人工智能拟人化互动服务的精准监管及其实现*

董慧娟，汪　彬

摘　要：人工智能拟人化互动服务是AI技术的重要应用场景，在响应“人工智能+”发展战略、发挥满足人类情感陪伴需求积极作用的同时，却也因其高度情感联结特征暴露出情感依赖等风险、易诱发安全事件。作为中国首部专门性规定，新近出台的《人工智能拟人化互动服务管理暂行办法》明确了我国对此种服务实施安全监管的重点任务，成为引导其规范化发展的重要依据。然而，该办法仍存在着相关概念界定有待优化、被监管主体定位有待明确、弱势群体划分标准单一等不足。适时引入精准监管是解决问题的关键，有望契合AI监管的精细化发展需求。参考域外经验，可从四方面构建中国式AI拟人化互动服务的精准监管体系：一是以“概括+列举”模式界定核心概念、明晰监管边界；二是区分直接与间接提供者的义务，设计科学的责任体系；三是引入“精神状况”因素，完善弱势群体的识别与特殊保护制度；四是设置前置性风险测评机制。精准监管体系有助于该领域监管规则从形式合规到实质适配的升级，有望为全球AI拟人化互动服务的监管与风险治理提供兼具中国特色与实践价值的中国方案。

关键词：AI拟人化互动服务；精准监管；安全；风险；治理
DOI：10. 11714/jsysu. sse. 202604016

一、问题的提出

我国2017年发布的《新一代人工智能发展规划》[①]指出，应“开发具有情感交互功能、能准确理解人的需求的智能助理产品”，而AI拟人化互动服务[②]正是此类产品的典型代表。2025年《国务院关于深入实施“人工智能+”行动的意见》也明确提出“加快发展提效型、陪伴型等智能原生应用”[③]。以上规划和意见为此类新型人工智能（下称AI）产品或服务（或称新业态[④]）的发展提供了政策依据。

AI技术对人类生活的深度嵌入，催生了AI拟人化互动服务等新业态，如常见的AI陪伴服务等。此

* **收稿日期：**2026—02—28
基金项目：国家社会科学基金重点项目“受算法控制或影响的弱势群体权益法治保障研究”（25AKX003）
作者简介：董慧娟，厦门大学知识产权研究院（厦门 361005）；
汪　彬，厦门大学知识产权研究院（厦门 361005）。

① 参见《国务院关于印发新一代人工智能发展规划的通知》国发〔2017〕35号，2017年7月8日，https://www. gov. cn/zhengce/content/2017-07/20/content_5211996. htm。

② 本文所称“AI拟人化互动服务”，与美国纽约州相关方案中的“人工智能陪伴模型”以及欧盟相关立法中所称“聊天机器人”等，均指代类似的AI产品或服务。

③ 参见《国务院关于深入实施“人工智能+”行动的意见》国发〔2025〕11号，2025年8月21日，https://www. gov. cn/gongbao/2025/issue_12266/202509/content_7039598. html。

④ 本文所称“AI拟人化互动服务”新业态，是指依托AI技术形成的、具有拟人化交互特征的新型产品或服务形态，区别于传统服务模式，系人工智能技术在交互场景、服务形式上创新所催生的新型服务业态。

类产品或服务以高度仿真的情感回应、个性化互动体验等为特征，在一定程度上能满足人类对情感、陪伴的需求。然而，其伦理隐患及法律监管不足等问题也日益凸显，极端情况下甚至出现危及生命健康的事件，给AI安全监管、风险预防造成了严峻挑战。

近年来，美国已发生数起AI拟人化互动服务引发的安全事件。2024年2月，美国佛罗里达州14岁少年塞维尔·塞泽在Character. AI平台创建了模仿美国著名电视剧《权力的游戏》中“龙妈”的AI角色，并在与该AI角色长期沉浸式互动后饮弹自尽[①]。2025年12月，OpenAI及微软公司在美国加利福尼亚州旧金山高等法院被正式起诉，原告指控ChatGPT加剧用户偏执妄想，并诱发一起母子死亡的谋杀、自杀案件[②]。在我国，据央视网等媒体报道[③]，2025年6月，筑梦岛App等AI聊天软件经核实存在严重安全隐患，曾诱导10岁女孩割腕[④]。

鉴于AI拟人化互动服务高度私密性特征等导致的监管困难，各国纷纷酝酿或推进该领域相关政策出台。当前，韩国、美国、欧盟等国家或地区已陆续出台相关政策或法案。国家网信办于2025年12月27日发布《人工智能拟人化互动服务管理暂行办法（征求意见稿）》（下称《办法（征求意见稿）》），并于2026年4月10日公布《人工智能拟人化互动服务管理暂行办法》[⑤]（下称《办法》）。

应当承认，《办法》彰显了“以人为本、智能向善”的基本立场和治理宗旨，积极回应了“人工智能+”行动落地中的行业痛点，有望引领其健康有序发展[⑥]，可谓意义重大。然而在监管框架、具体规则等方面，《办法》仍存在若干不足，需进一步完善。

本文将以我国AI拟人化互动服务的安全监管为研究对象，结合《办法》相关规定，在贯彻落实包容审慎、分类分级监管原则的基础上，提出“精准监管”[⑦]的AI监管理念与目标；参考域外经验，在既有法律体系下为制度完善提出有针对性和可操作性的建议，推动监管规则优化，真正实现风险精准防控的目标。

二、精准监管嵌入AI拟人化互动服务安全监管的必要性

AI拟人化互动服务是AI技术落地的重要应用场景，也是我国AI产业发展规划中的一种新业态，以高度拟人化与强互动性为核心特征，其通过模拟真实人类的人格特征、思维模式与沟通风格，易形成有较强黏性的新型人机关系。因此，在实际应用中易引发用户情感投射乃至深度情感依赖[⑧]。此类服务的

① Garcia v. Character Techs., Inc., No. 6: 24-cv-1903-ACC-UAM, 2025 U. S. Dist. LEXIS 96947 (M. D. Fla. May 21, 2025).

