面向风险的人工智能监管进展、创新与启示：基于欧盟视角的观察

王春晓，李怀胜

（中国政法大学　刑事司法学院，北京　100088）

摘　要：随着人工智能技术发展和市场化应用下的风险衍化，人工智能监管产生并独立于人工智能治理，成为化解人工智能风险的基础理念与新兴领域。欧盟首创面向风险的人工智能监管新范式，并借助国际合作、伦理约束与风险监管框架等方式实现欧盟人工智能监管的领导力和行动力。考虑到人工智能技术与社会需求的参差，欧盟优化人工智能监管主体、监管流程以及监管内容以达到人工智能动态风险的有效监管。面对复杂的人工智能风险态势，我国可借鉴欧盟经验，在政策决定层面，基于本土国情制定面向风险的人工智能法律政策与框架体系；在应用实践层面，构建司法机关、市场监督机关以及企业等主体的交互协同机制；在产业结构层面，通过市场赋能推动人工智能监管职能下沉和产业创新发展；在通识教育层面，培育人工智能素养作为防范人工智能风险的监管“安全阀”，实现公正、和谐和创新的未来人工智能发展。

关键词：人工智能监管；欧盟；人工智能法案；人工智能风险

中图分类号：TP18；D912. 1　　**文献标识码：**A　　**DOI：**10. 19358/j. issn. 2097 - 1788. 2024. 11. 015

引用格式：王春晓，李怀胜. 面向风险的人工智能监管进展、创新与启示：基于欧盟视角的观察［J］. 网络安全与数据治理，2024，43(11)：92 - 100.

Risk-oriented artificial intelligence regulation: progress, innovation and inspiration from the European Union perspective

Wang Chunxiao, Li Huaisheng

(School of Criminal Justice, China University of Political Science and Law, Beijing 100088, China)

Abstract: As artificial intelligence technology advances and market application risks evolve, artificial intelligence regulation has emerged and been independent of artificial intelligence governance, which has become the basic concept and emerging field of resolving artificial intelligence risks. The European Union has pioneered a new paradigm of risk-oriented AI regulation, and realized the leadership and action of EU AI regulation by means of international cooperation, ethical constraints and risk regulation framework. Considering the difference between artificial intelligence technology and social needs, the EU should optimize the main body, regulatory process and regulatory content of artificial intelligence to achieve the effective regulation of artificial intelligence risks. Faced with the complex risk situation of artificial intelligence, China can learn from the experience of the European Union, and formulate the risk-oriented legal policy and framework system of artificial intelligence based on local national conditions at the policy decision-making level. At the level of application and practice, an interactive and collaborative mechanism among judicial organs, market supervision organs and enterprises is constructed. At the industrial structure level, the market empowerment was used to promote the sinking of artificial intelligence regulatory functions and the development of industrial innovation. At the level of general education, it fosters artificial intelligence literacy as a Security Valve of regulation to mitigate risks associated with artificial intelligence, and realizes the fair, harmonious and innovative future development of artificial intelligence.

Key words: artificial intelligence regulation; European Union; Artificial Intelligence Act; artificial intelligence risk

0 引言

随着人工智能系统普遍嵌入数字经济产业，其引发的算法偏见、数据隐私、伦理失范等潜在风险也接踵而至。正如联合国教科文组织（United Nations Educational, Scientific and Cultural Organization, UNESCO）在《人工智能伦理建议书》所说，人工智能可以对人类大有裨益并惠及全球，但也可能加剧偏见而导致算法歧视、数字鸿沟和数字互斥，威胁文化、社会和生物多样性并造成社会经济失衡。与此同时，该文件将人工智能监管视为人工智能风险的解决之道，鼓励各国政府采用人工智能监管（Artificial Intelligence Regulation, AI Regulation）以预测后果、减少风险、避免有害后果、促进公民参与和应对社会挑战[1]。在此背景下，欧盟逐步意识到人工智能风险引起的监管需求，构建起人工智能监管范式并创新政策体系和实践举措，逐渐成为世界人工智能监管领域的重要一极。

镜鉴欧盟人工智能监管的先进经验对中国的人工智能监管改革、发展和变革具有重要的意义和作用。本文为明确欧盟人工智能监管的领域界分，首先在理论基础上对人工智能监管与人工智能治理（Artificial Intelligence Governance, AI Governance）进行区分和对比，明晰欧盟面向风险的人工智能监管的范式变革。其次，从欧盟最新出台的《人工智能法案》（下称《法案》）入手，分析欧盟面向风险的人工智能监管理论与实践动向，以展现欧盟在人工智能监管领域的先进举措。最后，结合欧盟人工智能监管理论创新和实践经验，为我国人工智能监管发展提供有益借鉴。

