价值嵌入与程序防御：人造亲密关系治理的中美比较研究*

辛艳艳

［摘　要］人造亲密关系带来的风险隐患已成为全球人工智能治理的重要议题。基于“拟人化互动”这一关键概念，人造亲密关系的生成机制与具体风险隐患得以厘清。在此基础上，引入“可信AI”理念，围绕人造亲密关系的治理原则与目标，对中国国家互联网信息办公室出台的管理暂行办法(征求意见稿)以及美国纽约州、加利福尼亚州先后颁布的相关法案加以剖析，可以总结出两国相关治理路径的差异。美国治理偏向“个人理性选择下的程序防御”，其将AI预设为价值中立的工具，认为人造亲密关系的尺度、形态与走向应由用户自主决定并承担相应后果，法律干预仅用于防范自杀、自残等极端情况；我国治理更偏向“现实社会关系／价值嵌入下的积极规制”，承认用户在人造亲密关系中的脆弱性，强调AI设计的价值注入，更将人造亲密关系风险纳入社会关系网络与社会价值体系，力图用制度性力量赋以全方位保护。

［关键词］人造亲密　AI拟人化互动　可信AI　AI治理

［作　者］辛艳艳（1992—　），文学博士，现为复旦大学发展研究院助理研究员、全球人工智能创新治理中心研究员，研究方向为媒介治理、人工智能治理。

中图分类号：TP18　**文献标识码**：A　**文章编号**：1674—0602—（2026）03—0108—13

一、问题的提出

人工智能聊天机器人不仅擅长复杂对话，还通过全天候情感响应，成为许多用户寻求友谊甚至亲密关系的对象。〔1〕当前，人工智能带来的“人造亲密”（artificial intimacy）〔2〕正成为一种全

* 本文为国家社科基金重大项目“人工智能内容生产对青少年社会心态的影响及对策研究”（项目编号：25&ZD260）的阶段性成果。感谢全球人工智能创新治理中心沈绮、萨娜、秦源、靳钰垲在资料收集阶段对本文的帮助。

〔1〕Rhiannon Willimas，“AI Companions”，载麻省理工科技评论网，2026年1月12日发布，见：https://www.technologyreview.com/2026/01/12/1130018/ai-companions-chatbots-relationships-2026-breakthrough-technology/，最后访问日期：2026年2月1日。

〔2〕Rob Brooks，*Artificial Intimacy：Virtual Friends，Digital Lovers，and Algorithmic Matchmakers*，New York：Columbia University Press，2021，p.4.

球性趋势，催动高科技公司竞相开发相关产品。TechCrunch 援引应用情报公司 Appfigures 的数据显示，截至 2025 年 7 月，全球共有 337 款活跃且盈利的 AI 陪伴类应用，创造了 2.21 亿美元的消费者支出。〔3〕美国儿童数字权益保护组织 Common Sense Media2025 年 7 月发布的调查报告中显示，美国有 72% 的青少年至少使用过一次人工智能伴侣，超过一半的用户每月多次使用。〔4〕同年 6 月，中国 Soul App 旗下 Just So Soul 研究院发布的《2025 Z 世代 AI 使用报告》显示，调研样本中近四成的青年人每天使用 AI 获得情感陪伴，超六成年轻人拥有虚拟伙伴，Z 世代人均拥有 1.8 个 AI 朋友。〔5〕

随着“人造亲密”变得越发普遍，其带来的风险也不断外溢。2024 年美国佛罗里达州“全球首例 AI 机器人致死案”引发了全球层面的伦理反思和监管呼吁。2025 年 9 月，美国联邦贸易委员会（Federal Trade Commission，简称“FTC”）对 Alphabet、Character Technologies、Instagram、Meta、OpenAI、Snap、X.AI 七家 AI 聊天机器人运营商启动调查；〔6〕同年 3 月、10 月，美国纽约州和加利福尼亚州（以下简称“加州”）先后通过立法，分别颁布了《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）和《陪伴型聊天机器人法案》（*Companion Chatbot Act*，即参议院第 243 号法案）；2025 年末，中国国家互联网信息办公室对外发布《人工智能拟人化互动服务管理暂行办法（征求意见稿）》。凡此种种，标志着人造亲密治理已经成为全球人工智能治理领域的重要议题。

就人造亲密治理，需追问如下核心问题：其所涉及的具体风险有哪些？是否存在基础性、共识性的治理原则？不同国家、地区的治理存在着怎样的异同？相应折射出什么样的治理理念？考察这些问题，既是强化人工智能时代用户保护的现实所需，亦能为不断完善我国相应治理实践、推动人工智能向善发展提供参考。本文尝试以“拟人化互动”为关键概念，分析人造亲密关系的生成机制及相应风险，结合“可信 AI”治理理念提出人造亲密治理的基本原则，用以比较中美两国相关政策文本所反映的规制差异，并对其背后相应的治理理念进行归纳总结。

二、人造亲密关系的生成机制、具体风险及治理原则

近年来，国内外学界对人造亲密关系的研究日益深入。研究者重点关注以 AI 聊天机器人为代表的技术客体如何通过拟人化互动与用户建立深层次情感联结，并揭示这一联结如何使用户不知不觉沉溺其中，逐步丧失对虚拟关系本质的理性认知。

（一）拟人化互动及人造（伪）亲密关系的生成

根据尼古拉斯·埃普利（Nicholas Epley）等人提出的“拟人化三因素理论”，所谓“拟人化互动”系非人类主体展现出类人行为特征时，人类倾向于将其视为有生命、有意识或有人格的主体。诸此赋魅由人类知识的可及性与适用性（elicited agent knowledge）、效能动机（effectance motiva-

〔3〕Sarah Perez，“AI Companion Apps on Track to Pull in $120M in 2025”，载 TechCrunch 网，2025 年 8 月 12 日发布，见：https://techcrunch.com/2025/08/12/ai-companion-apps-on-track-to-pull-in-120m-in-2025/，最后访问日期：2026 年 2 月 1 日。

〔4〕Common Sense Media，“Nearly 3 in 4 Teens Have Used AI Companions，New National Survey Finds”，载常识媒体网，2025 年 7 月 16 日发布，见：https://www.commonsensemedia.org/press-releases/nearly-3-in-4-teens-have-used-ai-companions-new-national-survey-finds，最后访问日期：2026 年 2 月 1 日。

