生成式人工智能“同事化”叙事的伦理后果：基于政策文本的商业美德治理分析

张旭峰[1*] 李涵[2]

（1.国家开放大学 实验学院，北京 100086；2.北亚利桑那大学，美国 86011）

摘 要：生成式人工智能在企业多场景应用中常被包装为“AI 同事/助手”，虽提升协作效率，却可能引发责任稀释、审慎弱化与正直受损。本文基于美德伦理与组织德性，采用政策分析与定性内容分析，对中国相关治理文本及 UNESCO、OECD、NIST AI RMF、欧盟 AI Act 等框架进行编码比较，考察透明告知、人类监督、风险评估与可追溯救济如何制度化促进审慎、责任与公平。研究发现政策总体强调“可控、可追责、可告知/可解释”，但对拟人化叙事导致的责任归因偏移关注不足。建议将拟人化设计风险纳入评估清单，强化人类最终责任提示与内部问责矩阵，并细化对话系统透明与不确定性沟通要求。

关键词：生成式人工智能；政策分析；美德伦理；拟人化；责任稀释；组织治理

一、引言

生成式人工智能（Generative AI）正以对话式交互、内容生成与检索增强等能力，嵌入企业的知识生产、客户服务、营销传播与决策支持链条。与传统信息系统以“规则—界面—流程”驱动不同，生成式系统更像一种语言化的协作装置：它通过自然语言把“信息整合—推理建议—行动提示”连续呈现，使用户在体验上更接近与“能沟通、能解释、能配合”的对象协作。正因如此，市场与组织内部叙事普遍将其包装为“AI 同事”“AI 伙伴”“第二大脑”等拟人化角色，用以降低采用门槛、提高使用频率，并在绩效压力与时间稀缺的现实环境中，快速建立可用性与信任感。

从商业伦理视角看，“同事化”叙事并非仅是传播策略，而是一种会改变责任结构与判断机制的组织性安排。首先，它可能触发责任稀释：当建议以“同事式”语气出现时，失误更容易被解释为“系统也参与了判断”，从而在心理与组织层面形成责任分摊，弱化个人的谨慎义务与组织的最终担责意愿。其次，它可能导致审慎弱化：拟人化通常会提升信任与依赖，进而降低复核、反证与二次验证的动机；在高节奏业务中，这种“信任捷径”会把效率优势转化为系统性风险。第三，它可能侵蚀组织正直：当组织以拟人化界面掩盖能力边界、数据依据与不确定性，透明义务就会被“拟似人格”遮蔽，用户对风险与权利的理解被削弱，进而损害诚信沟通与公平对待。

更重要的是，上述风险并不完全由技术属性决定，而在很大程度上取决于制度环境与组织治理方式。政策与标准通过透明告知、标识义务、人类监督、风险评估、审计记录与救济机制等制度装置，界定企业在“可解释、可追责、可控”的最低治理边界；但这些装置未必自动覆盖拟人化交互带来的心理归因偏移与组织问责扭曲。换言之，企业可能在形式合规上满足“标识—监督—记录”，却仍通过“AI 同事化”叙事制造过度信任与责任漂移，从而出现“合规满足但德性受损”的结构性空间。近期关于拟人化/拟主体化智能体的研究也提

作者简介：张旭峰（1998-），男，硕士研究生。
李涵（1993-），女，硕士研究生。

示：当系统被设计为更具诱导性与社会吸引力时，其“可被信任”与“更易被依赖”的一体两面会同时放大收益与风险[1]。

基于此，本文提出如下研究问题：现有生成式人工智能及相关数字治理政策，如何通过制度装置促进组织层面的商业美德（如审慎、责任、正直与公平）？又在哪些环节忽视了“同事化/拟人化叙事”引发的责任归因偏移与审慎动机下降？本文的分析目标并非评价某一政策优劣，而是解释政策工具如何在组织层面发挥“德性工程”效应，并识别其在拟人化交互风险上的治理断裂点，为后续政策细化与企业治理设计提供可操作的论证基础。

二、理论基础与文献述评

2.1 美德伦理与组织德性：从个体品质到制度化能力

美德伦理将道德评估的重心放在行动者的稳定品质及其养成机制上，强调“如何成为一个更好的行动者”，而非仅回答“某个行为是否合规”或“后果是否最大化”。在企业情境中，这一视角具有独特解释力：企业并非单一主体，而是由岗位分工、授权链条、激励约束与流程规范构成的制度化行动者。因而，商业伦理风险往往不是“某个个体不够道德”，而是组织在结构上把某些品质（例如审慎复核、诚实沟通、责任承担）变得昂贵或不必要，进而诱发系统性偏差。

