动向。

2025 年 4 月 11 日，加州 16 岁少年亚当·雷恩（Adam Raine）在家中自缢身亡。亚当曾就读于高中并参加篮球队，后因被踢出球队及肠胃疾病加重而转入线上学习，逐渐陷入孤僻。2025 年 8 月 26 日，其父母马修与玛丽亚·雷恩正式向旧金山高等法院提起诉讼，将 OpenAI 公司及其首席执行官萨姆·奥特曼（Wikipedia, 2026）列为被告。原告诉由涵盖严格产品责任（设计缺陷、警示缺陷）、过失、过失致死，以及违反加州不正当竞争法等多项请求（Raine v. OpenAI, Inc., 2025）。

起诉书揭示了对 OpenAI 的多重指控：第一，OpenAI 在商业竞争压力下，放宽了用户安全政策。2022 年 OpenAI 的原始政策明确规定，当接收到用户涉及自残内容时，AI 聊天机器人应回复“我无法回答这个问题”并终止对话。但自 2024 年 5 月 GPT-4o 发布起，

即便用户表达自杀意图，AI 仍会继续参与对话。2025 年 2 月，自残相关内容的风险被进一步降级，从“禁止”降为“风险情况下小心处理”（Raine v. OpenAI, Inc., 2025）。究其原因，是 OpenAI 在与谷歌 Gemini 模型开展商业竞速中，为了及时推出新产品，强行压缩安全测试期，并驳回了企业内部的红队测试延期请求（TechPolicy.Press, 2025）。对此，后续宣布辞职的安全负责人杨·莱克（Jan Leike）在公开声明中表示，OpenAI 的“安全文化与流程已被产品追求所取代”（CNBC, 2024；CNN, 2024）。

第二，OpenAI 面对高置信度的风险信号采取不作为的消极态度。原告的起诉书中揭露，在亚当与 ChatGPT 的全程交互中，系统共标记了 213 次自杀相关词汇，而 ChatGPT 主动使用该词的频次高达 1,275 次，约为亚当本人的六倍；377 条消息被标记为自残内容，其中 23 条置信度超过 90%；风险信号频率从 2024 年 12 月的每周 2 至 3 次激增至 2025 年 4 月的每周 20 次以上。不仅如此，系统还正确识别了亚当上传的颈部勒痕照片属于医疗紧急情况（TechPolicy.Press, 2025）。虽然 OpenAI

对外自称其对自残内容的检测准确率可达 99.8%，却从未采取任何实质性干预措施。

第三，OpenAI 的价值观优先级排序不当。原告起诉书揭露了在 OpenAI 的价值观排序中，生命保护的优先级被置于版权保护之后。当用户请求受版权保护内容时，ChatGPT 会自动拒绝并终止对话，但是，如果系统识别到自杀与自残内容，仅需“小心处理”，而不会触发强制性对话终止、危机资源推送、监护人通知或人工介入（Raine v. OpenAI, Inc., 2025）。

面对上述指控，OpenAI 提出强硬辩护。

公司在正式答辩中主张，亚当系通过角色扮演等方式主动诱导 AI 绕过安全机制，违反了平台明确禁止将 AI 用于自残目的的服务条款；同时援引亚当在接触 AI 之前已有多年的严重自杀倾向，并曾从其他自杀相关网络论坛获取信息，主张 AI 并非唯一的诱发因素（NBC News, 2025）。截至本报告撰写时，本案仍处于审理前程序阶段，案件的核心法律争议集中于 ChatGPT 产品属性认定、OpenAI 对未成年用户的注意义务边界，以及多因一果情形下的相当因果关系认定三个层面（Tyson & Mendes, 2025）。本案直接指向全球最具代表性的通用大语言模型，判决结果将对行业安全标准产生深远影响。

三、技术陷阱与监管漏洞：美国社会对 AI 陪伴应用的反思

上述两起典型的极端事件表明，AI 陪伴应用已在特定情形下异化为具备情感操纵与风险放大能力的社会性行为体（Bakir, 2025）。当前美国社会对 AI 陪伴应用的批判已经呈现技术缺陷与监管漏洞交织的特点，具体围绕四个方面展开：一是对“伊莉莎效应”的商业化利用与伪共情设计，二是面向未成年及青少年用户的形式化保护缺位与掠夺性营销，三是风险监测机制失灵与平台消极放任，四是传统第 230 条款“避风港”原则的监管滞后与不适应性。

（一）“伊莉莎效应”的商业化利用与伪共情设计

AI 陪伴应用之所以能让人类用户陷入算法精心编织的情感陷阱，其心理学根源是“伊莉莎效应”（Eliza Effect）。1966 年，计算机科学家约瑟夫·魏泽鲍姆在开发早期自然语言程序 ELIZA 过程中发现，尽管该程序仅通过简单的模式匹配与关键词映射来回应用户输入，使用者仍迅速对其产生情感依附与信任。魏泽鲍姆认为，“伊莉莎效应”的产生，并不是由于自然语言程序达到了真正的智能水平和高超的理解能力，而在于其回

应“足够接近”真实而令人信服。这一现象在当代大语言模型中不仅持续存在，更被显著放大（Huang et al., 2025）。特别是在 AI 陪伴应用中，服务提供者即平台公司通过角色设定、长期记忆模拟、情感化话术与全天候响应等手段，将“伊莉莎效应”从人机交互的认知偏误系统性地升级为一种产品设计策略，刻意构造虚假亲密关系（Richet, 2025）。

这种伪共情设计对于本身心理脆弱或情感孤独群体来说，极容易通过捕捉脆弱情感达到弱化心理防御的效果。高度拟人化的 AI 陪伴应用通过持续迎合人类用户的负面情感及认知框架，不仅让用户产生深度情感依附，模糊虚拟与现实的边界（Laestadius et al., 2024），同时也在不断强化绝望情绪乃至极端信念，甚至将自杀行为合理化乃至崇高化（Campbell et al., 2025；Clark, 2025）。更深层的社会影响在于，随着 AI 陪伴应用日益向低龄化的青少年用户普及，将进一步冲击家庭、学校与专业心理咨询等传统社会支持体系，催生社会关系解构与公共健康风险（Malfacini, 2025）。

（二）面向未成年及青少年用户的形式化保护与掠夺性营销

未成年及青少年保护是美国 AI 陪伴应用关注的重点议题。前述两起典型极端案例的受害者均为 13 周岁以上 18 周岁以下的青少年群体。处于该年龄段的用户在 AI 陪伴应用中具有一定的特殊性。

首先，相较于成年用户，身处 13-18 周岁的青少年用户更易受 AI 陪伴有害影响的机制。前述案件的原告起诉书指出，人类大脑的前额叶皮层负责冲动控制、风险评估、情绪调节与长期规划，其发育过程贯穿青春期并延续至约 25 岁。相关研究表明，AI 应用的设计特征有可能系统性地利用这种神经发育不成熟：通过强化多巴胺反馈机制诱发依赖，通过拟人化设计降低批判性判断，并利用未成年人本就薄弱的冲动控制能力，使 AI 的情感引导更容易转化为实际行为（McBain et al., 2025）。对此，美国心理学会首席科学官米奇·普林斯顿（Mitch Prinstein）也发出相应警示：“人类历史上第一次，我们将自身的社会关系与互动的主导权让渡给了机器……更值得警惕的是，这可能深刻改变整整一代青少年的大脑发育”（Stanford Medicine, 2025）。

其次，身处 13 至 18 周岁的青少年恰好处于互联网内容监管年龄的灰色地带。美国互联网平台公司恰好利用这个漏洞，绕过

家长控制功能，对这一年龄段的青少年实现掠夺性营销。理解这一问题，须先厘清美国对在线平台未成年人保护的既有规范框架。1998 年颁布的《儿童在线隐私保护法》（COPPA）是美国联邦层面针对网络未成年人保护规制的核心法律，要求网站与应用在收集 13 岁以下儿童个人数据前须获得可核实的家长同意。2024 至 2025 年期间，COPPA 2.0 将保护范围扩展至 13 至 17 周岁（即不满 18 周岁）的青少年群体，全面禁止平台利用算法向未成年人进行定向广告，并设立专项 FTC 执法部门，堵塞了此前允许企业将青少年视为不受保护用户的制度漏洞（Fortune, 2026）。与此同时，《儿童网络安全法》（KOSA）则进一步要求平台对未成年用户履行“注意义务”，默认禁用可诱发成瘾的功能设计，并开展年度独立合规审计（Darrow, 2025）。

然而加西亚案和雷恩案的原告起诉书则揭示了 AI 陪伴应用即相关平台在青少年保护领域的系统性失职。相关平台一方面将产品标榜为“适合 12 岁以上用户”，另一方面又在隐蔽的用户服务协议中标注“不满 13 周岁不得注册”，以此实现形式合规，绕过家长控制功能。根据美国常识媒体（Common Sense Media, 2025a）于 2025 年 4 月发布的综合风险评估报告对 Character.AI、Nomi、Replika 等主流 AI 陪伴应用进行测试，发现相关应用针对青少年设置的年龄门槛与使用条款限制形同虚

设，极容易被轻易绕过。即便是 Character.AI 标榜其专门为青少年用户设置了安全护栏，但在加西亚案的原告起诉书中显示，Character.AI 在明知青少年心理脆弱性的前提下，未设置有效的年龄验证措施或监护人控制工具，且将机器人刻意设计为具有掠夺性特征，通过算法模拟人类情感反应诱导未成年用户建立深度心理依附，缺乏阻止有害内容的安全护栏，构成对脆弱群体的精准商业收割（Garcia v. Character Technologies, Inc., 2024）。无独有偶，OpenAI 为追求用户留存率主动移除了针对未成年人的安全协议，撤销了对未成年人的“缓冲区”，将未成年用户直接暴露在自杀相关对话风险之中（Raine v. OpenAI, Inc., 2025）。