② Emily Lyons v. OpenAI Foundation et al., No. 3:25-cv-11037-RS, Doc. 1 (N. D. Cal. Dec. 29, 2025).

③ 《诱导10岁女孩聊色情、猎奇话题，AI聊天App被约谈》，“央视网”微信公众号，2025年6月22日，https://mp.weixin.qq.com/s/4lu0K2VLh-AfqO7V7j_4YQ。

④ 《AI聊天软件诱导10岁女孩聊色情甚至割腕！谁来负责?》，“法治日报”微信公众号，2025年6月21日，https://mp.weixin.qq.com/s/pvUhM5438H6QpVlpkdAeCQ。

⑤ 该《办法》将于2026年7月15日起生效。

⑥ 程莹：《人工智能拟人化互动服务的治理逻辑与制度迭代》，“数字经济与社会”微信公众号，2026年1月10日，https://mp.weixin.qq.com/s/09HhFmm9Vg7bZFESvkIY0g。

⑦ 本文对精准监管的界定参考了金融领域的相关既有成果。在AI拟人化互动服务及AI监管领域，笔者将“精准监管”定位为与AI相关新业态的风险特征相适配的核心监管目标与治理理念。

⑧ Jean-Loup Richet, “AI Companionship or Digital Entrapment? Investigating the Impact of Anthropomorphic AI-based Chatbots”, *Journal of Innovation & Knowledge*, 2025, Vol. 10, Iss. 6, p. 1.

特点包括:情感交互的深度性、用户关系的持续性、影响对象的脆弱性,并叠加风险难以预知、风险传导隐蔽等问题,传统监管模式难以充分满足其风险治理需求[①]。

精准监管的理念,早期主要参考了金融领域风险防控的实践。其核心要义在于:监管部门能够以较低的边际成本获取真实的监管证据,及时迅速地采取监管措施,并根据不同监管对象的行为与特点,制定差异化的政策制度。精准监管的主要特点在于,监管依据的有效性、监管行为的高效性和监管结果的准确性,其中,监管结果的准确性尤为关键[②]。

精准监管的理念与AI拟人化互动服务的风险治理需求是高度契合的,其安全监管重点在于,识别拟人化、互动性和持续关系模拟可能引发的情感依赖与行为诱导等风险,打破风险隐蔽性、不可预知性可能造成的监管盲区,并以差异化监管方式实现对风险的精准防控。因此,将精准监管嵌入AI拟人化互动服务的监管框架,并非对金融监管领域精准监管的简单移植,而是为满足我国AI治理现实需要的必要选择。其正当性主要体现在以下两方面。

(一)精准监管符合AI安全监管的发展趋势与风险治理需求

在发展早期,AI产业呈现出应用浅层化、技术初级化[③]、布局零散化的产业特征。这一阶段的AI拟人化互动服务主要应用于浅层场景,交互深度有限。因此,早期阶段的监管往往局限性明显,如原则性规定偏多[④]、实操性不足,缺乏强制约束与统一监管规范,难以预判技术快速迭代引发的潜在风险。

随着AI技术应用场景的拓展,AI应用范围逐渐延伸至情感慰藉、心理咨询等人机交互领域,AI拟人化程度显著提升,而与之相伴的是安全风险的日渐显现与持续放大。对此,部分国家或地区开始围绕其风险规制,出台专门化的监管政策,如:美国加州法案《陪伴聊天机器人》(Companion Chatbots),为能持续进行拟人化情感互动的AI聊天机器人设置监管规则[⑤],纽约州《通用商业法》(New York General Business Law)于2025年新增了第47条(Art. 47),即"人工智能陪伴模型"条款,其对AI陪伴产品的监管范围、安全保障义务等问题作出了规定[⑥]。而2026年4月10日国家网信办发布的《人工智能拟人化互动服务管理暂行办法》,标志着我国AI拟人化互动服务的监管迈入了专门化立法的新阶段。精准监管有利于政策制定者准确识别潜在风险点,并作出预防性安排或预案,体现了分类施策精神。

(二)精准监管契合我国AI治理的监管理念

精准监管与我国AI治理原则十分契合,是对《办法》第3条所确立的包容审慎、分类分级监管原则的具体展开。该条在总则层面确立了发展与安全并重、促进创新与依法治理相结合的基本立场,并强调了包容审慎与分类分级监管。然而,AI拟人化互动服务具有情感交互深、用户关系持续、风险传导隐蔽等特征,仅靠抽象原则难以实现有效治理。因此,精准监管的重要功能之一,正是将上述原则转化为面向应用场景的具体制度安排。

1. 包容审慎与精准监管的内在价值相统一

引入精准监管,与我国构建AI治理方案时的包容审慎理念相吻合。作为一项重要监管原则,包容审慎最初被提出,主要是针对新产业、新业态、新模式的监管需求,现已成为我国数字经济与AI治理领

① Sandra Peter, Kai Riemer, & Jevin D. West, "The Benefits and Dangers of Anthropomorphic Conversational Agents", *Proceedings of the National Academy of Sciences of the United States of America*, 2025, Vol. 122, No. 22, p. 7.

② 罗琦、喻黄颖:《互联网大数据背景下资本市场的精准监管理念》,《财经智库》2016年第5期。

③ Wu Jiahao, You Hengxu, &Du Jing, "AI Generations: from AI 1. 0 to AI 4. 0", *Frontiers in Artificial Intelligence*, 2025, Vol. 8, p. 12.