“组织德性”据此可被理解为一种能力：组织能否把关键美德嵌入制度与日常决策，使其不依赖个体偶然自觉，而成为可执行、可监督、可纠偏的常规实践。生成式 AI 的引入改变了判断与行动的分工方式——系统开始参与信息整合、生成解释、提出建议甚至形成默认选项——这意味着组织德性将面临再分配：审慎不再只是“员工是否小心”，责任也不再只是“谁签字”，而是必须通过制度设计把“谁能干预、谁应复核、谁对后果负责”重新固定。

2.2 责任稀释与“道德褶皱区”：自动化情境中的问责扭曲

在复杂自动化与人机协作系统中，责任往往沿着“研发—部署—使用—管理”跨主体流动，导致责任边界难以清晰定位。Elish 提出的“道德褶皱区”（moral crumple zone）概念指出：当自动化系统出错时，责任可能被误配给控制力有限的个体操作者，或在“系统看似自主”的叙事中被稀释，从而形成问责与控制能力不匹配的结构性扭曲[2]。

将该理论引入“AI 同事化”语境，可以更精确地理解责任稀释的发生机制：拟人化叙事一方面可能把系统包装为“能判断的行动者”，在事前促成对系统建议的默认采纳；另一方面在事后又可能把责任推回一线人员或分散于“人机共同体”，使组织难以建立稳定的担责与纠偏路径。此时，问题不只是“有没有人类监督条款”，而是组织是否拥有与其技术配置相匹配的问责结构与证据链，以避免责任在叙事与流程中被折叠变形。

2.3 拟人化、信任与责任归因：心理机制与商业伦理含义

拟人化之所以重要，根本原因在于它会改变信任的生成方式与归因框架。Waytz 等关于自动驾驶场景的研究表明，系统被拟人化后会显著提升信任水平，并影响个体对系统能力与可靠性的判断路径[3]。更进一步的人机协作研究显示，在团队任务中个体可能在行为层面更频繁接受 AI 的决策建议，而“系统身份的呈现方式”会影响合作与信任校准，并引出对“欺骗性身份呈现”的规范担忧[4]。这些发现共同指向一个关键事实：拟人化并非中性的界面装饰，而是会系统性改变个体在复核、质疑、追责时的心理阈值与行动倾向。

同时，拟人化还具有更强的“关系化”潜力：当系统被设计为更像可互动的主体时，用户更可能在自我概念与情感层面对其产生整合或依附，这会强化依赖并放大社会性影响效应[3]。对商业伦理而言，其后果并不止于“信任更高”，而在于：审慎（复核意愿）、责任（归

因与承担）与正直（透明沟通）的实现条件被重塑——组织若缺乏配套的透明义务、不确定性表达与责任锚定制度，就可能在效率提升的同时，把道德风险嵌入日常流程。

2.4 小结：从“合规”到“德性”的分析入口

综上，生成式 AI 的伦理问题不能仅被理解为内容风险或数据合规，而应被视为组织德性结构与责任分工的再配置问题。政策文本在其中具有关键意义：它不仅设定最低合规边界，更通过透明、问责与人类监督等要求，影响企业如何把“应当具备的品质”转译为组织化能力。UNESCO 的伦理框架明确指出透明与可解释性是权利保障与责任/责任制度有效运行的重要前提，这为从制度层面讨论“正直—责任—审慎”的组织化提供了规范支点[5]。因此，本文以政策分析与定性内容分析为方法，聚焦“政策工具如何塑形组织美德”以及“拟人化叙事为何可能成为制度盲区”，从而为后续的治理建议提供更具可操作性的理论与证据基础。

三、研究设计：政策分析与定性内容分析

3.1 方法选择的合理性

本文采用政策分析与定性内容分析，原因在于：其一，企业 AI 伦理风险高度制度化，关键约束来自法规、部门规章、标准与治理框架；其二，“同事化/拟人化叙事”的责任后果，往往以“透明义务、标识义务、人类监督与问责机制”等形式进入政策文本，适宜通过文本编码识别治理逻辑与制度缺口。

3.2 政策样本与纳入标准

（1）中国政策群：生成式人工智能服务管理暂行办法、互联网信息服务算法推荐管理规定、个人信息保护法、AI 伦理安全风险防范指引等，并补充纳入近期“生成合成内容标识”专门规范作为政策演进证据[6][7][8][9][10]。