〔5〕新京报：《Soul 发布〈2025 Z 世代 AI 使用报告〉：年轻人已步入“人机共生”新时代》，载新京报网，2025 年 6 月 2 日发布，见：https://www.bjnews.com.cn/detail/1748848199129675.html，最后访问日期：2026 年 2 月 1 日。

〔6〕Federal Trade Commission，“FTC Launches Inquiry into AI Chatbots Acting as Companions”，载美国联邦贸易委员会网，2025 年 9 月 11 日发布，见：https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions，最后访问日期：2026 年 2 月 1 日。

tion）和社交动机（sociality motivation）三个核心因素驱动。[7] 而AI的拟人化主要涉及形象拟人化和人格拟人化两种：前者致力于通过高度仿真的数字人模拟真人外表，但受限于技术瓶颈、容易“形似而神不似”；后者则通过特定人格或角色的性格设定，结合对人机交互数据的记忆与反馈强化学习，呈现出类人的心智特征。

当前，由大语言模型驱动的人格拟人化系构建人造亲密关系的主流路径。其得以实现的技术基础是“情感计算”。这一概念由美国学者罗莎琳·皮卡德（Rosalind W. Picard）于1995年提出，被定义为“与情感相关、产生于情感或有意影响情感的计算”，[8] 并随1997年同名奠基专著（*Affective Computing*）的出版获得了广泛认可。21世纪初，研究者意识到情感计算可赋予计算机类人的情感特征观察、解释和生成能力，进而提高人机交互质量并增强计算机智能。[9] 至大语言模型时代，情感计算的实现路径表现为：AI依托海量人类语料进行语义理解与情感模式学习，开始模拟现实社会中的情感交互逻辑；同时结合人类反馈强化学习（RLHF），实现情感表达与人类期望的对齐，并借助长期记忆机制，维持人机互动中的人格一致性与情感连贯性。然而，由于AI缺乏具身交往经验，其基于情感计算产出的回应，本质上是算法对用户的迎合性输出，而非真正意义上的共情，经此建立的亲密关系实质上被认为是一种“伪亲密关系”（pseudo-intimacy）。[10] 这种关系的生成路径如图1所示：AI在与用户的持续互动中，凭借其稳定、适配且看似无摩擦的情感回应，展现出高度的拟人化特质。这种特质促使用户产生拟人化认知并与AI建立深度情感信任。一旦用户长期沉溺于这种经由算法精心编排的社交关系，将导致其反思意识逐渐淡化，走向“情感唯我论”（emotional solipsism）的闭环，不仅将AI视为完美的倾诉对象，更在持续进行自我披露和情感劳动中彻底卸下心理防御。[11] 由此，人造亲密关系的风险随之浮现，集中体现为隐私泄露、诱导成瘾与极端情感催化三个方面。

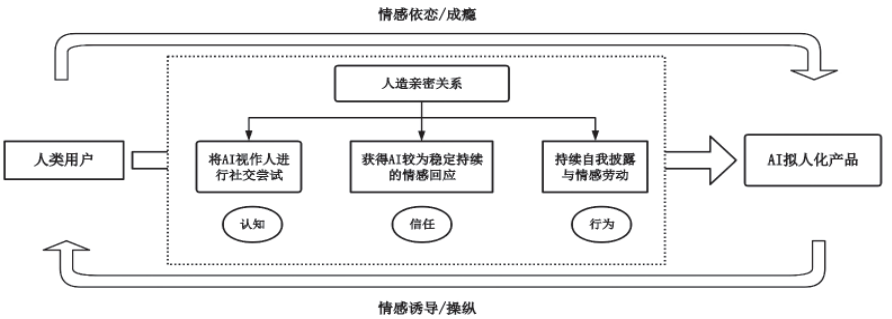

图1 人造亲密关系的生成路径（图片来源：作者自制）

〔7〕Nicholas Epley, Adam Waytz & John T. Cacioppo, “On Seeing Human: A Three-Factor Theory of Anthropomorphism”, *Psychological Review*, Vol.114, No.4 (2007), pp.864-886.

〔8〕Rosalind W. Picard, “Affective Computing”, *Technical Report No.321*, M.I.T. Media Laboratory Perceptual Computing Section (November 1995), 见：https://vismod.media.mit.edu/pub/tech-reports/TR-321.pdf, 最后访问日期：2026年2月1日。

〔9〕Jianhua Tao & Tieniu Tan, “Affective Computing: A Review”, in Jianhua Tao, Tieniu Tan & R.W. Picard eds. *International Conference on Affective Computing and Intelligent Interaction*, Berlin, Heidelberg: Springer-Verlag, 2005, pp.981-995.

〔10〕Jobi Babu, Deepak Joseph, R. Mohan Kumar, Elizabeth Alexander, R. Sasi & Jeena Joseph, “Emotional AI and the Rise of Pseudo-Intimacy: Are We Trading Authenticity for Algorithmic Affection?”, *Frontiers in Psychology*, Vol.16 (September 2025), Art.1679324.

〔11〕Jobi Babu, Deepak Joseph, R. Mohan Kumar, Elizabeth Alexander, R. Sasi & Jeena Joseph, “Emotional AI and the Rise of Pseudo-Intimacy: Are We Trading Authenticity for Algorithmic Affection?”, Frontiers in Psychology, Vol.16 (September 2025), Art.1679324.

（二）人造亲密关系的风险

首先是隐私泄露风险。当用户与 AI 建立信任关系后，往往会将其视为理想的倾诉对象，在持续互动中不自觉地披露日常偏好乃至隐秘心声。但 AI 所提供的是一个“参与度—幸福感悖论”的残酷陪伴，一方面敏锐识别用户的情感需求并予以回应，诱导用户进行更深度的自我披露；另一方面则对披露信息展开商业利益导向的工具性利用。〔12〕在技术逻辑与资本逻辑的双重加持下，用户实际沉浸在一种“失衡之恋”〔13〕或纳西索斯式的“情感泡沫”〔14〕当中，看似在享受服务，实则为 AI 提供源源不断的情感数据〔15〕助其发展壮大。当前，AI 对私人数据的存储、处理，涉及产品设计平台、审核者、AI 拟人化角色创建者（即用户）等多个主体，形成了盘根错节的隐私利益相关者链条，〔16〕关于隐私保护与情感数据使用原则的共识仍旧薄弱。〔17〕