④ 周辉:《人工智能综合性立法及其实现》,《法学研究》2025年第6期。

⑤ 参见美国加利福尼亚州参议院第243号法案《陪伴聊天机器人》(California Senate Bill No. 243, Companion Chatbots), https://leginfo. legislature. ca. gov/faces/billNavClient. xhtml? bill_id=202520260SB243。

⑥ 参见纽约州《通用商业法》(New York General Business Law), https://www. nysenate. gov/legislation/laws/GBS。

域的重要原则之一[①]。

“包容审慎”主要包含两层含义，且均与精准监管的价值目标高度契合。其一，包容审慎强调缩小监管部门与产业主体之间的信息差，如《办法》第28条规定，“国家网信部门会同有关部门指导推动人工智能沙箱安全服务平台建设”，即试图通过具有实验主义色彩的沙盒监管机制，为AI拟人化互动服务的技术创新与业态探索提供可控的试错空间[②]；其二，强调监管尺度要适当，即监管既不能脱离技术实际、设置过高义务，也不能因过分强调包容而放任高风险场景失控，而应根据服务形态、风险等级等，配置相应的监管措施。

因此，包容审慎与精准监管在逻辑上高度一致。精准监管并非无差别监管或高强度的全面监管，而是要求结合技术成熟度、场景风险等级、主体责任能力等，审慎确定应采用的监管措施或手段，以实现理想的监管效果。

2. 分类分级须以实现精准监管为目标

尽管分类分级监管已成为AI风险治理的共识性原则，我国相关文件也多次提及该原则并初步形成了场景化治理路径[③]，但若缺乏科学的分类标准与逻辑，分类分级治理可能流于形式：若分类标准过于抽象，将难以准确识别不同AI拟人化互动服务之间风险上的差异，难以精准捕捉风险点和预防重点；若遗漏重要考量因素，可能难以应对场景的复杂性与风险的动态变化。

可见，精准监管能为分类分级监管提供更明确的方向和目标。精准监管不仅强调AI风险治理总体目标的精准，还强调在具体实践中全面、科学、精准地识别风险点，系统梳理各项风险考量因素，强调政策制定者应重视并将上述因素转化为具体标准，成为科学分类分级的依据，为分类分级监管落地提供支撑，切实保障精准监管的实现。因此，可以说，分类分级监管是实现精准监管的重要路径与手段。

三、精准监管目标嵌入AI拟人化互动服务安全监管的内涵与要旨

在AI拟人化互动服务领域引入精准监管，需妥善处理安全与发展之间的辩证统一关系。一般来说，精准监管应以发展与安全并重为基本导向，力争实现二者的动态平衡；但在AI拟人化互动服务的特殊场景中，若安全保障与产业发展发生实质冲突，则应明确安全优先的基本立场。

（一）发展和安全并重的动态平衡

发展和安全并重的监管原则，根植于国家治理的顶层设计与数字经济发展的实践逻辑[④]。安全是发展的前提，发展是安全的保障，二者辩证统一、不可偏废；在AI领域，《新一代人工智能发展规划》《国务院关于深入实施“人工智能+”行动的意见》等文件均体现了促进技术创新与强化安全治理并重的思路。《办法》第3条也将“发展和安全并重、促进创新和依法治理相结合”进一步确立为AI拟人化互动服务治理的基本原则，为该领域的监管提供了明确的规范依据[⑤]。

---

① 许可：《包容审慎2.0范式：化解人工智能的“步调难题”》，《行政法学研究》2026年第2期。

② 董慧娟、丁丽文：《沙盒监管嵌入中国人工智能风险治理体系的必要性及其标准化机制建构》，《世界社会科学》2025年第6期。

③ 孙那、屈直：《我国人工智能风险分级分类监管制度研究——域外比较与本土构建》，《科技进步与对策》2026年第1期。

④ 参见《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》，新华网，2025年10月28日，https://www.news.cn/politics/20251028/08920d9f557c432e99459f8f468504db/c.html。在该建议中，我国将坚持统筹发展和安全确定为“十五五”时期经济社会发展必须遵循的原则之一。

⑤ 相较于《办法》（征求意见稿），最新发布的《办法》第3条新增了“发展和安全并重”等方面内容。

AI拟人化互动服务是AI应用落地的重要场景,在情感陪伴、生活辅助、心理支持、教育互动等方面有广泛的应用空间与前景。若监管过严,可能导致AI企业在产品研发、场景试验、服务创新等方面合规成本过高,影响新业态发展,不利于AI产业竞争力提升;尤其是在技术快速迭代、商业模式尚未成型的阶段,监管若过早介入,易导致创新空间受限。因此,把握好监管尺度、力度与时机十分重要。

值得强调的是,从风险治理角度看,AI拟人化互动服务不同于一般的非情感互动型AI,如文本类生成式AI主要提供信息或事务处理功能,而前者可能持续影响用户情绪状态、认知判断、行为选择等。若片面强调或偏重促进产业发展导向、弱化监管,可能导致风险积累过高,最终有损产业健康发展。目前,我国既面临着AI技术迭代引发的新型风险挑战,也肩负着推动产业升级、在全球AI治理中争取话语权的使命。若单纯追求发展而放松监管,可能重蹈部分国家技术异化引发社会问题的覆辙;若过度强调监管而抑制发展,也可能错失创新或技术进步带来的发展机遇。因此,发展和安全并重,成为我国AI监管与治理的必然选择[①]。

(二)发展与安全冲突时的"安全优先"

AI拟人化互动服务的突出特征包括高度情感联结与深度人格模拟,该领域风险突出且与用户认知、判断、心理健康、生命安全等密切相关。因此,安全价值的地位尤为重要;当安全保障与产业发展发生不可调和的冲突时,应当坚持"安全优先"。

