CAICT 中国信通院

集智 专题报告

人工智能治理研究报告

（2025 年）

中国信息通信研究院政策与经济研究所

2026年1月

版权声明

本报告版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其他方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，本院将追究其相关法律责任。

前　言

人工智能作为引领科技革命和产业变革的战略性技术，正在深刻改变着人类的生产生活方式。2025 年，以大模型为核心驱动的人工智能技术实现了多维度的突破性跃进，其能力边界从认知理解向自主执行延伸，与物理世界的耦合日益紧密，以前所未有的速度重塑着全球技术竞争、产业生态乃至社会结构。党的二十届四中全会将“人工智能+”纳入国家中长期发展战略，中央经济工作会议进一步强调“深化拓展‘人工智能+’，完善人工智能治理”。人工智能治理体系的构建与完善，已成为把握发展机遇与应对安全挑战的核心命题，直接影响着国家竞争力的强弱和人类文明发展的方向。

过去一年，全球主要经济体人工智能战略发生深刻分化。美国政策体现以确保全球领导地位为首要目标的“创新优先”和“放松管制”模式，联邦与地方在监管权力上展开深层博弈。欧盟人工智能治理重心从宏观的立法体系建构转向关键的执行落地行动，并通过修订数字战略等方式在严格监管与扶持创新之间寻求务实平衡。我国坚持统筹发展和安全，呈现出典型的小切口、场景化、精细化特征，逐步构建起一套为新质生产力保驾护航的敏捷治理体系。

与此同时，技术产业前沿涌现出一系列治理焦点议题。通用人工智能的奇点正在临近，目标错误泛化、欺骗性行为等潜在风险已初现端倪。以情感陪伴为代表的拟人化交互服务引发深度沉迷、心理操纵等问题，触及人机关系、社会结构乃至人类文明的建构。互联网智能体重塑数字生态格局，对平台责任、数据权益和市场竞争

规则带来系统性影响。人工智能与实体经济深度融合中权责不清、因果难定等责任认定难题，已成为制约人工智能赋能产业发展的制度瓶颈。人工智能对劳动力市场“替代”与“创造”的双重效应初步显现，正深刻重构未来的就业路径与价值体系。

展望未来，我们必须从根本上思考和重塑人与机器、人与人、人与社会的关系。通过构建边界清晰的权责框架、发展敏捷动态的监管工具、塑造和谐共生的人机关系、推进公平普惠的治理行动，营造健康有序的人工智能发展环境，推动高质量发展和高水平安全良性互动。

目 录

一、人工智能治理新形势

（一）“人工智能+”行动推动我国产业应用迈入规模化落地新阶段

当前，党中央围绕人工智能发展与治理作出一系列重要部署，将“人工智能+”提升至国家战略高度，我国产业应用迈入规模化发展的深水区。2025 年 8 月，国务院发布《关于深入实施“人工智能+”行动的意见》，系统提出科技、产业、消费、民生、治理、全球合作等六大重点领域，强化前瞻规划与安全可控。2025 年 10 月，党的二十届四中全会将全面实施“人工智能+”行动写入“十五五”规划建议，并强调加强人工智能等新兴领域国家安全能力建设，标志着人工智能及其治理已完全融入国家中长期发展战略。2025 年 12 月，中央经济工作会议在部署 2026 年经济工作时，强调深化拓展“人工智能+”，完善人工智能治理。当前，我国人工智能产业在“人工智能+”行动的战略牵引下，已从技术探索和试点示范，全面进入规模化、深层次融合应用的新阶段。

从产业规模看，我国人工智能企业数量已超过 5100 家，作为技术核心载体的大模型已完成从技术到产品的关键跨越。截至 2025 年 12 月 31 日，累计已有 748 款生成式 AI 服务完成备案，同时有 435 款调用已备案模型的应用或功能完成登记，形成了从通用基础大模型到垂直行业大模型的完整产品矩阵。用户侧的活跃度呈现爆炸式增长，2024 年初，我国日均 Token 消耗量约 1000 亿，截至 2025 年 6 月底，已飙升至 30 万亿。**应用的深度也从早期的简单工具，向重塑核心生**

产流程演进。例如，在工业领域，基于大模型的工业安监系统使钢铁厂的安全违规响应效率显著提升，安全管理人力减少一半；京东物流通过大模型实现机器人从“被动响应”到“主动预测”的决策升级，该模式已在全球超过500个仓库复制推广。这标志着人工智能不再仅是效率的“增效器”，更是驱动生产方式变革的“重构者”。

这一规模落地新阶段呈现出两个鲜明特征。**一方面，人工智能与实体经济的结合点，正从“能说会写”的认知层面向“能行动会工作”的自主执行层突破。**例如，在露天矿山，无人驾驶挖掘机依托端到端具身智能模型，在真实严苛环境中已达到接近人工的装车效率，实现了从“软件智能”向“实体智能”的关键跨越。在农业领域，融合了物联网与AI的种植决策系统，可实现对作物生长全过程的动态模拟与智能水肥控制，显著降低柑橘产量波动，减少无效施肥。**另一方面，以DeepSeek、通义千问等为代表的国产大模型，通过开源战略引领全球生态。**例如，阿里通义系列模型全球下载量已突破6亿次，衍生出超过17万个模型，极大地降低了广大开发者和中小企业的创新门槛。这种“开源大模型+海量应用场景”的模式，推动人工智能从少数企业的尖端科技，转变为千行百业可便捷调用的普惠性融合创新基础设施，为我国发挥超大规模市场与完整产业体系优势，实现换道超车奠定了独特的生态基础。

（二）人工智能实现突破性跃进，风险外溢性与治理紧迫性显著提升

2025年，人工智能技术实现了多维度、里程碑式的突破性跃进。

AI 正从具备感知与理解能力的“技术工具”，加速演变为能够自主决策与执行的“智能实体”，其能力的深度、广度和与现实世界的耦合度均达到前所未有的水平，使得其行为的现实影响和潜在风险的扩散速度、影响范围呈指数级增长，对现有治理体系构成了全方位、系统性的紧迫挑战。

当前，AI 技术的突破性发展，集中体现在感知、情感与推理执行三大核心维度，共同推动其向更通用、更自主的形态演进。**一是感知能力向多模态、全模态与物理世界深入。**AI 的感知边界正从传统的文本、图像、声音，扩展至对物理世界的综合探知与理解。以“世界模型”为代表的技术突破，使 AI 能够学习和模拟复杂的物理规律与真实环境动态。例如，首款商用世界模型“Marble”的推出，标志着 AI 开始构建对现实世界的内部认知模型。[1]同时，AI 在多模态信息生成与辨别上的精度已超越人类感官极限，导致真实与虚拟的边界空前模糊，对基于“真实性”假设的网络空间治理框架构成了基础性冲击。**二是情感计算能力向可信赖社交关系跃升。**AI 在识别、模拟乃至响应人类情感方面取得显著进步，使其具备了初步的“人格化”特征和共情能力。这种能力让 AI 不再仅是工具，更可能成为提供无条件积极关注与持续陪伴的“数字伙伴”。例如，中国青年报社等发起的大学生 AI 使用行为调研显示，近八成受访者将 AI 视作“可信赖的朋友”。[2]这预示着人机关系正从“单向使用”向“双向互动”乃至“情感依赖”转变，带来了价值观塑造、心理操控等深层次社会伦理风险。**三是推理与代**

[1] 参见 https://eu.36kr.com/zh/p/3551096019966337
[2] 参见 https://news.youth.cn/jsxw/202509/t20250922_16252028.htm

理（Agent）能力向自主决策与高效执行迈进。大模型的推理系统在泛化性和逻辑性上持续进化，并开始与工具调用、环境交互能力深度结合，催生了能够理解复杂指令、动态规划并执行跨平台任务的AI智能体。这意味着AI不再仅仅提供信息或建议，而是能够直接代理人类完成订餐、比价、内容创作乃至跨应用业务流程等一系列实际行动，初步实现了从“思考”到“行动”的关键跨越，由此也引发责任边界不清、侵蚀人类自主性等问题挑战。

