

统筹发展与安全：智能体治理的范式转换、路径探索与制度回应

摘要：2026年5月，国家互联网信息办公室、国家发展和改革委员会、工业和信息化部联合印发我国首部智能体专项政策——

《智能体规范应用与创新实施意见》，推动我国智能体治理进入制度化探索的新阶段。智能体作为以大模型为底座、具备自主感知、决策与执行能力的新型行动者，推动人工智能从被动应答向主动行动跃迁，引发交互方式、网络结构与应用形态的全方位变革，也带来越权操作、责任悬空、级联失控等新型风险，由此促使治理范式进一步转换。在此背景下，学者们围绕这一专项政策，展开多维解读：在基本理念上，主张统筹发展与安全，构建贯穿全生命周期、分类分级、多元共治的敏捷治理体系，实现从过往内容合规管控转向全链条行为管控的转变；在治理工具上，提出以身份管理筑牢责任根基、以权限管理划定行为边界、以协议治理防范系统性风险的递进式架构；在治理主体上，厘清政府在规则搭建、产业培育和需求创造中的职能范畴，以及社会主体参与协同共治的角色定位；在场景应用上，注重公共服务、城市治理等场景的多样性和差异化；在风险防范上，提出既要采取技术手段，也要建立人工监督机制、能力冗余备份体系等，确保用户知情权与最终决策权。学者们强调，智能体治理须在技术迭代的动态博弈中持续进化，融入预见性治理理念，方能在激发创新活力的同时筑牢安全底线，推动智能体产业健康有序发展，为全球智能体治理贡献中国方案。

关键词：智能体；智能体治理；智能治理；AI治理；人工智能

DOI：10.16582/j.cnki.dzzw.2026.07.001

从内容合规到行为管控：智能体治理的范式转换与路径探索

李强治 中国信息通信研究院政策与经济研究所

在人工智能技术代际跃迁的浪潮中，智能体的出现推动人工智能从“被动回答者”向“主动行动者”的角色跃升。2026年3月，智能体概念首次被写入《政府工作报告》，成为推动“人工智能+”行动向纵深发展的关键载体。然而，智能体高自主性、高权限性以及深度交互等特征，带来了越权操作、工具滥用、多智能体级联故障等新型风险，对现行人工智能治理框架提出了严峻挑战。2026年5月，国家互联网信息办公室、国家发展和改革委员会、工业和信息化部联合印发我国首部智能体专项政策——《智能体规范应用与创新实施意见》（以下简称《实施意见》），标志着我国智能体发展与治理进入制度化探索的新阶段。

一、智能体时代的范式变革与治理挑战

智能体是以大模型为底座，具备自主感知、记忆、

决策、交互与执行能力的智能系统。它突破了大模型“一问一答”的交互局限，不仅能够理解复杂的自然语言指令，还能在授权范围内，自主规划和拆解任务、调用外部工具并代为执行任务，实现了从响应式输出向自主性完成目标任务的转变。

（一）智能体推动AI革命范式升级

相较于通用大模型聚焦内容生成能力，智能体更强调目标导向的执行能力，也由此开启了人工智能落地应用的全新范式，为千行百业的智能化转型提供了更具能动性的技术载体。一是交互形态之变。智能体的崛起正在触发人机交互的底层范式革命。在传统互联网时代，用户必须适应机器逻辑，在层层菜单和APP矩阵中寻找功能入口。互联网经济的本质是注意力经济，平台通过内容投喂和算法推荐抢占用户时长。进入智能体时代，用户只需以自然语言陈述意图，由智能体自主完成意图拆解、工具调用和

者，而非治理的共同设计者。未来的方向在于，将隐私保护、公平、透明等价值要求，从抽象原则转化为可操作的参与流程和申诉机制，从控制论式的闭环管理转向开放的、持续的、协商性的民主治理。

