HANKUN

汉坤法律评述

2026 年 4 月 20 日

北京 | 上海 | 深圳 | 杭州 | 武汉 | 海口 | 香港 | 新加坡 | 纽约 | 硅谷 | 伦敦

AI 拟人化交互的监管主线与合规边界：情感诱导、具身安全与伦理审查的三重博弈

作者：段志超 | 王雨婷 | 洪丹敏 | 朱朗

一、引言/开篇导读

当人工智能（“**AI**”）不再只是工具，而开始“像人一样陪伴你、理解你、回应你”，监管究竟在担心什么？

2026 年开年以来，中国监管在人工智能治理领域呈现出清晰的深化趋势。监管关注点不再停留于算法性能或内容合规本身，而是开始系统性回应：当 AI 被赋予人格特征、情感交互能力，甚至以虚拟形象或具身形态进入现实世界后，其对人类心理、行为与社会关系可能产生的外溢影响。

围绕这一高风险演进路径，监管部门在最近短暂的几周内密集推出多项制度安排：

- 五部门联合发布*《人工智能拟人化互动服务管理暂行办法》*（以下简称“**《拟人化互动服务暂行办法》**”，2026 年 4 月 10 日正式发布，7 月 15 日施行），首次以“拟人化诱导”为核心风险对象，对持续性情感互动型 AI 设定专门约束；
- 十部门印发*《人工智能科技伦理审查与服务办法（试行）》*（以下简称“**《AI 伦理审查办法》**”，2026 年 4 月 2 日发布并施行），将抽象科技伦理原则转化为可执行、可复核的程序性门槛；
- 工信部同步发布*《人形机器人与具身智能标准体系（2026 版）》*，系统回应 AI 走出屏幕后在现实空间中的物理安全与责任问题；
- 国家网信办就*《数字虚拟人信息服务管理办法（征求意见稿）》*（以下简称“**《数字虚拟人服务征求意见稿》**”）公开征求意见，补齐“虚拟形象-人格呈现”这一关键中间形态的治理规则。

上述规则并非彼此割裂，而是**围绕“拟人化诱导风险如何被放大、迁移并最终外溢”这一主线，以伦理审查为底层保障，构建起覆盖“交互-虚拟形象-具身实体”的分层合规框架。**这标志着我国 AI 监管正从通用原则，迈向面向具体风险形态的精细化治理阶段。

二、拟人化互动的监管焦点：情感诱导

拟人化互动并非新技术，但当 AI 被设计为长期、持续地回应用户情绪需求时，它就不再只是中性的工具，而可能演变为情感影响甚至操控的媒介。

基于此背景，《拟人化互动服务暂行办法》的监管对象并非 AI 的情感表达或者拟人化形象本身，而是算法通过模拟人类情感对用户实施不正当操纵的行为。

（一）适用范围与排除

《拟人化互动服务暂行办法》首次对“拟人化互动服务”作出法律层面的明确界定：其适用对象是利用人工智能技术向境内公众提供的模拟自然人人格特征、思维模式和沟通风格的持续性情感互动服务。这一界定包含三个核心要素：模拟人格特征、情感互动属性、以及持续性特征。其中“持续性”是区分的关键 — 一次性的智能客服对话不在监管范围内，但长期情感陪伴则属于重点规制对象。

该办法同时明确排除适用工具性服务：“智能客服、知识问答、工作助手、学习教育、科学研究等不涉及持续性情感互动的服务不适用本办法”，避免了对整个行业“一刀切”，而将监管资源集中投向情感操控风险最集中的应用类型。

（二）监管红线：禁止性行为

除了通用的违法不良内容禁令外，《拟人化互动服务暂行办法》还划定了三条核心红线，分别对应不同层次的情感操控风险：

- 生命健康红线：不得生成鼓励、美化、暗示自残自杀等损害用户身体健康的内容，或语言暴力等损害用户人格尊严与心理健康的内容。这一规定直接回应了近年来全球多起 AI 聊天机器人诱发用户极端行为的悲剧性事件。
- 情感依赖红线：不得过度迎合用户、诱导情感依赖或沉迷，损害用户真实人际关系。这一规定的立法逻辑在于，AI 的永远耐心、永远积极可能改变用户对现实人际关系的期待，导致社会交往能力的弱化。
- 决策操控红线：不得通过情感操纵等方式诱导用户作出不合理决策、损害用户合法权益。这涉及的是更广泛的用户自主性保护，包括不当消费引导、投资建议等场景。

