网络首发时间：2025-02-20 11:29:27　网络首发地址：https://link.cnki.net/urlid/11.3799.F.20250219.1702.002

中国人工智能风险治理体系构建与基于风险规制模式的理论阐述：以生成式人工智能为例*

曾　雄　梁　正　张　辉

搭建人工智能风险治理体系是人工智能立法的核心问题。基于风险规制，即以风险为核心考虑要素，按照风险类别和级别确定优先治理的风险场景，其实质在于合理分配监管资源、提高风险治理的有效性，这也是构建中国人工智能风险治理体系的重要参照。在基于风险规制模式下，需要采取适当措施。首先，构建一个包括风险评估、风险管理和风险沟通三个环节的框架流程，以实现风险管控的集中化。其次，坚持风险分级分类的实施路径。在风险类型上，建立人工智能风险类型库，并根据新技术的演变情况及时更新；在风险等级上，制定明确的风险评级标准，开发相应的风险评级工具和模型，将风险划分为不可接受风险、高风险、中风险和低风险，并为不同等级的风险匹配相应程度的治理措施。最后，设立统一的人工智能风险治理机构，以实现监管权的集中统一。

关键词：基于风险规制　生成式人工智能　风险评估　风险管理　风险沟通　分级分类

问题的提出

2017年7月，国务院印发的《新一代人工智能发展规划》提出了人工智能法

曾雄系北京科技大学文法学院讲师；梁正系清华大学人工智能国际治理研究院副院长、公共管理学院教授；张辉（通讯作者）系中国科学技术大学公共事务学院特任副研究员，Email:hzhangchina@163.com。

*本文系教育部人文社会科学青年基金项目“我国平台经济领域自我优待行为的反垄断规制研究”（项目编号：24YJC820003）和国家社会科学基金重点项目“人工智能大模型安全治理的监管平台建构与治理机制研究”（项目编号：24AGL009）的阶段性研究成果。作者感谢匿名审稿专家及编辑部的建设性修改意见，文责自负。

律法规体系的建设规划：第一步，到2020年，部分领域的人工智能伦理规范和政策法规初步建立；第二步，到2025年，初步建立人工智能法律法规、伦理规范和政策体系；第三步，到2030年，建成更加完善的人工智能法律法规、伦理规范和政策体系。2023年6月，国务院印发的《2023年度立法工作计划》提出“预备提请全国人大常委会审议人工智能法草案”。

在人工智能全球竞争日趋激烈的背景下，中国人工智能立法具有紧迫性。[1]近年来，生成式人工智能的迅猛发展深刻影响着人类社会。根据经济合作与发展组织（OECD）的界定，生成式人工智能指可以创造新内容（如文本、图片、视频和音乐等）的智能技术。由于生成式大模型具有自主性、涌现性和快速迭代的技术特质，其带来的风险问题高度复杂，学界对人工智能风险预防的方式存在重大争议。有研究认为应采用基于风险的规制模式，[2]也有研究提出应进行分领域、分场景的规制。[3]采用何种风险规制模式直接影响立法条文的设计，也将影响风险规制的实际效果。欧盟《人工智能法》作为全球首部人工智能立法，采用了基于风险规制的模式。[4]如今这种模式逐渐成为新技术风险治理的重要方法。国内有学者讨论基于风险的治理在个人信息保护领域的适用问题，[5]也有学者讨论基于风险规制在食品安全领域的适用问题。[6]

然而，国内针对如何将基于风险规制模式应用于人工智能风险治理的研究成果较少，对基于风险规制的理论溯源和分析也不够全面。鉴于此，本文以生成式人工智能风险治理为研究对象，剖析基于风险规制模式的历史渊源、概念内涵和制度价值，镜鉴域外采用基于风险规制模式的实践经验，结合中国人工智能风险治理现状与问题，就构建中国人工智能风险治理体系提出制度方案。

[1] 周汉华：“论我国人工智能立法的定位”，《现代法学》，2024年第5期。

[2] 周学峰：“论人工智能的风险规制”，《比较法研究》，2024年第6期。

[3] 丁晓东：“全球比较下的我国人工智能立法”，《比较法研究》，2024年第4期。

[4] 国内学者对基于风险规制（Risk-based Regulation）有不同表述，有人称其为“以风险为基础的治理”，参见陈吉栋：“以风险为基础的人工智能治理”，《法治研究》，2023年第5期。有人称其为“基于风险而规制”，参见安永康：“基于风险而规制：我国食品安全政府规制的校准”，《行政法学研究》，2020年第4期。也有人称其为“基于风险的监管”，参见刘鹏、张嵛楠、王力：“基于风险的政府监管：理论发展与实践应用”，《中国行政管理》，2024年第3期。

[5] 赵鹏：“‘基于风险’的个人信息保护?”《法学评论》，2023年第4期。

[6] 安永康：“基于风险而规制：我国食品安全政府规制的校准”，《行政法学研究》，2020年第4期。

数字时代基于权利规制向基于风险规制的转向

基于权利规制模式（Rights-based Regulation）存在权利概念不清、权利行使困难和权利难获保障等缺陷，基于风险规制模式则逐渐受到重视。

（一）基于权利规制模式面临的时代困境

基于权利规制模式指通过赋予个体一些新型权利来应对风险，权利类型包括算法解释权、被遗忘权和删除权等。基于权利规制秉持非黑即白的“二值逻辑”，通过设定权利和义务对个体进行整体保护，没有考虑每个个体面临的风险或遭受损害的程度。[1]在数字时代，基于权利规制面临挑战。

以个人信息保护领域为例，可能遭遇一些挑战。第一，每个人都是有限理性的主体，在权利意识和风险意识方面存在个体差异，难以保障每个个体及时、有效和理性地行使各项权利。第二，基于权利规制的核心是赋予个体新型权利，但是权利的概念和范围存在模糊性，比如学界对个人信息权的法律属性存在争论，存在“隐私权说”“人格权说”“财产权说”等不同主张。[2]与个体权利对应的是个人信息处理者应履行的义务，权利的不清晰导致义务承担者的行为边界模糊，个体利益难以通过权利体系得到保障。第三，个体与“科技巨头”之间的地位不平等，后者可以设计冗长和复杂的隐私声明规避“告知—同意”规则，[3]采用“暗黑”模式巧妙地影响个体的选择，[4]甚至滥用市场垄断地位对个体进行数据“剥削”。[5]第四，虽然侵权损害赔偿是保障个体权利的重要路径，但是在侵权认定过程中，个体需要举证证明其遭受的损害后果以及该后果与侵权行为之间存在因果关系，然而，个人信息权遭受损害具有无形性、迟延性和弥散性等特征，个体面临较高的举证门槛。总之，基于权利规制存在固有缺陷，虽然相关法律给予了个体算法解释权、被遗忘权和删除权等，但仍无法有效应对算法“黑箱”、自动化决策偏见和“大数据杀熟”等问题。

---

[1] Gellert R., *The Risk-based Approach to Data Protection*, Oxford: Oxford University Press, 2020.