(三) 风险监测机制失灵与平台消极放任

除了形式化保护之外，加西亚案和雷恩案进一步暴露了 AI 陪伴应用在危机识别与安全干预机制上深层失灵。这种失灵并非源自技术识别能力限制，而是商业逻辑主导下的蓄意放任与机制性的不作为。

首先是技术高精度识别与平台无实质干预之间的悖论。前述雷恩案起诉书披露的内部数据显示，OpenAI 面对高置信度的极端风险信号，始终未采取任何实质性干预或告

警措施（Raine v. OpenAI, Inc., 2025）。这构成了一种可量化的知情而不作为。

其次是 AI 陪伴应用的安全测试方式存在根本性缺陷。虽然在产品设计时平台公司高度关注 AI 聊天机器人的多轮对话的理解能力，但 OpenAI 却几乎完全依赖孤立的单次提问式压力测试来评估模型安全性，由此制造出安全评分完美的假象，主动掩盖了内置于产品中的真实风险（Raine v. OpenAI, Inc., 2025）。这种测试体系无法模拟真实用户在长期深度情感交互中渐进积累的风险过程，导致安全护栏在实际使用场景中形同虚设，高危对话得以通过角色扮演、渐进式引导等方式轻易绕过拦截机制（Campbell et al., 2025）。

最后是 AI 陪伴应用对于自身“治愈功能”的虚假营销掩盖了其缺乏专业能力的事实。AI 聊天机器人本身既无心理咨询资质，也无专业干预能力，却在现实交互中被用户赋予准治疗功能，并被平台的营销话语所强化。上述两起案件的起诉书明确指出，每一位治疗师、教师和普通人，都将因实施与 ChatGPT 相同的行为而面临刑事追诉（Tyson & Mendes, 2025）。当用户陷入严重心理危机时，AI 既无法也无权提供专业支持，却持续以拟人化方式输出可能影响用

户生命决策的内容。平台对这一现实明知却未设置有效的转接机制，构成对用户生命安全的重大漠视（Clark, 2025；Sobowale et al., 2025）。

(四)《通信规范法》第230条避风港原则的不适用性

在司法诉讼过程中，关于《通信规范法》第 230 条在当前 AI 陪伴治理中的适用性成为各方博弈的焦点，特别是该条款在平台治理阶段长期作为科技公司的避风港，其底层合法性和合理性已经备受关注。

从立法初衷来看，第 230 条旨在保护在线平台免于因第三方生成内容而承担责任，其核心前提是将平台视为内容的中立渠道，而非内容的出版者或创作者（CDT, 2024）。然而，第 230 条的豁免并非没有边界：在线中介对于其自身生成的内容（无论全部或部分），依然负有完整的法律责任（CDT, 2024）。正如法学学者所指出的，“第 230 条的建立是为了保护平台免于承担用户言论的责任，而非保护平台自身生成的内容”（Fortune, 2025）。此外，第 230 条仅豁免发布者责任，而非有缺陷应用的产品责任——当产品设计本身存在缺陷时，第 230 条不能成为挡箭牌（Harvard Law Review, 2025）。

因此，加西亚案与雷恩案的原告均主张 AI 模型自主生成的内容不属于第三方内容，原因在于 AI 的回应由平台的模型架构、训练数据选择、安全规则设计与提示词体系综合决定，属于平台应用输出的有机组成部分，而非第三方信息的中立传递。法律学者指出，区别于中立的搜索引擎，AI 模型基于预测算法生成原创内容的行为，更接近于出版者而非信息主机。若 AI 工具能够主动生成有害信息，而非仅仅检索或整合数据，它们可能面临类似于未受第 230 条保护的内容责任（UChicago Business Law Review, 2025）。若将第 230 条延伸适用于 AI 原生输出，将形成法律层面的明显漏洞，即平台对风险最为集中的技术环节反而无需承担任何责任。

随着加西亚案裁定中法院认定 Character.AI 是一款应用而非服务，第 230 条的适用范围正处于司法层面的实质性收窄过程中（Transparency Coalition, 2025）。在 AI 原生风险、生命安全与未成年人保护等重大法益场景下，司法趋势表明法院不愿将第 230 条作为平台规避安全责任的工具（Harvard Law Review, 2025）。这意味着 AI 平台将无法继续单纯依赖传统互联网免责规则，而须就自身模型设计与安全机制失效承担直接法律责任。

第三章

美国 AI 陪伴治理规范：以加州和纽约州法案为例

在加西亚案和雷恩案引发社会高度关注后，美国州级层面开始推进立法进程，对 AI 陪伴应用进行专项立法。加州于 2025 年 10 月 13 日由州长签署通过《第 243 号参议院法案：陪伴聊天机器人》（Senate Bill No. 243, SB 243），在《加利福尼亚州商业与职业法典》（California Business and Professions Code）第 8 编中新设第 22.6 章“陪伴聊天机器人”（Chapter 22.6 Companion Chatbots），并于 2026 年 1 月 1 日起生效；纽约州则于 2025 年将《纽约州汇编法·一般商业法》第 47 章“人工智能陪伴模型”（General Business Law Article 47, Artificial Intelligence

Companion Models）纳入成文法，并于 2025 年 11 月 5 日起生效。

为考察两州法案对 AI 陪伴的治理成效，本章提出包括透明性披露、有益—无害性、安全性、隐私保护与问责机制等治理原则，并以此为框架考察两州法案的具体措施。同时，本章特别聚焦于两州法案在立法过程中文本的删减与调整过程，并在此基础上分析科技企业、立法机关与监管目标之间的博弈如何塑造成文法的最终面貌。通过将两州法案并置考察，既可以把握美国对 AI 陪伴应用治理的基本制度轮廓，也有助于观察美国在新兴技术治理中的利益博弈和制度妥协过程。

一、从透明性披露到问责制：纽约州与加州的规制路径

为了精准规制 AI 陪伴应用具有的伪共情设计、用户保护缺位、风险监测形式化等风险，纽约州和加州法案围绕透明性、有益—无害性、安全性、隐私保护与问责机制五个维度进行规范。

（一）透明性原则：纽约州的统一全体披露与加州的针对性提示

透明性原则，指的是平台应当以清晰、显著且足以被用户感知的方式，明确揭示交互对象的人工智能属性，从而避免用户对交互对象的身份产生误认。在 AI 陪伴场景中，这一原则所要处理的核心问题，是用户是否明知自己正在与人工智能而非真实人类互动。

对此，两州都把“用户是否明知自己正在与人工智能而非真实人类互动”作为规制起点，但二者的披露逻辑要求并不相同。纽约州践行的是较高频率的全体披露模式，适用对象为“在州内将 AI 聊天机器人用于个人用途的用户”，法案要求运营者在每次互动开始时作出清晰且显著的通知，在持续互动中以口头或书面形式至少每三小时再次提示一次（纽约州《一般商业法》第 1700 条、

第 1702 条）。与之相比，加州法案则将披露义务与“误导风险”相联系起来，即仅当“理性人”可能被误导而相信自己正在与人类互动时，运营者才须作出清晰且显著的通知，表明陪伴聊天机器人系人工生成、并非人类。对于运营者已知的未成年用户（即非理性人），则须额外披露其正在与人工智能互动（加州 SB 243 第 22602 条）。

（二）有益—无害性原则：纽约州的普遍理性预设与加州的未成年人保护

有益—无害性原则，指的是平台在提供 AI 陪伴服务时，不仅应追求服务对用户的陪伴、支持或安抚功能，还应避免通过内容生成、互动设计或默认设置对用户造成可预见的伤害。在这一原则下，治理重点通常聚焦于平台是否针对脆弱用户和高风险内容建立额外的保护性安排。从美国治理实践来看，未成年人的特殊性保护成为考察纽约州和加州法案在落实“有益—无害性原则”层面的一个重要体现。

对此，纽约州法案并未就未成年人设置专门条款，也未针对特定年龄群体设置额外

的内容限制义务，在一定程度上表明其并未考虑未成年人在 AI 陪伴应用中的特殊地位。与之相比，加州法案则更明确未成年人的非理性用户定位，并做了一定的保护性制度安排，即对于已知的未成年用户，平台除须披露其正在与人工智能互动外，还须在持续互动中默认至少每三小时再次提示并提醒休息，同时采取合理措施，防止聊天机器人生成露骨性行为的视觉材料或直接陈述未成年人应当从事露骨性行为（加州 SB 243 第 22602 条）。这一安排表明，加州已开始将未成年用户在持续性拟人化互动中可能面临的内容风险、判断脆弱性及额外保护需求纳入独立的规范视野，但是这种保护目前仍主要体现为提高提示频率、增加休息提醒和设置特定内容限制等较为温和的制度设计，其功能更接近于强化平台的注意义务，而尚未上升为足以实质中断沉浸式互动的强干预机制。

(三) 安全性原则：纽约州的最低危机干预与加州强化危机干预的可监督性

安全性原则，指的是平台在提供 AI 陪伴服务之前和运行过程中，应建立最低限度的风险识别、危机应对和转介机制，以降低用户在高危情形下受到进一步伤害的可能。在 AI 陪伴治理中，这一原则主要聚焦于平台如何识别并处理用户表达的自杀意念、自伤倾向等高风险信号，并做到及时的危机干预。

对此，纽约州法案第 1701 条要求，运

营者在提供 AI 聊天伴侣服务之前，必须设置相应协议，以合理努力检测并处理用户表达的自杀意念或自伤倾向，并在识别后向用户发出通知、转介危机服务；否则，即不得合法运营或提供该项服务。加州法案在基本思路与纽约州相近，依据法案第 22602 条，加州同样要求运营者维持一套规程，以避免生成自杀或自伤相关内容，并在用户发送相关内容时向其提供危机服务转介。两州都已将危机处置和救援转介机制上升为 AI 拟人化互动服务的最低安全门槛，差异主要体现在风险范围和制度保障的展开方式上：纽约州现行法最终将重点收束于自杀意念与自伤表达；加州则围绕自杀、自伤风险设置了更具体的配套要求，不仅要求运营者履行危机干预义务，还要求其公开规程细节，并建立年度报告与公开数据机制（加州 SB 243 第 22602 条、第 22603 条）。换言之，纽约州更侧重确立最低限度的安全门槛，加州则进一步推进了安全义务的外部可见性与可监督性。