其三，制约智能体需要差异化的治理路径。《实施意见》提出“分类分级治理”，要求根据应用场景风险高低划定“可为”与“不可为”的边界，明确决策权限，确保用户拥有知情权和最终决策权。这些制度设计，正是为了把治理意愿转化为可操作的规则，但制度不能只写在纸上，须落到实处。当前，多数城市在算法透明和问责机制上的制度建设仍处于起步阶段。尽管部分城市出台了地方性法规，但大多数条款尚在探索之中，实际效果有待实践检验和后续补充，更多地方仍依赖国家层面的统一规范，制度供给明显滞后于技术应用。就此而言，技术制度化是一个在时长沉淀、节奏调适与时机把握中的迭代演化过程。^[7]将项目试点的经验转化为稳定的制度规范，是当前亟待破解的现实问题。

给智能体立规矩，不是要给创新“踩刹车”，而是要为它在服务城市、服务市民的轨道上画好线、安好护栏。制度跟上，技术才能真正服务于人。人工智能不会停下脚步，但城市治理的方向，终究掌握在人的手中。

参考文献：

[1]Winner L. Do artifacts have politics? [J]. Daedalus, 1980,

109(01): 121–136.

[2]Gahnberg C. What rules? Framing the governance of artificial agency[J]. Policy and Society, 2021, 40(02): 194–210.

[3]Buhmann A, Fieseler C. Towards a deliberative framework for responsible innovation in artificial intelligence[J]. Technology in Society, 2021, 64: 101475.

[4]翁士洪. 城市数字治理中的人机物协同机理研究[J]. 上海行政学院学报, 2024, 25(01): 62–74.

[5]Degli Esposti S, Ball K, Dibb S. What's in it for us? Benevolence, national security, and digital surveillance[J]. Public Administration Review, 2021, 81(05): 862–873.

[6]Wu X. How does the government facilitate the co-production of digital public safety services?—Based on empirical evidence from Shenzhen[J]. Social Sciences in China, 2024, 45(02): 113–134.

[7]韩万渠，周倩倩. 技术制度化的时间机制：基于三个数字政府治理平台的比较案例分析[J]. 上海行政学院学报，2025, 26(02): 26–37.

作者简介：吴晓林，南开大学周恩来政府管理学院、国家治理研究院教授，南开大学前沿交叉研究院研究员。

基金项目：国家自然科学基金项目“特大城市数字化风险防控的组织体系重构：风险演化、技术赋能与韧性跃迁”（项目号：72574110）；中央高校基本科研业务费项目“城市治理现代化”（项目号：340–63263038）。

智能体治理中的政府职能：内涵、挑战与深化

高翔 浙江大学公共管理学院/浙江大学民生保障与公共治理研究中心

2026年5月，国家互联网信息办公室、国家发展和改革委员会、工业和信息化部联合印发了《智能体规范应用与创新发展实施意见》（以下简称《实施意见》）。这是我国首份针对智能体（Agent）领域的系统性顶层政策文件，标志着这一前沿技术应用正在从“野蛮生长”迈入“规范繁荣”的制度轨道。此次文件的出台传递出明确的信号：智能体已经成为深度嵌入经济社会运行和公共治理的新型技术应用，其巨大的发

展潜力、广泛而深刻的社会影响，均要求政府明确自身职能定位，以制度化的方式系统应对业已出现的各种挑战。

政府如何规范智能体应用，并促进其创新发展？经典理论从基础制度构建、产业政策驱动和市场需求创造等三个维度，界定了政府助推创新发展的三种角色。

《实施意见》延续了上述角色分类，同时顺应智能体的技术特性调整了政府职能的优先次序和政策内涵。然

而，面对高速发展的智能体技术及其应用，政府始终面临着“安全”和“发展”两大挑战：一是如何优化治理机制，最大限度平衡安全与发展之间的内在张力，守住安全底线；二是如何推动制度变革，超越“后发劣势”，有效促进创新产业的长期发展。这两种挑战表现形态不同，但都涉及政府在经济发展中的职责边界这一经典理论议题，尤其要求政府持续创新职能履行方式。本文运用政府职能的经典理论，首先分析《实施意见》中政府角色的类型与最新特征，进而剖析其面临的挑战，探讨可能的应对路径。