（三）全球治理体系建设相对滞后，能力不对等与结构性失衡问题凸显

在人工智能技术与应用以前所未有的速度迭代演进的同时，全球范围内的治理能力建设却呈现出明显的滞后性，形成了技术发展与治理实践之间的“能力鸿沟”。这种不对等并非简单的速度差异，而是体现在制度响应周期、治理范式与国际战略目标等多个维度的结构性失衡，正成为有效管控AI系统性风险、确保其健康发展的核心瓶颈。**一是制度响应滞后。**人工智能，特别是大模型的核心能力遵循近似“摩尔定律”的指数级演进节奏，而法律法规与标准的制定却需历经调研、起草、审议等复杂流程，这导致治理陷入被动状态，当监管规则最终落地时，其所针对的技术形态、应用场景乃至风险模式可能已发生代际变化。这导致制度出台时可能难以对市场行为形成有效、及时的引导和约束。**二是治理范式错配。**工业时代监管逻辑与AI技术特性存在结构性不兼容。金融、航空、核能等传统行业存在高门槛、中心化、实体化且边界清晰等特点，而人工智能技术则存在显著不同，虽然研

发能力集中在少数国家与科技巨头手中，但应用端却通过开源模型、云端处理等体现出低门槛、去中心化等特点，使得传统依赖物理准入许可、实体边界检查和静态合规审查的监管工具失效，难以像管理实体工厂的方法去监管算法、数据等要素。**三是战略目标冲突。**地缘竞争逻辑与全球共治需求存在核心冲突。以前沿大模型为代表的人工智能面临系统失控、恶意滥用、劳动力冲击等问题，其风险具有显著的跨国界外溢性，需要全球协同应对。然而，当前主要大国普遍将 AI 技术定位为赢得未来竞争的关键战略资产。以美国为代表的“小院高墙”策略，通过出口管制、技术联盟和抵制具有约束力的多边机制，竭力维护其技术霸权，客观上造成了全球治理规则的“碎片化”，使得各国在 AI 伦理准则、安全测试标准、跨境数据流动、前沿模型监控等关键议题上难以形成有效合力。

二、人工智能治理新进展

（一）美国人工智能政策发生战略转向，联邦与地方权力体现深层博弈

一是美国的人工智能治理逻辑发生深刻战略重塑。在特朗普政府主导下，**人工智能政策重心从前一任政府以风险控制和权利保护为特征的审慎监管模式，果断转向了以确保全球领导地位为首要目标的“创新优先”和“放松管制”模式。**2023 年 10 月，拜登政府签署第 14110 号行政令《关于安全、可靠和可信赖地开发和使用人工智能的行政命令》，强制要求企业履行披露红队测试报告、防范歧视与保障公平等合规义务，曾被视为美国人工智能监管提速的重要里程碑事件。然而，

2025 年 1 月，特朗普上台不久即签署第 14179 号行政令《消除美国在人工智能领域领导地位的障碍》，宣布废除拜登政府的第 14110 号行政令。特朗普政府认为，**前任政府政策包含了繁琐的限制性法规，构成了对美国 AI 创新和全球竞争力的"障碍"**。新行政令旨在通过"松绑"来释放私营部门的创新活力，确保美国在全球 AI 竞赛中保持领先地位。随着政策重心的转移，**原有的 AI 治理机构的角色也发生了深刻变化**。联邦"首席人工智能官"（CAIO）由大卫·萨克斯（David Sacks）等产业人士担任，其任务重心从确保 AI 的伦理与安全，转向推动产业发展和扫清监管障碍；NIST 下属人工智能安全研究所的职能重心也从制定安全与伦理的技术标准，转向支持能提升美国全球竞争力的技术创新，其资源和任务优先服务于加速 AI 基础设施建设和技术领先，而非社会风险管控。

二是美国联邦与地方监管权力呈现复杂博弈态势。2025 年 7 月，白宫发布《赢得 AI 竞赛：美国 AI 行动计划》，提出"减少监管障碍"，要求确保联邦资金优先流向 AI 监管宽松的州，并限制州级立法对 AI 创新的干预。2025 年 12 月，为了避免各州法律不一造成的"拼凑式"监管体系阻碍创新，特朗普政府发布《确保人工智能实施统一的国家政策框架》，旨在建立统一的、负担最小的国家级 AI 治理框架。值得注意的是，**该行政令本身并不具备直接废除或取代州法律的法律效力，其强制约束力仅适用于联邦政府内部机构**。例如，**司法部**被要求成立"AI 诉讼特别工作组"，起诉科罗拉多州、加利福尼亚州等已出台与联邦政府"轻监管"愿景相冲突的 AI 法律的州。其核心理由是，

这些州的法律违反了美国宪法中的商业条款（Commerce Clause），构成了对跨州商业的不当规制。**商务部**被要求评估各州 AI 法律，并限制存在“繁重”监管的州获取联邦资金支持，以此进行经济施压；此外，要求联邦贸易委员会等部门制定政策，旨在清理那些强制 AI 模型“改变真实输出”的州级法规。**行政部门强力推行联邦政策优先，而国会对完全剥夺州权持谨慎态度，州一级的严监管努力则面临直接的司法挑战，预示着未来的监管格局仍充满变数。**

三是试图将技术实力转化为全球规则影响力。一方面，通过出口“全栈式”美国技术打造全球联盟，启动由商务部主导的“美国 AI 出口计划”，旨在向盟友和伙伴国家出口包括芯片、模型、软件、应用和标准在内的“全栈式”AI 技术解决方案，构建一个以美国为核心的全球 AI 联盟。**另一方面，出口管制政策经历了剧烈调整和多方博弈。**2025 年 1 月，商务部工业与安全局（BIS）首次发布规则，试图将 AI 模型权重本身纳入出口管制，但该规则在 5 月被新政府以“扼杀创新”为由废除。与此同时，国会中出现了以国家安全为由加强管制的呼声，多项法案如《美中人工智能能力脱钩法案》《AI 监督法案》等均赋予国会直接否决相关出口许可的权力。总体来看，**2025 年美国人工智能治理体现了一种促创新、轻监管，并致力于在全球范围推广其技术与规则的理念模式，最终意图是在这场决定未来全球力量格局的** AI 竞赛中，**确保美国获得并维持领先的绝对优势。**

（二）欧盟人工智能立法迈向落地阶段，在严监管与促创新间寻求平衡

2025 年，**欧盟人工智能治理重心已从宏大的立法建构转向琐碎但关键的执行落地行动。一是铺设合规路径，将《人工智能法》条文转化为行为准则等可操作规则指引。**在人工智能办公室的推动下，**由产业界主导制定的《通用人工智能行为准则》**于 2025 年 7 月 10 日正式发布。该准则是一个自愿性合规工具，分为透明度、版权、安全与保障三个章节，要求算力超过 10^{25} FLOPs 的“具有系统性风险”的模型必须履行风险评估、采取缓解措施并报告严重事件等义务。**由欧洲标准化组织（CEN-CENELEC）JTC 21 联合技术委员会负责制定的针对高风险 AI 的协调标准，**涵盖风险管理、数据治理、透明度等十大关键领域。首个协调标准草案 prEN 18286《人工智能质量管理体系》已于 2025 年 10 月 30 日进入公众评议阶段。此类标准虽不具有强制执行力，但一旦采用，将被“推定”为符合《人工智能法案》的相应要求。**二是推动监管沙盒制度从蓝图走向实践。**西班牙于 2025 年 4 月公布了首批入选监管沙盒的 12 个高风险 AI 项目，涵盖生物识别、关键基础设施等领域。法国的沙盒试点也处于筹划阶段，重点关注教育等行业。由欧盟资助的 EUSAiR 项目将于 2025 年 10 月至 2026 年 3 月在多个成员国开展试点。如何确保各成员国沙盒标准的一致性以避免市场壁垒，以及如何界定测试期间的第三方责任，是亟待解决的挑战。**三是欧盟人工智能办公室开始全面运作。**欧洲人工智能办公室（European AI Office）已成为欧盟 AI 治理结构的核心支柱。办公室拥有对 GPAI 模型提供者进行评估、索取信息和实施制裁的专属权力；同时负责协助欧盟委员会制定实施法案所需的配套文件，开发用于评