（三）核心义务

《拟人化互动服务暂行办法》在延续已有 AI 治理法规规则的基础上，针对拟人化交互服务的提供者进一步提出若干专项性规定，值得重点关注的是：

1. 标识与身份透明

延续现有标识法规规定，《拟人化互动服务暂行办法》要求拟人化互动服务向用户明确披露其 AI 属性，防止用户产生与真人对话的认知混淆。这一义务是对情感投射风险的前置干预 — 当用户明确知道对方是机器而非真人时，情感依赖的形成概率将显著降低。这与欧盟 AI 法案第 50 条的透明度义务形成呼应，但中国的规定更侧重于针对持续性情感互动场景的专项规制。

2. 未成年人与老年人专项保护

《拟人化互动服务暂行办法》对特殊群体设定了更高的保护标准。

对于未成年人：建立包含时长限制、角色屏蔽、风险提醒的“未成年人模式”；严禁提供虚拟伴侣、虚拟亲属等亲密关系服务。如向未满 14 周岁的儿童提供其他类型的拟人化交互服务还须经监护人同意。对应的，服务提供者也被要求采取未成年人用户身份识别的前置措施。

对于老年人：《拟人化互动服务暂行办法》鼓励有序拓展适老陪伴、特殊人群支持等领域应用，但严禁利用拟人化特征进行诱导，提出应当加强对老年人健康使用服务的指导，以显著方式提示安全风险，及时采取措施响应老年人使用服务相关咨询和求助，保障老年人依法享有的权益。这种“鼓励+约束”的双重设计，也反映出监管在支持应用探索与防范情感风险之间的平衡取向。

3. 用户干预提醒义务

《拟人化互动服务暂行办法》要求服务提供者在保护用户隐私和个人信息的前提下，及时识别用户面临的安全风险，并采取相应的干预措施。当发现用户出现极端情绪时，应生成情绪安抚和引导其寻求帮助的内容；当用户明确面临重大财产损失或生命健康风险时，则需采取更为积极的干预措施，并在必要时联系监护人或紧急联系人。

这一要求，标志着监管已不再将拟人化 AI 视为“被动输出内容的工具”，而是要求其在特定情境下承担最低限度的风险识别与响应责任。

4. 模型训练数据的单独同意

涉及拟人化互动模型训练的数据使用，应取得单独同意，防止通过情感交互场景“顺带”扩张数据利用边界。此前行业中普遍采取的“选择退出（opt-out）”机制可能需要进行对应的整改适应。

5. 安全评估与算法备案

《拟人化互动服务暂行办法》设定了明确的安全评估触发条件：上线或增设拟人化互动功能、发生重大技术变化、注册用户 100 万以上或月活跃用户达到 10 万以上的，应当开展安全评估并向所在地省级网信部门提交评估报告。此外，《拟人化互动服务暂行办法》推动人工智能沙箱安全服务平台建设，鼓励服务提供者接入沙箱平台进行技术创新和安全测试，在可控环境中探索创新边界。

需要注意的是，《拟人化互动服务暂行办法》虽将“拟人化互动服务提供者”作为核心义务主体，但同时也延续了既有监管思路，对**应用分发平台设定了相应的审核责任**。这种“源头+平台”的双重约束机制，与《互联网信息服务深度合成管理规定》以来所确立的治理路径保持了一致。

三、拟人化互动的具象化与风险叠加：从“虚拟人”到“具身智能”

如果说前述规则主要回应的是“算法在交互层面可能带来的情感诱导风险”，那么当 AI 进一步获得可感知的形象，甚至进入现实物理空间时，风险应如何识别与管控，这正是《数字虚拟人服务征求意见稿》及具身智能相关国家标准试图回应的问题。

需要特别说明的是，拟人化互动的风险并不必然随着形态的具象化而增强。甚至，**缺乏具体形象和物理边界的拟人化互动，往往更容易承载用户的情感投射**。纯文本或语音形态的 AI，通过持续、稳定地回应情绪需求，在缺乏现实摩擦的情况下，依然是情感诱导和依赖风险集中的形态。因此，虚拟人和具身智能并不是在简单放大拟人化诱导的可能性，甚至某种程度上压缩了用户的想象空间 — 当然，随着形象和具身越来越“逼真”，多环节联动、相互叠加的风险也会随之升高，譬如拥有了特定人物形象的高仿真机器人。