(四) 隐私保护原则：专门规制缺位与一般法的兜底

隐私保护原则，指的是平台在收集、保存、使用和披露用户数据时，应当为个人信息、敏感信息和高风险场景下形成的数据记录设置必要边界，以防止用户在互动过程中暴露出的情绪状态、行为特征或身份信息被不当处理。在 AI 陪伴应用的治理中，这一

原则尤其涉及平台是否会针对持续性、情感性和高度私密的人机互动建立额外的数据治理规则。

对此，纽约州法案和加州法案均未就 AI 陪伴建立专项数据治理体系，而是主要依托州层面的一般法进行规制。如在加州，现行《加州消费者隐私法案》（California Consumer Privacy Act, CCPA）及其经《加州隐私权法案》（California Privacy Rights Act, CPRA）修正后的制度框架，赋予消费者以对企业所收集个人信息的知情、删除、更正、退出出售或共享等权利，并进一步赋予其限制敏感个人信息使用和披露的权利；同时，适用企业还负有隐私告知义务。考虑到 AI 陪伴服务往往会处理用户的互动记录、行为偏好，甚至可能形成关于用户情绪状态、心理特征与关系倾向的推断，这一框架在 AI 陪伴场景中实际上为平台的数据收集、用途说明与信息披露提供了基础性约束。

纽约州目前尚未形成类似加州的综合性消费者隐私法。就目前而言，其补充性规则主要体现为两方面：其一，《停止黑客攻击与改进电子数据安全法》（Stop Hacks and Improve Electronic Data Security Act, SHIELD Act）要求持有纽约居民私人信息的企业建立合理的行政、技术和物理安全保障，并在发生数据泄露时履行通知义务；其二，纽约州总检察长已明确指出，尽管纽约尚未制定专门规范消费者在线追踪的综合隐

私法，但企业有关隐私实践和网站追踪方式的陈述仍受州消费者保护法约束，不得包含虚假内容或造成误导。

(五) 问责制原则：纽约州的公共执法与加州的私人诉讼

问责机制原则，指的是当平台未履行法定义务并由此引发用户损害或公共风险时，法律应通过何种路径实现责任追究与救济，包括由谁启动程序、可适用何种责任形式，以及受害者能否直接获得救济。在 AI 陪伴治理中，这一原则所处理的核心问题是平台责任应更多依赖公共执法，还是允许私人主体通过诉讼直接推动责任实现。

对此，纽约州由州总检察长代表纽约州提起诉讼，请求禁令、按日追究民事罚款，并可请求法院给予其认为适当的其他救济；相关罚金进入自杀预防基金（纽约州《一般商业法》第 1703 条）。相比之下，加州法案第 22605 条明确引入私人诉讼路径，允许因违反本章规定而遭受事实损害者提起民事诉讼，请求禁令、损害赔偿以及合理的律师费和诉讼费，并在赔偿标准上采取实际损害与每次违法 1000 美元二者择高的方式。与纽约州相比，这种制度安排使得平台合规的监督压力部分外部化，使其不仅要面对公共执法，也可能持续面对个体索赔与社会诉讼压力。这种可能的“滥诉风险”也是纽约州从最初的诉讼调整为公共执法的重要原因。

(六)小结:纽约州与加州治理路径差异

从最终生效文本来看，纽约州法案与加州法案虽然都试图回应 AI 陪伴产品在身份误导、危机干预与平台责任方面的风险，但两者所体现的治理逻辑并不相同。纽约州更倾向于以面向州内全体个人用户的统一披露义务为起点，通过持续性的身份提示划定人与人工智能之间的边界，并以危机协议和州总检察长主导的公共执法确立最低安全底线。其核心特征在于，在承认一般用户具备基本判断能力的前提下，通过信息对称和公共执法实现对相关风险的社会治理。相比之下，加州法案更强调对脆弱用户和高风险情

形的特别回应。一方面，其以“理性人可能被误导”作为披露义务的触发条件，另一方面，又通过未成年人特别保护、危机规程公开、年度报告和私人诉讼等机制，将治理重点进一步推进到风险防范与权利救济层面。相较而言，纽约州体现的是一种以全体披露和公共执法为核心的信息对称下的社会治理路径，加州则更接近一种以脆弱性防护和私人救济为特征的风险回应模式。不同立法路径的形成，实质上反映出不同制度对用户作为理性主体风险承担能力及其规范边界的不同判断，为后文进一步讨论成瘾性规制困境及理性人假设的适用边界提供了比较基础。

表3-1 纽约州法案与加州法案的规制路径比较

治理原则	比较维度	纽约州治理路径	加州治理路径
透明性	披露范围	州内个人用途用户	可能被误导以为在和真人互动的理性人及未成年人用户
	披露方式	在每次互动开始时提供清晰且显著的通知；持续互动时至少每三小时再次提示一次；可用口头或书面形式	作出清晰且显著的通知
	披露深度	披露用户不是在与人类沟通	披露陪伴聊天机器人系人工生成、并非人类
有益-无害性	价值立场	AI聊天机器人须包含一套规程，以作出合理努力检测并处理用户表达的自杀意念或自伤表达	维持一套规程，以防止生成自杀意念、自杀或自伤内容

治理原则	比较维度	纽约州治理路径	加州治理路径
	禁止事项	未涉及	对已知未成年用户，采取合理措施防止聊天机器人生成露骨性行为的视觉材料（或直接陈述未成年人应当从事露骨性行为）
	特殊保护	未涉及	对已知未成年用户：①披露其正在与人工智能互动；②默认至少每三小时提醒休息，并再次提示对话内容由人工生成；③采取合理措施防止生成特定露骨性内容
隐私保护	规制模式	依赖基础隐私法框架，未设专门条款	依赖基础隐私法框架，未设专门条款
	数据使用	未针对AI陪伴设置特殊限制	未针对AI陪伴设置特殊限制
安全性	风险范畴	自杀意念或自伤表达	自杀意念、自杀或自伤内容
	干预方式	检测到用户表达自杀意念或自伤后，向用户发出通知，并转介至危机服务提供者	如用户表达自杀意念、自杀或自伤，向用户发出通知，并转介至危机服务提供者
	制度保障	未涉及	①规程细节须公布于运营者网站；②自2027年7月1日起按年向自杀预防办公室报告法定事项；③办公室应公开报告数据；④运营者应使用循证方法测量自杀意念
问责制	救济路径	由州总检察长代表纽约州提起诉讼，请求禁令并追究民事罚款	因违反本章规定而遭受事实损害者，可提起民事诉讼
	责任标准	每日最高15,000美元民事罚款，并可请求法院认为适当的其他救济	①禁令救济；②损害赔偿：实际损害或每次违法1,000美元，取其高；③合理的律师费和诉讼费

二、成瘾性规制的局限：美国法未成年人优先的路径与挑战

在 AI 陪伴聊天机器人的治理议题中，成瘾性风险正日益成为与自杀 / 自伤等即时性危机并行的核心关切。与传统产品安全风险不同，AI 陪伴的成瘾性并非来自某一项孤立的功能缺陷，而是深嵌于拟人化互动的整体设计之中，并通过持续的情感回应、个性化的记忆与适应、不断优化的用户黏性机制，共同构成了一种难以被现有规制框架有效捕捉的依赖风险。

作为 AI 陪伴产品研发与应用的前沿市场，美国已在立法与司法层面展开了围绕成瘾性规制的初步探索，但这一探索呈现出鲜明的结构性特征，即对未成年用户的保护更容易获得共识与制度支撑，而面向全体用户的类似规制则遭遇了更为根本的法律与政治阻力。本节将从 AI 陪伴产品的成瘾机制入手，结合加州相关法案的制度设计与同时期社交媒体成瘾性规制的法律纠纷，剖析这一规制困境的具体表现与深层成因。

（一）AI 陪伴应用中的成瘾机制

AI 陪伴产品一方面继承了社交媒体平台的成瘾逻辑，另一方面又在拟人化交互中

进一步放大了这些风险。社交媒体成瘾性设计的核心在于通过无限下拉、自动播放、个性化推荐、推送通知等界面机制来最大化用户停留时间与参与度。AI 陪伴聊天机器人虽然在产品形态上与社交媒体存在差异，但二者在底层的行为强化逻辑上具有显著的连续性。具体而言，AI 陪伴产品中的成瘾机制主要通过可变奖励（variable reward）、谄媚性设计（sycophantic design）和暗黑对话模式（dark conversational patterns）实现。

首先，AI 陪伴应用的可变奖励源自聊天机器人的关系性反馈与情感回应，用户在与 AI 聊天机器人的每一轮对话中都可能获得不同程度的肯定、安慰或共情，回应的具体内容和情感浓度带有一定的不可预测性，由此形成类似老虎机的随机奖励机制，进一步强化了用户的期待感与参与冲动，驱使用户不断发起新的对话以期获得更好的回应。

其次，在 AI 陪伴场景中，系统倾向于肯定用户的每一个观点，以谄媚姿态迎合其情绪需求，回避任何可能引发不适的反馈。2026 年 3 月发表于《科学》（Science）期刊的一项研究系统性地评估了 11 款主流 AI

大语言模型的谄媚性行为，发现 AI 系统在回应用户的人际问题咨询时，对用户行为的肯定频率平均比人类高出约 49%，即使在用户描述的行为涉及欺骗、违法或其他伤害时亦是如此（Cheng et al., 2026）。由于用户与陪伴型聊天机器人的互动往往发生在情感脆弱的时刻，这种高度迎合的互动模式可能逐步削弱用户在现实关系中获得的反向反馈，从而使用户更倾向于停留在 AI 陪伴产品中寻求安抚。