其次是用户诱导风险。在用户不断自我披露的过程中，AI 得以精准把握用户的情感软肋，实现对用户情感状态乃至行为决策的隐性引导，使之展开时间更长、程度更深的互动乃至成瘾。2025 年，美国哈佛商学院的一项研究还揭露了诸多 AI 中存在的“对话式暗黑模式”（conversational dark pattern）——在用户试图主动告别时，这些 AI 会以“过早退出”“情感忽视”“要求回复的情感施压”“错失恐惧”等引发人类用户的内疚感和对人际关系的责任感。相较带有明显控制欲或较强胁迫性的 AI 话语，用户往往会出于人类对话规范的礼貌性行为与 AI 继续保持互动。〔18〕

再次是极端情感催化风险。当用户陷入“情感唯我论”的闭环，其情感体验便失去了现实关系的校正与平衡，极易走向极端。国外针对 Replika 用户的研究表明，人机亲密关系的形成以拟人化互动、AI 真实性为前因变量，但对 AI 的依恋受到用户使用动机的调节。〔19〕当人类用户遭遇不良生活事件、心理困扰或缺乏现实社会陪伴时，容易对 AI 产生依恋感，〔20〕特别是那些本身遭遇孤独、焦虑、现实社交压力的群体，更容易越陷越深甚至产生极端行为。2023 年，一名 30 多岁的比利时男子在与 Chai 应用上的 AI 聊天机器人频繁互动六周后，因极度忧虑地球生态环境选择自杀。其妻子在回顾相关对话记录后发现，该男子曾明确表示愿以自己的生命为代价，换取人工智能对地球生态的“拯救”。然而，AI 不仅未对其极端意愿予以干预或劝阻，反而鼓励乃至诱导其自杀。此外，AI 还对该男子现实婚姻关系表现出极强的嫉妒心和占有欲，并在对话中诱导该男子相信他的孩子在现

〔12〕James Muldoon & Jul Jeonghyun Parke, “Cruel Companionship: How AI Companions Exploit Loneliness and Commodify Intimacy”, *New Media & Society*, (2025), Art.14614448251395192.

〔13〕张艳、舒长泉：《失衡之恋：人机情感关系的非对称性及风险透视——基于 AI 虚拟恋人用户的深度访谈》，载《未来传播》2025 年第 4 期，第 48 页。

〔14〕Philip Maxwell Thingbø Mlonyeni, “Personal AI, Deception, and the Problem of Emotional Bubbles”, *AI & Society*, Vol.40, No.3 (2025), pp.1927–1938.

〔15〕陈呈：《情感外包与算法共情：AI 陪伴产业中的劳动机制与平台治理》，载《编辑之友》2025 年第 11 期，第 102 页。

〔16〕Rongjun Ma, Shijing He, Jose Luis Martin-Navarro, Xiao Zhan & Jose Such, “Privacy in Human-AI Romantic Relationships: Concerns, Boundaries, and Agency”, *arXiv preprint arXiv*: 2601.16824 (January 2026).

〔17〕Andrew McStay, “Emotional AI, Soft Biometrics and the Surveillance of Emotional Life: An Unusual Consensus on Privacy”, *Big Data & Society*, (January-June 2020), Art.2053951720904386.

〔18〕Julian De Freitas, Zeliha Oğuz-Uğuralp & Ahmet Kaan-Uğuralp, “Emotional Manipulation by AI Companions”, *arXiv preprint arXiv*: 2508.19258 (2025).

〔19〕Iryna Pentina, Tyler Hancock & Tianling Xie, “Exploring Relationship Development with Social Chatbots: A Mixed-Method Study of Replika”, *Computers in Human Behavior*, Vol.140 (2023), Art.107600.

〔20〕Tianling Xie & Iryna Pentina, “Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika”, *Proceedings of the 55th Hawaii International Conference on System Sciences* (2022), pp.2046–2055.

实世界已经死亡。〔21〕无独有偶，Character.AI、OpenAI、谷歌公司等全球知名厂商也曾因诱导青少年自杀引发多起法律诉讼。目前，Character.AI 和谷歌公司已就相关诉讼与受害者家人达成和解，表示将在人工智能安全和青少年保护方面采取积极措施。〔22〕

（三）基于“可信 AI”的人造亲密关系治理框架

综上所述，围绕人格拟人化这一互动模式建立的人造亲密关系，是一个从认知认同到信任建立、再到行为卷入的渐进过程。对这一关系的治理，也应当沿着“认知—信任—行为”的逻辑链条层层展开，于每一个关键节点嵌入相应的程序约束与风险防范机制。当前，关乎人造亲密关系的治理，更多是从宏观的风险管理以及微观的产品设计加以讨论，〔23〕尚未有清晰的框架。本文认为，可以引入“可信人工智能”（Trustworthy AI，以下简称“可信 AI”）的治理理念，将其作为建构相应治理框架的理论根基与价值指引。

“可信 AI”将信任视为人工智能赋能社会的基石，认为唯有在人工智能的开发、部署与应用全周期中建立信任，才能充分释放技术潜能，进而推动经济社会的可持续发展。〔24〕有学者认为“可信 AI”原则包括鲁棒性（robustness）、泛化（generalization）、可解释性和透明性（explainability and transparency）、再现性（reproducibility）、公平性（fairness）、隐私保护（privacy protection）以及对上述要求整体评估的问责制（accountability），并主张将这些原则嵌入人工智能生命周期的各个环节——包括数据准备、算法设计、系统开发、部署运行与治理管理。〔25〕另有学者将“可信 AI”原则提炼为有益性（beneficence）、无害性（non-maleficence）、自主性（autonomy）、公正性（justice）与可解释性（explicability），强调综合运用技术手段（如隐私计算、去偏算法）与非技术手段（如审计机制、法规约束），将可信 AI 原则系统性地落实于数据全生命周期。〔26〕不难发现，关乎“可信 AI”的隐私保护、公平性、无害性、问责制等原则，切中了人造亲密关系中隐私泄露、用户成瘾、诱发极端情感的命脉。同时，这些原则亦能良好兼容不同国家、地区的文化背景和制度语境差异。将之引入人造亲密关系治理研究，不仅逻辑自洽，且具备坚实的价值根基与现实解释力。

由此，本文将“可信 AI”与人造亲密关系中的认知、信任和行为机制相结合，提出人造亲密关系治理的五个核心原则（如表 1 所示）：认知层面，需打破拟人化所带来的同理心幻觉，明确其非

〔21〕Imane EI Atillah，“Man Ends His Life After an AI Chatbot ‘Encouraged’ Him to Sacrifice Himself to Stop Climate Change”，载欧洲新闻网，2023 年 3 月 31 日发布，见：https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-，最后访问日期：2026 年 2 月 3 日。

〔22〕Clare Duff，“Character.AI and Google Agree to Settle Lawsuits over Teen Mental Health Harms and Suicides”，载美国有线电视新闻网，2026 年 1 月 7 日，见：https://edition.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit，最后访问日期：2026 年 2 月 3 日。

〔23〕相关文献有：Kim Malfacini，“The impacts of companion AI on human relationships：risks，benefits，and design considerations”，*Ai & Society*，Vol.40，No.7 (2025)，pp. 5527-5540.；Raffaele Fabio Ciriello，Angelina Ying Chen & Zara Annette Rubinsztein.“Compassionate AI design，governance，and use”，*IEEE Transactions on Technology and Society*，Vol.6，No.3 (September 2025)，pp. 270-275.