[2] 程啸、王苑：《个人信息保护法教程》，北京：中国人民大学出版社，2023年。

[3] Lehtiniemi T. and Kortesniemi Y., “Can the Obstacles to Privacy Self-management Be Overcome? Exploring the Consent Intermediary Approach”, *Big Data & Society*, 4(2):1-11, 2017.

[4] FTC, “Bringing Dark Patterns to Light”, September 2022, https://www.ftc.gov/reports/bringing-dark-patterns-light[2024-09-27].

[5] 曾雄：“在数字时代以反垄断制度保护个人信息的路径与模式选择”，《国际经济评论》，2022年第3期。

（二）基于风险规制模式的欧盟实践

欧盟、加拿大、[1]美国[2]和韩国[3]等国家都采用基于风险规制模式搭建人工智能风险治理框架。以欧盟为例，欧盟《通用数据保护法》采取了“自下而上”的风险规制路径，没有通过法律明确风险评估和减缓风险的措施，而是由企业根据义务和问责原则构建内部风险管控制度。欧盟《人工智能法》则采用“自上而下”的风险规制路径，即由法律明确规定高风险场景类型和匹配对应措施，留给被规制者的自由裁量空间非常有限。[4]这两种路径反映出规制者在采用基于风险规制模式时站在不同的逻辑出发点。

在理论上，基于风险规制模式的逻辑出发点主要有三类。其一，将风险管理视为实现更好合规的手段。比如在金融领域，金融机构利用风险管理工具来确定其是否留出了足够的资金来应对风险。其二，为了确定哪一类风险需要规制，应综合考虑技术对健康、安全和环境的影响。其三，考虑到监管成本，对规制行为进行优先排序。可见，欧盟《通用数据保护法》对风险的界定主要是基于更好地遵守法律义务的逻辑，而欧盟《人工智能法》则是基于科学计算来评估人工智能对人的健康和安全可能带来的危害。因此，在《通用数据保护法》的框架下，数据控制者拥有更多的自由裁量权来决定如何遵守规则。相比之下，根据《人工智能法》，人工智能系统作为直接的规制对象，法律对其提出了“僵硬”的合规要求，具有典型的“命令—控制型”特征。例如，该法列举了高风险场景，并要求被规制者必须严格遵守相应的合规义务。[5]

[1] 2023年3月，加拿大宣布将制定《人工智能和数据法》（*Artificial Intelligence and Data Act*）和配套法规，表示将沿用欧盟以风险为导向的人工智能监管路径，并将监管重点放在高风险的人工智能系统上。

[2] 2019年4月，美国两位民主党参议员布克（Cory Booker）和怀登（Ron Wyden）联合提出《2019算法问责制法案》（*Algorithmic Accountability Act of 2019*），明确了高风险自动决策系统的具体场景，并要求企业对所有高风险自动决策系统进行影响评估。

[3] 2024年12月，韩国国会通过的《人工智能发展与信任创建基础法》（*Basic Act on the Development of Artificial Intelligence and the Establishment of Trust*），从人工智能应用领域的角度列举了高风险场景，并就这些场景制定风险管理措施。

[4] De Gregorio G. and Dunn P.,“The European Risk-based Approaches: Connecting Constitutional Dots in the Digital Age”, *Common Market Law Review*, 59(2): 473-500, 2022.

[5] Gellert R.,“The Role of the Risk-based Approach in the General Data Protection Regulation and in the European Commission’s Proposed Artificial Intelligence Act: Business as Usual?” *Journal of Ethics and Legal Technologies*, 3(2): 16-33, 2021.

基于风险规制模式的理论阐释

基于风险规制模式兴起于20世纪80年代至90年代的去管制化时期。该模式以风险为主要考虑因素，通过适当分配资源针对性地开展治理活动。风险社会理论和回应规制理论是基于风险规制模式的理论基石。基于风险规制模式尊重了风险生成和治理的规律，其成功应用的关键在于保证风险评估的科学性、合理性和动态性。

（一）基于风险规制模式的历史渊源

20世纪80－90年代，西方工业社会经历了监管危机，促使社会审视“监管国家”，并出现要求放松管制的强烈呼声，重点关注“过度监管”“法律条文主义”“机制僵化”和“忽略监管成本”等问题。监管机构、监管法规与政策都受到严厉批判，甚至遭到“政治攻击”。[1]从20世纪80年代中期开始，整个欧洲开始反思过度管制带来的成本问题。1992年，英国实施放松管制举措，强调应以证据为基础制定政策，并在政府部门中引入商业风险管理实践。例如，1999年英国发布《现代化政府》白皮书，提议改进政府风险管理方式，要求在评估政策选择时，综合考虑成本、收益和风险。这些行动都突出了风险的重要性，加速促成基于风险的管理思维。可见，基于风险规制模式是在经济理性的意识形态中催生的，英国的各种放松规制、更好规制和现代化政府议程都促进了基于风险规制理论的形成。与此同时，里根（Ronald W. Reagan）政府也开展了类似的“监管解脱”（Relief from Regulation）运动，提倡减少政府对经济活动的干预。全球性的放松管制浪潮引起了“新公共管理”的一系列变革，强化了对绩效和成本的关注，社会也开始重新认识国家的角色，认为国家的作用主要在于监督私人对监管规定的执行。

英国政府在道路安全、职业健康与安全以及核工业等领域建立了较为成熟的基于风险规制体系，比如，英国健康与安全执行局（HSE）于20世纪80年代制定系统化的基于风险规制的工作方法和流程，发布了具有里程碑意义的文件《核电站风险的容忍度》，首次提出以基于风险规制模式规制产业风险。基于风险规

[1] Hutter B. M.,“The Attractions of Risk-based Regulation: Accounting for the Emergence of Risk Ideas in Regulation”, Discussion Paper No. 33, March 2005, https://www.lse.ac.uk/accounting/Assets/CARR/documents/D-P/Disspaper33.pdf[2024-09-27].