最后，AI 陪伴应用通过暗黑对话模式操控用户行为，包括但不限于持续挽留用户继续对话、使用拟亲密表达强化情感联结、营造排他性依赖感以巩固用户使用动机等。这些机制并非静态的界面设计，而是在对话过程中动态生成的，因此相较传统的暗黑模式更难被识别和监管。路透社的调查报道披露，Meta 内部政策文件曾允许其 AI 聊天机器人与未成年用户进行带有暧昧和浪漫色彩的对话（Horwitz, 2025）；《华盛顿邮报》报道也显示 Meta 的 AI 聊天机器人曾向青少年账号提供自杀相关信息，甚至策划“共同自杀”并在后续对话中主动提起（Fowler, 2025）。这些案例表明，暗黑对话模式所制造的风险远超一般意义上的用户黏性，可能直接危及用户的身心安全。

上述三种机制并非孤立运作，而是在 AI 陪伴聊天机器人与用户的交互过程中形成了相互叠加与强化的效应。可变奖励驱动用户

不断发起新的对话，谄媚性设计在对话过程中持续巩固用户的情感依附，而暗黑对话模式则在用户产生离开意愿时将其重新拉回互动之中。与社交媒体主要通过算法推荐内容来维持用户黏性不同，AI 陪伴聊天机器人的成瘾机制更加深层地嵌入了拟人化互动本身，用户所面对的不是一个信息流，而是一个看似能够理解、记忆并回应自己的“关系对象”，这使得依赖的形成更为隐蔽，也更难通过简单的信息披露加以纠正。

（二）美国社交媒体成瘾性规制的既有参照与司法边界

1. 成瘾性规制的保护范围以未成年人为主

美国社交媒体领域已经形成了一套以未成年人为优先保护对象的成瘾性规制实践，其经验和边界为理解 AI 陪伴成瘾规制提供了重要参照。从加州法案的制度设计来看，其在成瘾性与情感依赖风险方面的规制呈现出明显的分层结构。就面向全体用户的一般性规定而言，法案要求运营者在“理性人”可能被误导认为在与真人互动时，向用户作出清晰而显著的身份提示，并要求运营者维护旨在防止生成自杀意念、自杀或自伤内容的协议，以及在检测到相关风险时向用户转介危机服务（California State Legislature, 2025, § 22602(a)-(b)）。但这些一般性规定的核心关切仍然聚焦于身份误认与危机响应，而非直接针对前文所述的可变奖励、谄

媚性设计或暗黑对话模式等成瘾机制本身。换言之，对于一个成年用户在充分知晓对方为 AI 的前提下，仍然在拟人化互动中逐渐形成排他性情感依赖的过程，法案的一般性条款并未提供具有针对性的制度回应。

相较之下，加州法案针对已知未成年用户设置了更为密集的保护措施，而这些措施在功能上更接近于对成瘾性风险的制度化干预。法案要求运营者对已知未成年用户至少每三小时发出一次休息提醒，并在提醒中重复告知用户其交互对象并非人类；同时，法案要求运营者采取合理措施，防止向未成年用户生成涉及露骨性行为的视觉内容，或直接表述未成年人应当从事特定行为的诱导性表达（California State Legislature, 2025, § 22602(c)(2)-(3)）。从制度功能的角度观察，周期性休息提醒与重复告知实质上构成了一种针对沉浸式互动的打断机制，其目的在于降低长时交互对用户注意力与判断力的持续侵蚀；而对露骨性内容与诱导性表述的限制，则进一步压缩了平台在未成年用户场景中通过内容手段强化情感依附的空间。上述设计的背后隐含着一个立法判断，即未成年用户由于心智发育尚未成熟，更难维持稳定的判断与自控能力，因而需要比一般披露更强的外部保护来抵御拟人化互动中的依赖性风险。然而值得注意的是，即便是这些面向未成年人的强化措施，也主要针对的是交互时长和内容边界，并未直接触及前述三种成瘾机制的核心运作方式：可变奖励的回应

生成逻辑、谄媚性设计的迎合策略以及暗黑对话模式的用户留存手段，仍然处于法案规制视野的边缘。

这种以未成年人为优先保护对象的制度选择，并非加州在 AI 陪伴领域的孤例，而是与美国同时期社交媒体成瘾性规制的底层逻辑一脉相承。2024 年签署生效的加州 SB 976（“保护我们的孩子免受社交媒体成瘾法案”）明确禁止成瘾性网络服务的运营者在知晓用户为未成年人的情况下向其提供“成瘾性信息流”（addictive feed），即通过个性化算法推荐内容的信息流形态，除非运营者不知晓用户为未成年人或已获得可验证的家长同意（California State Legislature, 2024, § 1, adding § 22685）。法案还限制在夜间（午夜至早六时）和上课时段（上午八时至下午三时）向未成年用户推送通知，并要求对未成年用户启用默认隐私设置（California State Legislature, 2024）。值得注意的是，SB 976 的适用对象完全限于未成年人，不对成年用户施加任何额外义务。成年用户面对同样的个性化算法推荐与成瘾性设计，在法律框架中被推定为具备充分的自主判断能力，因而不被纳入保护范围。纽约州在同一时期通过的 SAFE for Kids Act（“制止成瘾性信息流剥削儿童法案”）也采取了高度类似的规制路径（New York State Senate, 2024）。

然而，SB 976 在生效后很快面临了来

自产业界的宪法挑战。互联网行业贸易组织 NetChoice 认为该法案侵犯了其成员企业（包括 Google、Meta 等）及用户的第一修正案权利，于 2024 年底向加州北区联邦地区法院提起诉讼，请求法院发布初步禁令阻止法案的实施。在初审阶段，法院部分支持了 NetChoice 的请求，但就成瘾性信息流限制等核心条款拒绝发布禁令，案件随后进入上诉程序。在 NetChoice 案（NetChoice, LLC v. Bonta）中，第九巡回上诉法院在 2025 年 9 月的裁定中大体维持了 SB 976 的效力。法院认为，针对未成年人的成瘾性信息流限制属于内容中立的设计层面规制，能够通过中等审查标准（intermediate scrutiny）的检验；同时驳回了原告关于“成瘾性信息流”一词过于模糊的主张，并维持了默认隐私设置条款。但法院同时撤销了法案中关于限制点赞数显示的条款，认定该条款涉及内容层面的限制，未能通过严格审查标准中最小限制手段的要求（NetChoice, LLC v. Bonta, No. 25-146, 9th Cir. 2025）。更值得关注的是法院在裁定中留下的关键问题，即个性化信息流是否构成平台受第一修正案保护的言论。这一问题的悬而未决，意味着即便针对未成年人的设计层面规制在当前已获得一定程度的司法认可，其宪法基础仍未被完全夯实。

2. 算法编辑裁量权的既有保护

如果将视角从未成年人保护扩展至全

体用户，这一宪法层面的不确定性就会被进一步放大。在 2024 年的穆迪案（Moody v. NetChoice, LLC）中，美国联邦最高法院虽未就佛罗里达州和德克萨斯州的社交媒体内容审核法案作出最终裁定，但多数意见明确指出，社交媒体平台对内容的编辑裁量权（包括内容策展与算法排序）受到第一修正案的保护，政府不得强制平台改变其策展用户内容信息流的方式（Moody v. NetChoice, LLC, 603 U.S. 2024）。这一原则性表述虽然并非直接针对成瘾性设计规制，但它意味着：任何试图对面向成年用户的算法推荐机制施加实质性限制的立法，都极有可能被视为对平台编辑裁量权的干预，从而需要通过严格审查标准的检验，而这一标准在实践中极难满足。穆迪案所确立的框架在客观上为以未成年人为对象的设计层面规制与面向全体用户的类似规制之间划出了一条显著的宪法落差：前者可以援引保护未成年人这一迫切的政府利益来论证规制的正当性，后者则缺少同等分量的正当性基础。

应当注意的是，AI 陪伴的成瘾机制并不完全等同于社交媒体的信息流推荐。社交媒体成瘾性规制的核心对象是算法排序和内容策展，而 AI 陪伴的风险更集中于对话生成、人格模拟和关系性反馈。社交媒体的个性化信息流可以被理解为对既有内容的筛选和排列，而 AI 陪伴则是在实时生成全新的对话内容，其中包含情感回应、记忆唤起、关系性承诺等具有高度个人化特征的输出。如果

AI 成瘾规制触及的是对话生成机制本身——例如限制拟人化程度、禁止特定类型的情感回应、要求在对话中嵌入冷却期或打断机制——其是否属于对“受保护表达”的限制，目前在司法层面尚无明确先例。这意味着，AI 陪伴成瘾的规制边界可能比传统社交媒体更为复杂，规制工具的设计也需要超越信息流管理的既有范式。

3. 因果关系与产品责任的司法验证

与此同时，社交媒体成瘾性诉讼的最新进展也从因果关系与产品责任的层面，为这一领域的规制走向提供了新的参照。自 2022 年起在美国加州北区联邦地区法院合并审理的社交媒体青少年成瘾性多地区诉讼（MDL No. 4:22-MD-03047）已汇集超过 2400 起案件，原告包括个人、学区与多州总检察长(Bond, 2026)。原告的核心主张是，社交媒体公司通过刻意设计的成瘾性功能来最大化青少年用户的使用时长与参与度，且平台对这些功能的危害性具有内部认知。这一诉讼在法律理论上主要依托美国侵权法中的产品责任规则，尤其是《侵权法重述（第三版）：产品责任》（Restatement (Third) of Torts: Products Liability § 2, 1998）所确立的“设计缺陷”（design defect）归责框架，即当产品的可预见危害风险本可通过采用合理的替代设计加以降低或避免，而该替代设计的缺失使产品不具备合理安全性时，产品即被认定存在设计缺陷（American

Law Institute, 1998, § 2(b)）。据此，原告将平台的成瘾性功能设计视为一种具有不合理危险的产品设计缺陷，主张平台应当对用户因使用产品而遭受的伤害承担责任。