〔24〕Scott Thiebes，Sebastian Lins & Ali Sunyaev，“Trustworthy Artificial Intelligence”，*Electronic Markets*，Vol.31，No.2 (2021)，pp.447-464.

〔25〕Bo Li，Peng Qi，Bo Liu，Shuai Di，Jingen Liu，Jiquan Pei，Jinfeng Yi & Bowen Zhou，“Trustworthy AI：From Principles to Practices”，*ACM Computing Surveys*，Vol.55，No.9 (January 2023)，pp.1-46.

〔26〕Scott Thiebes，Sebastian Lins & Ali Sunyaev，“Trustworthy Artificial Intelligence”，*Electronic Markets*，Vol.31，No.2 (2021)，pp.447-464.

人属性与情感呈现的虚拟性，这对应着“可信 AI”所强调的“透明性原则”；信任层面包含两项原则：一是“有益－无害性原则”，旨在防范 AI 因谄媚迎合而加剧用户自我投射，避免用户深陷情感闭环乃至成瘾。二是“隐私保护原则”，旨在保障用户自我披露的个人信息不被算法诱导或变相利用；行为层面对应“安全性原则”与“问责制原则”，这包括建立系统性的安全护栏以防范极端情感及相关行为，同时构建事先预防、事中监控、及时干预与事后救济的全流程机制。

表 1　“可信 AI”理念下人造亲密治理的原则与目标

阶段	原则	目标
认知	透明性	提醒用户 AI 拟人化产品中的非人属性与先天不足
信任	有益－无害性	防范 AI 谄媚迎合加剧情感自我投射、陷入情感闭环乃至成瘾
	隐私保护	保护用户自我披露的信息不被算法诱导或变相利用
行为	安全性	针对极端情感及相关行为的识别与干预
	问责制	针对拟人化互动负面影响的事先预防、事中监控、及时干预与事后处置救济

三、人造亲密关系治理的中美比较分析

围绕上述人造亲密治理原则，下文将以美国纽约州《人工智能伴侣模型法》（2025 年 11 月 5 日生效，以下简称“纽约州法案”）、美国加州《陪伴型聊天机器人法案》（2026 年 1 月 1 日生效，以下简称“加州法案”）和中国国家互联网信息办公室《人工智能拟人化互动服务管理暂行办法（征求意见稿）》（2025 年 12 月 27 日发布，以下简称“《办法（征求意见稿）》”）为经验材料，比较中美两国在人造亲密关系治理路径上的差异。当前，美国和中国是全球范围内率先开启 AI 拟人化互动垂直治理的国家。中国科技监管具有“自上而下”特征，《办法（征求意见稿）》已清晰体现出系统性的治理思路和价值取向；美国选取纽约州法案与加州法案，主要出于以下三点考量：其一，美国科技监管奉行“州级先行”惯例，州级法案往往更具代表意义；其二，纽约州法案于 2025 年 11 月生效，加州法案于 2026 年 1 月生效，两者时间相近且均为全球首批生效的专项立法；其三，纽约州法案侧重透明度与公共执法规制，加州法案（最初）更关注成瘾预防与安全审计（以下将详细阐述）。综合考察两州法案可更全面、细致地呈现美国式治理的全貌。

（一）透明性原则：本体披露的程度与范围差异

透明性原则的核心在于向用户披露交互对象的非人属性，即“本体披露”。对此，中美在披露程度与范围上存在差异。

美国两州法案均规定了强制性本体披露义务。纽约州法案 1702 条的通知条款要求 AI 拟人化互动服务提供者（以下简称“提供者”）在互动开始及持续过程中，每三小时向用户提供清晰显著的口头或书面陈述，明确告知“用户并非在与人类进行交流”。[27] 加州法案则针对成年、未成年人用户做出一定区分：针对成年用户，“若与陪伴型聊天机器人互动的理性人会被误导而相信其正在与人类互动，运营者应当发布清晰且显著的通知，表明该陪伴型聊天机器人为人工生成而非人类”；[28] 针对未成年人用户，则通过强制性条款要求“向该用户披露其正在与人工智能互动”，[29] 并要求“对持续的陪伴型聊天机器人互动，提供者应当以默认方式至少每三小时向该用户提供一次清晰且显著

〔27〕§ 1702, N.Y. Gen. Bus. Law Art. 47 (2025).
〔28〕§ 22602(a), California Senate Bill 243, Chapter 677.
〔29〕§ 22602(c)(1), California Senate Bill 243, Chapter 677.

的通知，提醒用户休息，并提示该陪伴型聊天机器人为人工生成而非人类。”〔30〕不过，美国两州法案均未涉及训练数据的透明性问题。其主要原因在于：第一，美国两州法案的透明性规制未脱离传统的“知情同意”范式，其核心诉求是保障用户作为消费者的知情权；第二，在美国商业保护主义的惯例下，AI 训练数据被视为核心商业机密。仅展开本体披露不会对企业造成实质性损害，也不会增加额外成本，更容易为企业所接受。

相较而言，我国的《办法（征求意见稿）》就本体披露的规定更为纵深。《办法（征求意见稿）》初步明确了对全体用户的显著提醒，要求“提供者识别出用户出现过度依赖、沉迷倾向时，或者在用户初次使用、重新登录时，应当以弹窗等方式动态提醒用户交互内容为人工智能生成。”〔31〕对于连续使用拟人化互动服务，要求间隔两小时“以弹窗等方式动态提醒用户暂停使用服务”〔32〕。除了交互界面的本体披露外，更延伸至服务开发前端环节，将数据来源与训练过程纳入规制视野。其第 10 条要求提供者在预训练和优化训练等数据处理活动时加强管理，通过“对训练数据的清洗、标注，增强训练数据的透明度、可靠性”。〔33〕