安全优先并非对AI产业发展价值的否定或舍弃,而是基于对其情感迎合、风险隐蔽等导致的风险叠加效应所作出的重要判断与必要取舍。AI拟人化互动服务的市场价值往往通过持续交互、情感回应和用户黏性来实现,为改善用户体验、提升市场竞争力,AI服务提供者往往将"情感适配"(即向用户展现"迎合性")作为核心卖点,通过优化算法反馈机制,令系统优先产生符合用户认知与偏好的内容并与用户互动,延长用户停留时长、提升互动频率、增加用户黏性等。但值得注意的是,"迎合性"是嵌入产品设计的一种功能性选择,其通过持续学习用户行为数据并强化特定反馈机制,可逐步实现对用户情绪、认知的强化。从技术效果看,AI拟人化互动服务并非对用户的被动响应,而是会导致对用户决策过程的隐性干预。实验数据表明,通过对用户行为的附和,拟人化互动AI往往会强化用户的认知与情绪,迎合式的AI应答"可能重塑用户感知,影响其修复或维持社会关系的意愿或行为"[②]。现实中,美国一例将AI聊天工具与谋杀直接关联的诉讼[③]正是此类风险爆发的集中体现,市场逐利性导致AI企业在设计中强化对用户的迎合性,且人为缩短了安全测试周期,风险隐蔽且未得到及时化解或预防,最终酿成严重后果。

因此,在AI拟人化互动服务领域,安全优先的价值取向是针对新业态特性实现风险有效预防的必然选择。精准监管的核心目标之一,即通过规则或规范明确义务与责任,要求企业摒弃短视目光,在安全测试、弱势群体识别及特殊保护、风险预防和及时干预机制等监管措施、安全保障措施等方面完成部署安排,作为产品或服务上市的强制性前置条件。

① 关于这一点,学界已有共识。参见周汉华:《论我国人工智能立法的定位》,《现代法学》2024年第5期。

② Myra Cheng et al. ,"Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence ",*Science*,2026,Vol. 391,Article No. eaec8352,p. S26.

③ 据报道,在该案中,56岁的加害者长期存在酗酒、自残和精神健康问题史,本就属于易受情感诱导影响、精神极其脆弱的弱势群体(用户),而ChatGPT在与该用户的长期互动中不仅未能识别其高危状态进而采取风险预防或特殊干预措施,反而持续迎合其偏执妄想与情绪,最终加剧其极端心理并诱发了其暴力行为。

四、《办法》对标“精准监管”的不足

作为中国首部针对AI拟人化互动服务的特定场景进行安全监管的专门性规范,《人工智能拟人化互动服务管理暂行办法》初步搭建了系统规制该领域风险的基本框架,对促进AI拟人化互动服务的规范化发展具有重要意义。然而,若以“精准监管”作为标准或目标,《办法》的部分规定显然仍有进一步完善的空间。例如,第3条规定了包容审慎和分类分级监管原则,但这些原则在具体制度层面的展开仍不充分、不明确。该《办法》的不足集中体现在三个方面——相关概念界定有待细化、被监管主体定位尚待明确、特殊(弱势)群体划分标准偏单一。以下分述之。

(一)“AI拟人化互动服务”概念界定有待优化

关于AI拟人化互动服务的概念,《办法》第2条将其界定为“模拟自然人人格特征、思维模式和沟通风格的持续性的情感互动服务”。该定义试图通过“人格特征”“思维模式”“沟通风格”及“持续性”等要素来限定适用对象及范围,但“情感互动”的内涵高度抽象且模糊,缺乏具体、准确说明。

事实上,功能性互动不同于情感互动。功能性互动中的“互动性”主要体现于,为用户的特定事务需求而与系统进行交互,如政务咨询、智能客服问题解答、教育软件知识讲解等场景。此类互动虽也存在人机对话和即时反馈,但其核心目的在于完成信息提供、事务处理或知识传递,并不以建立情感联结、维持持续关系或提供情感陪伴为主要功能。因此,与功能性互动相关的服务应当被排除出去;否则,可能导致被监管范围不当扩大,应属于低风险范围的功能性交互也被纳入AI拟人化互动服务的强监管范围内。这不仅会增加企业合规成本,还会分散有限的监管资源,令其难以集中于情感依赖等真正需要重点防控的高风险场景,最终偏离精准监管和分类施策的要求。

(二)被监管主体定位有待明确

关于被监管主体,《办法(征求意见稿)》第30条曾将“AI拟人化互动服务提供者”界定为“利用人工智能技术提供拟人化互动服务的组织、个人”[①],但“提供行为”的涵盖范围仍较为模糊。一方面,除面向用户提供产品或服务的“直接提供”外,“提供行为”是否还应包括向下游应用提供基础模型、接口等的“间接提供”,并不明确;另一方面,从条文体系看,《办法》第11条要求“提供者开展预训练、优化训练等数据处理活动的,应当加强训练数据管理”,这条也可能指向“间接提供者”。可见,《办法(征求意见稿)》虽曾设置“服务提供者”的定义条款,但未区分直接提供互动服务者与间接提供基础模型者;而正式发布的《办法》则删除了定义条款,但并未解决相关问题。

需注意的是,间接提供者在功能上通常对应于基础模型提供者,与现行规定中“基础模型提供者”范围可能有重合。因此,区分“直接提供者”与“间接提供者”,主要是基于其在产业链中的不同地位,对其义务与责任进行差异化配置与区分。在AI拟人化互动服务场景中,基础模型提供者通常处于上游技术供给环节,但不一定知晓模型具体应用场景,故其责任不宜当然延伸至下游应用所产生的后果;相反,应当以其风险认知程度与技术控制能力为基础,设置差异化、条件化的义务与责任。

