# 基于风险的分等级人工智能创新监管重点问题研究

陶盼盼 李佳

中检集团天帷网络安全技术（合肥）有限公司，安徽合肥，231200；

**摘要：**人工智能技术的迭代速度与应用广度正在重塑全球经济、社会与公共治理，也同步放大了数据泄露、算法歧视、技术失控等系统性风险。欧盟《人工智能法》和美国《国家人工智能倡议法案》分别用“统一立法”与“分层治理”两条路径探索风险分级监管，但我国尚缺乏与产业发展水平、技术复杂性和治理需求相匹配的整体框架。本文在梳理全球监管经验的基础上，提出以“风险识别—分级—校准—量化”四步闭环为核心的中国方案，为《人工智能法》草案提供可直接落地的制度设计，也为产业界提供清晰的合规指引。

**关键词：**人工智能；风险分级；动态监管；合规沙盒；技术主权

**DOI：**10.64216/3080-1508.25.08.035

过去十年，生成式人工智能专利申请量一路飙升，世界知识产权组织最新统计显示，2014—2023 年全球共计 5.4 万件，其中中国贡献了 3.8 万件，位列第一。专利繁荣背后，是技术嵌入千行百业后产生的巨大风险敞口：我国人工智能企业的数据合规成本年均增幅 37.6%，数据权属诉讼案件三年增长 215%；医疗诊断、自动驾驶等高风险场景仍有 38.2%的技术指标空白；技术出口备案量年增 45%，而现行《技术进出口管理条例》中与人工智能直接相关的管制条目仅占 3.7%，2023 年已出现 17 起核心技术违规外流事件。国务院《新一代人工智能发展规划》为立法设定了“三步走”时间表：2020 年初步建立伦理规范，2025 年初步搭建法规体系，2030 年建成更完备的制度体系，而 2023 年 6 月《人工智能法草案》已被列入年度立法预备项目，意味着制度窗口期已经到来[1]。然而，国内对“基于风险的规制”仍停留在理念层面，缺乏面向生成式人工智能的动态分级模型与制度落地研究。本文以生成式人工智能为切口，尝试回溯风险规制的理论源流，构建契合中国国情的分级监管框架，并给出可操作的实施路径。

## 1 人工智能监管模式跨国比较研究

### 1.1 欧盟统一立法模式特征

欧盟《人工智能法案》确立“禁止—高风险—有限风险—最小风险”四级体系，以健康、安全、基本权利为评判轴心[2]。配套三项机制：①欧盟人工智能办公室统筹跨境监管；②全生命周期追溯，开发日志存留≥10 年；③监管沙盒，中小企业可获 3 年测试期。实施后 AI 专利合规成本下降 23%，但市场准入门槛抬高，初创企业融资周期平均延长 4.8 个月。欧盟人工智能风险等级图如图 1 所示。


图 1 欧盟人工智能风险等级图

### 1.2 美国分层治理实践

美国联邦层面以《人工智能倡议法案》确立国家安全审查框架，商务部已发布 7 类核心技术出口管制清单；FTC 近三年 AI 反垄断案件年均增长 217%。州立法呈差异化：加州要求算法影响评估，德克萨斯聚焦能源 AI 安全。2024 年科罗拉多州《人工智能法》聚焦反歧视，锁定教育、就业、金融、医疗、住房等高风险场景。分