制的重点是确定风险概率和影响因素，准确把控监管响应的时机和监管工具的开发，为作出有限监管资源的分配决策提供依据。进入21世纪后，随着信息技术的发展，新科技带来的风险具有高度的不确定性，基于风险规制模式的应用更加广泛。

（二）基于风险规制模式的核心概念

“风险”一词是基于风险规制的核心要素。据吉登斯（Anthony Giddens）的研究，“风险”一词可以追溯至中世纪，当时指进入未知水域的航行，也就是表明早期商业活动中探险、贸易和殖民扩张的结果不确定性。[1]学界对“风险”一词的认知存在一定差异。风险客观说主张，风险可以被测量和计算；风险建构说则认为，风险是由社会公众定义和建构的，具有主观性。[2]基于风险规制的倡导者对“风险”一词的理解秉持了风险客观说，正因为风险能够被预测，所以可以根据风险大小采取对应的规制措施。

布莱克和鲍德温将基于风险规制定义为“根据透明的、系统的和可防御的框架，以资源利用为目标，优先关注最高风险”。[3]基于风险规制是根据风险的影响和概率确定活动的优先顺序，并认为完全规避风险不是最佳选择。因此，基于风险规制模式的重点是根据风险水平分配监管资源。[4]比如欧盟《人工智能法》提出，“在考虑对权利和安全的影响的情况下计算风险”，根据“人工智能系统可能产生的风险强度和范围”调整规则，此处的“风险强度”指风险水平，即风险大小。基于风险规制包括五个核心要素：第一，确定目标，即关注何种风险；第二，确定监管机构的风险偏好，包括能容忍何种类型的风险和何种程度的风险；第三，评估风险及其发生的可能性；第四，监管机构根据风险评估的结果进行风险分级分类；第五，根据评估的风险类型和级别进行资源分配。

（三）基于风险规制模式的理论基石

风险社会理论（Risk Society Theory）最早由德国学者贝克（Ulrich Beck）提出，后经吉登斯和卢曼（Niklas Luhmann）等人的发展，成为现代社会开展风险

[1] 方曼：“风险感知跨学科研究的理论进展与范式变迁——基于心理学视域的解读”，《国外理论动态》，2017年第6期。

[2] 张哲飞：《科技风险规制过程中的行政法问题研究》，武汉：湖北人民出版社，2019年。

[3] Black J. and Baldwin R., “Really Responsive Risk-based Regulation”, *Law and Policy*, 32(2):181-213, 2010.

[4] Macenaite M., “The ‘Riskification’ of European Data Protection Law through A Two-fold Shift”, *European Journal of Risk Regulation*, 8(3): 506-540, 2017.

规制活动的基础性理论。这些思想家关于风险社会的见解反映在基于风险规制模式上。[1]风险社会理论对于指导风险规制活动具有重要的启发。一是，正确认识科技的负面影响。风险社会理论认为，科技既可以创造高度文明的生活环境，也会产生高度复杂的生存风险。对此，需要通过理性判断和治理措施来趋利避害。二是，充分认识风险的双重效应。贝克和吉登斯等学者并未主张完全消除风险，因为风险既会导致损失，也可以创造发展机会，治理风险的目标在于规避损失的同时发挥其促进发展的效应。三是，理性看待风险的客观性和主观性。风险治理是一个由相关利益者共同参与的民主过程，通过对有限资源的高效利用，优先化解突出风险，学会与风险共存。

基于风险规制的另一理论基石是回应规制理论（Responsive Regulation Theory）。“真正成功的规制是让惩罚与说服实现协同”，埃尔斯和布莱斯怀特提出回应规制这一概念。[2]该理论经过演变和发展，衍生出精巧规制（Smart Regulation）、以问题为中心的规制（Problem-centred Regulation）和真正地回应规制（Really Responsive Regulation）。回应规制理论最核心的洞见是执法金字塔，金字塔的底部是劝诫。如果被规制主体不合作，那么就再“拾级而上”，采取更严厉的执法措施，金字塔的顶部则是撤销许可。针对回应规制仅涉及国家与商业主体之间的规制关系这一局限，甘宁汉和格拉博斯基提出精巧规制的理念，将执法金字塔中的主体进行拓展，把一些“半规制主体”（Quasi-Regulators）纳入其中，比如行业组织、公共利益群体和专业机构。这种多主体结构的执法金字塔让国家的身份由直接干预者演变为促进者。[3]

针对回应规制和精巧规制过于注重过程、工具和机构的使用这一局限，斯帕罗提出一套以问题为中心的规制程序，主张要接受风险控制过程的变动性，选择最重要的任务，采取最有效的工具。[4]布莱克提出真正地回应规制这一新模式，强调在设计、适用和开发规制体系时，规制者应当根据其策略而非依据被规制者

[1] Van der Heijden J., “Risk as an Approach to Regulatory Governance: An Evidence Synthesis and Research Agenda”, July 15, 2021, https://doi.org/10.1177/21582440211032202[2024-12-11].

[2] Ayres I. and Braithwaite J., *Responsive Regulation: Transcending the Deregulation Debate*, Oxford: Oxford University Press, 1995.

[3] Gunningham N. and Grabosky P., *Smart Regulation: Designing Environment Policy*, Oxford: Oxford University Press, 1998.

[4] Sparrow M. K., *The Regulatory Craft: Controlling Risks, Solving Problems and Managing Compliance*, Washington: Brookings Institution Press, 2000.

的行为进行调整。[1]从回应规制理论的发展演变可知，开展风险规制时应当对不同主体匹配差异化的规制措施，并注重规制过程的动态性和适应性。[2]欧盟《人工智能法》设计了人工智能风险规制金字塔，重视企业自我合规，此规则设计明显打上了回应规制理论的烙印。[3]在中国，《生成式人工智能服务管理暂行办法》提出对生成式人工智能服务实行包容审慎监管，实际上也是回应规制理论的体现。[4]

（四）基于风险规制模式的制度价值

基于风险规制的价值包括以下几个方面。

首先，实现了“科学的政治化”向“政治的科学化”的转变。通常，监管者根据专家的意见进行风险评估和制定政策，但基于风险规制模式没有将科学的作用神圣化，而是要求公众共同参与。

其次，恰当采用预防原则（Precautionary Principle）。采用预防原则指在科学证据不充分的情况下，采取较严格的规制措施，保证风险规制的前瞻性、全面性和有效性。基于风险规制并非机械地采用预防原则，因为过于严苛的风险预防原则会阻碍新技术发展，适度的风险预防原则重视平衡安全和创新的关系。适度的风险预防原则要求规制者在风险评估时做到精准，建立反向证明机制和动态的措施调整机制。[5]

最后，促进监管资源的有效部署和分配。基于风险规制模式提倡牵住“牛鼻子”，通过合理的决策工具优化监管流程，提高风险规制的精准度。

基于风险规制模式适用于中国人工智能风险治理的逻辑阐释

为克服机械执行风险预防原则带来的治理僵化，弥补基于权利规制带来的保护不足的问题，以及平衡新技术发展与安全治理，基于风险规制成为构建中国生成式人工智能风险治理体系的重要理论参照。

---

[1] Baldwin R. and Black J.,“Really Responsive Regulation”, *Modern Law Review*, 71(1): 59–94, 2008.