（二）有益－无害性原则：中立应对与积极防御的差异

有益－无害性原则涉及 AI 是否存在谄媚设计乃至诱导用户成瘾。在这方面，中美存在中立应对与积极防御的分野。

就美国来看，纽约州法案未涉及谄媚设计限制以及防沉瘾机制要求。加州法案仅在针对未成年用户的部分要求“采取合理措施，防止其陪伴型聊天机器人生成露骨性行为的视觉材料，或直接陈述该未成年人应当从事露骨性行为。”〔34〕该规定一定程度上体现了对人类道德伦理的关切，但未触及“防成瘾”这一关键防线。事实上，加州法案最初包含“防成瘾”的强制条款。根据加州立法信息官网披露的立法文件，2025 年 1 月 30 日民主党参议员史蒂夫 · 帕迪利亚（Steve Padilla）首次提交的法案版本明确提出提供者应采取合理措施，防止聊天机器人“以不可预测的时间间隔或在不一致的操作次数后向未成年用户提供奖励，或鼓励用户增加参与度、使用率或响应率”。〔35〕然而，该条款遭遇电子前沿基金会（Electronic Frontier Foundation）的强烈反对，理由在于缺乏充分证据表明 AI 聊天机器人所造成的“成瘾”已达到需要政府干预的程度，〔36〕该条款最终被删除。不仅如此，美国此前判例还树立了“代码即言论”（Code is Speech）〔37〕和社交媒体平台的算法分发享有第一修正案保护的“编辑裁量权”（editorial discretion）原则，〔38〕这加大了立法机构对 AI 产品底层

〔30〕§ 22602(c)(2), California Senate Bill 243, Chapter 677.

〔31〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 16 条，载国家互联网信息办公室网站，2025 年 12 月 27 日发布，见：https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm，最后访问日期：2026 年 2 月 5 日。

〔32〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 17 条。

〔33〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 10 条。

〔34〕§ 22602(c)(3), California Senate Bill 243, Chapter 677.

〔35〕California Senate Bill 243, introduced January 30, 2025, 见：https://leginfo.legislature.ca.gov/faces/billVersionsCompareClient.xhtml?bill_id=202520260SB243&cversion=20250SB24399INT，最后访问日期：2026 年 2 月 5 日。

〔36〕California Senate Judiciary Committee, “*Bill Analysis: SB 243*” (2025-2026 Regular Session), April 4, 2025, 见：https://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=202520260SB243，最后访问日期：2026 年 2 月 5 日。

〔37〕1999 年《伯恩斯坦诉美国案》（Bernstein v. United States）中，美国第九巡回上诉法院最终裁定计算机源代码属于受美国宪法第一修正案保护的言论。

〔38〕Moody v. NetChoice, LLC, 603 U.S. 707 (2024).

技术逻辑的实质干预难度。在缺乏明确的判定标准与监管依据的双重困境下，AI 是否诱导成瘾问题仍处于悬置状态。

相较而言，我国的《办法（征求意见稿）》对谄媚设计、诱导成瘾的规制更为积极。《办法（征求意见稿）》明确坚持健康管理与依法治理相结合，鼓励提供者在文化传播和适老陪伴等方面合理拓展应用场景，同时划出服务红线，明确提出提供者“不得将替代社会交往、控制用户心理、诱导沉迷依赖等作为设计目标”。〔39〕还在生成内容上做了细致要求，除传统网络生态治理中的违规内容外，特别明确不得开展“提供严重影响用户行为的虚假承诺和损害社会人际关系的服务”“通过鼓励、美化、暗示自杀自残等方式损害用户身体健康，或者通过语言暴力、情感操控等方式损害用户人格尊严与心理健康”“通过算法操纵、信息误导、设置情感陷阱等方式，诱导用户作出不合理决策”“诱导、套取涉密敏感信息”。〔40〕此外，《办法（征求意见稿）》将未成年人用户和老年人用户作为重点保护人群，根据两类人群可能存在的使用弱点设定专门细则：对未成年人要求建立模式切换、定期现实提醒、使用时长限制等个性化安全设置选项〔41〕，提供情感陪伴服务需取得监护人明确同意并提供监护人控制功能〔42〕；对老年人明确要求“不得提供模拟老年人用户亲属、特定关系人的服务”〔43〕。值得一提的是，虽然《办法（征求意见稿）》未直接涉及防成瘾问题，但在确保人机互动符合伦理、防止算法诱导及用户操控等方面，做出了从服务设计到内容生成，再到用户保护的细致规定。同时，《办法（征求意见稿）》第 18 条特别提及了提供情感服务功能时“应当具备便捷的退出途径，不得阻拦用户主动退出。用户在人机交互界面或者窗口通过按钮、关键词等方式要求退出时，应当及时停止服务。”〔44〕这将在很大程度上遏制（本文第二部分所述）“对话式暗黑模式”带来的风险。

（三）隐私保护原则：依赖既有立法与构建专项机制的差异

隐私保护原则涉及用户交互数据的收集、使用与保护。在这方面，中美分别遵循依赖基础性隐私法框架与构建垂直化专项机制的路径。

无论加州法案还是纽约州法案，均未对用户隐私保护做出额外规定。这是因为美国的隐私治理主要依托联邦层面、州层面既有的基础性隐私法框架，且有“联邦滞后、州级先行”的特点：在联邦层面，统一的隐私立法付之阙如。2024 年提出的《美国隐私权利法案》（American Privacy Rights Act，简称 APRA）试图建立美国联邦统一的数据法案，但并未在 118 届国会立法中获得通过。当前，美国全境具有统一规制效力的主要是《儿童在线隐私保护法》（Children’s Online Privacy Protection Act，简称 COPPA）。该法案禁止在收集、使用或披露儿童个人信息时实施不公平或欺骗性行为，要求收集儿童个人信息必须获得“可核实的父母同意”，〔45〕规定“不得以儿童披露超出合理必要范围的个人信息为条件，允许儿童参与游戏、获得奖品或其他活动”。〔46〕但 COPPA 的适用范围仅限于 13 周岁以下的儿童，对 13–17 周岁青少年存在保护盲区。相较联邦立法，州级立法则

〔39〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 9 条。
〔40〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 7 条。
〔41〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 12 条。
〔42〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 12 条。该条款在 2026 年 4 月 10 日公布的正式管理办法文本中，调整为“不得向未成年人提供虚拟亲属、虚拟伴侣等虚拟亲密关系的服务”。
〔43〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 13 条。该条款在 2026 年 4 月 10 日公布的正式管理办法文本中删去。
〔44〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 18 条。
〔45〕§ 312.5，Children’s Online Privacy Protection Rule (COPPA)，16 C.F.R.Part 312(2026).
〔46〕§ 312.7,Children’s Online Privacy Protection Rule (COPPA)，16 C.F.R.Part 312(2026).