一、经典角色的新内涵：《实施意见》中的政府职能界定

政府职能界定是理解经济发展、国家繁荣的核心议题。自20世纪30年代大萧条打破自由放任的政策共识以来，政府与市场的关系经历了多轮理论重构：凯恩斯主义为政府干预提供了宏观经济学基础，而东亚经济体通过产业政策实现赶超增长的经验则催生了发展型国家的相关理论，揭示政府如何系统性地塑造本国产业竞争力。2008年全球金融危机后，主要发达经济体产业政策的再度回归进一步表明，无论在理论还是实践层面，政府干预的边界、方式及其制度后果，已取代“是否干预”的二元争论，成为政治经济学与公共治理等领域研究的核心议题。

回答“如何干预”的问题，首先需要厘清政府在经济发展中承担的基本职能。其中，政府有三种经典角色：制度基础建构者、产业发展引导者和市场需求创造者。基础制度建构者的角色源于制度经济学的基本命题，强调制度是人为设计的约束，旨在降低交易成本、塑造激励结构。在这一角色下，政府需要制定法律法规，提供产权保护，构建标准体系和服务平台，建立市场机制有效运转的制度框架。产业发展引导者的角色延续了发展型国家的理论范式，认为政府可以通过干预资源配置、扶持战略性新兴产业来实现国家战略目标。市场需求创造者的职能界定则将政府视为新技术的领先用户，通过采购行为直接创造市场需求，降低创新产品进入市场的早期风险。相应地，政府的政策工具涵盖首购订购、示范应用等多种形式。

《实施意见》延续了政府职能的上述角色类型，但结合人工智能技术特性，对其优先次序和具体内涵作了如下调整：

第一，明确突出守底线的安全治理，将其作为制度基础构建的核心内容。制度构建是《实施意见》的重要组成部分，其内容主要集中在第二部分“夯实发展基础”和第三部分“守牢安全底线”。制度构建突出了政府是智能体发展和应用的规则制定者，需要为智能体技术划定行为边界和提供治理框架。值得注意的是，文件所指基础制度，尽管涉及标准体系、智能互联网架构等传统市场制度构建的必要内容，但其重心已大幅向安全治理框架倾斜。《实施意见》在“基本原则”部分将“安全可控”“规范有序”置于首要位置，更专设“守牢安全底线”，从明确产品准则、防范安全风险（供应链）、完善治理体系和强化行业自律等方面详述规范智能体发展、应用的治理框架。总的来说，《实施意见》中所涉及的基础制度更侧重防范智能体应用与发展的负外部性，旨在为其设定行为红线，折射出我国在前沿技术治理中“安全优先”的阶段性特征。

第二，产业政策仍然是政府促进发展的最主要路径。《实施意见》明确要求坚持创新驱动，内容集中在第二部分“夯实发展基础”、第四部分“强化应用牵引”和第五部分“建设创新生态”。这些产业政策具体指向基础技术研发攻关、智能体工具链完善，涵盖智能制造、金融服务等多领域的场景清单式引导，以及开源生态与产业协作平台的培育。审视其政策工具特征可以发现，文件更多运用信号型与协调型工具，即通过“研发”“探索”“推动”等表述指明技术方向与应用领域，未具体涉及专项补贴、税收优惠、信贷配给等经典发展型国家的资源型工具。作为由中央部委发布的纲领性文件，《实施意见》的政策工具选取延续了中央层面产业政策的一般模式，即以发布信号、组织协调为主，而较少直接参与资源配置。然而，在央地关系的制度结构下，中央文件列出的重点产业方向可能成为地方政府开展补贴竞争的合法性目录。相应地，地方政府在智能体发展与治理中的政策执行，将是观察、评估产业政策实践特征和落地实效的关键所在。

第三，需求创造型政策工具正在成为政府促进智能体发展的新兴路径。《实施意见》提出“基本原则”的第四条即为“坚持应用牵引”，要求“发挥典型应用场景示范效应”，第五部分则明确要求采取“公开招标、揭榜挂帅”等方式吸引智能体企业定制化开发，以及推动重点领域开放应用场景。此前，国务院办公厅在2025年11月发布的《关于加快场景培育和开放推动新场景大规模应用的实施意见》，同样将场景供给作为推动科技创新和产业创新融合发展的重要载体。考虑到政府是需求牵引的重要主体，文件更在第四部分“强化应用牵引”清晰列举了“政务服务”“司法服务”“公共安全”“城市治理”“招标投标”等公共部门的典型应用场景，“教育教学”“医疗健康”“人力资源”“信息服务”等领域也部分涉及公共部门的智能体应用。由此，政府采购及其推动的应用场景，可成为智能体应用发展的重要需求来源。