（2）国际政策群：UNESCO《人工智能伦理建议书》、OECD《人工智能委员会建议/原则》、NIST AI RMF 1.0、欧盟 AI Act（含人类监督与透明义务脉络）[5][11][12]。

纳入标准强调可复现性：文本需为公开可得、具有明确规范性（法律/规章/标准/权威框架），且与生成式 AI、算法治理、透明问责或伦理风险控制直接相关。

3.3 编码框架：从政策工具到组织美德

本文构建“政策工具—伦理能力—组织美德”三层编码框架（A－F），并将其映射为组织层面的关键美德机制：

编码维度	关键制度装置	对应伦理能力	组织美德
A 透明告知/标识	交互告知、生成内容标识、能力边界说明	信息对称与可理解沟通	正直/诚实
B 人类监督与最终责任	人工可干预/可停止、重要决策“辅助而非替代”	保留人类判断与承担	责任
C 风险评估与审计	事前评估、持续监测、复核验证、记录回溯	可验证的审慎流程	审慎
D 数据与隐私治理	合法性、最小必要、透明原则、权利救济	权利保护与不歧视	尊重与公平
E 伦理治理能力建设	伦理审查、培训、投诉举报、救济补偿	组织化治理与学习	制度性德性
F 误用防范与价值对齐	禁止诱导沉迷/过度消费、弱势群体保护	克制与公共利益导向	克制与公共精神

表 1 编码维度（政策工具）—制度装置—伦理能力—组织美德映射表

3.4 编码与信度控制

在实际写作中可采用“双人独立编码—协商一致”的路径：先对样本文本进行试编码以

校准口径，再对全量文本编码并记录分歧条目，通过协商形成一致编码表；必要时报告一致性指标与修订规则，以提升可复现性。

四、政策文本分析结果

4.1 中国政策群：以“主体责任—过程审慎—标识透明”为主轴

（1）生成式 AI 暂行办法：以提供者责任与内容标识构建可追责边界

生成式 AI 暂行办法将“提供者”置于责任轴心，强调安全与合规义务，并明确要求对图片、视频等生成内容进行标识（并与深度合成规则衔接），同时设置投诉举报机制与违规处置要求[6]。这一结构在美德伦理意义上，相当于通过外部制度把“责任”锚定在组织主体，而非让责任在“系统—员工—用户”的互动中漂移。

（2）算法推荐管理规定：把审慎转化为组织流程，并强化透明与用户权利

算法推荐规定提出算法服务应遵循公正公平、公开透明、诚实信用原则，并要求建立算法机制审核、科技伦理审查、安全评估监测与应急处置等制度；同时要求定期审核评估模型与应用结果，禁止诱导沉迷与过度消费等违背伦理的做法[7]。此外，规定要求以显著方式告知用户算法推荐服务情况，公示基本原理与运行机制，并提供关闭个性化推荐等选项，体现“透明—自主—救济”的权利保护取向。

（3）个人信息保护法：以“诚信+透明”设定数据治理的正直底线

个人信息保护法强调处理个人信息应遵循合法、正当、必要与诚信原则，并要求遵循公开透明原则，公开处理规则、明示目的方式范围，同时要求处理者对其处理活动负责[8]。在组织德性层面，这些条款构成生成式系统训练数据、日志记录与用户交互数据处理的“正直—尊重”基础约束。

（4）伦理风险防范指引：直接点名“责任边界模糊”等风险，并强调重要决策的人工可控

相关伦理安全风险防范指引将“模糊责任边界”等纳入风险谱系，并提出在公共服务、金融、健康等重要决策中应强调人工监督、紧急干预、信息回溯与救济保障等机制[10]。其治理语言更接近“把审慎与责任制度化”的政策表达，但对拟人化叙事触发的责任归因偏移仍缺乏专门条款。

（5）生成合成内容标识规则的细化趋势：从“内容真实”走向“全流程可追溯”

近期出台的生成合成内容标识规范明确区分显式标识与隐式标识，并将文本、图片、音频、视频、虚拟场景等纳入标识对象，强调生成—传播链条的可追溯治理[13]。这表明中国政策群对“可告知、可追溯”的制度工具持续加码，但其问题结构仍以“内容与传播风险”为中心，对“交互主体性误认”（把系统当作负责任的同事主体）尚未形成对等的治理颗粒度。