散式监管在灵活性与协调成本之间持续拉锯[3]。

### 1.3 中国人工智能风险治理面临的挑战

目前中国尚无统一基础性立法，依靠《算法推荐管理规定》《深度合成管理规定》《生成式 AI 服务管理暂行办法》等部门规章交叉规制[4]。主要痛点：

①规则碎片化，行业、专业监管重叠，治理成本高；

②工具单一，倚重事前审批与事后处罚，监管沙盒、伦理设计等新机制缺位；

③大模型决策不透明、产业链责任分散，传统线性规制难以适配动态创新。

结论：中国需尽快建立统一、动态、基于风险的分级监管框架，以平衡安全与创新发展。

## 2 构建中国人工智能分等级监管规范体系

人工智能风险分类应该遵循以下原则：其一，风险分类应该根据对个人、社会和生态系统的不利影响进行范围界定，包括对人的基础权利（比如安全和健康）的影响、使用程度、预期目的、受影响的人、替代方案的可获得性、危害的不可恢复性等方面。其二，应结合应用场景进行风险分类，风险维度包括伦理、性能和安全等方面，评估要点包括场景应用的目的、人工智能部件在整个系统中的功能以及应用所处行业的特性等内容。风险分类有以下目的：其一，识别风险类型后，便于查找风险来源、采取针对性风险应对措施；其二，风险类型可以用于判断损害后果的严重程度，让风险分级更加科学[5]。如上文所述，先构建生成式人工智能风险类型库（见表 1），再根据技术演变的情况及时调整需要重点关注的风险类型。

表 1 生成式人工智能风险类型库

| 风险因素 | 风险类型 | 具体内容 |
|---|---|---|
| 技术风险 | 算法风险 | 机制不可解释性、毒性、歧视、偏见、过拟合等 |
| | 数据风险 | 训练数据中的偏见（BiasinTrainingData）、数据投毒（DataPoisoning）、私密数据被窃取、个人隐私泄露、测试集被“污染”等 |
| | 模型风险 | 脆弱性、高能耗、保密性、训练过程不稳定等 |
| 应用风险 | 物理损害风险 | 基础设施被攻击、实施违法犯罪活动、开发和滥用自主性武器等 |
| | 个人安全、健康和基本权利风险 | 侵犯个人的隐私权、人身自由权、就业权、受教育权等 |
| | 政治与社会公共秩序风险 | 网络动员风险、干预选举、虚假信息泛滥、社会认知“割裂”、社会舆论控制等 |

资料来源：根据《人工智能安全治理框架》1.0 版进行整理。全国网络安全标准化技术委员会：《人工智能安全治理框架》，2024 年 9 月 9 日，https://www.cac.gov.cn/2024-09/09/c_1727567886199789.htm[2025-01-19]。

从敏感场景看，生成式人工智能应用对个人权益或社会公共利益有重大影响时都应被视为高风险，比如涉及税务、金融、教育、住房、就业等应用场景，或者涉及个人资格授予的政务场景，包括自动化行政、智慧司法等。从应用规模看，社交类平台、用户规模超大的平台和涉及内容分发业务的平台使用生成式人工智能等情景，都应该被视为高风险。然而，因应用环境、功能用途和人为因素等方面存在差异，人工智能的潜在风险程度不同。因而，除了由法律认定哪些场景或产品为不可接受风险或高风险外，还应秉持“场景驱动”的治理理念，依据风险评估标准对具体场景的风险等级进行灵活判断，常用工具包括风险矩阵法、风险指标法和风险模型法。风险矩阵基于两个交叉因素，即风险发生的可能性和损害后果的严重性，来认定风险等级。如表 2 所示，Ⅰ级为不可接受风险，即发生可能性高且损害后果严重；Ⅱ级为高风险，即发生可能性较高且损害后果严重，或者发生可能性高且损害后果较严重；Ⅲ级为中风险，即发生可能性低但损害后果严重，或发生可能性较高且损害后果较严重，或发生可能性高但损害后果轻；Ⅳ级为低风险。

表 2 风险矩阵

| 风险等级 | | 风险发生的可能性 | | |
|---|---|---|---|---|
| | | 高 | 中 | 低 |
| 损害后果的严重性 | 高 | Ⅰ | Ⅱ | Ⅲ |
| | 中 | Ⅱ | Ⅲ | Ⅳ |
| | 低 | Ⅲ | Ⅳ | Ⅳ |

资料来源：梁正、薛澜、张辉、曾雄：《人工智能治理框架与实施路径》，北京：中国科学技术出版社，2023 年。

领域基础分级作为人工智能动态分级监管体系的首要环节，是实现风险初步锚定的基础。该分级方式参照《关键信息基础设施安全保护条例》《生成式人工智能服务管理暂行办法》等相关法规的领域分类，依据应用领域的社会敏感性、公共利益关联度，将人工智能应用领域划分为禁止类、高风险、中高风险、中风险和低风险五个类别，为后续监管工作奠定初步的风险框架。