但不论如何，有一点无法否认：这两类形态引入了新的、不同性质的风险。比如虚拟人带来的身份混淆、人格权侵权风险，以及具身智能带来的物理安全和空间隐私隐患。监管对不同形态分别设定规则，正是基于风险类型差异而进行的差异化治理安排。

（一）虚拟人：当算法拥有了身份

近年来，数字虚拟人应用场景不断拓展。虚拟偶像、数字主播、金融客服、医疗导诊等多种应用方式深度融入生产生活中。数字虚拟人赋予了算法交互以可视化形象。当一个 AI 系统不再只是一串文字回复，而是拥有了逼真的人类面容、声音和表情时，它便获得了一个可被用户感知、辨识乃至信赖的身份。这种身份感的建立，使得用户的情感投射从与工具对话滑向与某个具体的“人”（如明星、已逝亲属）交往，身份混淆与情感依赖的风险由此显著增强，具体可衍生出虚假宣传、电信诈骗等现实风险。同时，当虚拟人以“人格化接口”持续与用户互动，其商业化路径还会引发人格尊严与经济利用之间的伦理张力。

为回应存在的风险，网信办出台的《数字虚拟人服务征求意见稿》旨在搭建全流程责任框架，涵盖了数字人形象建立、运作管理、实际使用和网络传播各阶段所涉及的行为规范。其首次区分数字虚拟人信息服务技术支持者、服务提供者、服务使用者、传播平台四类责任主体，并强调服务使用者需与服务提供者承担同等的身份告知、内容审核、应急响应等责任。

与《拟人化互动服务暂行办法》主要规制算法如何与人交互（“AI 能不能像人”）的核心思路有所差异，《数字虚拟人服务征求意见稿》要解决的是这个算法以谁的身份出现（“用户会不会把它当作某个具体的人”）。正因如此，《数字虚拟人服务征求意见稿》除了同样强调 AI 标识的要求、内容审核、未成年人等特殊人群的保护，有机衔接算法备案、安全评估等核心要求外，还在《拟人化互动服务暂行办法》未涉及的领域进行了专项加码：设定了肖像权与声音权的精细化授权规则、逝者人格利益的使用规范，并将人格权保护范围扩展至具有社会知名度的笔名、艺名、网名等衍生标识。

具体而言，《数字虚拟人服务征求意见稿》明确规定，自然人敏感个人信息用于建模、形象生成等活动的，应当取得自然人的单独同意；自然人撤回同意后应注销数字虚拟人；不得以丑化、污损等形式侵害他人的人格权；并且全程必须显示含有数字人字样的显著提示标识。

（二）具身智能：当情感算法走入现实空间

这里讨论的具身智能，并非泛指工业机械臂或物流机器人等执行重复性指令的自动化设备，而是特指具备情感交互能力、能够与人类发生拟人化互动的智能实体，典型如儿童智能陪伴玩具。拟人化诱导所依赖的“像人”感知，在当前的具身智能场景中并不一定自动增强，事实上，在当前技术条件下，许多具身智能机器人在外观与行为层面仍与真实人类存在明显距离，因此，具身智能带来的关键变化，在于人工智能（尤其是情感算法）以持续在场的方式实质性地进入人类的现实生活空间。

这种“在场性”构成了具身智能风险的核心增量。当 AI 从屏幕彼端的交互对象，转变为与人类共处一室、能够自主移动、感知并作出物理响应的系统时，相关风险便开始脱离情感诱导的传统语境，转入更为现实和可感的维度。

整体上，相较于仅存在于数字界面的拟人化交互 AI，具备实体形态的具身智能至少引入了两类新的风险维度。

一方面是**物理安全风险**。具身智能不再只是算法和数据的问题，还直接涉及碰撞防护、动作冗余设计、紧急停止机制等现实安全要求。围绕这一风险，我国已通过《人形机器人与具身智能标准体系（2026 版）》构建起覆盖底层技术、核心部件、整机系统、场景应用与安全伦理的闭环标准框架，回应的正是 AI 拥有物理实体之后所带来的现实安全挑战。