小结：中国政策群总体体现出三项治理取向：一是以提供者/服务主体责任为核心的可追责逻辑；二是以评估、审核、备案、日志与应急为核心的过程性审慎；三是以标识与告知为核心的透明治理。

4.2 国际政策群：以“价值原则—风险管理—高风险合规”递进

（1）UNESCO：以人类尊严、人权与责任问责为总框架

UNESCO《建议书》强调人权与尊严、隐私、透明与可解释、责任与问责、人类监督等原则，并要求在影响权利与自由的决策中提供可理解理由与复核纠错渠道，同时强调应告知用户产品或服务是否直接或在 AI 协助下提供[5]。其功能类似“德性方向盘”：规定组织应朝向何种公共价值与道德品质，但落实仍需更具体的治理工具承接。

（2）OECD：以可信赖 AI 的五原则与政策建议推动可操作化

OECD 的法律文书与原则体系将“透明与可解释、稳健安全、问责、尊重人权与民主价值、促进福祉”等作为可信赖 AI 的核心原则，并强调人类能动性与监督、风险管理与责任承担[11]。其对企业治理的启示在于：把“可信赖”拆解为可被治理与审计的维度，从而为组织德性工程提供结构化清单。

（3）NIST AI RMF：以生命周期风险管理将“审慎”转译为管理证据链

NIST AI RMF 以风险管理为中心，提出治理、映射、测量与管理等活动，强调持续性与证据化，将可信赖属性嵌入设计—开发—部署—使用全生命周期[12]。相较价值宣示型框架，RMF 更接近组织层面的“审慎工程”，其重点不是要求组织“更道德”，而是要求组织“能证明自己以可验证方式降低风险”。

（4）欧盟 AI Act：以风险分级与人类监督确立高风险合规底线

欧盟 AI Act 以风险分级为主线，在高风险系统中强调人类监督等制度安排，并关注因过度依赖等因素导致的风险控制失灵[14]。对企业而言，这相当于以强制合规方式要求保留最终可控与可追责结构，但对“拟人化交互如何改变责任归因”的规制仍主要通过一般性的透明与监督条款间接覆盖。

4.3 比较与解释：共同的“可控可追责”，差异化的“问题中心”

总体上，中国与国际框架在“透明、监督、评估、可追溯、救济”方面呈现趋同：均试图把责任从技术话语中拉回组织治理结构。 但两者问题中心存在差异：中国政策群更突出内容治理与提供者责任锚定，并通过标识、备案与安全评估强化可追溯；国际框架更突出人权与社会技术风险治理，并以风险管理框架与高风险合规将原则落地。

五、讨论：政策如何“制度化塑形”组织美德，以及拟人化盲区

5.1 从合规工具提炼德性机制：政策的“德性工程”效应

基于编码结果可以看到，相关治理文本的功能并不止于“禁止违法”，而是在更深层面以可执行的制度装置把组织应具备的伦理能力固定下来，从而对组织德性产生“工程化”的塑形效应。其一，围绕透明告知与标识的条款，通过要求在内容或交互界面呈现可感知的提示、并辅以技术性可追溯标记，降低信息不对称与“能力神话”，使组织在对外沟通中更难以通过模糊身份与夸大能力获利，进而把“正直/诚实”转化为可被检查的合规事实。中国对生成合成内容标识采取“显式+隐式”的制度设计，并明确标识可进入交互场景界面，体现了这种把正直制度化的治理路径。

其二，人类监督与最终责任相关要求，通过“可监控、可干预、可停止、可复核”的制度化安排，实质上是在组织结构中保留一条“人类判断—人类承担”的责任链，防止在高影响决策中形成事实上的道德外包。欧盟 AI Act 将人类监督作为高风险系统的核心要求之一，并明确其目的在于预防或尽量降低对健康、安全与基本权利的风险，且强调对可预见误用情境的控制；欧盟官方释义进一步突出应对“过度依赖”的意识与可覆盖全流程的监督能力。

其三，风险评估、审计与记录回溯等条款把“审慎”从个人品格转换为组织流程与证据链：风险被要求被识别、被测量、被持续管理，组织需要留下可核验的治理痕迹。NIST AI RMF 以“治理—映射—测量—管理”的功能结构，提供了将可信赖性与风险控制纳入生命周期管理的操作框架，使审慎不再依赖个体自觉，而成为跨部门可协同、可复盘的管理活动。