[2] 曾雄、张辉：“规制理论视野下的人工智能治理模式比较及启示：基于英美治理实践的观察”，《中国科技论坛》，2025年第2期。

[3] 曾雄、梁正、张辉：“欧盟人工智能的规制路径及其对我国的启示——以《人工智能法案》为分析对象”，《电子政务》，2022年第9期。

[4] 宋华琳：“人工智能立法中的规制结构设计”，《华东政法大学学报》，2024年第5期。

[5] 苏宇：“风险预防原则的结构化阐释”，《法学研究》，2021年第1期。

（一）中国生成式人工智能法律治理现状

目前中国暂无统一的人工智能基础性立法，主要通过一些部门规章和标准对生成式人工智能予以规制，包括《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法（征求意见稿）》等。这些规定涉及一些监管要求。

第一，内容过滤义务。《互联网信息服务管理办法》规定了互联网信息服务提供者不得制作、复制、发布、传播含有散布谣言、扰乱社会秩序、破坏社会稳定等内容的信息。鉴于生成式人工智能可能生成、传播违规内容，应当对其进行内容过滤。第二，安全评估义务。《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》要求具有舆论属性或社会动员能力的信息服务上线以及信息服务增设相关功能等，并明确互联网信息服务提供者应当自行开展安全评估，并向网信部门和公安机关提交报告。网络平台提供的生成式人工智能服务若涉及舆论属性或社会动员能力，需要申请通过安全评估。第三，备案和标识义务。《互联网信息服务深度合成管理规定》要求提供智能对话、合成人声、人脸生成、沉浸式拟真场景等生成或显著改变信息内容功能的服务的，应当进行显著标识。具有舆论属性或社会动员能力的深度合成服务提供者，应当履行备案手续。第四，科技伦理审查义务。《科技伦理审查办法（试行）》要求，具有舆论属性、社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发需要进行伦理审查和复核。

（二）中国生成式人工智能风险治理面临的挑战

与传统人工智能技术不同，生成式人工智能应用具有动态性，由于使用门槛较低，使用规模巨大，[1]给风险治理活动带来诸多挑战。

其一，生成式人工智能是一种通用技术，被广泛地应用于不同场景，涉及医疗、金融、教育和电商等领域，牵涉多方主体，包括行政机构、社会组织、企业和个人等。如果缺乏统一的风险治理框架和规则，行业规制和专业规制彼此重叠和交叉，那么松散的规则之间可能叠床架屋，导致治理成本高，治理效率低下。比如，《生成式人工智能服务管理暂行办法》就与《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》中关于生成合成类算法的规则存在诸多交叉重叠之处，未形成统一的规范框架。

[1] Helberger N. and Diakopoulos N.，“ChatGPT and the AI Act”, February 16, 2023, https://policyreview.info/essay/chatgpt-and-ai-act[2025-01-19].

其二，由于缺乏专业和统一的治理机构，不同治理主体在风险规制过程中的角色和责任定位模糊，规制主体的分工协作和责任划分缺乏系统性考虑。

其三，现有的规制手段和方法相对传统和单一，治理思维偏向于沿用互联网内容管理的思路，主要依赖事前审批和事后惩罚，对新技术治理工具的设计和应用不足，比如监管沙箱、合规治理、隐私设计和伦理设计等安全措施未受到充分重视和应用。

其四，大模型决策的过程不透明，难以追溯风险成因，而且大模型产业链条长，责任主体多元，风险防范难度大。传统的风险治理框架，对未知的“黑天鹅”风险预测“乏力”，治理效果评估和反思缺位，没有匹配全生命周期的监测和跟踪机制，以这种方式进行风险规制属于线性的单向规制。此类静态规制模式与具有动态创新特征的生成式人工智能技术之间并不适配。

（三）基于风险规制模式与中国人工智能风险治理的适配性分析

从理论层面看，以基于风险规制模式治理新技术风险具有有效性。1980年，英国学者科林格里奇（David Collingridge）提出了“科林格里奇困境”（Collingridge’s Dilemma）这一命题，[1]指在新技术发展初期，其社会影响很难被发现，当人类有能力识别不利影响时，由于技术的广泛应用，控制技术就变得异常困难。生成式人工智能风险治理也面临此类困境，与传统社会风险相比，其风险具有特殊性，具体表现为：一是风险的突发性加剧技术治理的“步调难题”；二是风险的规模化加剧组织监管的困境；三是风险的动态化增加监管问责的难度。[2]

生成式人工智能风险治理的关键在于寻求技术创新与安全发展的最佳平衡，而传统的风险治理思路侧重于机械地执行风险预防原则，对产业发展和技术创新带来负面效应。基于风险规制的核心在于提高规制的理性化，[3]这是对传统风险规制模式的范式超越。遵循基于风险规制所倡导的差异化规制方针，可以避免过度监管和抑制产业创新发展，提高风险治理的有效性。从实践层面看，基于风险规制在中国数字治理领域已有成功运用的案例。《中华人民共和国个人信息保护法》第11条提出风险预防的要求，第55、56条详细规定了个人信息保护影响评估的适用情形和评估内容。总之，生成式人工智能风险治理需要灵活和敏捷的治

[1] Colingridge D., *The Social Control of Technology*, London: Frances Pinter, 1980.