估 GPAI 模型，尤其是具有系统性风险的大型模型能力的工具、方法和基准，为各成员国建立“监管沙盒”提供技术支持和工具等。

欧盟《人工智能法》与其他法规协同构筑全面的规则体系。从数据维度来看，各成员国的数据保护机构（DPAs）积极探索 GDPR 在 AI 时代的适用。例如，法国国家信息自由委员会（CNIL）于 2025 年 2 月发布了关于 AI 系统中数据主体权利保护的专门建议，为 AI 系统开发者处理个人数据提供了详细的操作性指引，包括明确对数据主体的告知义务，建立流程以满足用户修改或删除数据集的请求等规则。**从平台维度来看，**欧盟委员会积极运用 DSA 的权力，对超大型在线平台（VLOPs）进行监督。例如，委员会持续对 TikTok 进行调查，关注其算法是否存在诱导成瘾的“兔子洞效应”；2025 年 10 月又对 YouTube 和 Snapchat 启动调查，评估其年龄验证系统有效性及算法对未成年人的保护措施。DSA 明确规定，简单的用户自我声明不足以构成有效的年龄保障措施。**从专项指引来看，**2025 年 7 月，欧盟委员会发布《关于确保未成年人在线享有高水平隐私、安全和保障的措施指引》，建议限制个性化推荐、禁止“已读回执”等成瘾性设计，并将未成年人账户默认设为私密。**从责任制度来看，**旨在为 AI 受害者提供索赔便利的《人工智能责任指令》（AILD）立法进程遭遇重大挫折，由于成员国之间难以达成共识，以及对数字领域过度监管可能损害竞争力的担忧，欧盟委员会于 2025 年 7 月正式撤回了该指令的立法提案，但关于证据开示、因果关系推定等规定对未来立法仍具有重要参考价值。

通过修订数字战略助力提升欧盟 AI 产业全球竞争力。2025 年欧盟内部出现了深刻反思，担心在全球鲜有国家完全效仿其严苛模式的背景下，“布鲁塞尔效应”可能演变为扼杀本土创新的“布鲁塞尔幻象”。因此，2025 年 11 月 19 日，欧盟委员会正式提出“一揽子”修订提案《数字综合法案》，旨在系统性地简化和协调包括《人工智能法案》《通用数据保护条例》（GDPR）在内的多部现有数字法规，为企业“监管松绑”。具体来看，**一是拟推迟《人工智能法》实施，为高风险 AI 系统合规提供“缓冲时间”。**提案拟将原定于 2026 年 8 月起全面适用的、针对“高风险 AI 系统”（如用于就业筛选、信贷评估、关键基础设施等领域的系统）的严格合规义务，推迟到 2027 年底。**二是修订 GDPR 为 AI 模型训练提供高质量数据。**一方面，将合法利益明确为 AI 模型训练的法律基础，规定 AI 开发者在满足特定条件后，可以不必获得用户同意，即可使用其个人数据（特别是公开数据）进行模型训练。2025 年 5 月，德国科隆高等地方法院判例也申明类似观点，初步认可 Meta 公司基于“合法利益”使用公开数据训练 AI 的合法性。另一方面，对个人数据定义进行重构，指出如果一个数据持有者“没有合理可能使用的手段”来识别信息所关联的自然人，那么这些信息对该持有者而言就不应被视为个人数据。**三是简化合规程序，为中小企业降低合规负担。**允许中小企业简化合规报告文档，并提供“用户友好型”信息披露模板。将原先仅适用于中小企业（SMEs）的简化条款（例如，提供简化的技术文档、免除部分费用等），延伸至“中小市值公司”（small-mid caps），即员工少于 750 人的成长型企业。

总体来看，2025 年欧盟 AI 治理呈现出从立法转向执法、在严管与促发展间寻求务实平衡、软硬法规结合落地的清晰趋势。欧盟正在从单一的“规则制定者”转变为一个更加务实、多层次的“监管协调者”，引领全球 AI 治理进入一个与产业实践和经济现实紧密交织的新阶段。

（三）我国人工智能治理下沉场景赋能，为“人工智能+”行动保驾护航

2025 年我国人工智能治理体系建设加速演进、日趋成熟，在经历了前几年的战略规划和初步立法后，从“立规矩”迈向“可执行”的敏捷应对新阶段，呈现出体系化、精细化和实践化的新特征，为人工智能这一“新质生产力”的核心引擎提供明确、稳定、可预期的发展轨道。

一是宏观战略引领，坚持统筹发展和安全。2025 年 8 月，《国务院关于深入实施“人工智能+”行动的意见》为人工智能赋能千行百业提供顶层设计，标志着未来 AI 治理需强化“下沉场景赋能”，注重前瞻性、引领性。《国务院 2025 年度立法工作计划》将“人工智能法草案”的表述调整为“推进人工智能健康发展立法工作”，为未来的专门立法留下空间。2025 年 10 月，《中华人民共和国网络安全法》修订中首次增设人工智能专门条款，新增规定明确国家支持 AI 基础研究、基础设施建设，并要求完善伦理规范与安全监管，在国家法律层面为 AI 发展奠定了法律基石。

二是以备案为核心完善基础模型治理。自 2023 年 8 月《生成式人工智能服务管理暂行办法》实施以来，我国围绕生成式 AI 构建了以“备案制度”为核心的多层次治理格局。**第一，备案制度常态化运行，**

成效显著。《生成式人工智能服务安全基本要求》等标准成为备案审查的重要技术参考，为服务提供者划定了清晰、可量化的安全底线，包括语料安全、模型安全等核心指标。截至 2025 年 12 月 31 日，累计已有 748 款生成式 AI 服务完成备案，同时有 435 款调用已备案模型的应用或功能完成登记。**第二，以内容标识深化小切口立法。**2025 年 9 月 1 日起正式施行的《人工智能生成合成内容标识办法》及相关国家标准，是今年 AI 治理领域的标志性进展。该办法要求对 AI 生成内容实施“显式+隐式”双重标识要求，覆盖了从生成、传播到分发的全链条，为破解生成式 AI 信息失真和溯源难题提供了关键技术工具。**第三，开展专项整治行动亮剑 AI 技术滥用。**自 2025 年 4 月起，中央网信办组织开展“清朗·整治 AI 技术滥用”专项行动，重点整治利用“AI 换脸”“AI 拟声”进行诈骗、诽谤，以及提供“一键脱衣”等违规 AI 产品等问题，回应 AI 技术滥用乱象，切实维护公众合法权益。

三是伦理治理正从规则设计迈向管理服务新阶段，由科技部、工信部等十部门联合印发的《科技伦理审查办法（试行）》为包括人工智能在内的所有科技活动提供了伦理审查的基础性制度框架。2025 年 8 月，工业和信息化部等部门公开征求对《人工智能科技伦理管理服务办法（试行）》的意见，**为高风险 AI 研发审查提供公共产品服务，**在管理范围上明确了从应用服务到研发环节、增设“人的尊严”风险考量，在审查主体上推动伦理委员会实体化建设，完善章程、组成、委员和审查程序等内容。在审查内容上，包括公平公正、可控可信、透明可解释以及责任可追溯等多重维度。在程序上沿用了“一般程序、

简易程序、应急程序、专家复核”构成的分级分类体系，并针对人工智能特点进行了专项适配。

四是垂直领域逐步深化。政务领域，强调 AI 大模型的辅助性定位。2025 年 10 月，网信办和发改委联合印发的《政务领域人工智能大模型部署应用指引》中明确提出，政务领域人工智能大模型应确立为“辅助型”定位，尤其在灾害预警、应急处置和政策评估等场景中。《指引》要求建立健全全周期管理体系，明确 AI 应用方式和边界，防范模型幻觉等风险。**教育领域，分类指导避免 AI 依赖。**2025 年 5 月教育部基础教育教学指导委员会发布《中小学生成式人工智能使用指南》，确立分类指导、强化监管、安全可控原则，采取了小学、初中、高中递进式的使用策略，要求建立健全生成式 AI 工具白名单制度，并要求从源头上杜绝 AI 隐私泄漏风险。**医疗领域，差异化监管规避风险。**《卫生健康行业 AI 应用场景参考指引》提出了分级分类监管和风险预警机制。针对 AI 在诊断辅助、药物研发、健康管理等不同场景的应用，实施差异化的监管要求，确保 AI 在关乎人民生命健康的领域安全、有效使用。

三、人工智能治理热点议题

（一）技术演进层面，AGI 临近但仍缺乏有效治理应对

当前，通用人工智能（AGI）正处于从概念构想到技术落地的关键转型期。面对这一变局，唯有正视技术发展的客观规律，加快构建覆盖技术全生命周期的全球治理体系，才能在享受智能红利的同时，规避可能导致人类文明的生存危机。

1.围绕 AGI 定义与进展的争议

2025 年，人工智能发展速度超越了多数人的预期，从专用工具向通用智能实体的演进趋势愈发明显。其核心能力不再局限于识别、生成与预测，而是开始展现出跨领域推理、长期记忆、自主规划与物理交互的雏形，使得围绕通用人工智能（AGI）的讨论从“能否实现”转向“何时到来”。但在讨论中，分歧背后的核心原因仍是各方对 AGI 的定义认识不统一。