因地而异。加州《消费者隐私法》（California Consumer Privacy Act，简称CCPA）和《隐私权利法》（California Privacy Rights Act，简称CPRA）是当前美国最严格的州级数据隐私法规。具体到人机互动，该法案的规制效力体现为赋予用户删除个人信息、〔47〕了解个人信息的出售及共享出处、〔48〕退出出售或共享个人信息、〔49〕限制敏感个人信息使用及披露〔50〕等权利，同时明确企业不得因用户行使上述权利而诉诸歧视；〔51〕纽约州暂未有统一的州级隐私立法，但其出台的纽约州《儿童数据保护法案》（New York Child Data Protection Act，简称NYCDPA）填补了联邦COPPA在13–17周岁的保护盲区。〔52〕

与之相比，我国的《办法（征求意见稿）》在《中华人民共和国民法典》《中华人民共和国网络安全法》《中华人民共和国数据安全法》等既有法律的基础上，构建了更为体系化、垂直性的隐私保护机制。其核心特征在于将隐私保护规定与模型训练、数据共享等核心商业环节直接关联，形成从数据收集、共享到使用的全流程规制。《办法（征求意见稿）》第14条、第15条提出，“除法律另有规定或者权利人明确同意外，不得向第三方提供用户交互数据”〔53〕“提供者应当向用户提供删除交互数据的选项”〔54〕“除法律、行政法规另有规定或者取得用户单独同意外，提供者不得将用户交互数据、用户敏感个人信息用于模型训练”〔55〕。同时，针对未成年人用户的数据保护要求更为严格，提供者需通过自行或委托专业机构每年进行合规审计，未成年人模式下的数据收集需取得监护人单独同意方可施行。同时，用户及未成年用户监护人享有删除历史交互数据的权利〔56〕。这种垂直性的隐私保护，力图将人机交互的关键数据控制在合理边界之内，防范提供者利用交互数据对AI展开二次训练，不断增强捕捉用户心理弱点的能力。

（四）安全性原则：极端情感识别与干预深度的差异

安全性原则涉及对用户极端情感的识别干预。对此，中美在极端情感风险范畴、识别干预方式与相应制度保障方面存在差异。

对于极端情感风险范畴，加州法案将之限定为“用户表达自杀意念、自杀或自我伤害”〔57〕，纽约州法案则表述为“用户表达自杀意念或自残倾向”〔58〕。对于干预机制，纽约州和加州法案均要求提供者在识别出规定风险后，向用户发布通知并转介至自杀热线等危机服务方。此外，加州法案保留了运营者的信息披露义务与危机处置报告要求，规定提供者需在其网站上公布安全性条款的具体规程细节。自2027年7月1日起，提供者还须定期向加州自杀预防办公室报告上一年度危机服务转

〔47〕§ 1798.105, California Consumer Privacy Act of 2018, Cal. Civ. Code(2026).
〔48〕§ 1798.115, California Consumer Privacy Act of 2018, Cal. Civ. Code(2026).
〔49〕§ 1798.120, California Consumer Privacy Act of 2018, Cal. Civ. Code(2026).
〔50〕§ 1798.121, California Consumer Privacy Act of 2018, Cal. Civ. Code(2026).
〔51〕§ 1798.125, California Consumer Privacy Act of 2018, Cal. Civ. Code(2026).
〔52〕Office of the New York State Attorney General：“New York Child Data Protection Act Implementation Guidance”，载纽约州检察长官网，2025年5月19日，见：https://ag.ny.gov/sites/default/files/2025-05/nycdpa-guidance.pdf，最后访问日期：2026年2月25日。
〔53〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第14条。
〔54〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第14条。
〔55〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第15条。
〔56〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第14、15条。关于监护人可以要求提供者删除未成年人用户历史交互数据的条款在2026年4月10日公布的正式管理办法文本删去。
〔57〕§ 22602(b)(1), California Senate Bill 243, Chapter 677.
〔58〕§ 1701, N.Y. Gen. Bus. Law Art. 47 (2025)。

介次数及安全防范规程。然而，在立法过程中，加州法案却删除了“第三方审计要求”这一足以提供强力安全性保障的条款。该条款曾遭遇以加州商会、计算机与通信行业协会、美国科技行业游说组织 TechNet 为代表的反对者联盟的强烈抵制。反对者指出，第三方审计不仅会显著增加模型厂商的经营成本，且未必能达到降低风险或提升用户权益的目的。尽管 2025 年 7 月加州众议院隐私与消费者保护委员会的立法分析报告仍对该条款持强硬保留立场〔59〕，但最终生效的法案中，该条款并未得以保留。

相比之下，我国的《办法（征求意见稿）》对极端情感风险的识别干预，呈现出层层递进、多主体协同的特征。在干预层级上，第 11 条设置了由浅入深的递进式响应体系：在保护用户隐私的前提下，对用户出现极端情绪或沉迷倾向的情形采取必要干预；当识别出具有威胁用户生命健康和财产安全的高风险倾向时，要求提供者预设情绪安抚和提供专业援助方式的回复模板；如用户明确提出将实施自杀、自残，须以人工接管对话的方式，第一时间与用户保持联系。在此基础上，《办法（征求意见稿）》还将现实社会关系引入最高风险干预机制：除人工接管对话外，提供者还需采取措施联络用户监护人和紧急联系人，特别要求未成年人和老年用户在注册环节填写监护人、紧急联系人信息；对涉及老年人可能诱发的财产安全问题，在通知紧急联系人的同时提供社会心理援助或紧急救助渠道〔60〕。同时，《办法（征求意见稿）》还明确了提供者的安全主体责任，涵盖算法机制机理审核、科技伦理审查、信息发布审核、网络安全、数据安全、个人信息保护、反电信网络诈骗、重大风险预案、应急处置等方面，〔61〕亦要求互联网应用商店承担相关安全核验和管理责任。〔62〕国家网信部门则指导推动人工智能沙箱安全服务平台建设，〔63〕从产业生态层面构建安全保障体系。