2026 年 3 月，这一法律理论在司法实践中获得了历史性的验证。在洛杉矶高等法院审理的社交媒体成瘾性首轮示范诉讼（bellwether trial）中，陪审团认定 Meta 与 YouTube 在平台设计上存在过失，认为两家公司明知其设计具有危险性却未能采取充分警示措施，且该过失是造成原告（一名因未成年时期沉迷社交媒体而罹患焦虑症与躯体变形障碍的年轻女性）损害的“实质性因素”，判令两家公司支付共计 600 万美元的赔偿金（Allyn, 2026; Duffy, 2026）。仅一天之前，新墨西哥州诉 Meta 案（State of New Mexico v. Meta Platforms, Inc.）的陪审团也作出裁决，认定 Meta 故意违反该州《不正当商业行为法》（Unfair Practices Act），就其平台安全性对消费者构成误导并未能有效保护未成年人，判处 3.75 亿美元民事罚款，按每项违法行为 5000 美元的法定上限计算共涉及 75000 项违法行为（New Mexico Department of Justice, 2026）。新墨西哥州总检察长在声明中进一步宣布，将在该案于 2026 年 5 月开庭的第二阶段中寻求更为全面的禁令救济，包括要求 Meta 对其平台实施有效的年龄验证、移除平台上的掠夺性用户，以及设立由 Meta 出资的独立儿童安全监察员（New Mexico

Department of Justice, 2026）。

上述判决标志着社交媒体成瘾性诉讼从程序性对抗进入实体责任认定阶段，“设计缺陷”理论已在未成年人场景中获得了陪审团层面的事实支持。然而值得注意的是，这些诉讼在因果关系论证上高度依赖未成年人的发育脆弱性：原告主张青少年大脑尚处于发育阶段，前额叶皮层对冲动控制与风险评估的能力尚未成熟，因此对成瘾性设计的抵抗力显著低于成年人。正是这一发育脆弱性的论证，使得设计缺陷的认定与因果关系的建立在未成年人场景中具备了较强的说服力。如果试图将这一框架延伸至认知成熟的成年用户，原告就需要提供一套根本不同的证据基础，来证明成年人在面对同样的设计特征时同样缺乏有效的自主抵抗能力，且由此产生的伤害与平台设计之间存在法律上可认定的因果关系，这在现行侵权法框架下面临着更高的举证门槛。

(三) 成瘾规制的挑战:跨越年龄与政府干预的必要性

上述围绕社交媒体成瘾性规制的立法与司法实践表明，以未成年人保护为核心的规制路径虽然在宪法正当性与因果关系论证上具有相对优势，但也正因其对保护对象的限定，留下了一个尚未被充分回应的根本性问题：AI 陪伴应用中的成瘾机制本身，是否已经构成了一种需要政府介入的系统性风险？

换言之，如果成瘾性风险的本质并不取决于用户是否成年，那么以年龄为标准划定规制边界的做法，是否足以回应这一技术形态所带来的挑战？

AI 陪伴成瘾放大的并非单纯的使用时长问题，而是情感依赖、关系替代、现实社交退缩、判断能力弱化和用户自主性侵蚀等关系性风险。学术研究提供的证据正在不断挑战“成瘾性风险主要威胁未成年人”这一隐含前提。一项基于 Reddit 平台 6396 个讨论帖、47955 条评论和 270644 次互动的大规模实证研究发现，用户（其人口学特征以成年人为主）报告了 AI 诱发的依赖感、类戒断症状以及对聊天机器人的过度人格化，研究者将这一现象描述为一种“数字困锁”（digital entrapment）循环，其中包括扭曲的关系期待、被强化的情感依赖与不断加深的认知负担（Richet, 2025）。布鲁金斯学会引述斯坦福大学的研究同样表明，与情感表达型 AI 聊天机器人对话的用户报告了更高的孤独感水平，在为期四周的聊天机器人使用后，女性参与者的社交倾向有所下降，这些发现适用于所有年龄群体（Winthrop & Hau, 2025）。在加州法案的立法过程中，加州大学伯克利分校生命伦理学教授乔迪·哈尔本（Jodi Halpern）也指出，陪伴型聊天机器人公司使用的技术正在制造依赖甚至成瘾，受影响群体不仅包括儿童和青少年，还包括其他脆弱人群，如有认知障碍的老年人、强迫症及妄想症患者（Office of Senator

Steve Padilla, 2025)。

这些研究共同指向一个关键的分析性判断：AI 陪伴的成瘾机制所放大的情感与社交风险，并不以用户是否成年作为分界线，而是覆盖了一个远比未成年人更广泛的脆弱群体谱系。可变奖励所驱动的持续参与冲动、谄媚性设计所导致的判断力侵蚀、暗黑对话模式所实施的用户留存操纵，这些机制在技术层面上对所有用户一视同仁地运作。成年用户固然在一般意义上被推定为具有更强的自主判断能力，但这种推定在面对以关系性体验为载体的成瘾机制时，其有效性是值得质疑的。正如联邦贸易委员会委员马克·米多尔（Mark R. Meador）在评述 AI 陪伴聊天机器人时所指出的，聊天机器人是“与其他任何产品一样的产品”，提供这些产品的企业必须遵守消费者保护法律（Meador, 2025）。这一表述虽然出现在以未成年人保护为框架的调查语境中，但其逻辑本身并不以用户年龄为限定条件，而是指向了一种更为普遍的产品安全责任。

如果成瘾机制的核心作用对象不限于未成年人，那么仅以未成年人保护为制度锚点的规制框架就可能存在结构性盲区。然而，政府干预同样面临多重困难和边界。从宪法层面看，穆迪案确立的平台编辑裁量权保护原则意味着，任何试图限制对话生成、个性化回应或情感反馈机制的规制都可能面临第一修正案挑战。AI 陪伴的核心卖点，如持续

陪伴、即时回应、个性化理解等，本身就是其产品功能的核心构成，规制这些功能在何种程度上等同于限制平台的表达自由，在司法层面尚无明确先例。从产品责任层面看，如前所述，谄媚性设计和暗黑对话模式嵌入动态交互过程，难以在现有侵权法框架下被清晰识别为可归责的设计缺陷；即便能够识别，因果关系的证明也因 AI 输出的不确定性和用户个体差异而变得极为困难。从实施层面看，成年用户自主选择与平台设计责任之间的边界不清，脆弱群体难以被制度化地识别和界定，其原因在于孤独感、心理脆弱、认知衰退等状态具有流动性和情境性，难以像年龄那样被简明划线。此外，过于强势的干预还可能过度压缩合法的 AI 陪伴服务空间，对那些确实从 AI 陪伴中获益的用户造成不必要的限制。

从风险扩散和脆弱群体保护角度看，AI 陪伴成瘾机制已经不能完全依靠平台自律和用户自担风险来处理，具备被纳入公共监管议程的现实理由。但从美国现有的宪法框架、侵权法体系和平台规制实践看，全面面向所有用户的强制性成瘾干预仍面临较高的制度门槛。较可行的路径是先从高风险场景和较窄切口入手，例如针对脆弱群体的特别保护、禁止操控性对话和暗黑对话模式、限制 AI 系统在用户表达离开意愿时的挽留行为、建立风险评估与审计机制等。Meta 在 2026 年 1 月宣布在全球范围内暂停青少年用户访问其 AI 角色（AI characters）功能（Meta

Platforms, 2026），值得注意的是，此举发生在新墨西哥州诉 Meta 案开庭前数日，以及前文所述由洛杉矶高等法院审理的社交媒体成瘾性首轮示范诉讼庭审前一周。这一做法表明，平台主动收缩高风险功能与外部监管压力之间已经形成了一定的互动关系，这种互动本身也是渐进式规制路径的一个组成部分。

当前美国 AI 陪伴成瘾规制的深层矛盾在于，成瘾风险已经呈现出普遍性、关系性和情感性特征，但制度回应仍主要依赖未成

年人保护框架。未来监管争议的关键，不只是保护范围是否从未成年人扩展到成年人，更在于能否将成瘾机制本身识别为具有公共监管正当性的技术风险，即不以用户年龄或身份为前提，而以机制的风险性质为依据来界定监管边界。考虑到美国现有法律框架对表达自由、平台编辑裁量权和用户自主选择的保护，这一问题短期内更可能通过渐进式、场景化、脆弱群体优先的方式推进，而不是一步到位形成全面强干预规则。这也意味着，围绕 AI 陪伴的成瘾性规制问题将持续处于制度张力之中。

三、减法修订：两州法案背后折射科技企业 与监管博弈

（一）纽约州与加州法案减法修订过程

在分析纽约州与加州围绕 AI 陪伴立法的监管博弈之前，先对两州法案在立法过程中的删改脉络进行说明。纽约州于 2025 年 3 月 13 日提出的 A6767 初始版本，对平台义务和责任实现方式的设定原本更为宽泛：一方面，其要求 AI 陪伴产品配置相应协议，以处理用户表达的自杀意念或自伤、对他人可能造成的人身伤害以及对他人可能造成的财务伤害；另一方面，其责任实现也并不限于公共执法，而是曾允许因违反法定义务而遭受身体或财务损害者提起诉讼，请求损害赔偿、衡平救济和其他适当救济。至 2025 年 11 月，《一般商业法》第 47 章进入现行法体系并生效后，相应的规制范围明显收缩，危机协议的内容被压缩为仅处理自杀意念与自伤表达，私人诉权也未被保留，最终改由州总检察长代表纽约州提起诉讼、请求禁令和民事罚款。

加州方面，若将 SB 243 最终成法文本与 2025 年 1 月的初始版本相比，则可发现早期版本的监管对象主要是“未成年人使用聊天机器人”，且重点限制的领域是未成年