1 欧盟面向风险的人工智能监管概述

1.1 概念区分：人工智能监管（AI Regulation）与人工智能治理（AI Governance）

欧盟面向风险的人工智能监管范式，其基础来自于人工智能监管的兴起。在政治经济学角度，人工智能监管可视为人工智能治理的基础。人工智能监管较人工智能治理出现较晚，但存在与人工智能治理长期混淆的现象。在当前 AI 技术飞速发展与风险动态变化的背景下，深入理解和区分二者，有助于更好明晰欧盟人工智能监管的职能范畴。

早在二十世纪八十年代，人工智能治理就受到了研究关注，起初主要解决 AI 技术在军事领域的应用问题[2]。2015 年之后，有关人工智能治理的研究激增，并扩展到了公共卫生、技术开发等多类领域[3]。目前比较公认的治理概念由 Almeida 提出，将数字时代的治理定位为不同参与者制定政策，并建立正式和非正式规范，构建社会使用的数字基础设施、平台、服务和应用程序所涉及的各层次结构[4]。其目的是使机构、社会和其他利益相关者能够在动态演化的风险环境中共同努力并实现政策目标，而不会对社会造成严重干扰或损害[5]。人工智能监管的出现起源于 AI 技术迭代与市场化下产生的新型风险与旧有风险的叠加。它不仅是对 AI 技术的理解、评估和监测，更是一种促进 AI 创新和实践，激励技术发展和社会进步的知识体系。现多用来代指政府、国际和政府间组织、非国家和个人为规范 AI 系统的开发和部署而制定的规则、流程和决策程序，包括软性准则和硬性法规，前者如 Microsoft 等跨国人工智能企业的内部道德准则，后者如欧盟已经通过的《人工智能法案》[6]。

事实上，人工智能监管可视为人工智能治理的基础。在政治经济学角度下，市场经济领域内 AI 的负面影响和干扰可以被视为市场失灵的一种体现，这一外部效应导致公权力机关在利益相关者的要求下介入，进而将 AI 相应应用和服务作为监管的核心。监管机构通过满足不同利益相关者的期望，实现公共利益与个体利益的平衡，进而构建人工智能治理[7]。

概言之，人工智能监管与人工智能治理在宏观处有着众多相似之处[8-11]：（1）背景源起相似。风险社会下 AI 在社会各领域广泛应用，其本身的先进性、复杂性和动态性引发了前所未有的风险和社会挑战。（2）宏观职能相似。二者均关注 AI 的对人类领域的应用并有着风险最小化的功能取向。一是强调 AI 的发展需考虑人身、环境和公共安全等个人和公共利益等前提要素。二是主张使 AI 技术的发展与人类社会和道德价值观相协调，调整 AI 与人类伦理观的适配程度。（3）约束体系相似。人工智能治理与人工智能监管均存在落地需要，这决定了二者均重视提升 AI 技术标准的内在约束以及培养必要的人工智能技能和素养等外部防护，体系性把控人工智能应用产生的风险与不当后果。

二者在宏观层面较为相似，但在微观领域则界限分明[8-11]：（1）规范范畴不同。人工智能治理的范畴发展具有渐进性，由军事领域的人工智能应用规范逐渐囊括人类群体交互中与 AI 相关的领域中的全部规范和非正式规范。人工智能监管出现较晚，主要以公权力主体施加的政策文件、市场规则、技术标准等正式性规范约束为主要范畴。（2）作用效果不同。人工智能治理整合 AI 风险应对和机遇拓展下的两极导向，与人类群体的伦理观和价值观演进保持一致，指导人工智能政策法规、社会规范等的制定、完善和维护。人工智能监管侧重于微观层面下现有规范、技术架构如何有效适用于 AI 并消除其风险，强调社会对 AI 开发者、部署者、使用者等特定群

体的制裁力，维持最低的道德伦理格局，使其合于规定之下促进 AI 的发展和创新。

1.2 面向 AI 风险：欧盟人工智能监管的范式变革

随着 AI 技术在市场化运用中深入到千家万户、各行各业，带来个人风险的同时，其固有的不确定性结构性风险点可能因为控制不当产生多米诺骨牌效益，甚至对人类共同利益产生威胁[12]。欧盟网络安全机构发布的 2023 年威胁态势报告结果显示，欧盟境内一年内发生 2 580起涉及 AI 的风险事件，对象涉及公共行政部门、公共卫生、制造业等核心领域。