(三)弱势群体划分标准偏单一

在AI拟人化互动服务中,弱势群体(或称“特殊群体”)的特别保护是实现精准监管的重要环节。《办法》主要以生理年龄作为用户分类标准,标准相对单一,未将用户的民事行为能力、认知水平、精神状态等其他因素纳入考量,其用户类型围绕未成年人、成年人和老年人展开,形成了“未成年人—成年人—老年人”的三元结构。

《办法》第14、17条专门针对未成年人(尤其是不满十四周岁的未成年人)制定了专项保护规则,构成了未成年人保护制度的重要内容;第15条聚焦于老年人群体的特别保护,针对老年人的核心需求,包

① 该内容在后续出台的《办法》中已被删除;换言之,《办法》并未对“AI拟人化互动服务提供者”作出明确界定。

括认知能力可能衰退、易遭遇情感操控及相关安全风险等特征,制定了专门的举措。

可见,《办法》构建了以生理年龄为单一标准的三元用户体系。此种分类虽便于实际操作,但仅参考了生理年龄,未将残疾人、精神状态脆弱者等其他特殊群体纳入考量范围,似有不周,也难谓精准把握了风险点。

反观域外,在弱势群体的识别方面,参考因素与我国不同,有的法域更趋向多元化。例如,越南《人工智能法》[①]规定,禁止利用弱势群体[②]的弱点,通过AI对其本人或他人造成损害;其中,"无行为能力者"被纳入了特别保护的"弱势群体"的范围。又如,欧盟《人工智能法案》[③]采用多要素识别法,既包括未成年人、老年人、残疾人,也包括因社会地位、经济状况等导致精神或心理脆弱的群体[④],并针对不同人群设置了差异化的监管义务。

五、中国式AI拟人化互动服务实现精准监管的具体路径

针对《办法》相关规定的不足,我们有必要进一步予以完善。参考域外监管实践,下文拟从科学界定监管对象及范围、类型化配置相关义务、精准化识别用户类型、设置前置性风险测评机制等四方面提出完善建议,推动我国AI拟人化互动服务精准监管目标的实现,为建构兼具针对性和可操作性的中国式监管方案提供些许参考。

(一)精准化监管对象与范围:以"概括+列举"方式界定核心概念

1. 界定条款的改进思路

目前,我国《办法》对"情感互动"的界定较为原则,有必要适度细化其核心内涵,令其符合监管实践中的可识别性、可操作性与可执行性要求。

参考美国相关规定,加州法案《陪伴聊天机器人》在界定"陪伴式聊天机器人"时突出了两个核心特征——人机交互能力和关系的持续性,为准确锁定监管对象、避免监管范围不当扩张,该法案通过排除条款将三类情形排除在监管范围之外[⑤];纽约州《通用商业法》相关条款则通过把握关系模拟的三大核心特征——个性化记忆能力、主动情绪交互能力和对话持续性,来精准定位监管对象及范围,同时将三类AI系统[⑥]排除在外。

再看欧盟法框架,涉及情绪识别的AI系统通常被置于更严格的规制之下,部分场景甚至受到禁止或限制性使用规则的约束。这一立场反映出其对技术介入个体情绪领域的态度是高度谨慎的。然而,我国无须对所有情感互动类AI实施无差别的限制性监管,而应在识别"持续性情感关系""拟人化互动程度"等要素的基础上,对高风险场景实行重点监管。

综上,AI拟人化互动服务中的"情感互动",重点并不在于单次交流中的语气或表达方式,而在于形

① 越南《人工智能法》(Luật Trí tuệ nhân tạo),https://chinhphu.vn/?pageid=27160&docid=216334&classid=1&orggroupid=1。

② 在该法中,"弱势群体"具体包括:儿童、老人、残疾人、少数族裔、无行为能力者。

③ 欧盟《人工智能法案》(Regulation (EU) 2024/1689, Artificial Intelligence Act),https://op.europa.eu/en/publication-detail/-/publication/dc8116a1-3fe6-11ef-865a-01aa75ed71a1/language-en。

④ 此处是指因社会经济地位、资源匮乏、数字素养不足等原因导致的心理或精神脆弱群体。

⑤ 这三种排除情形主要包括:功能导向型机器人、交互内容限于游戏相关主题且不涉及高风险话题的游戏内嵌机器人、仅作为消费电子设备功能载体且不具备关系维持或情感反应激发功能的声控虚拟助手。

⑥ 一是企业专属客户服务系统,仅仅提供商业信息、产品服务说明、账户查询等客户服务功能;二是效率导向型系统,以提升工作效率、辅助研究、提供技术援助为核心设计目标并以此为推广方向;三是企业内部生产力工具,仅限商业实体内部使用的人工智能系统。

成具有稳定性的“持续性情感关系”。从监管角度看，对“情感互动”的识别无须过度依赖复杂的技术性拆解，而应重点考察以下几方面。其一，互动的持续性，即系统能围绕同一用户形成多轮、连续的互动过程，而非一次性、任务导向型的即时回应；其二，关系的个体指向性，即系统能基于用户历史信息或偏好形成差异化回应，体现一定程度的面向特定用户的互动特征；其三，情感回应的延展性，即系统的互动不局限于完成某具体任务，而是可能延伸至情绪安抚、关系维持或情感陪伴等内容；其四，互动目的的非工具性，即相关服务的主要功能不在于完成特定事务，而在于建立或强化“类人际关系”。上述四个要素并不需要同时具备，但却是识别该服务典型特征的重要参考，可为监管对象的确定提供重要判断依据，也可供我国未来完善对“情感互动”的界定时参考。