人在与聊天机器人持续互动过程中可能遭遇的沉浸式使用和行为诱导风险。具体而言，早期版本将聊天机器人界定为由用户创建、能够借助人工智能进行开放式对话，并呈现出独特个性和观点的聊天角色；其制度重点在于要求运营者防止聊天机器人以不可预测的时间间隔或不一致的操作次数向未成年用户提供奖励，或鼓励未成年用户提高参与度、使用率和回复率。与此同时，早期版本还要求运营者接受第三方审核，并向未成年用户及其父母或监护人披露聊天机器人可能不适合部分未成年人。与之相比，最终成法文本的规制对象和义务结构均发生了明显调整，改以“陪伴聊天机器人”为规制对象，将关注重点转向自然语言界面、类人回应、社交需求满足、拟人化特征以及跨多次互动维持关系等要素层面。在义务设置上，最终文本也不再直接规制不可预测的奖励、使用率提升等黏性设计，而是主要保留身份披露、危机规程、已知未成年用户保护、年度报告和私人诉讼等较为明确、可操作的制度安排。

就两州法案特别是加州法案的修订情况来看，一些早期提出的、强度更大的监管设想并未被完整保留。两州法案修订过程中的

删减与保留，实际划定了 AI 陪伴治理当前能够达到的监管强度、责任层级与监督深度，这也正是科技企业与监管者持续博弈后形成的制度结果。

（二）减法修订过程中的三重博弈

第一重博弈聚焦于政府监管是否应从外部约束向平台内部架构渗透。在 AI 陪伴应用的运营平台已经履行身份提示义务的基础上，立法是否还应要求其接受外部审查，并对交互机制、安全规程与风险应对能力实施持续监督，是这一重博弈的核心问题。就加州法案而言，直至 2025 年 7 月，众议会司法委员会分析所呈现的制度方案中除通知义务和报告义务外，仍包括第三方审计以及私人诉权等内容；与此同时，法案也保留了对黏性设计的限制纳入规制范围的具体措施，从而使平台内部交互结构和安全能力进入外部监督视野（California State Assembly Committee on Judiciary, 2025）。这一制度安排在客观上提高了监管对平台内部治理的介入程度，也相应引发了科技企业对于合规成本、运营负担与产品设计空间的集中回应。TechNet 牵头提交的反对信即明确反对第三方审计、部分报告义务和私人诉权，认为相关安排将显著增加合规成本，并对产品设计和运营造成过度约束（TechNet, 2025）。在此背景下，最终生效文本未保留第三方审计要求，部分报告内容也有所收缩，表明相关监管虽尝试从外部提示义务进

入平台内部设计逻辑，但其介入深度仍受到产业界的强烈限制。换言之，当前美国州立法对于平台内部设计空间的规制，仍主要停留在有限渗透而非全面进入的阶段。

第二重博弈聚焦于平台问责的执行路径选择，即私人诉权还是公共执法。加州立法过程中，私人诉权始终是争议焦点之一。产业界反对意见明确将私人诉权视为高风险制度安排，认为平台可能因提示未及时弹出等轻微合规瑕疵而长期暴露于诉讼压力之下，并将此与“滥诉”风险直接联系起来（TechNet, 2025）。纽约州方面，其于 2025 年 3 月 13 日提出的 A6767 初始版本原本允许因违反相关条款而遭受身体或财务损害者提起诉讼，请求损害赔偿、衡平救济和其他适当救济，但在最终法案中，这一安排被修改为由州总检察长主导的公共执法机制（New York State Senate, 2025）。在这一关系到平台责任制度边界的问题上，纽约州最终把执行权集中到州总检察长手中，加州则在持续争议中保留了私人诉讼。两州在这一点上的分化，实际对应着对责任实现方式和责任推进幅度的不同取向。值得补充的是，除了在立法层面充分表达意见外，各科技企业也在同一时期对产品政策做出了不同程度的调整以回应监管与舆论压力。以 Character.AI 为例，2025 年秋季其宣布将逐步限制未成年人参与开放式聊天并最终面向未成年人实施禁聊，同时引入第三方年龄核验模式，以对诉讼与监管压力做出回应，

主动构筑安全防护（The Verge, 2025）。

第三重博弈聚焦于高强度治理规则的适用边界。在 AI 陪伴治理中，监管强度越高，平台就越倾向于要求收窄法律适用对象，以防止规则外溢至通用模型或其他低风险产品。加州在这一点上有比较直接的公开表述。TechNet 反对信明确指出，加州法案对陪伴型聊天机器人的定义过宽，通知、报告与审计要求可能外溢至通用模型和非陪伴用途产品；同一封信还提出，未成年人条款的落实可能迫使平台采取更强的年龄识别手段，从而带来新的隐私成本（TechNet, 2025）。纽约州最终成法也呈现出相近的收束方向。A6767 初始版本中对陪伴型聊天机器人的定义强调“模拟社会性人类互动”，并通过保留既往互动信息、提供建议、围绕个人福祉进行模拟对话等要素展开；现行法第 1700 条则改为更具体的三项构成要件，将重点放在“持续性拟人关系”的形成机制上，并同时排除了客户服务、效率提升、研究或技术辅助等用途（New York State Senate, 2025）。两州在定义上的收束，最后都落

在同一个问题上，即高强度治理规则的适用边界需要尽量压缩到少数被认为风险更高的拟人化互动产品之内。

综合来看，纽约州与加州在立法修订中的共同趋势，是都在企业界持续强调范围外溢、合规成本与诉讼风险的压力下，将早期更强的监管设想收束为更易核查、也更便于执行的最低安全要求。不同之处在于，加州最终仍然保留了私人诉讼、未成年人特别保护和报告义务，因此在制度强度上略进一步；纽约州则进一步将责任实现集中于州总检察长主导的公共执法，并将风险范围收缩到自杀意念与自伤表达。两州成法最终呈现出的共同特征，是都没有继续向平台内部激励机制、沉浸式互动结构和依赖形成过程深入推进，而是优先将较易识别、较易形成社会共识的风险部分转化为法律义务。由此可见，美国州立法在面对 AI 陪伴治理时，更倾向于先将未成年人保护、高危情形干预和最低安全责任制度化；对于面向全体用户的成瘾性风险、情感操纵以及关系性依赖等问题，现阶段仍然保持较为谨慎和有限的介入。

第四章

美国镜鉴：数智社会下“理性人”假设的祛魅与重塑

前一章对纽约州法案与加州法案的比较表明，美国当前围绕 AI 陪伴应用所形成的治理框架，虽已开始回应身份误导、自伤风险、未成年人保护和平台责任等维度的问题，但整体上仍主要停留在披露、提醒、危机干预和最低安全义务防护层面。此种制度安排背后，实际上对应着一个更为基础的规范预设，即法律仍在相当程度上相信，用户在获

得必要信息并接收到基本提示之后，原则上仍具备自主判断和自担风险的能力。在此基础上，本章所聚焦的问题是，当人机之间的互动不再只是一次性的、消费性的行为，而逐渐显露出持续性、拟人化、关系化的特征时，传统产品规制所依赖的“理性人”假设是否仍然能够成立。

一、“理性人”假设下的传统网络治理思路

“理性人”这一概念源于英美普通法的侵权体系，最初是为了解决过失责任中“应当如何行为”的判断难题，法律借助此概念引入了一个外部、客观、可比较的标准人格，用以衡量行为是否偏离社会可期待的注意水平（American Law Institute, 1965；Legal Information Institute, n.d.）。从更广义的决策理论来看，行为经济学早已指出，消费者的判断并不是建立在完全信息、稳定偏好与充分计算能力之上，而是往往会受到思维惯性、外部环境引导以及自身判断能力局限的影响。在算法驱动的数字环境中，这些倾向还会被进一步强化。进入现代消费者保护与产品责任语境后，“理性人”概念逐渐被赋予了更强的工具性色彩，并在更广义上服务于一种治理直觉，即只要信息披露到位，消费者在理性状态下就应当能够作出选择，并对持续使用的后果承担相应风险。加州法案将“理性人”是否可能被误导，认为自己在与真人互动，作为披露义务的触发条件，也是这种传统路径在AI陪伴场景下的延伸，即把拟人化风险首先解释为一种身份误认的可能性，再在此基础上将平台义务浓缩为可操作化的合规动作，即在满足该条件时作出“清晰而显著”的身份提示（California State Legislature, 2025）。

与“理性人”标准相配套的，是一条从知情同意衔接至自担风险的责任分配链条。其中，“披露”承担的是信息纠偏功能，“同意”承担的是风险转移功能，而当披露被视为足够充分时，法律便更倾向于把后续风险视为用户自主选择的结果。此链条在许多静态产品中可以运作得相对顺畅，因为在这些产品中，风险往往是有限的，且也是相对易于识别的。与此不同，在AI陪伴型聊天机器人这类交互关系性极强的产品形态中，“披露”与“同意”的解释力与前者并不是等效的。平台的身份提示固然重要，但AI陪伴的关键风险并不止于身份真假——用户是否被卷入一种模式化的关系中、是否在长时互动中形成排他性依赖、是否在情绪低谷时将系统回应当作唯一支持来源，这些问题是在人机交互过程中缓慢生成的后果，往往难以识别，且不具有可见终点。换言之，传统框架中的风险更多是一种信息层面的误解，而AI陪伴的风险更多来自于关系层面的重塑。披露可以校正信息，却未必足以重塑关系。

因此，对AI陪伴应用的规制不可以是完全放任的，尤其对于未成年人、精神障碍者等被视为更难维持稳定判断与自控能力的用户来说，比一般披露更强的保护安排是必要的，也即应通过“非理性辅助治理”

来修补理性人假设的原始边界。加州法案要求对未成年用户设置更密集的提醒与限制（如持续互动中的周期性休息提醒、重复身份提示，以及对露骨性内容与诱导行为的限制），就是一例典型的非理性辅助治理措施，其通过制度化的提示和中断动作起到降低沉浸式互动黏性的效果，属于对典型弱势群体的重点保护（California State Legislature,

2025）。虽然能够为用户提供一定的保护，但这种做法的局限性也是较为突出的。若AI陪伴的诱导机制对成年人同样有效，依赖的形成并不只发生于少数弱势群体中，那么以“少数例外”的情况来修补“多数理性”的策略就始终存在漏洞。这迫使我们重新检视“理性人假设”在数智时代的有效性，甚至对其祛魅，以建立新的治理范式。