二、智能体治理中政府保障安全、促进发展的双重挑战

《实施意见》对政府职能的界定，充分表明安全与发展是我国政府在智能体治理中的两个关键目标，尽管两者存在内生张力。智能体的产业变革潜力要求政府加速推动其创新发展，而它因自主执行能力而衍生的系统性风险，又迫使政府必须同时承担安全守门人角色。在国际层面，安全与发展的紧张关系已分化为两种治理范式：欧盟以权利保护与社会风险防控为优先，美国则以市场机制主导产业发展。相比之下，我国仍试图兼顾两者：既在发展语境下建构有效的安全治理体系，又要维持创新驱动的发展动能。两者均属高难度政策目标，要求政府在恰当界定职能范围的同时，持续优化履职方式。

守住安全底线，要求政府超越技术遏制路径，在智能体应用推进中建构适配的治理架构。然而，智能体治理面临三重难题。第一，其技术特性消解了传统治理的基本前提。智能体具备强自主性与跨域执行能力，可在无人实时干预下感知环境并自主决策。多智能体交互更易涌现未曾预设的行为模式，使传统问责机制面临失效风险，按行业或场景划分的监管框架亦难以适配。第二，技术迭代与制度供给之间存在步调差异。法律法规的制定需要严格遵循立法程序，治理规则虽更加灵活，也难充分跟上智能体等

人工智能技术的指数级演进速度。第三，智能体应用可能引发更加广泛的社会风险，远超单一部门的应对能力。就业替代、算法歧视、数字鸿沟等衍生后果涉及价值判断与利益分配，对治理体系提出了更高要求。

对此，《实施意见》初步建构了面向智能体应用的多维监管治理架构。在技术维度上，要求发展安全治理工具，提升对非合规行为的发现、干预、阻断与恢复能力，并通过规则内嵌、行为围栏等技术手段实现合规性前置。在法律规范维度上，明确决策权限边界，强化行为管控与可追溯机制，要求建立全生命周期安全规范与供应链安全管理制度，以刚性制度约束划定行为边界。在行政管理维度上，确立分级分类治理框架，依风险等级与应用领域采取备案管理、行业自律等差异化措施。此外，引入信用评价等软性约束机制作为补充。总体而言，《实施意见》通过技术标准、法律规范、行政管理与行业自律的协同配置，形成了初步的制度回应。

然而，既有治理框架在应对不同类型风险时表现出了较为明显的制度强度差异。针对技术内生风险，《实施意见》的制度设计最为充分：内生安全能力要求、供应链安全管理、行为管控机制与分级分类治理等多重工具叠加，旨在应对数据投毒、算法篡改、系统漏洞与运行失控等技术性问题。相比之下，针对社会衍生风险的制度供给明显薄弱，相关表述多停留于原则性宣示，尚缺乏可操作的机制设计。这一监管强度差异揭示了智能体治理的深层困境：技术风险具有可工程化处理与可标准化评估的特征，易于纳入刚性制度框架；社会衍生风险的回应要求复杂社会互动，难以通过技术标准、行政规制即有效应对，在政策文本中往往被边缘化或笼统化处理。

政府在智能体应用发展中面临的第二重挑战，在于传统发展型国家模式在前沿技术创新领域的适用性边界。传统产业政策的有效实施多处于“追赶型”发展阶段。其成功往往依赖明确的技术对标、可复制的产业路径以及集中投入的要素。换言之，工业化后发国家可以通过模仿、借鉴发达国家业已被证明成功的技术、组织和产业模式来谋求发展。然而，智能体所代表的的前沿人工智能技术具有高度的不确定性，缺乏既定的技术、组织和治理轨道可供跟随。当技术前沿从已知转向未知，单纯