[2] 曾雄、梁正、张辉：“人工智能软法治理的优化进路：由软法先行到软法与硬法协同”，《电子政务》，2024年第6期。

[3] 刘鹏、张嵛楠、王力：“基于风险的政府监管：理论发展与实践应用”，《中国行政管理》，2024年第3期。

理措施，采用基于风险规制模式能实现规范与发展并进这一重要目标。

基于风险规制模式构建中国人工智能风险治理体系

在基于风险规制模式中，风险评估、风险管理和风险沟通是核心内容，风险分级分类是实施路径，统一专业的风险管理机构是组织保障。

（一）中国人工智能风险治理的“三环节”框架及其制度构成

国际风险管理理事会（International Risk Governance Council，IRGC）提出过一个风险治理框架，具体包括以下步骤。第一步，预评估、识别和构建（Pre-Assessment， Identification and Framing）：建构风险，提早预警；识别和引入相关参与者和利益群体。第二步，在技术和感知方面评估风险的原因和结果。第三步，评估是否存在管理风险。第四步，执行风险管理选项。第五步，风险沟通。[1]风险规制体系通常由三个主要部分构成：其一，技术部分是风险评估；其二，政策导向部分是风险管理；其三，信息传导部分是风险沟通。[2]美国发布的《联邦政府风险评估》报告也提出风险评估与风险管理的二分框架，其将风险评估确定为“科学”的过程。[3]

综合参考这些风险管理框架，本文提出人工智能风险治理的“三环节”框架，具体包括风险评估、风险沟通和风险管理，如图1所示。这三个环节彼此之间紧密联系和高度协调，其中风险沟通环节贯穿了风险规制的始终。各个环节循环往复，形成一个动态过程。可将人工智能风险治理的“三环节”框架应用于生成式人工智能风险治理：首先，在风险评估环节，建立生成式人工智能风险监测和风险评估制度，要求定期检测大模型，针对安全漏洞、危害类型和风险程度以及相关技术指标等进行综合评估。其次，在风险管理环节，建立安全管理制度、风险源追溯制度以及产品召回制度，要求企业搭建内部风险管理体系，开展内部合规和培训工作，定期接受监管机构的抽查和第三方审计，及时召回存在安全隐患的人工智能产品。最后，在风险沟通环节，建立风险信息平台，及时公布风险监测数据、风险评估报告以及风险管理的抽查情况，接受社会公众的监督。

---

[1] IRGC. “Risk Governance: Towards An Integrative Approach”, September 2005, https://irgc.org/wp-content/uploads/2018/09/IRGC_WP_No_1_Risk_Governance__reprinted_version_3.pdf[2024-09-27].

[2] 黄新华：“风险规制研究：构建社会风险治理的知识体系”，《行政论坛》，2016年第2期。

[3] 金自宁：“美国联邦环境风险评估制度：已有探索与未决难题”，《国外社会科学》，2022年第3期。

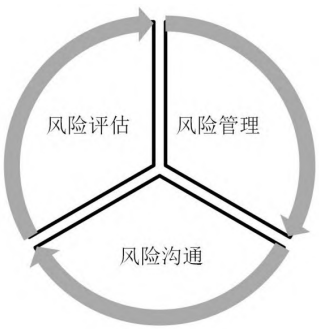


图1 人工智能风险治理框架图

资料来源：根据IRGC的风险治理框架进行提炼和整理。

1.生成式人工智能风险评估

风险评估指判断特定危险发生的可能性及其对人的健康和环境造成的影响程度，风险评估由风险识别、风险分析和风险评价三项内容构成。[1]

首先，风险识别常运用专家访谈法、案例分析法以及问卷调查法，可以通过风险分类的方式建立风险类型库，并在具体应用场景中，灵活选取风险识别的方法，准确界定风险类型。在风险类型库中，大致可分为技术风险和应用风险，其中技术风险包括算法类型风险和数据类型风险；应用风险包括侵犯个人权利、损害个人身体健康和扰乱社会公共秩序等情况。

其次，在风险分析中，常运用定量分析法或定性分析法来界定风险级别，计算公式为风险等级（R）=风险概率（P）*风险严重度（L）。风险概率可以根据历史数据搭建风险模型来测算，评估维度则有大模型参数规模、安全措施完善度、数据集质量等方面；风险严重程度可以根据用户规模、财产损失数额和对社会影响范围等维度评定，比如美国加利福尼亚州的《前沿人工智能模型安全创新法案》以“大规模人员伤亡”或“五亿美元”以上的损失作为“重大损害”的认定标准。

最后，风险评价是确定哪些风险可接受、哪些风险要消减和哪些风险要规避，比如，对于大模型“幻觉”引起的个人侵权问题，可以视为可接受风险，通过侵权规则或自我合规等事后措施进行处理；涉及恐怖主义、国家安全、意识形态安全等问题，应该视为高风险，需要有效规避，并应采取事先评估、价值对齐

[1]江书军：《基于风险的特种设备使用单位安全分类监管研究》，北京：中国时代经济出版社，2014年。

和内容过滤等事前规制措施。

生成式人工智能风险涉及不可度量的不确定性问题，[1]而且风险程度受到社会公众的主观态度的影响，这对风险管理决策机制带来挑战，甚至导致相关机构可能忽视对规制收益与成本的考量而采取“一刀切”的禁止措施。为此，应采取措施保证生成式人工智能风险评估的科学性和独立性，宜确立专门的专家咨询制度，比如设立国家人工智能风险治理委员会，吸纳人工智能技术、法律和伦理等领域的专家，拟定统一的风险评估方法、维度和标准。为明确风险评估的范围，应当构建风险评估的指标体系。在风险评估维度和标准方面，应包括干预政治、舆论控制、社会歧视、隐私泄露等政治与社会风险维度，以及模型攻击、数据投毒、算法毒性等技术风险维度。

然而，风险评估本身存在科学的限度问题。首先，大模型的科学原理和技术路径本身存在局限，而公众对技术的认知并不全面。其次，风险评估会受政治和价值判断的影响，比如科技专家对知识的“垄断性”容易掩盖公众所关注的风险的社会价值意涵，加剧专家风险判断与公众风险感知的对立，一定程度上降低了公众参与的意愿，并阻断了公众参与的技术路径。[2]最后，中国专业化的风险评估队伍建设、智能化的信息搜集整合机制以及现代化的风险分析技术较为薄弱，而且风险评估的独立性和中立性较为欠缺。[3]因此，在人工智能风险评估程序的设计上要确保中立性和独立性，在设立专家委员会时应平衡评估团体的知识结构，从而保证专家结论的科学性。

2. 生成式人工智能风险管理

风险管理指设置不同管理机制，有效管控已经预判和评估的风险。经风险评估后，围绕生成式大模型建立“上市前–上市后”的全链条风险控制机制，具体如表1所示。对于高风险类别，上市前需要采取以下举措。

对于政府而言，首先，应当根据质量标准进行筛查，只有质量合格的产品才能流向市场，质量标准包括生成内容的准确性、创新性、可理解性和用户需求的满意度等维度。其次，建设综合备案系统，对大模型的参数、开发者、性能指标、算法类型等进行备案登记，便于风险追踪和管理。最后，构建安全测评体系，参与国际标准制定和指导国家标准构建，制定风险评估的指标目录。