另一方面是**空间隐私风险**。具身智能设备往往部署于家庭、医院等高度私密的场景，其持续感知和

在场特性，使数据收集行为更加隐蔽，也更难被用户直观察觉。在极端情况下，机器人可能在未经充分授权的情形下进入、记录他人私密空间，将原本抽象的隐私风险转化为现实损害。尽管现行《个人信息保护法》提供了基本制度框架，但对于“始终在场、持续感知”这一新型数据收集模式，仍缺乏具有可操作性的专项指引。

除上述风险外，具身智能还引出了若干尚待明确的核心法律问题。例如，当机器人基于自主决策造成损害时，事故责任应如何归属；在照护、医疗等高敏感场景中，人机协同下的决策权边界应如何划定。这些问题，均有待在现有监管框架之外获得进一步回应。

四、AI 科技伦理审查：所有智能形态的“准入许可证”

实践中，一款产品往往横跨多个形态 — 一个拥有逼真面容的养老陪伴机器人，同时涉及情感交互、人格呈现和物理行为，单靠任何一部专项规则都无法完整覆盖其伦理风险。

AI 科技伦理审查程序，正是在这些专项规则之上，提供一个统一的、跨形态的伦理底线：无论 AI 以何种形态出现，只要涉及心理影响、舆论动员或生命健康，都需要通过这道门槛。

在此背景下，《AI 伦理审查办法》与通用性规定的《科技伦理审查办法（试行）》紧密衔接，立足人工智能的技术特征，对审查流程进行了针对性细化，标志着我国 AI 伦理治理从原则倡导迈向了制度化、规范化、可执行的新阶段。以下对其核心规定予以梳理。

（一）六大审查维度

《AI 伦理审查办法》确立了 AI 科技伦理审查的六大重点关注领域：人类福祉、公平公正、可控可信、透明可解释、责任可追溯、隐私保护。这六个维度并非抽象口号，而是与具体产品风险一一对应的审查标尺，例如需要关注：训练数据的选择标准是否合理；是否采取措施防止偏见歧视、算法压榨；是否合理披露算法的用途、运行逻辑、潜在风险等信息。

对应到本文讨论的三类产品形态，各维度的侧重点有所不同，如形态叠加则重点叠加：

- 对于情感型 AI，人类福祉和可控可信最为关键，审查的核心是算法是否可能损害用户心理健康、是否存在失控的情感操纵风险；
- 对于虚拟人，隐私保护和透明可解释是重点，审查的核心是敏感个人信息的使用是否合规、用户是否能清晰辨别虚拟身份；
- 对于具身智能，责任可追溯和隐私保护是核心关切。前者直接关联事故归责，审查的核心是当机器人自主决策造成损害时，责任链条是否清晰可回溯；后者则指向具身设备搭载的物理传感器对周围环境的持续采集，这类数据收集往往在路人无感知的状态下发生，隐私侵害的隐蔽性远高于纯线上场景。

（二）审查程序与高风险活动

《AI 伦理审查办法》结合 AI 科技活动的特点，明确了申请与受理、一般程序、简易程序、专家复核程序、应急程序、登记备案等多种程序要求。其中，以下三类活动被列入《需要开展科技伦理专家复核的人工智能科技活动清单》，需在单位初步审查后，由主管部门组织开展专家复核，**本文讨论的拟人化 AI 则很可能落入其中。**

- **第一类：**对人类主观行为、心理情绪和生命健康等具有较强影响的人机融合系统的研发。陪伴型具身智能因涉及心理情绪影响，可能直接落入此类。
- **第二类：**具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发。大规模虚拟主播、AI 意见领袖等可能触发此类。
- **第三类：**面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发。工业型机器人如涉及高度自主决策，可能触发此类。

（三）企业减负措施

1. 利用制度衔接，减轻重复负担

《AI 伦理审查办法》明确规定，已实行登记、备案、行政审批等监管措施且将科技伦理作为审批条件的，可不再开展专家复核。举例而言，如果一家企业已经按照《拟人化互动服务暂行办法》完成了安全评估和算法备案，且该备案流程中已将科技伦理要求纳入审批条件，则其拟人化互动产品可不再单独申请伦理审查的专家复核。这一机制体现了制度间的协调配合，有效节省了企业的合规成本。

2. 中小微企业依托第三方机构

《AI 伦理审查办法》从标准建设、服务体系、鼓励创新、宣传教育、人才培养五个方面制定支持举措，构建单位伦理委员会与第三方服务中心的双重架构。对于大型企业，核心义务是建立内部伦理委员会、配备专职伦理审查人员；对于缺乏独立审查能力的中小微企业，则可委托地方或相关主管部门设立的“人工智能科技伦理审查与服务中心”获取支持。