最后，数据与隐私治理、治理能力建设与救济、以及误用防范等条款，共同将“尊重与公平”“制度性德性”“克制与公共精神”落到组织可执行的责任边界上：一方面通过权利保护与救济机制降低伤害扩散，另一方面通过禁止诱导沉迷、过度消费等行为约束技术带来的道德风险外溢。换言之，政策工具在组织内部形成一套“可告知—可监督—可评估—可追

溯—可救济”的治理闭环，其效果并非简单增加合规成本，而是在外部制度压力下推动企业把关键美德嵌入流程与岗位责任之中。

5.2 拟人化“同事化”叙事的制度盲区：从内容真实到主体性误认

尽管现有政策体系在透明、追溯与监督方面已形成较强的治理取向，但仍存在一个关键断裂：多数规则将主要风险聚焦于“内容真实性、数据合规与决策可控”，而对交互层面的主体性误认缺乏同等颗粒度的专门治理——即用户或员工把对话系统当作具备意向、判断与可担责地位的“同事主体”，从而改变其信任校准、复核行为与责任归因方式。以中国标识制度为例，其显著进步在于把标识扩展到文本、图片、音视频乃至虚拟场景，并明确标识既可在内容层呈现，也可在交互界面呈现；但其问题中心仍以“合成内容提示与溯源”为主，未必足以覆盖“同事化叙事”带来的心理与组织后果[9]。

这一盲区会带来至少两种相互强化的机制性后果。第一，责任归因偏移：拟人化往往提高信任与社会性联结，使个体更倾向把系统输出视为“可靠同事建议”，当结果失误时又更容易在“系统—使用者—组织”之间移动责任位置；在复杂自动化系统中，这种偏移可能与“道德褶皱区”现象叠加，出现责任被不当挤压到控制能力有限的一线人员、或被技术表象稀释的扭曲问责[2]。第二，审慎动机下降：当“AI 同事”被配置为权威口吻、确定性措辞与强拟人身份（昵称、头像、人格化表达）时，员工在时间压力与绩效情境中更可能直接采纳建议、减少复核，从而在组织层面形成“合规看似满足、德性实际受损”的结构空间。需要强调的是，这并不等同于政策缺少透明或监督，而是现有制度多以“内容为合成”“系统需可监督”为主要对象，尚未把“交互身份告知、能力边界与不确定性表达的组合义务”明确上升为与内容标识同等可检查、可审计的治理单元。

六、结论与建议

6.1 结论

本文基于政策文本的编码分析表明，中外治理框架普遍通过透明与标识、人类监督、风险评估与审计、可追溯记录与救济机制，强化“可控、可追责”的制度取向，并在客观上推动企业将审慎、责任、正直与公平等关键商业美德转译为可执行的组织能力与流程证据。然而对生成式 AI 的“同事化/拟人化叙事”可能引发的责任归因偏移与审慎弱化，现有政策关注仍不足，导致从合规要求到组织德性塑形之间出现断裂：组织可以在满足“标识—监督—记录”的形式合规后，仍在交互层面通过拟人化策略制造过度信任与责任漂移。

6.2 面向监管与标准制定：政策建议

监管与标准层面可将治理重心从“是否标识合成内容”进一步推进到“如何防止主体性误认与过度依赖”。一方面，应把拟人化设计作为风险源显性纳入评估与审计口径，将昵称/头像/拟人语气/权威化表达与高确定性措辞等界面与叙事要素，纳入可检查的风险项，要求组织对其可能触发的责任归因偏移进行论证与缓释；另一方面，应细化对话式系统的透明义务，形成“交互身份明确告知—能力边界说明—不确定性提示”三位一体的最低要求，并与现有“显式/隐式标识+可追溯”的治理链条衔接，使透明不仅发生在内容层，也稳定发生在交互决策层。同时，对于招聘、授信、医疗辅助等高影响业务，应以证据链底线固化审慎与责任：明确记录、回溯、申诉与人工复核的最低标准，并将“防过度依赖”的监督措施纳入高风险系统使用条件。

6.3 面向企业内部治理：管理建议

企业治理的关键在于把“AI 同事”重新锚定为“可控工具”，并在组织结构中固化责

任承担。建议以责任矩阵（如 RACI）对模型团队、产品团队、业务负责人、合规与一线使用者的职责边界进行可操作化配置，使任何关键决策都能对应到明确的最终责任人，避免“是 AI 建议的”成为卸责语言。其次，应将 UX 与文案纳入伦理审查与变更管理：对拟人化叙事设定红线（例如限制权威化与确定性承诺），并对关键建议输出强制呈现依据、适用条件与不确定性信息，以制度手段恢复复核动机。再次，将美德导向的行为准则嵌入培训与绩效：把审慎复核、透明披露、可解释沟通、对弱势群体敏感等要求转化为岗位可评价指标，使德性成为组织激励的一部分，而非只停留在合规宣讲。

