HANKUN

汉坤法律评述

2026 年 8 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

汉坤具身智能数据合规与治理系列（四）：AI 治理

作者：陈文昊 | 原舒仪 | 李诗

摘要

本系列第一篇将具身智能治理划分为隐私保护、重要数据、AI 治理与生态治理四个层次，并将 AI 治理的对象界定为基础模型、世界模型、动作模型与智能体（Agent）的全生命周期，核心是判断模型与 Agent 在具体产品中是否安全、可信、可控。本文沿用这一框架，讨论具身智能 AI 治理如何从“输出内容是否安全”进一步延伸至“系统行为是否安全”。

模型偏差、模型幻觉、Agent 越权调用以及身份误认与拟人化，并非具身智能独有。其特殊性在于，模型判断可能经由 Agent 规划、工具调用和设备控制，直接转化为移动、抓取、开门、支付或生产控制等现实行动，并由此引发产品质量、消费者权益、合同履行、侵权责任、科技伦理及行业监管等问题。

本文首先说明四类风险如何沿“感知—判断—规划—调用—执行”链条传导；继而从模型、Agent 和人机交互三个层面，讨论训练数据、版本溯源、场景化评测、工具与动作权限、长期记忆、人工确认、身份标识及拟人化边界；最后分析上线、变更、事件处置与停用阶段的持续治理。

一、从内容输出到现实行动：具身智能 AI 治理的特殊性

具身智能 AI 治理的对象并非单一大模型，而是基础模型、世界模型、动作模型与 Agent 共同构成的决策和行动体系。在部分系统中，基础模型承担语言、视觉和多模态信息理解；世界模型理解周围环境并预测行动后果；动作模型将任务目标转化为设备可以执行的动作方案；Agent 则通过记忆、规划和工具调用连接上述能力。本文讨论的四类风险，正是这些治理对象在具体场景中的主要风险表现。

（一）风险传导方式的改变

单纯的大模型产品同样可能出现偏差、幻觉、越权调用或拟人化问题，这些风险通常先表现为错误、失实或不当内容，现实后果还需经过用户采信、平台传播或人员执行。具身智能则将模型嵌入传感器、Agent、业务接口和设备控制器，模型输出可能直接进入执行环节。

二者真正的分野，不在于是否具有类人外观，而在于判断与行动之间的距离被显著缩短。同一句错误指令，由聊天机器人给出，可能只是需要纠正的信息；由能够控制机械臂、门锁或支付账户的 Agent 执行，则可能造成财产损失、人身伤害或生产中断。法律关注由此从内容治理延伸至产品缺陷、合理注

意义务、侵权责任、生产安全和行业准入。碰撞、误抓取或错误启停等“危险动作”，则是需要单独控制的行为后果。

（二）四类风险应当放在行动链中理解

模型偏差，是指模型受训练数据、标注方式、算法设计、优化目标或部署环境等因素影响，对特定人群、对象或场景形成系统性、持续性的识别或决策倾向。在具身场景中，偏差可能进一步改变设备的服务对象、行动路径或安全距离。

模型幻觉，是指模型在缺少充分事实、感知信息或知识依据时，仍生成虚构、错误或相互矛盾的判断或行动方案。在具身场景中，它还可能表现为误判障碍物、误认任务已经完成，或在信息不足时继续生成动作指令。

Agent 越权调用，是指 Agent 超出用户指令、任务目的、角色权限或系统预设边界，访问数据、调用工具、使用资源或执行动作。越权既包括突破技术权限，也包括虽有接口权限，却将其用于未经授权的对象、目的或范围。例如，用户要求机器人“完成采购”，不等于授权其选择任意供应商、承诺任意价格并直接付款。

身份误认与拟人化风险，是指产品因名称、外观、声音、长期记忆或情绪回应等设计，使用户对其人工智能属性、实际能力、专业资格或情感关系形成错误认识。对于能够持续陪伴、主动行动的设备，这种错误信赖更容易延伸至行动层面：用户不仅相信其“说得对”，还可能误以为它“知道该怎么做”。

四类风险并非彼此孤立。偏差和幻觉可能形成错误判断，Agent 越权会扩大判断的可执行范围，身份误认又可能降低用户核验和制止系统行为的意愿。治理的关键，是识别风险在任务链中的连接方式，并在损害发生前设置阻断点。