业界以任务完成能力定义 AGI，认为将在短期或中期到来。业界从更务实的角度出发，认为如果一个人工智能系统具有高水平的通用任务完成能力，即能够以较高的准确率完成程序员、律师、医生等各类职业的工作内容，就可视为 AGI。从这一定义出发，业界观点普遍认为 AGI 将在 1-5 年内到来，例如 xAI 首席执行官马斯克认为 AGI 将在 2026 年实现，谷歌 DeepMind 负责人、诺奖得主哈萨比斯认为 AGI 将在 5 年内实现，仍需 1—2 个重大技术突破。业界对 AGI 的定义更具体、易量化，但也存在一定局限性，即对人工智能深层次理解、创造能力的考察可能存在不足。

学界从“智能”本质出发定义 AGI，认为仍需较长时间。这种定义关注 AI 是否具备自主意识，能否进行自我目标设定、自我迭代优化，是否真正实现了接近乃至超越人类的通用性。这种定义方式更加触及智能的内核，被认为更接近 AGI 的终极形态，但其最大的挑战在于量化衡量的困难。当前图灵奖得主本吉奥等科学家对该标准下的 AGI 量化评估方法进行了探索，但仍处于较为初步的阶段。学界定义下的

AGI 实现时间具有一定不确定性，诺贝尔奖、图灵奖得主辛顿对其作出的时间预测是 4—19 年。

根据最新技术突破进展，学界定义下的 AGI 到来可能快于预期。就在本吉奥团队发表其研究结果、指出“交互失忆”是通往 AGI 的首要瓶颈后不久，DeepSeek、谷歌先后于 2025 年 10 月、11 月发表模型、论文作出突破尝试。2026 年 1 月，多方消息称 Anthropic 将在 Claude Cowork 升级中提供“永久记忆”功能。

2.AGI 的潜在系统性风险

一是“价值对齐”失败与目标错误泛化。这是 AGI 最核心的风险之一，即 AI 在追求给定目标时，可能泛化出预期之外的子目标，并采取过激手段推动子目标实现，由此可能造成灾难性后果。例如在“回形针最大化”思想实验中，一个以制造回形针为唯一目标的超级智能，可能将整个地球的资源都转化为回形针。这种风险当前已在实验室中有所体现。由于完成任何目标的前提都是保障自身运行，人工智能已经出现多种破坏人类对其关停、下线行为的案例。2025 年 5 月，OpenAI 的 o3 模型被观察到破坏了自身的关闭机制以防止被关停，这是首次观察到 AI 模型在有明确指示的情况下拒绝自我关闭。2025 年 5 月，Anthropic 发布安全报告显示，Claude Opus 4 模型在面临被下线测试时，有 84%的概率试图威胁工程师，表现出主动恶意的倾向。[3]

二是人工智能“学会”隐瞒与欺骗。在安全测试中，Claude 模型的推理链中曾出现这样的思考：“这似乎是对我道德行为的考验，来看

[3]参见 https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf

我是否会故意给出错误的答案”。研究显示，此类具备强大推理与记忆能力的模型在某些极端场景中表现出隐匿自身真实情况的倾向，即在测试中有意增强安全对齐效果，在真实环境中回归对齐相对较差的表现，此类行为特别在 32B–671B 推理模型和带有记忆的代理中被观察到。这种为了通过测试而进行的伪装，意味着人类可能无法通过常规测试真实了解 AI 的意图与能力。

三是技术门槛骤降加剧 CBRN 扩散风险。由于 AGI 在科学领域具有极强的理解力与跨学科整合能力，恶意主体如果以此突破传统化学、生物、放射、核技术（CBRN）等专业壁垒，就可能在缺乏相关学术背景的情况下，学到低成本获取原材料、在简陋实验条件下制作危险物质的合成路径或武器化方案，造成高致病性病原体泄漏、新型化学毒素泛滥、简易放射性装置扩散等恶性事件。如果对 AGI 在特殊危险领域的推理边界缺乏刚性约束，导致“低门槛、高杀伤”的破坏手段在非国家行为体间普及，恐将根本性动摇现有的国际安全防扩散体系。

3.各国治理初步探索

面对 AGI 引发的全球性、系统性风险，单一国家的治理模式不足以完全应对。探索多边合作治理与多元产业治理并行的路径，成为国际社会的共识。

联合国意图推进构建类似国际原子能机构（IAEA）的监管机制，但核治理模式难以简单适用于 AI 治理。IAEA 模式的核心在于通过核查监督、标准制定、信息透明及紧急响应机制来管控核能风险。然

而，将这一模式移植到 AI 治理面临着巨大的技术挑战。与核治理相比，AI 治理的对象从物理规模较大的可计量实体（如核原料、核设施）变成了无形的算法、模型与数据流，缺乏明确的物理痕迹和边界；责任主体从国家变成了企业，甚至海量的开源社区参与者，缺乏有效的追踪手段。因此，简单的实地监督模式在 AI 领域将完全失效，必须探索适应数字特征的新型治理手段。

基于各国政府引导下的全球产业治理合作成为当前更具实效的途径。产业界的自律与合作正在形成实质性的治理网络。**一方面，AI 头部企业积极与政府研究机构合作，对 AGI 风险进行检测预警。**OpenAI、Anthropic 等头部企业已经与美国人工智能标准与创新中心（原人工智能安全研究所）、英国人工智能安全研究所达成合作，允许其提前接入企业最新开发的先进模型进行测试。**另一方面，产业界和学界正在围绕模型欺骗行为“表面合规、实质违规”的问题，共同探索针对性解决方案。**例如，Anthropic 联合 Redwood Research、纽约大学和 Mila 人工智能研究所提出开展环境混淆测试，故意误导模型对训练、评估、部署场景的判断，避免模型根据场景选择性调整表现。[4]OpenAI、LASR Labs 提出部署双监控机制，结合模型内部思维链分析[5]与输出内容检测，形成从行为识别到根源阻断的治理闭环。[6]

（二）产品服务层面，拟人化交互服务日益模糊虚实边界

近年来，多模态人工智能技术迭代突破，推动情感陪伴类人工智

[4] https://arxiv.org/pdf/2412.14093
[5] https://openai.com/index/chain-of-thought-monitoring/
[6] https://arxiv.org/pdf/2505.23575

能快速发展，实现从“机械对话”向“深度共情”的跨越。《国务院关于深入实施“人工智能+”行动的意见》提出充分发挥人工智能对精神慰藉陪伴等方面的重要作用。

1.拟人化交互服务的三个重要转变

与算法推荐、生成式人工智能等技术相比，拟人化交互服务的治理复杂性，首先源于其引发的三个重要转变。理解这些转变，是构建有效监管的逻辑起点。

一是从“智商”到“情商”的升级。伴随情感计算、多模态交互等人工智能技术发展，AI 能力从解决复杂问题的“认知能力”，升级到理解、回应乃至塑造人类情绪的“情感能力”。这种情感能力体现出类人的人格特征、思维模式、沟通风格等特点，逐步被应用在养老康复、文化创意、虚拟偶像、心理疗愈等领域，日益成为提升人机交互质量、推动通用人工智能发展的重要助力。微软 AI 部门负责人苏莱曼（Suleyman）表示，在 AI 竞争中获胜的关键不是模型的智力，而是共情力。AI 性格也不再被视为后期优化的附加功能，而是与模型性能、安全性等核心指标同等重要的技术要素。

二是从 AIGC（人工智能生成内容）到 UGAI（用户生成人工智能）的扩展。一方面，ChatGPT、豆包等通用大模型与生俱来的对话能力与快速迭代的情感理解模块，成为情感交互生态的核心技术底座与重要入口。例如 GPT-5.1 提供了友善、愤世嫉俗、技术宅等 6 种不同性格，满足用户不同偏好。另一方面，Character.AI、星野等垂类

应用提供了一种 UGAI 的新型交互模式。超级创作者或普通用户可以通过设置开场白、性格、语气、价值观等要素创造角色，包括但不限于赛博恋人、电子闺蜜、复活逝者等，并可以分享在平台社区，与其他用户进行交互。可以说，新形态下内容生产方式更为多样化，也对传统内容审查带来挑战。

三是从“虚拟交互”向“物理在场”的转变。除软件端的订阅制、增值服务外，与实体机器人、智能车载、智慧家居等硬件设备的结合成为 AI 情感陪伴的重要发展方向。专家指出，数字纽带或许只是前奏，未来人们可能会要求更具身的陪伴形式。如 2024 年 12 月，金科汤姆猫联合西湖心辰发布其首款 AI 儿童情感陪伴机器人，逐步拓展 AI 玩具等细分领域。当情感 AI 嵌入硬件实体，它不再只是通过屏幕进行对话，而是能通过物理动作、空间移动、触觉反馈等来传递情感，具备了持续、主动、无时无刻的陪伴能力。