禁止类领域主要包括社会煽动、军事杀戮、违法监控等场景。此类应用严重违背社会公共利益、伦理道德和法律法规，可能对社会稳定、国家安全和公民权利造成不可估量的危害，因此被明确列为禁止应用，从源头遏制其发展与传播，体现了监管的底线思维。高风险领域涵盖医疗诊断、自动驾驶、司法裁判等典型场景。这些领域与公民的生命健康、财产安全以及社会公平正义紧密相关。以医疗诊断为例，人工智能在疾病诊断中的应用直接关系到患者的治疗方案和生命安全；自动驾驶技术的失误可能导致严重的交通事故；司法裁判中的人工智能应用则影响着司法公正的实现。因此，这类领域的基础风险等级被定为 4 级（极高风险），需要实施严格的监管措施。中高风险领域包括金融风控、教育评估、招聘筛选等场景。在金融风控中，人工智能模型的准确性影响着金融机构的信贷决策和金融市场的稳定；教育评估和招聘筛选中的人工智能应用则涉及到教育公平和就业公平。该领域的基础风险等级为 3 级（高风险），监管力度需介于高风险领域和中风险领域之间。中风险领域涉及内容推荐、工业质检、公共服务等场景。内容推荐虽然可能影响用户的信息获取和价值取向，但相对而言危害程度较低；工业质检中的人工智能应用主要影响产品质量，公共服务中的应用则关系到服务效率和质量。该领域基础风险等级为 2 级（中风险），采取适度的监管措施即可。低风险领域包括娱乐生成、办公助手、智能家居等场景。这些领域的人工智能应用主要为用户提供娱乐、办公便利和生活便捷，对社会公共利益和公民权利的影响较小，基础风险等级为 1 级（低风险），监管以引导和规范为主。

在领域基础分级确定初步风险等级后，还需根据人工智能应用的技术特性与应用方式进行功能应用校准，实现风险等级的±1 级调整，动态调整参照示例表如表 3 所示。这一环节使得监管更加灵活和精准，能够充分考虑不同应用方式下风险的差异性。

**表 3 动态调整参照示例表**

| 校准因素 | 风险上调情形 | 风险下调情形 |
|---|---|---|
| 决策自主性 | 完全自主决策（无人工干预） | 人类最终决策权（仅辅助建议） |
| 影响不可逆性 | 导致人身伤害/重大财产损失 | 错误可低成本修正（如推荐内容） |
| 数据敏感性 | 处理生物识别、医疗健康等敏感数据 | 仅使用公开非敏感数据 |
| 应用规模 | 国家级基础设施/亿级用户覆盖 | 企业内限定场景使用 |
| 系统透明度 | 黑盒模型（不可解释决策逻辑） | 白盒模型（全程可追溯） |
| 示例：<br>医疗领域（基础 4 级）+辅助诊断工具（医生最终决策）→降为 3 级<br>教育领域（基础 3 级）+AI 自动录取评分（无人工复核）→升为 4 级 | | |

决策自主性是重要的校准因素之一。当人工智能应用具备完全自主决策能力且无人工干预时，其决策失误可能带来难以挽回的后果，风险相对较高，需上调风险等级；而当人类拥有最终决策权，人工智能仅提供辅助建议时，能够有效降低决策风险，故风险等级可下调。例如，在医疗领域，若人工智能完全自主进行疾病诊断并制定治疗方案，无医生干预，其风险等级应上调；而辅助诊断工具在医生最终决策的前提下，风险等级可适当下调。影响不可逆性也是关键的校准因素。当人工智能应用的错误可能导致人身伤害或重大财产损失，且影响难以逆转时，风险等级需上调；若错误可通过低成本修正，如内容推荐出现偏差可及时调整，风险等级则可下调。数据敏感性方面，处理生物识别、医疗健康等敏感数据的人工智能应用，一旦数据泄露或被滥用，将严重侵犯公民的隐私权和个人信息安全，风险较高，应上调风险等级；仅使用公开非敏感数据的应用，风险相对较低，可下调风险等级。应用规模同样影响风险等级调整。覆盖国家级基础设施或亿级用户的人工智能应用，