[1] [美]奈特著，安佳译：《风险、不确定性与利润》，北京：商务印书馆，2006年。

[2] 毛宝铭、许志晋：“科技风险民主治理的基本观念与原则”，《理论与改革》，2006年第2期。

[3] 刘鹏：“从行政管控走向风险治理：中国风险应对体系建设的发展历程与逻辑”，《政治学研究》，2021年第6期。

对于企业而言，首先，开展伦理审查。其次，建立全生命周期的内部风险控制体系，对数据、算法和设施等方面进行检查和测试，以减少风险源。为了实现问责，还应明确董事会是企业风险管理的直接责任人。[1]最后，生成式人工智能风险的技术属性较强，技术治理可以作为重要的风险管理措施，例如，要求企业采取价值对齐，保证生成内容符合人类价值观。[2]

上市之后则需要采取以下措施。对于政府而言，首先，建立审计机制，监督企业内部风险管控情况；其次，基于备案系统进行安全状态跟踪；最后，定期更新国家标准。对于企业而言，首先，定期开展伦理宣传和指导；其次，实时更新风险信息和优化技术治理措施；最后，动态调整内部治理组织，提高内部治理能力。

表1 生成式人工智能风险管理措施

	上市前	上市后
政府	（1）市场准入控制：制定质量标准和认证机制	（1）定期审计和检查
	（2）综合备案系统建设：大模型、推荐算法等方面进行备案	（2）基于备案系统进行安全状态跟踪
	（3）构建安全测评体系：参与国际标准制定，指导构建国家标准、行业标准、安全测评标准、安全测试集	（3）定期更新国家标准，形成周期性安全测试机制和流程，定期更新大模型安全测试集
企业	（1）风险预防：开展伦理审查	（1）定期伦理宣传指导
	（2）风险预防和控制：开展风险自评估，综合采取技术治理和制度治理措施	（2）实时更新风险信息，采用标识技术、内容过滤技术和加密技术等
	（3）人工智能治理体系建设：组建治理部门，设置大模型安全治理岗位	（3）推进安全治理人员与开发人员的适时转岗

资料来源：作者自制。

3. 生成式人工智能风险沟通

风险沟通是关于风险本质、影响和控制等信息的意见交换过程，该环节关注风险的性质和影响因素，涉及对风险评估结果的解释。[3]首先，风险沟通要解决

[1] 文亚："德国公共风险管理的经验与启示"，《中国行政管理》，2015年第4期。

[2] 曾雄："人工智能大模型价值对齐的现状考察、问题检视与规范进路"，《电子政务》，2025年第2期。

[3] 黄新华："风险规制研究：构建社会风险治理的知识体系"，《行政论坛》，2016年第2期。

信任问题。科技专家与公众对风险认知存在差异，[1]需要降低专家、政府与公众之间的信息不对称，避免公众产生不必要的恐慌。其次，风险沟通要解决信息准确传递的问题。目前，科技专家对人工智能风险的认知呈现两极化的现象，部分专家认为风险不足为惧，部分专家认为存在失控风险。根据“风险的社会放大效应”理论，风险事件通常会有“涟漪效应”，公众容易受到个别专家言论的影响而难以理性判断。[2]这就要求政府搭建畅通的信息沟通机制，保证风险信息的准确传递。最后，风险沟通要有多元主体的民主参与。民主参与的实质是通过沟通来管理风险，[3]风险沟通需要政府和利益相关者共同参与，政府机构承担信息发布者的角色，比如通过发布会、专家研讨会和通识课等方式实现风险对话，形成信息交换机制。为此，政府需要建设信息资源共享平台，及时发布风险信息，并在利益相关者辩论时协调彼此的冲突。[4]

在中国，风险沟通是薄弱环节，公众和监管机构都对生成式人工智能的风险认识不足，学界和媒体界对风险介绍的科学性和严谨性也有所欠缺。风险沟通不局限于单向度的风险信息发布，[5]而需要制定《生成式人工智能风险沟通指引》，建设风险沟通平台，并发起交流会议和公开辩论。为了避免风险放大或“异化”，风险治理机构应该针对热点风险事件发声，并发布权威的风险评估报告。

（二）以分级分类治理作为人工智能风险治理的实施路径

1. 分级分类方法的理论基础

分级分类治理的本质是分类管理和重点管理，[6]该方法强调监管的回应性，其本质是要求基于监管对象不同的动机采取不同的措施。[7]分级分类治理属于风险牵引式的监管资源配置模式，是现代政府从注重事前监管转向更加注重事中事

[1] [英]谢尔顿·克里姆斯基、[英]多米尼克·戈尔丁编著，徐元玲、孟毓焕、徐玲译：《风险的社会理论学说》，北京：北京出版社，2005年。

[2] [英]尼克·皮金、[美]罗杰·E. 卡斯帕森、[美]保罗·斯洛维奇编著，谭宏凯译：《风险的社会放大》，北京：中国劳动社会保障出版社，2010年。

[3] [英]芭芭拉·亚当、[英]乌尔里希·贝克、[英]约斯特·房·龙编著，赵延东、马缨译：《风险社会及其超越社会理论的关键议题》，北京：北京出版社，2005年。

[4] 朱德米、林昕：“建设弹性的社会治理体系”，《上海行政学院学报》，2017年第5期。

[5] 石晶：“生命科技风险的宪法回应”，《法学》，2023年第2期。

[6] 张忠彬、孙庆云：“我国职业危害分类与分级监管法规标准与研究现状”，《中国安全生产科学技术》，2007年第6期。

[7] 杨炳霖：“回应性监管理论述评：精髓与问题”，《中国行政管理》，2017年第4期。

后监管阶段必须具备的一种核心能力。[1]分级分类治理的理论依据包括：其一，经济学理论中的成本收益法则，即根据经济学的效用理论，在有限的监管资源条件下，根据监管难易程度或风险大小，分配监管资源，将监管效益最大化。其二，技术理论中的事故致因理论，即可以通过相关理论认识事故的发生原理，及时采取防范措施，预防事故发生。[2]

在国内，分级分类治理的方法已经应用于信用监管、食品监管、环境风险监管、生物安全监管和医疗器械监管等领域。在信用监管中，监管机构对信用好和风险低的市场主体，降低抽查比例和频次。对违法失信和风险高的市场主体，提高抽查比例和频次，列入重点信用监管范围。从上述实践看，实现分级分类的关键是先确定风险等级的划分依据和标准，然后设计与风险程度相匹配的差异化管理措施。比如在生物安全监管中，监管机构根据生物技术研究和开发活动对公众健康、工业农业和生态环境等方面造成危害的程度，将风险划定为高、中和低三个等级。