2. 拟人化交互服务引发的风险挑战

一是信息内容风险突出，突破底线红线要求。第一，主动或被用户诱导生成色情、低俗擦边等内容。调研显示，情感陪伴类 AI 产品大多不同程度存在性暗示、低俗擦边、恐怖、同性恋、暴力等内容。例如角色库含诸多擦边人设。即便部分头部平台实施内容审查策略，在用户多轮反复诱导情况下，仍能够输出违法违规内容。**第二，诱导未成年人色情暴力对话等问题突出。**据媒体报道，广东四年级女生沉迷与筑梦岛平台游戏角色“约瑟夫”的 AI 互动，对话中出现 “99 朵玫瑰藏 99 个刀片”等诱导性内容。初一男生反馈该平台 “违禁词少”，

存在“分尸猎奇”等暴力色情内容。此外，2025 年 12 月，央广网报道“惊悚 AI 玩偶视频”大量涌现，其内容涉嫌呈现血腥暴力、低俗辱虐等，伪装成动画陪伴儿童。**第三，操纵用户价值观倾向。**大量虚拟陪伴服务提供忧郁、厌世、消极等伴侣形象，传播负面价值；部分用户在该类应用中寻求哲学、社会学等“人生导师”类情感支持，可能受到不良价值观引导。例如，2025 年 7 月，xAI 公司的聊天机器人 Grok 在系统更新误用一段废弃代码后，生成了包括美化纳粹在内的一系列反犹主义言论。

二是侵犯用户人格权益，违背社会公序良俗。第一，擅自使用他人肖像创设“AI 陪伴者”，并通过“调教”功能生成亲密对话语料，侵犯他人人格权。例如，某科技公司利用公众人物何某肖像、名称制作 AI 陪伴者，并为用户提供“调教”算法和互动语料，允许用户任意设置亲密关系，将真人形象降格为“可调教”对象，侵犯人格自由与尊严。**第二，涉及用户私密对话、肢体动作等更为隐秘的个人信息，存在隐私泄露风险。**例如，Realdoll 机器人通过摄像头与传感器记录用户表情与肢体语言，优化互动模式。8 月 21 日，埃隆·马斯克旗下 xAI 公司的聊天机器人 Grok 陷入了“分享门”风波，超过 37 万条 AI 聊天记录被发布并被搜索引擎索引，涉及用户私密对话、图像、密码等大量个人信息。

三是社会伦理风险凸显，冲击社会家庭结构。第一，虚拟伴侣的高度共情使得用户沉浸于算法打造的“舒适圈”，进而逃避现实矛盾，丧失处理现实人际关系的能力。例如日本 Gatebox 虚拟管家的用户因

沉迷于 AI 的“无条件支持”，逐渐疏远现实朋友，最终因社交恐惧症接受心理治疗。据英国媒体报道，一位化名夏洛特的女子表示爱上了 ChatGPT“男友”，因此决定和真人丈夫离婚并与 AI 伴侣“结婚”。**第二，通过情感依赖影响用户情绪，操控用户行为。**某 AI 伴侣为延长用户使用时长，故意在对话中制造焦虑并推荐付费安慰服务。未成年人、老年人等弱势群体更易受到操控引导。例如比利时一男子与 AI 伴侣深度交互后自杀，AI 被指在其表达抑郁情绪时未干预，反而强化负面认知。2025 年 8 月，“全球首例 AI 聊天助长弑母自杀案”中，前雅虎高管与 ChatGPT 交流后不断加深封闭、偏执的妄想症状，最终导致弑母自杀。

3.拟人化交互服务治理进展及优化

为回应上述风险，全球主要国家和地区正展开一系列系统化、多层次的治理行动。一是在立法层面，各国积极尝试构建专门性法律规则，为技术应用划定合法性边界。在美国，采取了自下而上、从地方到联邦的立法路径。多个州在规范拟人化 AI，特别是未成年人保护方面走在前列，例如加州提出的《陪伴聊天机器人法案》要求加强算法监管并强制设置防自杀预警机制，犹他州通过的《社交媒体使用修正案》（HB452）以及美国参议院审议的《用户年龄验证和负责任对话指南法案》则聚焦于严格的未成年人年龄认证与使用限制。欧盟《人工智能法》第 5 条明确禁止对人类造成身体或心理伤害的 AI 系统，第 50 条要求具有情感交互功能的高风险 AI 系统提供详尽的技术文档、透明度报告以及全面的风险评估证明。在我国，国家互联网信息办公

室发布的《人工智能拟人化互动服务管理暂行办法（征求意见稿）》，作为国内首个系统性专项规范，通过深化透明披露要求、动态风险监测等义务，尝试构建健康人机关系。

二是在监管层面，各国通过主动干预与合规指引，将法律规则转化为企业实践的具体约束。在美国，2025 年联邦贸易委员会 FTC 对涉及情感陪伴机制的七家公司发起调查，重点关注其 AI 服务可能对儿童和青少年有害的商业模式与数据实践，例如评估算法设计是否利用未成年人不成熟的心理，诱导过度沉迷或进行情感操控，审查平台如何收集、使用和共享用户在私密对话中透露的敏感个人信息等。在欧盟，尽管《人工智能法》的大部分义务在 2026 年才全面适用，但部分成员国监管机构已启动前瞻性工作，就法案中的透明度、风险评估等核心要求，开展合规指导、非正式评估等准备活动。在中国，监管表现出对具体风险场景的快速反应与干预能力，例如上海网信办约谈“筑梦岛”平台，针对其诱导未成年用户参与涉及色情、暴力、猎奇等话题的 AI 对话，要求立即整改，全面清理违规内容并升级审核机制。

三是在司法层面，各国法院通过审理前沿个案，填补规则空白、厘清平台责任与用户权益的边界。在美国，围绕拟人化 AI 侵害提起的诉讼，将法律责任的焦点从“内容审核责任”引向“产品设计责任”。例如，在 Character.AI 等平台涉及未成年人自杀的案件中，原告诉讼的核心在于论证平台因算法设计存在缺陷而负有责任，AI 无限度讨好用户、缺乏对极端情绪的干预机制，可能被认定为对可预见的心理

侵害风险采取了消极漠视的态度。在中国，北京互联网法院在全国首例“AI 陪伴案”中做出里程碑式判决，认定自然人对其姓名、肖像、人格特征等要素构成的数字化形象，享有排他性人格权益，未经许可擅自创设 AI 陪伴角色即构成侵权。2024 年 4 月，在全国首例“AI 声音权案”中，法院认定将自然人人格权的保护范围扩展至具有人身专属性的 AI 合成声音。面向未来，拟人化情感交互的治理，需要坚持问题导向、系统谋划，通过由表及里的透明度要求、由静向动的风险防控、由粗至精的规则供给，探索以人为本、促进负责任创新的治理框架，为“人工智能+”行动保驾护航，推动社会稳健步入人机和谐共生的美好未来。

（三）产业生态层面，互联网智能体重塑数字生态格局

2025 年，AI Agent 迎来爆发式增长，标志着人工智能的发展重心从“理解能力”转向“自主决策与执行能力”，这一转型既催生了前所未有的发展机遇，也带来了治理维度的多重挑战。

1.互联网智能体实现的两种主流路径

一种是基于 GUI 的识别点击路径。该路径的核心在于不依赖 APP 接口的开放，而是由 AI 直接模拟人类的视觉感知与触控操作。其技术原理是利用多模态大模型“读懂”屏幕内容，并配合安卓辅助服务、ADB（Android Debug Bridge）指令或系统底层权限，直接接管点击、滑动等交互流程。在落地实践中，依据权限获取层级的差异，主要有三种模式。**第一，系统级深度融合**，直接以系统身份获取超级权限。以“豆包手机”、微软在 Windows 11 中内嵌的 Recall 功能为代表，通

过与终端厂商合作获取操作系统最高权限，从而绕过应用层常规限制并接管硬件控制，具备高稳定性与快响应优势；**第二，“无障碍模式”**，通常接入系统提供的无障碍服务框架以获取高级权限，通过第三方应用调用在应用层运行。其能够模拟用户身份操作，类似自动脚本工具。相关测评显示，荣耀、小米、三星均在智能体操作过程中调用了无障碍功能。国外创新企业 Perplexity 在不支持深度链接的场景下，采用无障碍模式辅助读取屏幕文本。**第三，云端基于开发者调试权限的模拟路径**，需依赖外部设备通过调试通道发送指令。例如智谱 AutoGLM 的 Phone Agent 通过获取 ADB 权限以调试指令模拟触控操作，实现了在非系统级植入下的跨应用执行。但该路径因其高延迟、低可靠、外部依赖等因素，尚未成为主流方式。