6.4 研究局限与展望

本文局限在于以政策文本为主要材料，尚未直接检验企业执行差异与员工心理归因机制。后续研究可结合企业案例、访谈与对照实验，比较不同拟人化策略（身份呈现、语气权威化、确定性表达等）对审慎复核、责任承担与组织正直的影响，并进一步评估“交互透明—不确定性提示—责任锚定”组合工具在不同业务场景中的治理有效性。

参考文献：

[1] Peter S ,Riemer K ,West D J . The benefits and dangers of anthropomorphic conversational agents.[J].Proceedings of the National Academy of Sciences of the United States of America,2025,122(22).

[2] ELISH C M . Moral Crumple Zones Cautionary Tales in HumanRobot Interaction[J].Engaging Science, Technology, and Society,2019,540-60.

[3] Amani A ,Ana J ,Diana G . AI anthropomorphism and its effect on users' self-congruence and self–AI integration: A theoretical framework and research agenda[J].Technological Forecasting & Social Change,2022,182.

[4] Guanglu Z ,Leah C ,Kenneth K , et al.Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation[J].Computers in Human Behavior,2023,139.

[5] UNESCO. Recommendation on the Ethics of Artificial Intelligence[EB/OL]. (2021-11)[2025-12-19].

[6] 国家互联网信息办公室. 生成式人工智能服务管理暂行办法[EB/OL]. (2023-07-13).

[7] 国家互联网信息办公室, 工业和信息化部, 公安部, 国家市场监督管理总局. 互联网信息服务算法推荐管理规定[EB/OL]. (2022-01-04).

[8] 全国人民代表大会常务委员会. 中华人民共和国个人信息保护法[EB/OL]. (2021-08-20).

[9] 最高人民检察院网站. 关于印发《人工智能生成合成内容标识办法》的通知[EB/OL]. (2025-03-14).

[10] 全国信息安全标准化技术委员会秘书处. 网络安全标准实践指南—人工智能伦理安全风险防范指引（v1.0-202101）[S/OL]. (2021-01).

[11] OECD. Recommendation of the Council on Artificial Intelligence[EB/OL]. (2019-05).

[12] TABASSI E, et al. Artificial Intelligence Risk Management Framework (AI RMF 1.0)[R/OL]. Gaithersburg: NIST, 2023.

[13] 新华社. 国家互联网信息办公室有关负责人就《人工智能生成合成内容标识办法》答记者问[EB/OL]. (2025-03-14)

[14] European Commission AI Act Service Desk. Human oversight & over-reliance (AI Act explanatory resources)[EB/OL]. (2024-2025).

The Ethical Consequences of the "AI-as-Colleague" Narrative in Generative Artificial Intelligence: A Business-Virtue Governance Analysis Based on Policy Texts

ZHANG Xufeng[1]*, LI Han[2]

(1. Experimental College, The Open University of China, Beijing 100086, China; 2. Northern Arizona University, AZ 86011, USA)

Abstract: In multi-scenario corporate deployments, generative artificial intelligence is frequently packaged as an "AI colleague/assistant". While such framing can increase collaboration efficiency, it may also trigger responsibility diffusion, weaken prudential judgment, and erode organizational integrity. Grounded in virtue ethics and the concept of organizational virtue, this study employs policy analysis and qualitative content analysis to code and compare China's relevant governance texts with international frameworks including UNESCO, OECD, the NIST AI Risk Management Framework (AI RMF), and the EU AI Act. We examine how institutional mechanisms—transparent notice, human oversight, risk assessment, and traceable remedies—are institutionalized to promote prudence, responsibility, and fairness. We find that policies generally emphasize "controllability, accountability, and noticeability/explainability", yet devote insufficient attention to attributional shifts caused by anthropomorphic narratives. We therefore recommend incorporating anthropomorphic-design risks into risk-assessment checklists, strengthening cues that reinforce human ultimate responsibility and internal accountability matrices, and refining requirements for transparency and uncertainty communication in conversational systems.

Keywords: Generative artificial intelligence; Policy analysis; Virtue ethics; Anthropomorphism; Responsibility diffusion; Organizational governance;