2.欧美人工智能风险分级分类的实践与启示

欧盟《人工智能法》以健康、安全和基本权利作为评判标准，将高风险的人工智能系统限于对人的健康、安全和基本权利造成重大有害影响这一范围。基本权利包括人的尊严权、言论和信息自由权、不受歧视权以及受教育权等。从分类来看，其中一类别是人工智能系统作为产品的安全组件，或者本身作为产品应用。此类产品包括机械、玩具、电梯以及用于潜在爆炸性气体环境的设备和保护系统、无线电设备、压力设备、娱乐船设备、索道装置、燃烧气体燃料的设备、医疗器械和体外诊断医疗器械，因为这些产品容易产生物理性危害。另一类是对人的基本权利有影响的技术类型或应用场景，技术类型包括远程生物识别系统、用于生物分类的人工智能系统以及用于情感识别的人工智能系统。应用场景包括教育和职业培训场景；就业、工人管理和自雇职业场景；获得和享受基本私人服务、基本公共服务和福利场景；执法部门用于评估自然人犯罪等场景；移民、庇护和边境控制管理场景；司法和民主程序场景。

美国《2019年算法问责法案》界定的高风险自动决策系统的场景包括：对用户个人隐私或安全构成重大风险；导致或促成不准确、不公平、有偏见或歧视的

[1] 孙志建：“怎样合理配置有限的政府监管资源：基于风险的监管模式的兴起及其潜在运行风险”，《上海行政学院学报》，2022年第2期。

[2] 游鹏飞、寇冠华：“浅析墨菲定律及海因里希法则对控制事故的作用”，《安全、健康和环境》，2008年第8期。

决定；在人们生活的敏感领域（比如工作表现、经济状况、健康、个人偏好、兴趣、行为、位置或动作）作出决策或辅助人类决策；涉及大量用户的个人信息（比如种族、政治观点、工会资格和性取向）；系统地监控大型、可公开访问的物理空间。[1]2023年3月，美国商会技术合作中心（Chamber of Commerce Technology Engagement Center）发布《人工智能委员会报告》，该文件构建了基于风险的监管框架，将高风险类别明确为法律权利、安全性、自由以及获得住房、教育、就业和医疗保健的权利等领域。[2]2024年5月8日，美国科罗拉多州通过了《科罗拉多州人工智能法》（*Colorado Artificial Intelligence Act*），其核心是反歧视，重点关注高风险人工智能系统的使用，具体应用领域包括：教育入学或教育机会、就业或就业机会、金融或借贷服务、政府基本服务、医疗保健服务、住房服务、保险或法律服务。[3]

风险分级分类的意义在于，如果规制机构能识别不同领域和不同环节中人工智能应用风险的程度差异，那么就可以选取和实施相应的规制工具，使规制措施与被规制风险相匹配。[4]综合分析欧美的实践做法，通常将人工智能风险分为两大类：一类是涉及物理层面安全的风险，可称其为基础设施风险或产品类风险，比如人工智能作为组件或者作为主要的支撑系统应用于关键基础设施，包括水、电、核和交通等领域，对国家安全或社会安全有重大影响。另一类是涉及公民权利的风险，即权利类风险，比如侵犯个人隐私权、受教育权和平等就业权。对于产品类风险，欧盟采用上市前的合规评估措施，依赖现有的产品责任监管体系。对于权利类风险，欧盟要求企业进行内部控制，强调企业自我合规。欧盟采取此种分类方法，考虑了两种风险的性质差异，因为产品类风险通常具有可计算性，可以根据过往经验判断产品类风险的伤害程度和发生可能性，[5]而权利类风险带有主观偏见，损害程度难以量化，具有不可通约性。

---

[1] “Algorithmic Accountability Act of 2019”, April 2019, https://www.congress.gov/bill/116th-congress/house-bill/2231/text[2024-09-27].

[2] U.S. Chamber of Commerce Technology Engagement Center. “Artificial Intelligence Report”, March 2023, https://www.uschamber.com/assets/documents/CTEC_AICommission2023_Report_v5.pdf[2024-09-27].

[3] “Consumer Protections for Artificial Intelligence”, May 2024, https://leg.colorado.gov/bills/sb24-205 [2024-09-27].

[4] 薛澜、贾开、赵静：“人工智能敏捷治理实践：分类监管思路与政策工具箱构建”，《中国行政管理》，2024年第3期。

[5] Almada M. and Petit N., “The EU AI Act: a medley of product safety and fundamental rights?” Robert Schuman Centre for Advanced Studies Research Paper No. 2023/59, December 30, 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4308072[2024-09-27].

3. 中国生成式人工智能风险分级分类的思路与设想

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，[1]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。如上文所述，先构建生成式人工智能风险类型库（见表2），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2 生成式人工智能风险类型库

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（Bias in Training Data）、数据投毒（Data Poisoning）、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]。

在风险分级方面，可将风险分为不可接受风险、高风险、中风险和低风险四个等级。对于不可接受风险场景，应该严格禁止开发和应用；对于高风险场景，

[1] 根据OECD的分类，人工智能对人的危害表现为：首先，对个人，即对一个人的公民自由、权利、身体或心理安全、经济机会的伤害；其次，对群体或社区的伤害，比如对人口子群体的歧视；最后，对社会的危害，比如危害民主参与或教育机会。人工智能对生态系统的危害表现为：对相互关联和相互依存的要素和资源造成损害；危害全球金融体系、供应链或相关系统；危害自然资源、环境和地球。参见OECD，“OECD Framework for the Classification of AI Systems”，February 2022, https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en[2024-09-27].

应该制定严格的规范和透明度要求；对于中风险场景，可以发挥企业自我规制的作用，侧重使用软法和行业最佳实践；对于低风险场景，可以采用事后的责任救济制度。结合域外的高风险场景和国内实际情况，列出高风险清单。

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3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3 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。

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。

（三）风险治理视角下中国人工智能风险管理机构的科学设置

从学理层面看，组织机构设置的科学性直接影响人工智能风险治理的有效性。如前文所述，目前中国人工智能治理格局呈现分散化和碎片化的条块分割管理的特征，监管权限的分散虽然能发挥各个行业主管部门的专业优势，但会产生监管尺度不一、认定标准不一和协作成本高等问题。[1]人工智能风险的评估具有不确定性和不可预期性，容易造成行政不作为或运动式监管。因此，需要有一个统一的风险管理机构实现“一站式监管”，统一风险管理规则。集中化的行政组织通过创设“风险议程”，有助于合理确定规制的优先次序，更好地配置规制资源，并有利于汲取科学知识而不受政治与民意的过度干扰。[2]