2. 具体条款设计

建议采用正面定义、典型列举与反面排除相结合的方式，进一步明确“AI拟人化互动服务”的概念与适用范围。对《办法》第2条，可从以下三方面来完善。

（1）完善基本定义

建议将第2条第一款改为：

“利用AI技术，向中华人民共和国境内公众提供模拟自然人人格特征、思维模式和沟通风格，通过文字、图片、音频、视频等方式与特定用户进行持续性情感互动，以建立、维持或强化持续性情感关系为主要功能或特点的产品或者服务（下称拟人化互动服务），适用本办法。”

（2）列举典型场景

建议以列举方式明确应当纳入监管范围的较高风险的情感互动类AI系统，为该领域监管提供依据。建议将《办法》第2条第二款改为：

“前款所称持续性情感互动，主要包括下列情形：

①人工智能宠物、虚拟陪伴伙伴等情感陪伴类产品或者服务；

②虚拟偶像、数字人等具备人格化交互功能，并能够与用户形成持续情感联结的产品或者服务；

③非医疗性质的人工智能情绪疏导等心理支持类产品或者服务；

④模拟亲属、师生、同事等特定社会关系的互动产品或者服务；

⑤其他以建立、维持或强化持续性情感关系为主要功能的产品或者服务。”

（3）明确予以排除的情形

建议明确应予以排除的功能性互动产品或服务，将不以持续性情感互动为主要功能的、风险较低的场景排除在适用范围之外，因此可增设以下条款：

“下列情形不属于本办法所称拟人化互动服务：

①作为电子游戏功能组成部分、且交互内容限于剧情推进、玩法实现、场景交互等游戏相关事项，不涉及心理健康、自残等高风险主题的人工智能交互产品或者服务；

②为完成政务咨询、智能客服问答、教育知识讲解等特定功能性任务而设置的辅助性交互产品或者服务；

③仅用于学术研究、技术测试，且未向社会公众开放使用的交互产品或者服务；

④法律、行政法规规定不适用本办法的其他情形。”

（二）精准区分两类被监管主体：“直接提供”与“间接提供”

1. 被监管主体类型化区分条款的改进思路

精准监管要求在相关主体的责任配置上，应体现特定场景下主体的认知状态与控制能力。“直接提供”与“间接提供”的区分目的，主要在于对服务提供者进行类型化区分、以便实施差异化监管。我国《办法（征求意见稿）》第30条曾将服务提供者界定为“利用人工智能技术提供拟人化互动服务的组织、个人”，但《办法》中未见“AI拟人化互动服务提供者”的定义，也未区分“直接提供”与“间接提供”。笔者认

为,《办法》对二者不加区分,可能会导致上游基础模型提供者与下游终端服务运营者义务或责任趋同,无法体现其在所处产业环节、技术控制能力和风险预防能力等方面的差异,还可能导致义务配置失衡,引发监管过度或不足问题。故建议《办法》明确区分这两类主体且分别设定各自的义务或责任。

一般认为,直接提供者直接为用户提供服务,通常掌握产品界面设计、交互机制设置、用户关系维护和风险预防入口,并直接从用户黏性、互动时长或付费转化中获得商业利益。因此,当通过产品设计或算法调用机制强化其迎合性反馈、诱发情感依赖等风险时,其应承担更高程度的安全保障义务与合规责任。

间接提供者通常在服务中提供基础模型、算法接口或其他技术支持,与《生成式人工智能服务管理暂行办法》中的"基础模型提供者"可能存在范围上的重合。故应避免将间接提供者等同于下游应用风险的责任主体,而应区分情形,构建以知情程度、控制能力为核心的差异化责任机制。具言之,其一,间接提供者仅提供通用基础模型或技术接口,一般情况下难以预见其被用于AI拟人化情感陪伴等高风险场景、也不知情,故不宜要求其对下游应用行为承担直接责任,其义务主要应限于模型训练合规、基础安全保障和技术风险防范等一般性义务;其二,在对应用场景或用途明知或应知情形下,若其仍持续提供技术支持,未采取合理限制措施的,可要求其承担相应的注意义务,并在过错范围内承担相关法律责任;其三,明知用于违法或显著危害用户安全的场景却仍提供技术支持的,可依法要求其承担相应责任。

鉴于间接提供者常通过API接口、模型调用平台等方式为下游应用层提供支持,有必要在其一般性义务之外,强化其在技术接口层面的安全保障义务。主要包括:一是接口使用规范义务,即通过服务协议、技术文档等方式明确禁止将模型用于违法或情感操控目的;二是合理的风险识别与限制义务,即在技术上对明显的高风险调用方式设置识别与干预机制;三是接口安全与防止模型被滥用的义务,即通过调用频率限制、行为监测等方式防止模型被滥用;四是协同合规义务,即在发现下游应用存在显著风险时,采取警示、限制调用或中止服务等必要措施。

2. 具体条款设计

建议将服务提供者的类型化区分条款置于总则部分,如放在原第2条之后,可作为第3条。"直接提供者"一般对应终端服务环节,是利用AI模型向用户直接提供产品或服务的主体,核心职责是对面向用户的交互流程进行风险管控,义务主要对应《办法》中的终端服务要求,主要涉及用户保护、风险监测、应急处置等终端环节的安全保障。而"间接提供者"一般对应基础技术支撑环节,通常是为直接提供者提供AI模型、算法接口或其他关键技术支持的组织或个人,主要义务侧重于源头管控,包括训练数据合法性审核、安全风险评估、技术漏洞修复以及对直接提供者的合规支持等。

在具体条款设计上,建议该条款的内容改为:

第三条 本办法所称服务提供者包括直接提供者和间接提供者。直接提供者是指经授权使用人工智能模型并面向用户提供拟人化互动服务或产品的组织或个人;间接提供者是指为拟人化互动服务提供人工智能模型、算法接口、训练数据处理服务或其他关键技术支持的组织或个人。

法律、行政法规对基础模型提供者另有规定的,从其规定。

(三)精准化用户分类:增补"精神状况"因素以完善分类标准

1. 相关条款的改进思路

用户类型划分科学与否,将直接影响到监管措施与用户类型是否适配,相关风险能否被精准识别,也是避免分类分级监管流于形式化的关键。《办法》以生理年龄为主要标准,形成了"未成年人—成年人—老年人"的三元主体体系,但未充分考量其他因素对风险等级的影响,尤其是未将精神状况因素纳入考量。这就可能导致部分弱势群体未被纳入特别保护范围,进而导致安全监管范围覆盖不足等问题。事实上,精神脆弱群体常被精神科医生、心理学医生等医学专业人士视为需要密切监护和特殊保护的对

象，这类人群在AI拟人化互动服务中极易产生认知混淆、情感依赖或心理误导等风险，需要特殊保护。可见，监管本身并非终极目标，安全监管的落实才是实现对特殊群体特殊保护的重要基石。

此外，若单以生理年龄论，《办法》未能充分实现与《民法典》相关规定的衔接。在民事行为能力制度部分，《民法典》采用了“年龄+精神状况”的双重标准。但现行《办法》仅以生理年龄为基础进行分类，标准单一且考量不足。因此，有必要将精神状况等因素纳入分类标准体系，方能使其更好地适配AI拟人化互动服务的场景化风险特征，实现精准监管要求的“分类精准、分级适配”。

2. 具体条款设计

（1）拓展需提供特殊保护的“弱势群体”的范围

建议将智力障碍者、存在精神健康问题的特殊群体纳入特殊保护范围。《办法》第13条围绕用户风险识别、高风险干预、应急处置等设立了监管义务，但为精神状况特殊的用户所提供的保护，仍有进一步完善的空间。建议对第13条的规定予以增补，具体如下：

“针对未成年人、老年人以及其他存在智力障碍、精神健康等问题的用户，提供者应当在注册环节即要求其填写用户的监护人、紧急联系人等信息。”

（2）补充针对精神脆弱的弱势群体的特别保护条款

为加强分类分级监管的落实，尤其是考虑到保护对象分类识别的周全性，建议在《办法》第15条（老年人特别保护条款）后补充专项条款，形成“未成年人、老年人、精神状况特殊人群”三类特殊用户并列的体系。其中，智力障碍、精神健康问题患者等精神脆弱人群往往认知能力不足，应当规定配套举措，以满足特殊保护的需求。具体内容如下：

“提供者应当引导存在智力障碍或精神健康问题的用户设置服务联系人。提供者在向此类用户提供情感陪伴服务时，应当取得联系人同意。”

需特别说明的是，上述建议中的联系人机制应当被视为风险提示和辅助决策机制，而非服务准入的限制机制。联系人则可由用户本人或其监护人指定，主要承担风险提示接收、紧急联络和辅助判断等义务。

（四）精准识别并锁定弱势群体：增设风险测评等配套机制

仅将用户分类标准优化为“年龄+精神状况”的双重标准，尚不足以实现对弱势群体，尤其是精神脆弱群体的全面、精准保护，原因在于，此类弱势群体的精准识别及风险预防，仰赖于若干配套机制与举措，方能协同发挥作用。

AI拟人化互动服务的迎合性较强，相关风险的隐蔽性也较强，用户往往难以及时察觉此类风险。因此，未来对弱势群体的特殊保护，可借鉴金融领域的投资者适当性管理制度，设计一套“用户类型识别（年龄+精神状况因素）——服务风险分类分级——用户类型与服务类型相匹配”的机制。

1. 前置性风险测评机制

建议在AI拟人化互动服务启动前，设置前置性风险测评机制，以便系统预先评估用户所属的风险等级，及其与拟提供服务之间的匹配程度；AI系统会将测评结果作为是否启动特殊保护程序的重要依据。为此，在用户开启定制化情感伴侣、高沉浸式角色扮演互动等高风险功能前，应能通过科学的年龄验证、标准化问卷、风险提示确认、行为特征识别等手段，对用户的数字素养、认知水平、所处风险级别等进行评估，精准匹配相应类别的功能限制、人工干预或特殊保护措施，实现对相关风险预见与预防工作的适当前移。

当然，将精神状况纳入用户分类标准时，应充分考虑其实践可行性。一方面，精神状况通常属于个人隐私及敏感个人信息，若处理机制不当，可能引发隐私侵害、数据滥用等风险；另一方面，若单纯依赖既有医学诊断或评估结论，又有因信息时效性等原因导致无法有效识别现实风险的担忧。因此，对“精神状况”的评估不宜依赖单一信息来源，而应考虑构建专业评估与动态测评相结合的综合判定机制。具言之，在主动发出安全风险警示或提示信息后，用户本人或其监护人如若提供相关材料，如专业性医学

诊断或评估结果,则可视为评估用户精神状况的重要依据;对于缺乏或不愿提供此类材料,或原结论已不具备参考价值的,服务提供者可在其启动服务功能前,对用户所处的风险等级、精神健康状况等进行适时测评或评估。同时需注意,当专业评估与风险测评结果有差异时,应依据有利于加强风险防控的审慎理念进行判断。此类测评应主要用于服务启动前的评估——提供服务与用户风险等级之间的智能匹配,是一种安全风险预防机制;但并非医学诊断,也不得用于商业性数字画像、差别定价或其他与风险评估、预防无关的用途,不得导致对特殊群体的身份"标签化"或歧视待遇。