（三）场景是合规判断的基本单位

同一套模型用于家庭陪护、医院辅助或工厂作业时，任务、环境和错误后果并不相同。家庭机器人误认家庭成员，可能涉及隐私、消费者权益和人身安全；医院设备作出类似误认，还可能影响诊疗秩序和医疗器械管理；工业机器人在人员与设备混行区域误判，则需同时考虑产品安全与生产安全。因此，场景是确定适用规则、注意义务和控制强度的事实基础。

因此，使用已备案的大模型或算法，并不能当然证明整机产品已经履行全部合规义务。备案覆盖的主体、模型版本、服务功能和测试条件，未必与具体部署方式一致。模型提供者、Agent 开发者、设备制造商、系统集成商和现场运营者控制哪些环节，仍需结合实际产品链条判断；合同分工也不能排除各主体依法应承担的责任。

场景评估应贯通产品事实与控制措施：系统使用哪些模型、知识库、插件和控制组件，可以接触哪些数据、调用哪些工具、实施哪些动作；信息不足、权限冲突或设备异常时由谁停止；发生事件后能否还原判断和执行过程。下文将这些问题嵌入四类风险的治理逻辑。

二、模型偏差与幻觉治理：证明模型适合预期场景

偏差与幻觉的成因虽不相同，在合规审查中却指向同一问题：企业凭什么相信某个模型适合其承诺的场景，并允许模型判断进入后续任务链。数据审查、模型溯源和评测，正是回答这一问题的证据方法。

（一）产品的预期用途决定审查边界

法律通常不会要求模型绝对无偏、永不出错。企业是否尽到合理注意义务，需要结合产品用途、错误的可预见性、可能损害和现有技术条件判断。《生成式人工智能服务管理暂行办法》有关防止歧视和提高训练数据质量的规定提供了基本依据；模型嵌入实体产品后，还需叠加审查产品质量和行业规则。

审查不能只看整体准确率。例如，仓储机器人主要依据照明良好、货架整齐的样本完成测试，进入人员频繁穿行、光线变化较大的仓库后，识别和避障能力可能明显下降。此时未必是数据来源违法，而是现有证据不足以支持企业对该场景作出的性能承诺。

法务与合规部门无需代替研发选择算法，但应将产品承诺转化为可验证的条件：系统面向哪些对象，在何种光照、噪声、网络和空间条件下运行，错误可能影响何种权益。未经验证的能力，不宜宣传为现实环境中的稳定表现；已知限制也应进入产品说明、销售材料和交付文件。

（二）溯源和评测的价值，在于支撑法律判断

为说明模型适合特定用途，企业首先要知道系统实际使用了什么。除基础模型的提供者、名称和版本外，还应识别对结果有实质影响的数据集、微调方法、提示词、检索知识库、长期记忆、插件和安全规则。溯源不是为了建立庞杂档案，而是为了发现可能改变系统行为的关键要素，并为缺陷调查和责任划分保留依据。

数据来源合法与数据足以支持预期用途，是两项不能相互替代的审查。前者关注个人信息、知识产权、保密义务和数据安全，后者关注样本能否覆盖实际对象和环境、标注规则是否一致。《网络安全技术 生成式人工智能预训练和优化训练数据安全规范》（GB/T 45652—2025）和《网络安全技术 生成式人工智能数据标注安全规范》（GB/T 45674—2025）可以为训练数据处理、标注和质量核验提供技术参照，但符合相关标准不会自动证明产品适合某一具体场景。

外购模型也不能成为责任调查的中断点。整机企业即使无法获得全部训练语料和底层参数，仍可通过尽调和合同确认接入版本、许可用途、已知限制、数据回传、更新通知和事件协查机制。备案信息和供应商评测报告可以作为材料，但应核对其版本、功能和测试条件是否与实际产品一致。这里的“版本”还包括数据集、知识库、提示词、插件和控制参数在特定时点的状态。

评测范围应从模型单点延伸至完整任务链。偏差测试要观察不同人群、物体和现场条件下是否存在持续性差异；幻觉测试要检验感知缺失、信息冲突或工具异常时，系统是否仍会生成确定性判断并继续规划。只检查语言输出，无法说明错误会不会被 Agent 转化为工具调用；只检查末端动作，又难以还原错误来源。