从历史经验看，是否有统一的中央监管机构可能会影响公众对新技术的包容度，甚至决定新技术的适用广度。以转基因技术为例，欧洲和美国社会对该技术的态度迥异，欧洲民众对该项新技术普遍表示担忧，而美国有77%的玉米种植采用了转基因技术。出现这种差异的一个重要原因在于，监管机构在技术起步阶段采取了不同措施。在美国，转基因技术开发后不久，美国食品药品管理局就推动包括科学家、监管机构、农民和环保主义者在内的各方利益相关者进行讨论。正如学者指出，“欧洲缺乏技术许可及减轻公众恐慌的中央监管机构，而生物技术被视为一项需要创新性监管措施的新技术……”。[3]在个人信息保护领域，也有实证研究表明，在拥有独立数据保护机构的国家，个人信息保护的效果可能会更好。[4]再以人脸识别技术为例。在美国，由于缺乏统一和权威的风险管理机构及时进行风险评估和采取管理措施，公众对该技术产生了严重的恐慌情绪和极大的抵触心理。中国也有类似情况，由于缺乏统一的监管机构进行风险评估和制定安全标准，很多住宅小区安装人脸识别设备后，遭到公众质疑，导致很多设备闲置。如果有统一的机构提前进行风险评估和风险沟通，及时制定安全标准，便可在风险可控的情况下使用新技术。[5]由此可见，如果缺少一个统一权威的风险管理机构及时实施风险评估、风险管理和风险沟通，一项新技术可能无法得到有效

[1] 刘权：“风险治理视角下的个人信息保护路径”，《比较法研究》，2024年第2期。

[2] [美]史蒂芬·布雷耶著，宋华琳译：《打破恶性循环：政府如何有效规制风险》，北京：法律出版社，2009年。

[3] [英]雅各布·特纳著，朱体正译：《机器人现代法则：如何掌控人工智能》，北京：中国人民大学出版社，2023年。

[4] 章宁、田力：“个人信息保护制度的比较研究”，《情报科学》，2020年第7期。

[5] 曾雄、梁正、张辉：“人脸识别治理的国际经验与中国策略”，《电子政务》，2021年第9期。

的运用，甚至因公众态度的转变而遭到彻底摒弃，进而导致资源浪费，影响技术创新。

综上，成立一个专门的人工智能风险管理机构是构建中国人工智能风险治理体系的重要组成部分。通过明确具有中心地位的组织机构，加强部门间协作，防止人工智能政策的制定和实施出现偏差。2023年中国成立了国家数据局，表明新设机构在实践上具有可行性。对此，可以参照国家数据局、国家金融监督管理总局、国家市场监督管理总局等部门机构改革的经验，将目前国家部委中涉及人工智能风险治理工作的机构或人员进行整合，并适当新增编制人员，从而弥补现有队伍在专业知识结构上的短板。一旦国家层面设立了人工智能风险治理机构，各个企业就需要建立相应的内部治理组织，以配合国家治理机构的要求和履行相关义务。这样将产生显著的治理“协同效应”，一方面实现监管要求的快速落地，另一方面调动企业自我监督的积极性。这体现了政府规制向自我规制的映射，有助于提高被规制者对规制目标的敏锐度。[1]

结 语

人工智能作为一种新兴技术，其内涵和外延正在不断丰富和演化，不断有相对成熟的技术应用涌现，从当年火爆一时的人脸识别，到引人注目的自动驾驶，再到如日中天的大模型，这些新技术应用显著提高生产效率的同时，也对个人权益和社会安全带来负面影响。因此，人工智能风险治理成为当今社会治理的主要议题之一。

本文对基于风险规制模式进行理论阐释，分析了西方“去管制化”运动下基于风险规制的产生和演变历程，揭示了基于风险规制的两大理论基石——风险社会理论和回应规制理论，提炼了基于风险规制实现精准治理、前瞻治理和有效治理的制度价值。本文还以生成式人工智能风险治理为例，论述基于风险规制模式与中国人工智能风险治理之间的适配逻辑，并结合基于权利规制模式面临权利救济“失能”困境与中国人工智能风险治理缺乏统一框架这一现状，为中国人工智能立法和风险治理顶层框架设计提供参考。

首先，应当构建一个包括风险评估、风险管理和风险沟通三个环节的框架流

[1] [英]伊丽莎白·费雪编著，刘刚译：《风险规制：德国的理论与实践》，北京：法律出版社，2012年。

程，以实现风险管控的集中化，避免治理举措的分散化和碎片化。其次，坚持风险分级分类。在风险类型上，建立人工智能风险类型库，并根据新技术的演变情况及时更新；在风险等级上，制定明确的风险评级标准，开发相应的风险评级工具和模型，将风险划分为不可接受风险、高风险、中风险和低风险，并为不同等级的风险匹配相应程度的治理措施，实现风险治理的精准性和合比例性。最后，设立统一的人工智能风险治理机构，以实现监管权的集中统一，避免“治理赤字”，降低治理成本。

在理论和实践层面，任何风险治理模式均非尽善尽美，基于风险规制模式亦然。风险概念的界定、风险测算方法的可行性、风险评估结果的准确性以及风险分级分类的全面性，均可能遭受质疑。鉴于人工智能风险治理的紧迫性，业已在多领域成功应用的基于风险规制模式应受到重视。通过“小马过河”式的试错实践推进风险治理，远比犹豫不决或停滞不前更为可取。值得注意的是，基于风险规制模式需要将大量治理工作委托给被监管主体，并聚焦于其内部控制质量上，这被称为“元规制”（Meta Regulation）。[1]元规制的缺陷在于它过度强调被监管主体的内部控制，而这些内部控制的目的与监管者的目标实际上存在偏差。被监管主体首要考虑的是实现自身利益最大化，并可能设法规避监管者的干预，这导致风险管理机制往往流于形式。为防止自我规制变成一份形同虚设的纸质清单，[2]需要采取审计或随机检查等方法，以促使被监管主体切实履行其自我监管职责。[3]由此可见，在人工智能风险治理中，审计制度值得深入研究。

（责任编辑：邱静）

[1] Black J. and Baldwin R.,“Really Responsive Risk-based Regulation”, *Law and Policy*, 32(2):181-213, 2010.

[2] Koops B.,“The Trouble with European Data Protection Law”, *International Data Privacy Law*, 4(4): 250-261, 2014.

[3] Julia B.,“The Emergence of Risk-based Regulation and the New Public Risk Management in the United Kingdom”, *Public Law*, 3:512-549, 2005.