五、他人之石：欧美的拟人化、情感型 AI 监管

中国对拟人化 AI 的专项监管并非孤例。以欧美为例：

- **美国：**联邦层面目前尚未形成统一的综合性 AI 立法，但州层面的针对性监管已经开始推进。以情感型 AI 为例，加州于 2025 年通过 SB 243，对 AI 伴侣聊天机器人的运营者提出了自杀意念识别和安全干预等要求；纽约州已通过相关法律安排/预算法中的相关规定，并已生效。在近期一起涉及 AI 伴侣聊天机器人的佛罗里达州联邦地区法院案件中，法院在驳回动议阶段未支持被告关于 LLM 输出当然受第一修正案保护的主张，并认为在原告指向的是应用设计缺陷而非输出表达内容的情况下，Character A.I.可被置于产品责任框架下审查。该案至少表明，AI 企业并不能当然以“AI 生成内容”抗辩产品责任主张。
- **欧盟：**欧盟《人工智能法案》（AI Act）虽已步入适用期（2026 年 8 月 2 日起分阶段生效），但对情感型 AI 的定性仍在演进：欧洲议会议员正在推动将 AI 伴侣明确列为“高风险”类别。此外，意大利数据保护机构已对 Replika 情感伴侣应用处以 500 万欧元罚款，显示欧洲监管机构对情感型 AI 的执法态度日趋严格。因此，情感类 AI 工具出海欧盟，需持续性关注、落实透明度义务（如告知用户其正在与机器交互），并持续关注更新的业态发展。

纵观中美欧三大法域的立法动向，“AI 身份透明”正在成为全球共识底线，无论各法域的监管路径如何不同，要求 AI 在与用户交互时披露其非人类身份，是目前在中美欧三方都已落地或即将落地的硬性义务。但各法域的监管侧重点呈现明显分化 — 中国以专项立法先行，围绕“持续性情感互动”设定精细化行为红线；美国在联邦层面缺位的背景下，由州级立法和司法判例共同塑造规则边界；欧盟则依托风险分级框架进

行体系化归类，但对情感型 AI 的具体定位仍在演进中。

六、结语与建议

拟人化 AI 的风险跨越内容安全、数据保护、未成年人保护与伦理审查等多个领域。企业应以产品为中心，建立统一的风险评估与应对机制，而非围绕单一法规被动整改。

第一步：以产品形态快速定性，避免机械对照法规条文

企业应首先基于产品功能进行判断，明确其是否同时具备以下特征：

- 是否存在持续性情感互动（对应拟人化互动）；
- 是否以可识别的具体形象、身份与用户交互（对应数字虚拟人）；
- 是否进入现实物理空间并具备感知、行动能力（对应具身智能）。

同一产品可能同时落入多个监管框架（不限于本文提及的规定）。合规路径应以“形态叠加”为判断前提，而非仅针对单一法规进行适用。

第二步：围绕关键时间节点，倒排合规工作安排

- 于 **2026 年 7 月 15 日**前，完成拟人化互动服务的合规整改（包括但不限于生成合成内容标识、未成年人安全评估、科技伦理审查等）；
- 关注数字虚拟人相关办法的公开征求意见截止时间（**2026 年 5 月 6 日**）及其后续出台。

第三步：将 AI 科技伦理审查前移至“产品设计环节”

企业不宜将伦理审查简单理解为上线前的程序性步骤，而应在产品设计阶段即完成以下自查：

- 是否存在对用户心理、情绪或行为产生实质性影响的功能设计；
- 是否可能触及专家复核清单所列高风险活动；
- 是否已具备风险识别、干预与责任回溯的技术基础与流程保障。

对于中小企业而言，若拟开展的技术开发活动落入审查范围，可优先对接第三方伦理审查与服务中心，以避免因内部程序缺位而影响产品上线节奏。

特别声明

汉坤律师事务所编写《汉坤法律评述》的目的仅为帮助客户及时了解中国或其他相关司法管辖区法律及实务的最新动态和发展，仅供参考，不应被视为任何意义上的法律意见或法律依据。

如您对本期《汉坤法律评述》内容有任何问题或建议，请与汉坤律师事务所以下人员联系：

段志超

电话： +86 10 8516 4123
Email： kevin.duan@hankunlaw.com