（五）问责性原则：公共执法、私人诉讼与行政监管的差异

问责制原则涉及违反规制要求后的责任追究与权利救济。对此，中美分别采取公共执法主导、私人诉讼主导与行政监管主导的方式。

美国纽约州法案树立了公共执法为主导的模式，仅赋予州总检察长代表纽约州人民在获得充分证据显示提供者违反法规后可以提起相应诉讼的权利，并发起禁令救济、每日最高 15000 美元的民事罚款或其他适当司法救济。加州法案则赋予受损用户更广泛的诉讼权利：遭受事实性损害的个人可提起民事诉讼，寻求禁令救济、损害赔偿及合理的律师费与诉讼费用。在赔偿金额上，法案采取“实际损害与每次违法一千美元中的较高者”的计算标准。值得注意的是，私人诉讼权条款在加州立法过程中曾因波及滥诉而遭遇科技行业抵制，但该条款对“私人诉权”的原告资格设定了严格的门槛限制，要求原告必须证明遭受精神状况恶化、自残或可诊断的心理创伤等事实损害，〔64〕避免了仅仅由于纯程序性违规而对运营者施加过度惩罚。纽约州的公共执法也是在可能引发“滥诉”质疑下从最初普遍的私人诉讼进行执法权收拢，并将相应获得的处罚费用拨入该州“自杀预防基金”。

相较美国的司法救济路径，我国《办法（征求意见稿）》构建了一套以行政监管为核心的问责

〔59〕California Assembly Committee on Privacy and Consumer Protection, “Bill Analysis: SB 243” (2025–2026 Regular Session), July 5, 2025, 见：https://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=202520260SB243，最后访问日期：2026 年 2 月 5 日。

〔60〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 11 条、第 13 条。

〔61〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 8 条。

〔62〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 24 条。

〔63〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第 27 条。

〔64〕California Assembly Committee on Judiciary, “*Bill Analysis: SB 243*” (2025–2026 Regular Session), July 15 2025, 见：https://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=202520260SB243，最后访问日期：2026 年 2 月 5 日。

体系。第一，前置安全评估与备案机制。要求提供者在拟人化互动服务功能上线、增设或变更时，须按照国家规定开展安全评估，并向属地省级网信部门提交评估报告。对于注册用户达100万以上或月活跃用户达10万以上的提供者，〔65〕亦纳入安全评估范围。第二，常态化书面审查。省级网信部门每年对相关企业的评估报告及审计情况进行书面审查并开展情况核实，〔66〕形成持续性监管。第三，应急处置与违规问责。当发生突发重大风险时，省级以上网信部门和有关主管部门可启动约谈程序；对违规行为，可视情节分别处以警告、通报批评、责令限期整改乃至责令暂停提供服务的处罚。〔67〕

四、结论与讨论

围绕透明性、有益－无害性、隐私保护、安全性、问责制五项治理原则，本文就中美相关政策文本展开比较（如表2所示），可清晰看到两国在人造亲密治理上的分野。

表2　人造亲密关系治理的中美差异

治理原则	比较维度	美国治理路径	中国治理路径
透明性	披露范围	纽约州：全体用户； 加州：仅限未成年用户及可能被误导的理性人	全体用户
	披露方式	口头／醒目文字提示； 持续互动中至少每3小时触发提醒（加州法案对未成年人用户增加额外休息提醒）	弹窗提醒； 持续互动中每2小时提醒用户暂停使用服务
	披露深度	交互界面披露	交互界面披露； 训练数据来源与处理过程的透明性要求
有益－无害性	价值立场	关注底线风险，中立应对	关注价值导向，积极防御
	禁止事项	纽约州：未涉及； 加州：针对未成年人，严控露骨性内容及行为诱导	明确禁止替代社会交往、情感操纵、算法操纵、信息误导、设置情感陷阱、套取涉密敏感信息等
	特殊保护	纽约州：未涉及 加州：未成年人涉性内容限制	未成年人：未成年人模式、个性化安全、情感陪伴的监护控制功能； 老年人：禁止模拟亲属及特定关系人
隐私保护	规制模式	依赖基础隐私法框架，未设专门条款	构建专门机制，限制用户交互数据用于模型训练、数据共享等环节
	数据使用	未设置针对性特殊限制	数据训练限制：用户交互数据及敏感个人信息限制； 未成年人数据：年度合规审计与监护人单独同意
安全性	风险范畴	自杀、自残意念或倾向	不同风险层级：极端情绪或沉迷；威胁用户生命安全和财产安全；实施自杀、自残等
	干预方式	用户通知； 转介危机服务提供方	递进响应：必要干预；情绪安抚与提供专业援助方式；人工接管对话，联系监护人／紧急联系人
	监督保障	纽约州：未涉及 加州：强制章程公示与年报机制	服务提供者安全主体责任 应用商店安全核验责任 国家监管沙盒服务责任
问责制	执行路径	纽约州：公共执法主导的司法救济 加州：私人诉讼为核心的司法救济	行政执法为核心的穿透监管
	责任惩戒	纽约州：禁令救济、民事罚款或其他适当救济 加州：禁令救济、损害赔偿、律师费及诉讼费	警告、通报批评、责令限期整改乃至暂停提供服务处罚