依靠资源动员已经难以确保持续的竞争优势，后发国家常用的行政主导模式就会面临深刻的制度适应性危机。

面对这一挑战，《实施意见》已经呈现出与传统发展型政府模式的显著差异。该文件以“应用牵引”为核心导向，强调通过典型场景示范、标准体系建设和基础架构搭建来降低创新门槛，体现出政府试图从资源动员者转向生态构建者的转型意图。然而，上述政策的执行实效仍然有待观察。一方面，该文件较少涉及需求驱动的机制设计，如尚未系统引入创新采购、首购订购等成熟的需求侧政策工具。政府作为领先用户的功能还需要各部门、各地方的配套文件，以推动其从信号释放向市场购买的跨越。另一方面，地方政府仍有可能延续传统路径制定产业政策。如《实施意见》中列举的应用场景，极有可能异化为地方政府间的补贴竞争。换言之，中央转型意图与地方执行惯性之间的张力，可能构成了政府助推智能体创新发展的重要结构性障碍。

三、智能体治理的深化路径

整体而言，《实施意见》已为智能体治理中的政府角色确立了清晰的框架。围绕安全与发展的双重目标，该文件调适了政府在制度基础构建、产业发展引导与市场需求创造等方面的职能内涵。在安全维度上，确立了风险防控的底线优先原则，搭建了涵盖技术标准、法律规制、行业自律与行政监管的复合型治理框架。在发展维度上，产业政策正逐步超越传统的要素动员逻辑，转向以场景开放与生态营造为核心的新路径。

当前，政府的智能体治理仍面临两重深层挑战。其一，拓展安全治理边界。既有政策多聚焦技术风险的法律与技术规制，尚未有效覆盖就业替代等社会衍生风险，亟需从单一技术约束向社会风险综合治理延伸。其二，防范地方发展型政府的路径依赖。在推动智能体创新发展中，地方政府易受惯性影响，继续采用补贴等传统工具，迫切要求有效引导地方摆脱短期招商逻辑，转向创新生态营造与新型市场制度供给。

深化智能体治理，关键在于优化政府职能的履行机制。在安全维度，现有监管框架已超越消极合规逻辑，呈现出多主体协同的敏捷治理特征。不过，这一协同网

络仍高度集中于行政部门、技术专家与企业等供给侧主体，需求侧的公众参与明显不足。然而，智能体的自主性与交互性决定了其社会风险并非由技术单向预设，而是源于人机复杂互动的动态建构。相应地，仅靠供给侧划定技术与制度边界难以根治问题，有效的安全治理必须要求凸显公众的主体性，推动其从被动的风险承受者向积极的治理参与者转变。对政府来说，这意味着要将培育负责任使用者纳入制度体系，打造具有“合供”特征的智能体治理框架：通过普及数字素养教育，提升公众理解智能体逻辑与识别潜在风险的能力；通过凝聚技术伦理共识，使规范体系从自上而下的强制命令转化为社会协商的内生秩序；通过引导负责任的实际应用，确保用户在充分应用技术的同时保持批判性反思。

聚焦发展，政府行为模式转变的关键层级在地方，高层级政府亟需引导地方政府从资源配置向制度创新转变。当前，在构建全国统一大市场的制度语境下，地方政府的经济促进行为正受到日益严格的约束。不过，《公平竞争审查条例》将“增强国家自主创新能力”列为法定例外情形。尽管该例外附有严格的比例原则与期限要求，但在实践中常被泛化适用，异化为地方以创新之名行补贴竞争之实的制度通道。事实上，前沿创新的高度不确定性决定了其无法依赖传统的要素补贴或行政指令，而更需场景开放、生态营造与规则供给。人工智能等产业发展所必需的各类制度创新，如数据产权界定、责任归属与跨域协作等，也需要在市场竞争、互动与试错中逐步演化。因此，高层级政府必须强化对创新领域地方行为的引导与规制，推动地方政府从资源分配者转向生态维护者，通过搭建协商平台、保障信息透明与规则公平，为适应性制度在互动中涌现提供基础条件。

作者简介：高翔，浙江大学公共管理学院教授，浙江大学民生保障与公共治理研究中心副主任。

基金项目：浙江省哲学社会科学规划重大培育专项“公共数据支撑的数字政府效能提升路径与机制研究”（项目号：26NDPY010ZD）；国家自然科学基金重点项目“数字政府驱动的治理范式变革研究”（项目号：72434004）。