（三）未经验证的判断不得无条件进入执行环节

具身智能无法彻底消除模型错误，合规控制的现实目标是避免错误判断直接触发不可接受的行动。问答、提醒等低风险任务可以通过提示和事后纠正处理；涉及人身安全、重大财产、医疗辅助或生产控制的任务，则需设置更严格的执行条件。即使系统能够给出识别概率或判断把握程度，也不能据此认定结论可靠，更不能将其作为自动执行高风险动作的唯一依据。

企业可以对关键事实交叉核验，在信息不足时拒绝执行或转交人工，并对动作对象、空间、速度和力度设置独立于模型的安全规则。传感器冲突或运行环境超出适用范围时，系统应进入预定的安全状态。相关措施能否阻止可预见的模型错误转化为现实损害，是评价其合规意义的关键。

评测结论还应约束上线决定。发现特定场景存在显著偏差或高频幻觉后，可以补充数据、调整模型、限制用途、增加人工复核或暂不开放相关功能；决定接受剩余风险的，应说明理由、批准权限和持续监测方式。研发结论、产品承诺与实际控制保持一致，才能形成可供监管检查和争议处理的证据。

三、Agent 越权与危险动作治理：限制错误判断的现实后果

模型判断通常要经过 Agent 的任务分解、工具调用和控制组件放行，才能影响现实。因此，Agent 治理既要审查它“能做什么”，也要审查它“在什么条件下有权做”；危险动作则需在执行层被独立控制。

（一）技术权限不等于法律授权

接口调用成功，只能说明系统具备技术能力，不能证明调用具有合法授权。账户能够读取全部客户资料，不意味着 Agent 查询一份订单时可以调取其他客户信息；设备可以打开所有房门，也不意味着一句“帮我收拾房间”已经授权其进入任何区域。判断是否越权，应同时审查用户指令、任务目的、企业自身权限和产品预设边界。

权限设计不宜只有“允许”与“禁止”两个状态。数据访问、工具调用、资源使用和物理动作所依据的授权及可能后果不同，应分别设置适用对象、目的、参数范围、期限和撤销条件。维护调试、正常运行和紧急处置也应采用不同权限，任务结束后临时授权应及时失效。

涉及个人信息、商业秘密、客户系统或资金支付时，还要核对企业自身是否有权将相关数据或处分能力交给 Agent。用户的概括同意不当然覆盖所有后续调用。技术权限矩阵只有与法律依据、合同安排和内部职责相对应，才能说明授权从何而来。

（二）工具调用与物理动作需要两道边界

高风险工具可以设置允许清单，并采取金额或参数上限、调用频次和双人批准等控制。网页、用户输入或插件返回值可能夹带指令，诱导 Agent 调用不相关接口，因此还要明确哪些外部内容可以成为任务依据、工具结果能否继续作为模型指令，以及调用失败后的默认处理。第三方插件对输入数据的留存和利用，也应纳入尽调与合同审查。

即使工具调用获得授权，物理动作也不应当然放行。护理机器人可以读取用药提醒，不意味着它无需再次核对人员、药品和剂量就能递送药物；工业 Agent 可以生成调节方案，也不意味着控制器应立即改变生产参数。工具权限解决“能否接入”，动作权限则解决“能否在此时、此地、对这一对象执行”。

对移动、抓取、开门、医疗辅助和生产控制等动作，安全阈值、空间限制和设备联锁原则上应独立于模型判断。人员进入危险区域、传感器相互冲突或网络中断时，系统应停止或降级运行，而不是让模型继续猜测。这些限制也是判断产品是否存在不合理危险、企业是否尽到安全保障义务的重要事实。

（三）人工确认是控制措施，不是免责工具

人工确认只有在确认人能够看到关键信息、理解后果并拒绝执行时，才构成真实控制。只显示“是否继续”的弹窗，如果没有说明动作对象、异常状态和风险，也没有修改方案的空间，难以证明人工真正参与判断。需要专业资质才能决定的事项，更不能交给普通用户机械确认。

确认方式还应与风险匹配。金额较小、容易撤销的操作可以一次确认；涉及人身安全、重大财产或不可逆动作的，应展示关键依据，并由有权限的人员核验。即使用户点击同意，企业仍可能因产品缺陷、提示不足或控制设计不合理承担责任，不能把确认按钮当成免责工具。

与确认相配套的是现场和远程停止能力。谁可以按下急停、远程运营方何时中止、通信中断后设备进入何种状态，都应在上线前确定并测试。异常发生时能够真正切断“规划—调用—执行”链条，人工介入才不只是文件中的原则。