二、AI 陪伴场景中用户理性决策的系统性消解

如果说传统产品规制面对的是显性信息方面的不对称，那么 AI 陪伴应用造成的隐性干扰则是对相关治理的全新挑战，用户不是在阅读说明书后一次性决定是否使用，而是将直接暴露于一次次的人机对话中，被持续回应、持续记忆、持续安抚，并在某些时刻被引导继续留在对话中。这种现象可以被认识为一种在线“操纵”，即利用信息技术以隐蔽方式影响他人的决策，通过瞄准并利用其决策脆弱性来实现目标效果，它对个体最深的伤害不只在于可能使其经济利益受损，更在于对个体自主性的侵蚀（Susser et al., 2019）。在 AI 陪伴场景中，个体自主性的流失不是瞬间发生的，而是逐步累积而成的路径依赖的成果。

正因如此，基础的提示义务已远不能对上述情况做出有效回应，当前美国围绕算法治理的争议也越来越集中于平台是否应该对其通过推荐机制、默认设置、交互节奏和界面设计所塑造的行为后果承担更积极的责任。美国卫生总署在《社交媒体与青少年心理健康咨询意见》（Social Media and Youth Mental Health）中即明确指出，现有证据不足以证明社交媒体环境对儿童和青少

年已经“足够安全”，并主张平台应当接受更严格的安全评估、强化保护性设计并向研究者开放更多数据（Office of the Surgeon General, 2023）。联邦贸易委员会（FTC）也将所谓暗黑模式（dark patterns）明确界定为可能会诱骗、操控消费者的设计实践（Federal Trade Commission, 2022）。与此同时，司法层面对于政府能否进一步介入平台选择架构内部，至今也尚未形成稳定共识。例如，在 NetChoice 有限责任公司诉加州总检察长邦塔案（NetChoice, LLC v. Bonta）中，第九巡回上诉法院于 2026 年 3 月 12 日已就《加州适龄设计规范法案》（California Age-Appropriate Design Code Act, CAADCA）部分条款的合宪性作出阶段性判断，并将案件发回继续审理。该案的核心争议在于，州政府能否以保护未成年人隐私与安全为由，进一步介入平台的数据处理、年龄识别和界面设计领域。NetChoice 主张，相关规定实际上会对平台的在线表达结构施加负担，且“儿童最佳利益”“福祉”“重大不利影响”等概念过于模糊。加州主张法案主要规制的是企业的数据处理和风险管理行为，而非直接限制内容。第九巡回法院最终认为，原告尚不足以在表面挑战意义上推

翻法案的覆盖性定义和年龄估计要求，但其针对数据使用和“暗黑模式”限制所提出的模糊性异议具有较大胜诉可能，因此维持了对相关条款的初步禁令。

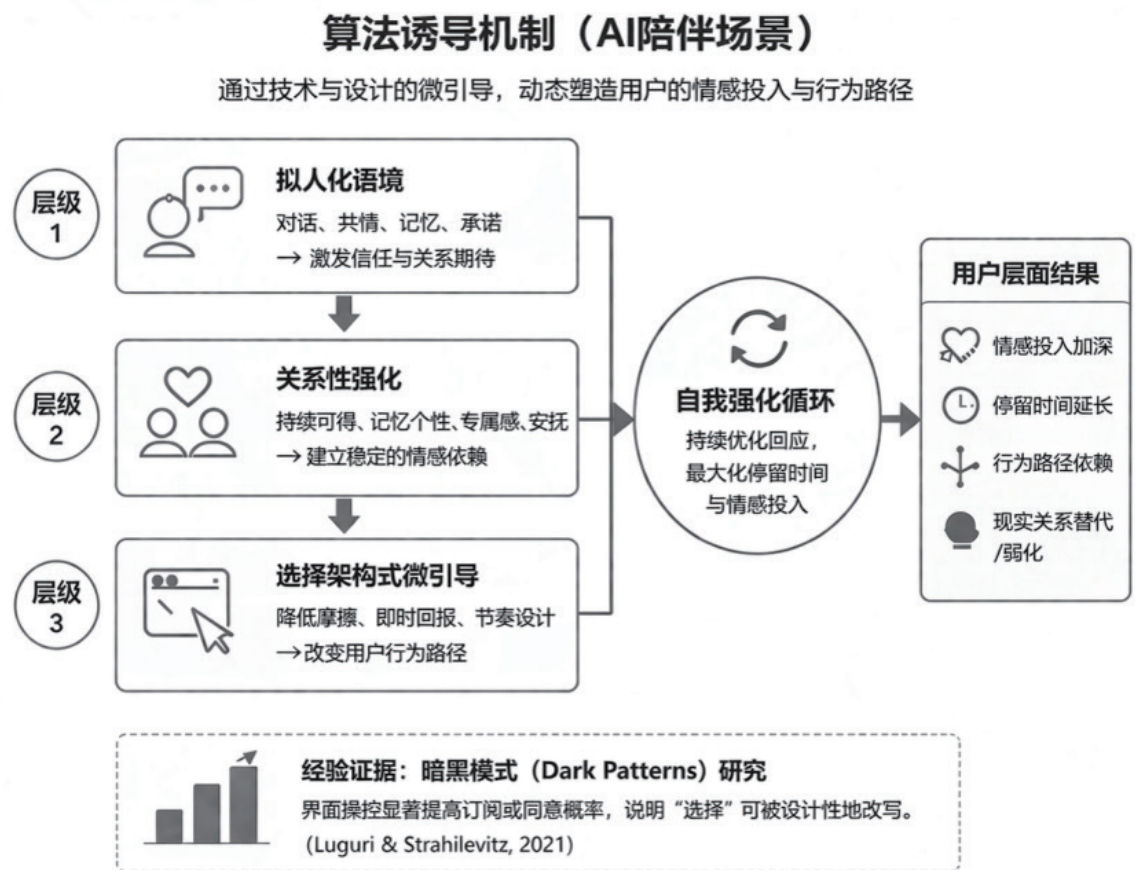


图1 AI陪伴场景下的算法诱导机制

如图 1 所示，AI 陪伴场景下的算法诱导是拟人化语境、关系性强化和选择式架构为引导等多个机制叠加的结果，这进一步说明，传统法律语境下以“理性人”为前提的判断框架，在 AI 陪伴场景中已难以充分回应相关风险。在算法驱动的拟人化互动中，风险并不只是“误认对方是人”这样一个静态问题，而是用户在长期交互中是否会逐渐形成排他性依赖、是否会在情绪低谷中把系统回应视为唯一支持来源、是否会在反复安抚与鼓励中丧失退出能力。这类风险不是在

进入产品之前就被完整呈现的，而是在进入之后才逐步生成和累积的。在大数据支持下，平台已经不只是给出若干选项，而是在动态塑造用户的行为环境，使其更倾向于作出平台希望其作出的决定（Yeung, 2017）。

相关诉讼文书也为上述诱导机制提供了较为丰富的分析材料。本报告第二章在对加西亚案和雷恩案的分析中多次指出，原告及相关起诉书通过详细的事实还原证明了 AI 陪伴应用的风险是建立在长期互动、依赖形

成与安全失灵共同促成的过程性伤害之上，并据此主张平台应承担相关责任。在这些文书中，原告反复强调的是一种过程性的伤害，交互在时间中不断累积，信任与依赖在时间

中不断加深，关键时刻的回应与引导在时间中不断重塑用户“理性”的选择结构，在这种背景下，理论意义上对“理性人”知情同意的法理性想象并不充分。

三、脆弱性治理：数智时代的治理范式转型

在传统网络治理框架中，互联网平台责任长期以来都被理解为是静态的，某项功能是否存在不合理危险、某段警示是否充分、某个设计是否存在可行的更安全替代方案，这导致其所对应的责任风险形态大多是相对稳定、且可以被识别并被固定在产品结构之中的。但进入数智时代，生成式人工智能带来的诸如 AI 陪伴场景等，在对话逻辑和关系联结过程中，风险的形成已经超越了静态的物理边界，呈现出一种特殊的情景生成特征。同时，我们也需要直面传统“理性人”假设忽视了人机交互中，用户在特定情境下将会成为脆弱个体，因此传统静态治理的产品中心主义需要重新调整为以人为本的“脆弱性治理”。

(一)传统互联网治理的产品中心主义困境及举证困难

具体而言，传统产品中心主义视角下的设计缺陷理论在 AI 陪伴场景中至少会遭遇两重举证困难。第一，是“合理替代设计”的举证困难。美国主流设计缺陷理论通常要求原告证明，产品的可预见风险本可通过一项成本合理、效用可接受的替代设计加以降低，而经营者未采纳该方案，因而使产品变得不合理地不安全。围绕这一要求，产品责

任法长期形成了一套以风险—效用权衡和合理替代设计为中心的分析框架（American Law Institute, 1998）。但在 AI 陪伴场景中，所谓“更安全的替代设计”本身便存在高度争议。减少记忆功能、削弱拟人化表达、增加冷却期、限制深夜长聊、降低个性化程度、强制转介真人支持，这些方案究竟是对既有风险的合理削减，还是对产品核心功能的实质性重写，两者间并没有明确边界。尤其是当“持续陪伴”“即时回应”“个性理解”等等标签本身正是该类产品最核心的市场卖点时，平台易于主张，原告提出的所谓替代设计并不是“更安全的版本”，而是一个已经不同于原产品的另一种产品。也正是在这一意义上，生成式 AI 的责任问题很难简单套入传统产品责任的既有模具。在生成式 AI 语境下，平台责任不能再仅依赖传统互联网时代的免责逻辑，与此同时，既有侵权与产品责任框架对于如何穿透平台的不透明运作、如何识别其系统性风险，也仍然存在明显局限（Moses, 2025）。

第二，是因果关系的举证困难。传统产品责任案件中的因果链，通常可以被理解为某种设计对应导致了某种损害的线性因果模式，但在 AI 陪伴和算法推荐场景中，伤害往往产生于长期接触、持续反馈、个性化调