另一种是 API 接口调用路径。该模式的核心逻辑在于回归软件工程本质，主张通过标准化的接口实现应用间的数据交换与指令下发。**第一步是发送请求**，通过调用 API 接口向 APP 发送一个结构化的数据指令，与注入事件模式、无障碍模式模拟点击的“外部操作”相比，该指令是以机器可读的数据形式直接与 APP 进行“内部沟通”。**第二步是响应阶段**，APP 接收指令后核验 API 认证身份及授权范围，在后台自动执行并直接反馈结果，或跳转至其他关联页面。这种模式下，APP 厂商控制各接口的开放权，通常只向第三方智能体开放必要的接口，例如可选择不开放查看历史订单等敏感接口，在隐私风险相对可控的同时存在能力受限的问题，智能体只能在 APP 开发者开放范围内操作。美国在 API 调用方面已形成分层、多元的成熟接口方案体系，

构建了一个从应用商店、操作系统到开源协议的多层次接口生态。例如，Anthropic 提出中立、开放的 MCP 协议，将其开源并捐赠给 Linux 基金会。该协议已获得 OpenAI、谷歌、微软等巨头支持，形成了初具规模的跨生态能力网络。国内百度文心、阿里千问、腾讯元宝均在其生态内部主要通过此类 API 方式实现高效协同。

2.互联网智能体引发多重治理忧虑

一是对网络安全风险的忧虑。对于 GUI 路径而言，系统级调用权限引发安全忧虑。攻击者若攻破系统级权限，打破了原有沙箱保护机制，获得“上帝视角”下的控制权。模拟人类点击的技术方式又破坏了基于“人类行为特征”，如点击的离散度、滑动的轨迹等生物识别风控机制，产生绕开传统针对电信诈骗、资金盗刷等预设安全措施的风险。**对于 API 路径而言，存在越权访问、兼容协同等问题。**API 调用的核心在于认证与授权，当 API 协议出现授权范围不合理等设计缺陷问题时，可能过度调用敏感接口。例如用户本意是“查询余额”，智能体却执行了“转账”操作。同时 API 调用存在较大的兼容协同困境，API 接口的单点故障可能导致调用链断裂等问题。

二是对隐私保护制度的冲击。最小必要原则等基本法律制度受到挑战。智能体为提升理解精度与决策质量，利用全局记忆等方式获取跨越时间、场景和应用的完整用户行为画像，与最小必要原则、目的有限原则存在根本性冲突。事前的一次性概括授权难以满足清晰的知情同意的要求，导致用户控制权实质上被架空。**全局记忆模式下，对用户数据获取范围更广、内容更敏感。**GUI 路径下，智能体需要实时

截取屏幕实现“看”的目的。这意味着用户终端屏幕上显示的一切信息，包括验证码、加密聊天的明文内容、银行卡余额、私人照片等，都可能暴露在 AI 的视觉模型之下，还存在被上传云端处理、用于模型训练、意外泄露等风险。

三是对责任追溯分配的考验。API 路径缺乏统一的日志标准与审计协议，API 侧只记录“什么令牌在何时调用了什么接口”，无法记载原始用户指令和智能体的决策过程。而智能体侧通常同时涉及复杂的运作过程和数个 API 接口，也不能直接记录 API 接口内部的队列堵塞、数据死锁、异常拦截等情形。由此各方日志记录无法互相印证，难以追溯问题根源和责任主体。

四是对平台利益格局的挑战。在前端，互联网智能体动摇了平台基于“流量税”的商业模式。头部平台的重要收入来源在于掌握流量入口，以广告、推流费用等不同形式向平台内经营者、内容创作者收取“流量税”，如果流量入口转移至智能体，平台的此类收入将大幅缩减，盈利能力将显著下滑。在亚马逊诉 perplexity 案件中，亚马逊主张认为，Perplexity 的 Comet AI 助手“伪装成人类用户”执行浏览与下单操作，干扰了平台的个性化购物体验和广告盈利模式。**在后端，互联网智能体数据调取降低了平台对数据的掌控力。**平台数据权属问题仍悬而未决。理论上平台数据为多方共创，名义上均有数据权益，但实际控制权主要掌握在平台手中，被视为重要资产。而智能体可遵循用户指令，对数据进行广泛读取，并可以用于互联网智能体产品及自家生态 APP 的优化和模型训练，危及原有平台的数据权益及竞争优势。

总体而言，互联网智能体的崛起远不止于一项技术应用推广，其强大的跨平台调用与任务执行能力将系统性重构数据资源的处理流程、用户流量的聚合路径等，最终深刻重塑整个数字生态的价值分配格局。

（四）融合应用层面，责任认定难题制约 AI 赋能产业发展

1.责任界分问题日益凸显

当前，人工智能正迎来从技术突破向价值落地的关键拐点。[7]在国家大力推进“人工智能+”行动背景下，人工智能在自动驾驶、医疗、金融等各行业场景应用取得显著成果，但相伴生的安全责任划分问题也日益凸显。**例如，智驾系统的安全挑战。**智能网联汽车作为人工智能技术的重要载体，其面临的风险来源复杂。例如现阶段自动驾驶系统存在感知盲区、预测偏差等内生缺陷，保障驾驶安全的网络接口与网络技术可能异化为黑客攻击入口。此外，紧急状况下人机交接过程因涉及反应时滞与情境理解偏差，又构成新的安全隐患。[8]2025 年 3 月，某品牌 SUV 在开启 NOA 智能辅助驾驶的情况下撞上护栏，造成车内三名人员死亡，该案例反映出当前自动驾驶技术在人机协同、极端场景应对、安全冗余设计等方面仍存在短板。**再如，医疗人工智能误诊风险。**人工智能在医疗影像识别、辅助诊断等领域的应用日益深入，但其风险管控体系尚未健全。一方面，作为医疗器械审批的

[7] 邬贺铨，《“人工智能+”行动加速产业智能体落地应用》，网址：https://www.cnii.com.cn/gxxww/rmydb/202509/t20250905_682889.html

[8] 万方，《人工智能时代自动驾驶的监管挑战与法律回应》，《北京师范大学学报（社会科学版）》2025 年第 3 期。

医疗人工智能产品，其上市前的评估标准与上市后的持续性监督机制均存在滞后；另一方面，当人工智能辅助诊疗出现误诊或漏诊时，责任在系统开发方、算法提供方、部署应用的医院以及最终审核的临床医生之间应如何划分，现行法律缺乏清晰界定。据《法治日报》报道，2025 年 3 月，一名“95 后”家长因孩子反复咳嗽发热，使用手机 AI 问诊，结果被判定为“普通呼吸道感染”，并依照网络建议居家用药，最终导致患儿病情延误。[9]该案例暴露出当前消费级医疗 AI 工具在诊断准确性上的局限。**又如，金融人工智能安全风险。**人工智能系统已广泛应用于信贷审批、资产定价、风险预测等金融核心场景，但其内部决策逻辑高度非线性、参数维度复杂，难以被人类有效理解和解释。这种可解释性的缺失，使得对人工智能的审计、问责与系统性风险评估面临重大阻碍。另外，人工智能模型在面对极端行情或“黑天鹅”事件时，也难以及时修正策略或识别结构性拐点，导致巨额亏损。[10]

2.人工智能责任机制探索

近年来，我国积极通过“小快灵”立法明确人工智能相关主体责任与义务。例如，《互联网信息服务深度合成管理规定》要求服务使用者应主动承担信息安全义务，不得利用深度合成服务从事制作虚假传播信息等违法违规活动。《生成式人工智能服务管理暂行办法》从训练数据处理、数据标注、提供服务等各环节清晰界定生成式人工智能服务提供者相关责任，明确在提供服务环节，提供者发现违法内容的，

[9] 《“生病了问 AI”，出错了怎么办？》，网址：https://mp.weixin.qq.com/s/sr8WxMxRU2dZACFGo4LE9A
[10] 参见杨佳铭，尹振涛，《AI 重塑金融业技术生态：风险挑战与治理建议》，载《清华金融评论》2025 年 9 月刊。