（四）记录应当能够还原授权和行动链

发生投诉或事故后，仅保留最终动作结果通常不足以判断责任。运行记录应在合理范围内还原当时使用的模型和 Agent 版本、关键输入、权限取得、工具调用、人工干预及设备反馈。日志并非越多越好，还要兼顾个人信息保护、商业秘密和保存期限；关键是能够解释决策过程及控制措施是否生效。

长期记忆也是授权治理应当关注的对象。错误、过时或跨用户串用的记忆，可能在多个任务中反复影响 Agent 判断。企业应明确记忆的来源、写入条件、适用对象、有效期限和纠正删除方式；身份、健康、权限和设备状态等足以改变行动的重要事实，不宜仅凭模型推断自动写入。

四、身份误认与拟人化治理：校正实体交互中的错误信赖

身份误认虽发生在用户认知层面，却会反过来影响行动风险。面对具有实体外观、自然声音和持续记忆的机器人，用户更容易高估其知识、专业能力和对现实后果的理解，进而在未经核验时扩大授权。

（一）身份提示与内容标识解决不同问题

人工智能身份提示旨在使用户知道正在与 AI 而非自然人互动，并理解系统的能力边界；生成合成内容标识则用于说明文字、图片、音频、视频等内容的生成来源和传播属性。前者回应交互对象及能力误认，后者回应内容溯源，不能以一处“AI 生成”标记替代对设备身份和行动能力的说明。

《人工智能生成合成内容标识办法》及配套强制性国家标准已于 2025 年 9 月 1 日起施行。具身智能设备向用户生成或对外传播文字、语音、图片和视频的，应按具体服务和传播方式判断显式、隐式标识义务；模型仅在设备内部用于路径规划或控制计算，并不当然产生内容标识要求。语音播报、屏幕显示、文件导出和平台上传等功能，应分别判断。

设备身份提示还要适应实体场景。公共场所、医院或工厂中的人员未必是注册用户，也不会都阅读用户协议。设备是否由 AI 控制、何时记录信息、移动或执行动作，可以通过语音、灯光、屏幕和现场标识持续呈现，使现场人员能够保持距离、提出异议或寻求人工帮助。

（二）拟人化互动应当守住能力与情感边界

仅仅提示“这是 AI”，未必足以消除错误认识。产品名称、实体外观、声音语气、职业装束、长期记忆和情绪回应，会共同塑造用户对系统身份与能力的理解。机器人如果以“医生”“律师”“老师”或真实工作人员的身份和口吻作出判断，容易使用户误认为其具有相应专业资格。通过拟人化表达强化信赖，却未如实说明能力边界和使用限制，还可能引发误导宣传、合同说明不足及侵权责任等问题。

《人工智能拟人化互动服务管理暂行办法》（以下简称《拟人化互动办法》）已于 2026 年 7 月 15 日起施行。其调整的并非所有具有人形外观、自然声音或拟人化名称的产品，而是向境内公众提供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。智能客服、知识问答、工作助手等不涉及持续性情感互动的服务，一般不适用《拟人化互动办法》。企业应结合互动目的、持续程度、情感属性和服务对象判断其适用性。

对于进入适用范围的服务，合规要求不止于表明人工智能身份。《拟人化互动办法》要求提供者具

备过度依赖风险预警、情感边界引导和心理健康保护等能力，不得以替代社会交往、控制用户心理或诱导沉迷依赖为服务目标，也不得通过情感操纵诱导用户作出不合理决定。企业还应审查产品是否利用用户的孤独、焦虑或情感依赖，促使其持续付费、扩大授权或损害自身权益。

身份提示应随风险变化持续发挥作用。《拟人化互动办法》要求提示用户正在与人工智能而非自然人互动；发现过度依赖或沉迷倾向时，应显著提醒；连续使用每超过两小时，还应提示使用时长。用户要求退出的，提供者应及时停止服务，不得以持续对话或情感挽留阻碍退出。对于能够主动接近、跟随用户或在家庭空间活动的设备，退出机制还应与停止交互、关闭长期记忆和限制动作相衔接。

未成年人和老年人更容易受到拟人化表达和情感反馈的影响。《拟人化互动办法》禁止向未成年人提供虚拟亲属、虚拟伴侣等虚拟亲密关系服务；向不满十四周岁的未成年人提供其他拟人化互动服务，应取得监护人同意并设置保护机制。面向老年人的服务则应加强健康使用指导并显著提示风险。陪伴、照护类机器人还应结合用户年龄、认知能力和使用环境设置身份提醒、时长与付费限制、风险预警及紧急联系机制。