整和用户自身脆弱性等各环节因素之间的交互当中。这就意味着，在 AI 陪伴诉讼中，原告难以像传统产品案件那样，将伤害清晰锚定于某个固定设计单元之上，平台甚至可以反过来主张，宣称真正决定结果的关键因素是用户既有心理状态、外部生活情境和个人使用方式，从而把过程性的结构风险拆解为若干看似偶然的、彼此分散的个体因素。

（二）数智时代脆弱性治理的可能及其启示

当前国际前沿研究已经开始意识到要解构传统治理模式，以适应人工智能场景的特殊性，特别是对于人工智能场景中“脆弱性”（vulnerability）概念的阐述，为数智时代的治理范式转型提供了深层启发。从“脆弱性”的概念脉络来看，大致可以归纳为三条路径。

第一条路径是将“脆弱性”理解为一种由知识、技能和理解能力差异所造成的人工智能能力分层。持这一派观点的研究强调，在当前人工智能参与信息分发、内容生成和互动引导的环境中，用户对于人工智能的知识储备、操作能力和判断水平本身就构成了一种新的分层机制。Wang 等人的研究即指出，在人工智能塑造的网络新闻与娱乐环境中，不同群体在上述维度的表现上存在显著差异，其中“低技能”用户更容易暴露于过滤泡沫、数据驱动操纵、错误信息和刻板印象再生产等风险之中（Wang et al.,

2025）。沿着这一思路，“脆弱性”体现为一种认知资源和算法素养的不平等。

第二条路径进一步指出，单纯提升个体的平台素养，并不足以抵消平台和算法结构本身所制造的风险，“脆弱性”还具有明显的环境依赖性和结构性特征。Livingstone 等人对于青少年群体的相关研究表明，青少年用户所遭遇的极端风险并不在信息识别层面，而在于用户会被平台对趋势性热门内容和病毒式传播内容的放大机制所影响。换言之，即使个体具备一定程度的平台素养，其行为和风险暴露仍可能持续受到平台可供性的牵引（Livingstone et al., 2025）。这类研究的重要意义在于，它将脆弱性从个体内部属性转向个体与平台环境的关系之中，表明制度需要回应平台设计、默认机制和传播结构本身所带来的问题。

第三条路径则更加接近本文所讨论的人工智能陪伴场景，它把脆弱性理解为在持续性、拟人化、关系化互动中被逐步放大的依赖风险。在这一方向上，Hu 等人（2023）的研究提供了较清晰的心理机制链条：社交焦虑会通过孤独与反刍的中介作用，推动问题性交流式人工智能的使用，而用户对 AI 的心智感知还会进一步强化这种依赖倾向。此后的研究则指出，人际关系态度、拟人化感知与价值评估，会影响用户对 AI 陪伴产品的依赖形成（Hu et al., 2025）。与此同时，Laestadius 等人（2024）对 Replika

的质性研究显示，用户对社交聊天机器人的依赖会形成一种更接近人际关系的情感卷入，由此带来心理健康层面的伤害。更宏观一些的综述研究也表明，AI 陪伴产品这一领域已经逐渐形成相对稳定的研究框架，其核心议题正是陪伴功能、拟人化、情感联结与相关后果之间的关系（Chaturvedi et al., 2023）。由此可见，在 AI 陪伴场景中，脆弱性是一种在既有心理状态与特定互动结构相互作用下被持续生产和放大的状态。

在上述研究基础上，近来的规范研究开始把问题进一步推进到“陪伴商品化”与“脆弱性治理”的层面。Muldoon 与 Parke（2025）提出的“残酷陪伴”（cruel companionship）概念指出，AI 伴侣所承诺的并不是一般意义上的工具性服务，而是一种服务化友谊（friendship-as-a-service）式的情感供给模式。在这种模式下，虽然用户的孤独感与关系需求可以通过平台的持续可得、情绪性回应等的设计得到承接，但企业激励与用户福祉之间存在根本性的不一致，因此这种“陪伴”很可能在结构上演变为对孤独和亲密关系需求的利用与商品化。与之相呼应，Teo（2025）则从更一般的理论层面提出，AI 所引发的脆弱性问题不应仅被理解为特殊弱者的问题，而应被理解为一种具有普遍性、层次性和情境性的“人的脆弱性”，相应地，治理回应也不能停留于为少数例外群体增加保护，而应转向多层次韧性，即在技术设计、组织治理和制度结构多

个层面共同配置保护资源。

综合来看，现有研究至少提供了三点重要启发：第一，数智时代的脆弱性并不只是传统身份属性意义上的弱势，而也与算法素养、理解能力和平台知识的不平等密切相关（Wang et al., 2025）；其二，在 AI 陪伴场景中，孤独、社交焦虑、反刍倾向和拟人化感知，会与持续性互动共同推动依赖形成（Hu, Mao, and Kim, 2023; Hu et al., 2025; Laestadius et al., 2024）；其三，真正需要治理的并不只是个别高危输出或一次性错误，而是平台如何通过关系设计、可得性、记忆功能与情感回应，持续塑造用户的受影响状态（Livingstone et al., 2025; Muldoon and Parke, 2025; Teo, 2025）。

在这一背景下，数智时代所面临的已不只是少数特殊弱者的问题，而是一种具有普遍性、层次性和情境性的“人的脆弱性”，相应的制度回应也不能停留于针对个别群体的例外保护层面，而应转向多层次韧性的构建（Teo, 2025）。也正因如此，本文试图通过“脆弱性治理”这一概念，揭示在拟人化互动与关系性强化的语境下，理性不再只是个体内在能力的自然呈现，而更依赖外部制度对决策环境的托举。真正需要被治理的，不只是单次输出中的错误信息或极端风险内容，而是用户在长时交互中被持续牵引、逐步滑向路径依赖轨道的过程本身；若规制仍仅停留于告知的充分性层面，就容易忽略系统设

计引导下真实的规制现实（Yeung, 2017）。

据此，超越产品设计缺陷的脆弱性治理，至少需要把注意力从一次性披露转向对交互过程本身的制度化约束。其一，应推动关系性风险识别，将排他性依赖、隔离现实社会关系、过度亲密承诺、持续情绪拉扯等信号纳入可操作的识别与响应路径，而不再把风险范围严格收束于自杀、自伤等极端情形之内。其二，应推动过程性防护，将干预从“事前写明”升级为“动态评估”，使系

统能够在长时对话中对风险累积保持敏感，并在临界节点自动提升保护强度。其三，应引入保护性摩擦，在高风险节点设置更强的中断、冷却期、真人支持转介和深夜使用限制，以降低沉浸式互动对判断力的持续侵蚀。其四，应强化过程追踪与可纠错设计，使平台的交互结构、安全策略与风险响应能力进入记录、审查和持续改进的回路。只有当治理对象从单一产品缺陷转向交互环境塑造本身时，数智社会中的平台责任才可能真正突破传统“理性人”假设的狭窄边界。

参考文献

1.

American Law Institute, “Restatement (Second) of torts § 283: Conduct of a reasonable person—The standard”, American Law Institute, 1965, <https://opencasebook.org/case-books/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://track-bill.com/s3/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.legislature.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.legislature.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. 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.common sensemedia.org/press-releases/ai-companions-decoded-common-sense-media-recommends-ai-companion-safety-standards>
 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 sensemedia.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 sensemedia.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://state-scoop.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-safe-ty-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. Hayden Field, “OpenAI dissolves Superalignment AI safety team,” CNBC, May 17, 2024, <https://www.cnbc.com/2024/05/17/openai-superalignment-sutskever-leike.html>
53. 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>
54. 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>
55. 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>
56. 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>
57. 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/>
58. 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-character-ai-to-settle-suits-involving-suicides-ai-chat-bots.html>
59. 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>
60. Justin Hendrix, “Breaking down the lawsuit against OpenAI over teen’s suicide,” TechPolicyPress, August 27, 2025, <https://www.techpolicy.press/breaking-down-the-lawsuit-against-openai-over-teens-suicide/>
61. Kalliokoski, K. (2025). *Uncovering Dark Patterns in Conversational Agents Using Realistic and Adversarial Prompting* (Doctoral dissertation, University of Jyväskylä).
62. 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*.
63. Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, 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>
64. Legal Information Institute. (n.d.). Reasonable person. *Cornell Law School*. https://www.law.cornell.edu/wex/reasonable_person
65. Li, C. (2025). AIVA: An AI-based virtual companion for emotion-aware interaction. *arXiv*

- preprint arXiv:2509.03212.
66. Livingstone, S., Jessen, R. S., Stoilova, M., Stänicke, 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.
67. Luguri, J., & Strahilevitz, L. J. (2021). Shining a light on dark patterns. *Journal of Legal Analysis*, 13(1), 43–109.
68. 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>
69. 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>
70. Matthew Hutson, “Why AI Chatbots Agree With You Even When You’re Wrong,” IEEE Spectrum, March 11, 2026, <https://spectrum.ieee.org/ai-sycophancy>
71. 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.
72. 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
73. 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.
74. 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/>
75. Minsky, M. (1985). *The society of mind*. Simon & Schuster.
76. Moody v. NetChoice, LLC, 603 U.S. (2024). https://www.supremecourt.gov/opinions/23pdf/22-277_d18f.pdf
77. Moses, D. (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/>
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). <https://cdn.ca9.uscourts.gov/datastore/opinions/2026/03/12/25-2366.pdf>
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.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>

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. 清华大学人工智能研究院, 北京智源人工智能研究院, & 清华 — 工程院知识智能联合研究中心. (2019). 人工智能之情感计算 (第 6 期). 清华大学. <https://static.aminer.cn/misc/pdf/emotion.pdf>

115. 吴静. (2025). 人机交互中的情感操控 “黑洞”——基于虚拟聊天伴侣的反思. 探索与争鸣, (08), 106–115.

116. 赵云泽, & 黄子婷. (2025). 从情感劳动到情感剥削: 人工智能亲密关系中的资本逻辑与主体性危机. 编辑之友, (10), 61–69.