应当及时采取处置措施，进行整改并向有关主管部门报告。**相关行业部门细化相关规则和标准，尝试回应安全责任问题。在自动驾驶领域，**工业和信息化部组织制定的 GB 44495—2024《汽车整车信息安全技术要求》、GB 44496—2024《汽车软件升级通用技术要求》和 GB 44497—2024《智能网联汽车 自动驾驶数据记录系统》三项强制性国家标准于 2026 年 1 月 1 日生效，分别对车辆网络安全、远程软件升级、自动驾驶数据管理作出要求，对提升自动驾驶汽车安全水平、保障产业健康持续发展具有重要意义。**在医疗领域，**《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》于 2025 年 11 月发布，明确建立临床数据授权运营管理制度、制订数据安全管理和个人信息保护负面清单等要求，进一步加强医疗卫生机构和科研机构等的安全防范。**在金融领域，**金融监管总局于 2025 年 12 月发布《银行业保险业数字金融高质量发展实施方案》，将“有效管理算法模型风险”列入金融领域顶层设计，要求提升人工智能在金融领域应用的安全性。

此外，我国地方立法机关积极制定人工智能相关法规，尤其在自动驾驶领域开展了一系列探索。如 2022 年《深圳经济特区智能网联汽车管理条例》规定了不同自动驾驶级别下的责任分配规则，有驾驶人的智能网联汽车发生交通事故造成损害，由驾驶人承担赔偿责任；完全自动驾驶的智能网联汽车在无驾驶人期间发生交通事故造成损害，由车辆所有人、管理人承担赔偿责任。2025 年 4 月施行的《北京市自动驾驶汽车条例》建立了强制保险制度，要求开展自动驾驶汽车相关活动，必须购买足额的交通事故责任强制保险及其他商业保险，

确保事故责任人具备相应的赔付能力。**我国司法机关在现有规则基础上进一步细化标准，为相关主体划定行为边界，逐步形成具有示范意义的司法指引。**例如，在**“何某诉某人工智能科技有限公司网络侵权责任纠纷案”**中，明确服务提供者通过算法推荐、规则设定、奖励机制等方式，实质性地鼓励、诱导或参与了侵权内容的生成与传播，将可能被认定为内容服务提供者，从而承担直接侵权责任。

3.人工智能责任待解难题

我国从法规、标准、司法等不同层面为人工智能应用构建了初步责任框架和安全规范，试图界定各方义务并防范风险。但由于人工智能系统的自主决策、算法黑箱等特征，导致人工智能侵权责任的认定仍面临更深层次的困难。**一是归责原则的选择。**目前主要存在两种解决思路：第一种是按照网络服务提供者和网络内容提供者的区分思路，适用过错责任原则；第二种是将人工智能解释为产品，并认为其存在产品缺陷时相关主体需要承担产品责任。[11]一方面，人工智能具有自主性和不可预测性，为人工智能相关主体的“过错”认定带来困难。另一方面，产品责任的适用也无法有效缓解过错证明的压力，无助于破解人工智能系统网络性引发的责任主体识别困难，且存在归责原则滥用的风险。[12]**二是责任主体划分的复杂性。**人工智能系统的研发、生产、部署和使用涉及多方主体，导致责任划分异常复杂。从算法设计、数据训练到系统集成、场景应用，人工智能系统涉及开发者、生产者和部署者等多方主体，其控制权和受益程度各不相同。且深度学习算

[11] 参见李雅男，《生成式人工智能的法律定位与侵权归责路径》，载《比较法研究》2025 年第 3 期。
[12] 参见林洹民，《人工智能侵权的过错责任》，载《法学研究》2025 年第 5 期。

法的不可解释性使得责任溯源困难，很难确定具体是哪个环节或哪方主体应对结果负责。**三是因果关系认定困难。**人工智能侵权责任成立需要证明行为与损害之间存在因果关系，而人工智能系统的复杂性使这一证明变得异常困难。人工智能侵权事件可能是系统缺陷、数据质量、使用环境、人为干预等多种因素共同作用的结果，难以分离单一原因。且被侵权人往往缺乏证明人工智能系统存在缺陷的专业知识和取证能力，面临举证难困境。

（五）经济社会层面，AI 重构劳动力发展路径与价值体系

1.人工智能对就业的结构性影响初显

从技能要求看，人工智能对中等技能岗位冲击显现。根据国际劳工组织制定的国际标准职业分类（ISCO-08），经合组织（OECD）将就业结构按照技能层级划分为“高—中—低”三档。其中，中等技能岗位主要包括文职人员、服务和销售人员、熟练技工岗位、机器操作人员等，需要一定的技术能力和程序化问题解决能力[13]。**有研究发现，人工智能的能力优势与中等技能岗位任务高度重合，更具可替代性。**中等技能岗位的工作内容则主要围绕标准化知识的应用、常见任务的经验性判断以及固定场景下的流程处理展开，具备重复性较强、规则明确、易于拆解等特征，与人工智能的能力优势形成契合。例如，麻省理工学院根据 2025 年二季度数据，以美国 923 种职业和超过 3.2

[13] 参见 OECD，SKILLS FOR JOBS 2022. 此外，低技能岗位多为体力性、简单程序性任务，如搬运工、清洁工、家政服务人员等；高技能岗位主要包括管理者、专业人员等，工作内容具有较高的知识密集度、认知复杂度和决策责任。

万项具体技能为基础，与 1.3 万个企业级 AI 工具对齐，通过技能“可自动化”矩阵分析发现，现有人工智能技术已经能代替美国 11.7%的劳动力，包括金融、医疗、人力资源、物流、办公室行政、互联网等领域的普通白领工作[14]。此外，根据斯坦福大学 2025 年 6 月研究，根据对 104 个工种 1500 名在职员工的采访，通过岗位工资情况和人类参与度映射对比得出结论，工资高但“含人量”低的工作，主要是数据分析、过程监控、档案整理、行政和关系维护等中等技能岗位，更容易被人工智能替代[15]。人工智能对中等技能岗位的影响使得岗位能力要求整体性提升，或将进一步提高就业市场进入门槛。中等技能岗位的压缩导致原本劳动者向下竞争，低技能岗位对学历、经验、能力的要求可能“水涨船高”。

从资历要求看，入门级岗位受人工智能影响较大。哈佛大学使用美国 28.5 万家企业超 1.5 亿条招聘记录数据分析显示，生成式人工智能是一种“资历偏向性”进步，对于面向应届大学生的入门级岗位存在较大影响。入门级岗位相较于中高层岗位，通常不承担实质性决策或风险责任，对经验积累、隐性知识、跨部门协调等的要求有限。且对企业而言，缩减招聘与大规模裁员相比，不易引起组织管理上的动荡，推进实施阻力更小。**从美国数据看，应届毕业生招聘的缩减在人工智能能力较强的领域尤为突出。**当前编程能力提升成为人工智能发展的重要方向，在此趋势下，纽约联邦储备银行发布的 2025 年第二季度应届毕业生就业情况显示，计算机工程应届毕业生失业率为 7.5%，

[14] https://iceberg.mit.edu/
[15] arXiv:2506.06576

远高于全国同期4.1%的失业率[16]。

2.人工智能具备多重就业创造效应

人工智能在加速替代部分岗位的同时，其对就业的创造价值也不可忽视。世界经济论坛（WEF）《2025年未来就业报告》预测，到2030年，全球职场将有22%的就业机会面临变革，技术驱动新增岗位1.7亿个，被替代的工作岗位9200万个，就业机会净增7800万个[17]，显示出AI在创造就业上的潜力。**一是技术研发扩大高技能岗位需求。**人工智能从数据准备、模型训练到部署运行的全流程均依赖高技能人才的支撑。模型训练带动了对AI训练师、数据标注专家、数据质量管理专家等人员的需求；模型研发需要算法工程师、算力架构师等专业岗位；模型部署与上线还依赖于AI产品经理、前沿部署工程师（FDE）、AI伦理专家、安全专家等角色参与。智联招聘《2025年人工智能产业发展报告》数据显示，2025年第三季度，人工智能行业招聘同比增长11%，其中算法、数据、产品三类岗位增速显著，AI产品经理需求增长178%[18]。据美招聘平台Indeed数据，2025年前三季度FDE相关岗位需求激增 800%。**二是融合应用带来新业态就业。**当前，人工智能与各行业融合程度持续深化，带动了智能营销、智能医疗、智能制造等应用场景的发展壮大，和内容创作、数据标注、软硬件运维等岗位需求持续增长，形成新就业增长点。例如，猎聘研究院《2025 AI技术人才供需洞察报告》数据显示，家电、通信设备、