其一旦出现问题，影响范围极广，风险等级需上调；而仅在企业内限定场景使用的应用，影响范围有限，风险等级可下调。系统透明度也是校准的重要考量。采用黑盒模型，决策逻辑不可解释的人工智能应用，其潜在风险难以预测和把控，风险等级应上调；采用白盒模型，决策过程全程可追溯的应用，便于监管和问题排查，风险等级可下调。以教育领域为例，其基础风险等级为 3 级，若应用 AI 自动录取评分且无人工复核，决策自主性高且影响不可逆，风险等级应升为 4 级；而若人工智能仅作为教学辅助工具，由教师掌握最终决策，风险等级则可能下调。

危害量化定级是人工智能动态分级监管体系的最终环节，通过三维度评估矩阵（如下表 4 所示）确定最终风险等级，实现对风险的量化评估和精准界定。

**表 4 三维度评估矩阵**

| 评估维度 | 评估指标 | 量化方法 |
|---|---|---|
| 危害可能性(L) | 系统故障率、对抗攻击脆弱性、数据偏差程度 | 概率模型（低/中/高/极高） |
| 危害严重度(S) | 人身安全、财产安全、社会秩序、公民权利 | 影响分级量表（1-10 分） |
| 影响范围(R) | 个体/群体/区域/国家级 | 辐射范围权重系数 |

该评估矩阵包含危害可能性（L）、危害严重度（S）和影响范围（R）三个维度。危害可能性主要通过系统故障率、对抗攻击脆弱性、数据偏差程度等指标进行评估，采用概率模型划分为低、中、高、极高四个等级。系统故障率高、易遭受对抗攻击、数据偏差程度大的人工智能应用，其发生危害的可能性相对较高。最终风险值通过“最终风险值=L×S×R”的公式计算得出，并根据分值划定为 5 个风险等级：90 分及以上为 5 级（禁止类），70-89 分为 4 级（严格许可），50-69 分为 3 级（专项备案），30-49 分为 2 级（基础备案），30 分以下为 1 级（自主合规）。

总之，风险分级分类是方法而非目的，其落脚点是针对具体场景快速识别风险类型和判断风险等级。风险分级分类的标准和目录不是一成不变的，需要根据技术发展状况和应用规模灵活调整，特别是引入技术专家定期评估关键领域的风险类型和风险等级，动态调整风险分级标准和风险类型库。

## 3 结语

构建基于风险的分等级监管体系，是我国人工智能立法进程中的关键一步。这一体系不仅能为《人工智能法》草案的制定提供理论参考，更能为产业界提供清晰的合规指引，推动生成式人工智能在安全可控的前提下实现创新突破。在全球人工智能竞争日趋激烈的背景下，唯有建立既符合中国国情又与国际接轨的风险治理框架，才能真正实现“规范与发展并进”的治理目标，为人工智能技术的可持续发展筑牢制度根基。

## 参考文献

[1]周汉华.论我国人工智能立法的定位[J].现代法学,2024,46(5):17-34.

[2]周学峰.论人工智能的风险规制[J].比较法研究,2024(6):42-56.

[3]丁晓东.全球比较下的我国人工智能立法[J].比较法研究,2024(4):51-66.

[4]赵鹏.“基于风险”的个人信息保护?[J].法学评论,2023,41(4):30-45.

[5]安永康.基于风险而规制:我国食品安全政府规制的校准[J].行政法学研究,2020(4):40-55.

第一作者简介：陶盼盼（1990.11-），男，汉族，本科，研究方向：人工智能技术。

第二作者简介：李佳（1988.08-），男，汉族，本科，研究方向：人工智能技术。