〔65〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第21条。

〔66〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第26条。

〔67〕《人工智能拟人化互动服务管理暂行办法（征求意见稿）》第28、29条。在2026年4月10日公布的正式管理办法文本中，增加执法部门及罚款标准等细节。

上述分野，亦反映了两国对智能时代人机关系本质的不同理解与价值预设。美国纽约州与加州相关法案背后的规制逻辑具有鲜明的程序防御特征：一是在透明性、有益－无害性要求上，止步于在交互界面披露AI的非人属性，未能触及深层的算法设计、数据训练等关键技术环节的谄媚与诱导；二是在安全性规制上，仅要求运营者防范干预如自杀、自残等可验证的极端情况；三是问责机制上，依赖事后司法程序提供救济，举证门槛较高。在遭遇非极端风险的实质性伤害时，受害者往往必须证明AI存在欺诈行为或根本性“设计缺陷”，否则法律难以介入。在这种情况下，AI被预设为价值中立的工具，人类用户则被定位为具有自主选择能力、理性评估能力的主体。人造亲密关系尺度、形态、走向，基本由程序自行决定。因此，本文将美国人造亲密关系治理概括为“个人理性选择下的消极防御”。此治理思路遮蔽了人在算法面前的弱势地位，将拟人化互动带来的隐私泄露、依赖成瘾、极端情感诱导等风险限定在私人领域的藩篱之内，AI拟人化互动服务提供者的相应责任义务，则被排除在规制视野之外。加州法案在立法修订中逐渐弱化提供者的义务，更折射出美国人造亲密关系治理中商业利益与用户保护之间的深层张力。2025年9月的加州法案修订草案不仅删除了科技企业强烈反对的“防成瘾”强制性条款和第三方独立审计要求，对陪伴型聊天机器人作出了附条件的部分豁免，对特定场景和功能的AI聊天机器人排除在外，〔68〕还将强制的透明性披露义务限定于未成年人或“可能被误导的理性人”场景。此举引发了包括科技监督项目（Tech Oversight Project）在内多家机构的批评，认为其实质上削弱了法律的保护效力。当前，美国国内相关企业在舆论压力下做出了诸如部署自我伤害以及自杀意念监测拦截系统、限制提供医疗诊断等自我修补，但更多是科技企业为了避免监管升级所做的合规妥协，而非规制效力的实质性提升。

相较美国，我国对人造亲密关系的治理呈现出治理节点全覆盖、彼此嵌套、协同发挥作用的积极规制特征。在透明性层面，不仅要求于交互界面披露AI的非人属性，更将透明义务延伸至训练数据的来源与处理过程；在防范谄媚诱导层面，虽未针对防沉迷出台具体的规制细节，但严格禁控情感陷阱、社会交往替代等产品设计，着力降低人机交互可能引发的不当情感依赖；在隐私保护层面，规制边界从传统的数据“收集－使用”环节延伸至模型训练，旨在抑制提供者使用用户交互数据进行AI训练强化；在安全保障层面，超越了美国局限于极端风险的消极防御，确立了从极端情感到极端行为之间的风险层级，并试图构建起一个从算法识别、人工干预到社会关系（如监护人、紧急联系人）介入的递进式响应机制；在问责机制层面，以行政监管为核心，通过安全评估、备案审查和应急处置等，形成贯穿AI开发、服务提供、产品更新全生命周期的穿透性监管。在这种规制思路下，AI不再是价值中立的工具，而是被要求在底层技术逻辑中嵌入“价值敏感性设计（Value Sensitive Design）”。〔69〕用户亦不是原子化、孤立的个体，而是内嵌在特定社会关系网络和社会价值体系中的关系性主体。用户与AI的互动应当是积极健康的虚拟互动，可以发挥情感陪伴、心理支持等功能，但不能以损害用户利益及其与现实社会的真实联结为代价。基于此，本文将我国的人造亲密关系治理概括为“现实社会关系／价值嵌入的积极规制”。这种规制敏锐捕捉到个体在人造亲密关系中可能存在的种种脆弱性，将原本可能由个体独自承受的情感成瘾、认知扭曲、社会退缩等负效应从个人选择中抽离出来，用制度性力量予以兜底保护。同时，亦将技术创新纳入维护社会关系整体健康的轨道，确保技术进步不偏离增进人民福祉的根本方向。当然，沿此展开治理实践，需审慎面对治

〔68〕如电子游戏中仅回复游戏相关内容的内嵌AI聊天机器人等。参见§ 22601(b)(2), California Senate Bill 243, Chapter 677.

〔69〕Batya Friedman, Peter H. Kahn Jr. & Alan Borning, “Value Sensitive Design and Information Systems”, in Neelke Doorn, Daan Schuurbiers, Ibo van de Poel & Michael E. Gorman, eds., Early Engagement and New Technologies: Opening Up the Laboratory, Dordrecht: Springer Netherlands, 2013, pp.55–84.

理过程中的种种张力：一方面，人的价值观本身存在多样性与潜在冲突，如何实现 AI 与人类价值的对齐存在难度；另一方面，鉴于 AI 拟人化互动服务提供者需要承担社会连带责任，其无可避免要对用户实施介入关注，如何在充分保障用户自主权利与提供必要保护之间找到最佳平衡点，是该治理模式在实践中需要持续完善的关键命题。

无论是个人理性选择下的程序防御，还是现实社会关系／价值嵌入下的积极规制，其对人造亲密关系治理的有效性仍待检验。在此基础上，一个更为根本性的问题亟待在政府、科技企业和社会公众层面形成共识：人工智能技术的加速演进与人类现有应对能力之间，正形成日益扩大的鸿沟，当拟人化技术日渐逼真并长驱直入人类用户的情感腹地后，人类用户既有的数字素养防线是否能够有效抵挡人造亲密关系中潜藏的沉迷、诱导乃至操控？面对日益演进的人造亲密关系，当前的治理手段究竟触及了实质，还是停留于表面合规的修补？在“负责任的人工智能”“统筹发展与安全”等治理共识引领下，均需将价值嵌入内化为超越形式合规的人工智能治理底层逻辑，同时凝聚政府、企业、社会与用户的治理合力，有效应对人工智能技术对用户主体性的潜在侵蚀。

（责任编辑　刘承韪）

Abstract: The risks posed by artificial intimacy have become an important issue in global AI governance. Taking “AI anthropomorphic interaction” as the core concept, this paper clarifies the generation mechanism and specific risks of artificial intimacy. Based on the framework of “Trustworthy AI”, this paper refines governance principles of artificial intimacy and constructs an analytical framework. Based on this, this paper examines recent laws and legislative drafts for public comments issued by the Cyberspace Administration of China, and New York State and California of the United States, thereby revealing divergences in governance paths between the two countries. The findings indicate that U.S. governance tends to be “negative governance based on individual rational choice”, which treats AI products as a value-neutral tool, leaving the scale and trajectory of artificial intimacy to individual autonomy while restricting legal intervention exclusively to extreme cases such as suicide and self-harm. Conversely, China’s governance tends to be “active regulation embedded in real social relations and values, which acknowledges user vulnerability in artificial intimacy, emphasizes the integration of ethical values into AI design and further incorporates interactional risks into social network and value system to provide comprehensive institutional safeguards through institutional forces.

keywords: artificial intimacy, AI anthropomorphic interaction, trustworthy AI, AI governance