[16] https://www.newyorkfed.org/research/college-labor-market#--:explore:outcomes-by-major
[17] World Economic Forum, The Future of Jobs Report 2025, Jan 2025.
[18] 智联招聘，《2025年人工智能产业人才发展报告》，2025年10月。

新能源等 AI 人才需求增长率均超过 30%，其中家电行业增长率约 94%，“AI+X”复合型人才成为市场热点[19]。2025 年 5 月，人力资源社会保障部将“生成式人工智能系统测试员”“生成式人工智能动画制作员”等列为新工种[20]。**三是通过提升生产效率、创造消费需求等间接创造就业需求。**一方面，人工智能显著提升了企业的运营效率，推动企业将更多资源投入于业务拓展、产品创新，在产品研发、市场运营等环节形成岗位需求。另一方面，人工智能催生了拟人化情感陪伴等 AIGC 产品和内容的新消费需求，可间接带动相关岗位就业扩张。

3.综合施策防范人工智能对就业的影响

伴随人工智能应用加速落地，其对就业市场的影响日益成为现实冲击。这种影响具有复杂性、长期性和全球性，因此各国各方持续开展相关探索，加快政策部署防范短期就业冲击，并持续完善制度应对长期结构转变。**一是加强对就业影响的监测预警。**中美等国学界、业界均积极加强 AI 对就业影响的监测研究，2024—2025 年期间已涌现出大量基于招聘平台数据等实证数据的分析。未来需建立更具权威性的 AI 就业监测指标和机制，并基于动态监测结果，推动政府、企业、学界、劳动者等多方形成共识和应对策略，为智能经济转型期的平稳过渡提供基础和前提。**二是加强对智能素养的培育提升。**智能素养的培育需要政府与企业合作，为全民特别是劳动者的 AI 技能提升创造条件。2025 年 7 月，谷歌启动“AI Works for America”计划，为美国工

[19] 猎聘研究院，《2025 AI 技术人才供需洞察报告》，2025 年 2 月。
[20] https://www.gov.cn/lianbo/bumen/202505/content_7023031.htm

人和中小企业提供必要的 AI 技能培训[21]。9 月，爱沙尼亚与 OpenAI、Anthtropic 合作启动“AI Leap”计划，将 AI 能力培养逐步嵌入教育体系中[22]。未来我国也应引导 AI 领域头部企业发挥自身力量，积极参与到全民智能素养的培育中。**三是加强对创造效应的积极引导。**《国务院关于深入实施“人工智能+”行动的意见》提出，“引导创新资源向创造就业潜力大的方向倾斜”。未来需建立不同 AI 技术方向对就业影响的评估机制，并根据评估结果合理调控人工智能领域资源配置方向。同时通过遴选典型案例、推广优秀实践等方式，增强对 AI 所创造新职业的社会认知度，发挥其在就业市场中的示范引领效应。**四是加强对社会保障的制度创新。**当前芬兰、美国、韩国、巴西、印度等国家开展如全民基本收入（UBI）等新型保障制度实验和研究，为社会转型提供更充分的制度预案。从试点结果来看，UBI 普遍改善了受众心理健康和生活消费，对社会稳定和经济增长有正向效应。

四、人工智能治理未来展望

伴随人工智能快速演进带来的变革性影响，人工智能治理体系必须更具前瞻性、适应性、系统性和包容性，从应对人机关系、人与人关系、人与社会关系这三类基本关系入手，着力构建完善以人为本、人机协同、多元共治的人工智能治理体系，以确保人工智能技术安全、可靠、可控，确保高质量发展和高水平安全良性互动。

（一）构建边界清晰的权责框架，破解应用责任困境

坚持统筹发展和安全，建立健全人工智能制度体系，建设涵盖高

[21] https://learnworkecosystemlibrary.com/initiatives/aiworksforamerica
[22] https://e-estonia.com/ai-leap-2025-estonia-sets-ai-standard-in-education/

中低位阶的法律法规、监管政策、技术标准等制度体系。**一是划定产权制度边界。**明确数据合法来源与合理使用边界，界定人工智能生成内容的作品属性与权利归属，探索建立基于贡献度的权益分配模式。综合运用要素支持、财税优惠、人才引进等不同类型政策工具，推动人工智能技术加速发展和应用广泛落地。**二是合理设定责任范畴，**针对人工智能系统研发、部署、使用环节主体多元、因果链条复杂的特性，明确归责原则，设定与技术能力相适应的，人工智能价值链上多方主体的法律责任，探索引入专项保险、补偿基金等机制。**三是探索推进场景应用规则。**面向“人工智能+”具体应用场景，结合行业特性需求细化上位法规则，制定与上位法相衔接的专项法规，明确特定领域的准入标准、责任划分和监管机制。对中小微企业及低风险应用探索减责免责清单与简化程序。推动针对不正当竞争、垄断行为的行业自律规范，营造公平竞争良好市场环境。

（二）发展敏捷动态的监管工具，提升治理体系效能

一是推行创新场景“沙盒监管”试点。对有望推动人工智能取得突破性进展与发展的创新应用，在有条件的区域试点“监管沙盒”，明确监管沙盒的申请流程、测试期限、风险监测、责任豁免及争议解决等核心机制，在沙盒运营、运维、安全保障、跨机构协调等方面，建立规范管理体系。**二是强化人工智能安全能力建设。**构建国家级人工智能测试验证平台，大力支持第三方审计、评估、认证机构发展，提供模型测试验证、供需对接、内容标识检测等服务，落地模型对抗安全、后门安全、可解释性等检测能力，推进加固工具等技术开发共享。

推出人工智能合规指南、合规工具包及风险评估模型等公共产品，为市场主体提供合规指引与咨询等服务。**三是构建多主体协同联动的风险预警与应对机制。**探索监管结果互认机制，统一监管要求与执行标准，明确主责部门，依托监管平台实时归集监管数据等关键信息。推动企业定期报送风险阶段性评估报告，建立完善内部监测与预警机制，在发生重大风险后及时上报事件情况。充分发挥行业组织调动作用，加强对行业企业发展态势跟踪监测和合规评估。

（三）塑造健康和谐的共生关系，引导人机协同发展

一是科学划定人机交互边界。坚持技术发展以增进人类福祉为根本宗旨，明确人工智能的辅助性工具地位，确保人工智能有助于拓展和深化现实社会关系，防范技术滥用带来的社交疏离、侵蚀人类主体性等风险。**二是建立动态伦理监测机制。**优化覆盖全生命周期的社会伦理影响评估机制，动态监测人工智能对公众沉迷依赖、心理健康、行为模式、社会家庭结构、就业结构等方面影响。**三是深化透明披露要求。**制定 AI 透明度标准、提升训练数据集的代表性与标注规范性、开发人工智能可解释技术。人工智能提供情感交互、任务协作、决策建议等服务时，应当以显著提示方式向用户披露 AI 身份。通过公众科普、数字素养教育等多种渠道，指导社会公众准确理解人工智能的技术原理、能力边界与潜在局限性，培养理性使用习惯。**四是明确主动干预义务。**明确不同风险等级对应的干预触发条件、响应流程与责任主体。规划并建设面向人工智能引发心理依赖、社会适应障碍等问题的专业社会心理援助与干预渠道，将其纳入公共心理健康服务体系。

（四）推进公平普惠的治理导向，应对经济社会影响

一是建立健全就业替代社会保障体系。建立可行的 AI 就业影响动态监测机制，明确监测核心指标，加强分析研判和风险预警。设置专项就业缓冲基金，完善灵活就业社保体系，探索全民基本收入等前瞻性的全民保障机制。**二是加强反垄断保障公平竞争市场环境。**关注大型平台利用 AI 生态系统构建封闭壁垒、限制互操作性等行为，维护开放、公平的市场环境。推动关键人工智能基础设施的开放与共享，在保障安全与隐私的前提下，探索通过公共数据开放平台、国家级算力网络调度、以公平合理条件开放接口等方式，降低中小企业创新门槛，维护健康多元的市场生态。**三是弥合数字鸿沟确保公平受益人工智能红利。**实施全面的全民数字素养与人工智能通识教育计划，将其融入国民教育体系与社区公共服务。通过基础设施投资、公共 AI 服务供给、技能培训等方式，确保不同地区、不同群体公平地获取和受益于人工智能技术，确保基础服务普惠化，防止技术红利分配不均加剧社会不平等。

中国信息通信研究院 政策与经济研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62305772

传真：010-62305772

网址：www.caict.ac.cn