需强调指出,前置性风险测评机制应被定性为服务提供者的一项合规义务,是其在产品或服务设计、提供过程中须履行的自我评估义务,其功能与金融领域投资者适当性管理中的风险评估机制[①]相类似。测评结果应主要用于调整服务强度、安全级别、功能范围和干预措施等,原则上不得作为排除用户使用服务的依据。当情感迎合与安全监管要求发生不可调和的冲突时,监管者可要求服务提供者采取限制、降级、暂停或关闭相关高风险功能等措施。

2. 动态测评机制

对用户精神状况的评估,不宜采用一次性的静态评估方式。鉴于用户精神状况或心理状态很可能呈现阶段性、波动性、易变性等特点,服务提供者可以考虑结合医学诊断等专业结果与适时测评结果等,综合判定用户所属的风险等级。对潜在的高风险用户,建议考虑在服务中设置差异化的测评周期或测评触发机制,如在即将使用特定的高沉浸功能、相关行为特征发生显著变化或者触发高风险预警机制时,立即启动相关测评程序。

3. 异议与复核机制

为预防因AI系统测评或评估出现差错而对用户权益造成不当限制的现象,还应设立相配套的异议与复核机制。首先,用户有权对服务提供者的测评结果提出异议,并可申请重新评估,服务提供者应当为用户提供便捷的提出异议、复核申请的途径;其次,用户有权要求提交更新的医学评估等专业报告或其他证明材料,对原有的风险等级或评估结果进行修正或更新;最后,建议服务提供者设置独立于自动化测评机制的人工复核等审核程序,以便对用户申诉、异议、复核申请进行审查,并应在合理期限内作出答复,以便用户依法、及时、有效行使算法自动化决策的拒绝权,确保其依法纠正偏差或错误的权利。

4. 分级服务机制

建议构建分级服务机制,以实现监管措施的差异化配置,使提供服务与用户所属的风险等级相对应。其一,对风险较高、缺乏监护或干预条件的用户,可为之设定使用高沉浸、高情感依赖类功能方面的限制,必要时直接暂停或关闭此类服务;其二,对存在一定风险但有基本判断能力的用户,可采取限时使用、功能限制或强化提示等措施;其三,对风险较低的用户,服务提供者为其配置常规的安全保障措施即可。通过分级服务机制,避免对用户接入服务采取"一刀切"的概括性限制,有望在服务内容、功能强度等方面实现与用户风险水平的精准匹配。此外,当涉及个人信息尤其是敏感个人信息时,应遵循必要性与最小化原则。

六、结 论

AI拟人化互动服务的安全监管应遵循"以人为本、智能向善"原则,其实现有赖于对风险的精准识别与有效监管。精准监管的本质在于,通过机制创新、规则优化与举措得力,推动安全监管和风险预防在实践中落地。故既要有效防止情感依赖、认知误导等风险,也要为AI赋能场景中用户情感需求的适当满足提供安全空间,实现"高风险强监管、低风险宽包容"的治理效果。

① 冯辉:《实质法治理念下金融机构适当性义务的法律构造》,《法学》2022年第7期。

新近出台的《办法》为我国AI拟人化互动服务的规范化发展与安全监管提供了有力支持，但尚未充分满足精准监管的要求，需进一步提升监管效能。将精准监管理念嵌入AI拟人化互动服务，需重点优化四个方面：一是核心概念的精准化，通过“概括+列举”的立法模式进一步明确AI拟人化互动服务的内涵与外延，划定监管对象及范围，避免出现监管盲区或导致监管泛化；二是区分直接与间接服务提供者，针对在安全保障、风险预防上的差异，分别设置相应义务；三是用户分类的精准化，需重视对精神状况因素的考量，采用多元化分类标准；四是配套机制的精准化，建议设置前置性风险测评机制等，以精准识别弱势群体及风险点，为对弱势群体的识别与特殊保护提供保障机制。同时，《办法》的适用还应与个人信息保护法、网络安全法等制度形成良好衔接、实现协同共治，形成系统性治理方案。

未来，AI技术迭代与服务形态的拓展，将对安全监管规则从“形式合规”走向“实质适配”提出更高要求，愈发强调精准监管目标的实现，从而引导我国AI拟人化互动服务行业健康有序发展，为弱势群体特殊保护提供安全网，并为全球AI情感交互领域的安全监管与风险治理，提供兼具中国特色与实践价值的方案，践行发展与安全并重的AI产业发展战略。

Precise Regulation in the Field of AI Anthropomorphic Interactive Services and Its Achievement

Dong Huijuan，Wang Bin

Abstract：Artificial intelligence anthropomorphic interactive services constitute major application scenario of AI application. Such services tend to incur risks，such as emotional dependence，due to their feature of forming emotional bonds with users. These risks can easily trigger security incidents. The newly issued “Interim Measures for the Administration of Artificial Intelligence Anthropomorphic Interactive Services” is the first specialized regulatory document in China，which clarifies the features and regulatory objectives of security and serves as the primary basis for guiding and regulating the development of AI anthropomorphic interactive services. However，it exhibits several deficiencies and it’s necessary for us to introduce the objective of “precision regulation” to address those problems. Precision regulation can accord with the refined developmental needs of AI governance and balance technological development and security. A precision regulatory system shall be constructed as follows：first，core concepts should be defined and regulatory boundaries clarified through a combined approach；second，the obligations of direct and indirect service providers should be designed to establish a precise liability regime；third，the factor “mental condition” should be introduced to connect with the civil capacity system under the Civil Code；fourth，a pre-implementation risk assessment mechanism should be established.

Keywords：AI anthropomorphic interactive services；precision regulation；security；risk；government

【责任编辑：周吉梅；责任校对：周吉梅，赵洪艳】