即使产品不属于《拟人化互动办法》规定的持续性情感互动服务，消费者权益保护、广告宣传、个人信息保护、产品质量和侵权责任等一般规则仍然适用。企业还应审查产品名称、销售话术、交互脚本和设备行为是否共同制造了超出真实能力的印象。拟人化程度越高，能力说明、情感边界和停止方式越应清晰；与整体设计相矛盾的免责声明，通常难以消除用户已经形成的合理信赖。

五、从上线审查到持续治理：合规结论应随系统变化更新

偏差、幻觉、越权调用和拟人化风险，不会在产品上线时停止变化。具身智能的实际行为取决于模型、知识库、插件、长期记忆、权限配置、控制参数和部署环境，任何关键要素变化都可能使原有评测和法律判断失去基础。因此，场景评估不应是上线前的一次性程序，而应成为决定产品能否发布、何时重新审查以及异常后如何处置的持续治理机制。

产品上线前，企业应将前述审查汇总为明确的场景合规结论，说明经过验证的预期用途、禁止或限制使用的场景、需要保留的人工控制及主要剩余风险。未经相应场景验证的能力，不应仅凭模型备案、供应商承诺或通用评测结果上线。涉及科技伦理审查、算法备案、内容标识或行业准入的，还应确认相关程序覆盖的版本、功能和部署条件与实际交付状态一致。

产品上线后，更换关键模型、新增知识库或高风险插件、扩大长期记忆、提高动作权限、启用持续学习，或进入新的场景，都可能改变原有合规结论。企业可以实施分级审查：一般修复由常规测试确认；涉及模型能力、工具权限、安全阈值或适用场景的重大变化，则重新开展专项评测和法律审查，并判断是否需要履行备案、科技伦理审查、产品测试或客户通知等程序。新版本未经验证或出现重大风险时，应能够停止使用并恢复到此前经过测试和批准的稳定状态。

对于同时处理重要数据的项目，AI 治理还应与数据安全风险评估衔接。《网络数据安全风险评估办法》有关定期评估、重大变化及时评估及报告保存的要求，可能与模型、Agent 或部署场景变更同时触发。两类评估可以共享材料，但不能相互替代：前者关注模型和 Agent 是否安全、可信、可控，后者关注网络数据处理活动的安全风险及控制措施。

系统异常时，处置不能止于关闭模型接口。企业还应根据任务链暂停工具调用和设备动作，撤销临时权限，隔离可能继续影响 Agent 判断的长期记忆，并保全必要记录。涉及人身、财产或生产安全的，还应衔接用户通知、监管报告、现场警示和必要的召回。模型或 Agent 正式停用后，也应验证旧版本不能继续连接客

户系统和设备，并依法处理相关权限、数据、记忆和缓存。

企业最终需要形成的，不是与产品流程分离的一套文件，而是能够说明风险如何进入经营决策的证据链。场景评估说明企业在上线前知道并判断了什么，发布和变更记录说明相关风险由谁决定接受，运行日志和事件材料则用于还原控制措施是否生效。只有合规结论随产品版本、权限和部署环境持续更新，前述措施才不会停留在一次性评测或原则性文件中。

结语：具身智能的可信与可控，最终要落在行动上

具身智能的四类 AI 风险并非全新的法律概念。特殊之处在于，偏差和幻觉会经 Agent 规划、工具调用和设备控制进入现实世界；身份误认和拟人化又可能使用户放松核验、扩大授权。治理不能停在内容是否正确、模型是否备案，而要继续追问系统为何能够行动、谁有权放行以及何时必须停止。

数据与模型溯源、场景化评测、权限矩阵、人工确认、内容与身份标识，共同服务于同一目标：证明模型适合预期用途，将 Agent 限制在合法授权内，并阻断未经验证的判断转化为危险动作。只有相关控制随版本和部署环境持续更新，AI 治理才能从原则性的“安全、可信、可控”，转化为可验证、可追责的产品能力。

****

本系列最后一篇将讨论更外层的生态治理：当机器人厂商、基础模型公司、云平台、插件提供者、OEM、集成商和客户共同参与产品交付时，企业如何通过第三方管理、接口规则、数据分享和合同安排维持上述控制。

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：

陈文昊

电话： +86 21 6080 0359
Email： wenhao.chen@hankunlaw.com