欧盟面向风险的人工智能监管新范式基于传统的人工智能监管理念，吸纳风险社会的特殊背景，特别强调对人工智能应用过程中可能产生的各种风险进行识别、评估、预警和应对。面向风险的人工智能监管要求根据新的情况进行动态调整和优化，不断跟踪和评估新的风险点，及时调整监管策略和措施，确保监管工作的有效性和针对性。欧盟首先通过指引制定各项政策法律文件防范与消除 AI 风险。2016 年，《欧盟机器人民事法律规则》从法律和道德角度深度解析民事领域的 AI 风险，针对性提出未来的欧洲民事领域的 AI 技术财产伤害赔偿的监管原则[13]。随后出台各种规章制度、法律法规等政策文件以强调人工智能监管下的风险应对。2023 年，《法案》历经数次研讨后一致通过，欧盟自此在世界上率先拥有了统一的面向风险的人工智能监管法律法规。

2 欧盟面向风险的人工智能监管进展

2.1 国际协同合作：欧盟竞争 AI 话语权的重要议题

全球化背景下，人工智能监管领域的国际间协同合作趋势也随人工智能风险张力不断加强[14]。欧盟积极组织、参与国际协同合作活动介入全球性 AI 监管体系，参与大国竞逐，表现欧盟对国际 AI 监管的渗透性发展和话语权争夺。总体上，欧盟以区域合作为依托，积极参与构建 AI 监管领域的跨国双边或多边合作机制。在国际双边合作领域，2021 年 9 月，于匹兹堡举行美国 - 欧盟 TTC 首次部长级会议之后，欧盟与美国共同发起了三个针对可信赖 AI 监测评估、技术开发和理论研究的项目，致力欧美两国在人工智能发展上达成深层次共识。在区域合作层面，欧盟境内就 AI 的前沿研究和金融投资等方面进行交流，并提出包括《关于推动人工智能发展的欧洲路径的交流》《欧盟委员会：关于发展欧洲制造的“人工智能”的交流》等在内的建设性合作倡议和交流计划。在多边国际合作领域，欧盟与 28 个国家共同发布了《布莱奇利宣言》，特别关注网络安全和生物技术等关键领域的 AI 风险。这一宣言促使欧盟与各国在建立人工智能监管底线等方面达成共识，所署文件也为人工智能监管的国际化标准奠定基础。此外，欧盟与国际标准化组织和国际电工委员会两个国际标准机构合作制定新型 AI 标准，并研究包括 IEEE 在内的他类 AI 标准，与国际上人工智能标准领域保持互联互通、合作与共。综合来看，欧盟希望利用更广泛、更平等的全球性、区域性、渐进性的合作机制，为其在更广阔范围内主导 AI 监管秩序重构寻找合法性，从而谋求、巩固欧盟在现行人工智能监管体系中的领导地位。

2.2 重视 AI 伦理：宏观原则与具体举措的双重约束

关于人工智能的伦理反思已成为全球范围内的重点议题。面对人工智能技术发展下的伦理困境，目前至少有 84 个公私合作倡议发表了声明，描述了指导 AI 伦理发展、部署和治理的高级原则、价值观和其他原则。其中，UNESCO 制定了《人工智能伦理建议书》，在 2021 年 11 月得到了 193 个成员国的通过和采纳[15]。欧盟同样高度重视 AI 发展的伦理约束，在平衡风险最小化目标与社会公共需求的同时，确保 AI 朝着有益于人类的方向可持续发展。“欧盟标准”下的人工智能伦理观以“人本性”为基本伦理取向，将伦理约束内嵌于《发展欧洲制造的人工智能》《可信赖的人工智能伦理准则》等相关的人工智能监管政策中[16]。如 2018 年《可信赖人工智能道德指引评估清单》中重点阐述了 AI 系统整个生命周期中的三大原则要求，其中伦理原则被描述为确保 AI 的开发进程中能够遵守道德责任和以人为中心的核心价值观。2024 年《法案》则从政策制定者角度重申 AI 应以人为本的价值取向，要求利益相关者遵循《人工智能伦理准则》中的七项伦理原则，并不得妨碍国家或者欧盟法律可能要求的伦理审查活动。此外，欧盟还将人工智能伦理嵌入利益相关者的义务规范与技术标准约束中，进而应对不断迭代的人工智能技术带来的动态应用风险。综合而言，欧盟提出的人工智能伦理观念在理念基础上建立扎实的保障措施，强化了实践层面“人本性”的意识凝聚力和道德约束性，赋予人工智能利益相关者应对人工智能风险的切实行动力和实际效果。

2.3 风险监管框架：实现“去风险化”与基本权利保障

仅靠原则性的理念论述无法对 AI 风险实施有效监管，需要通过合理的风险监管框架促使利益相关者在 AI 生命周期内遵守相关规范，促进实践中 AI 生命周期活动的“去风险化”和个人基本权利保障[17]。国际学术界中不同学者从特定视角出发，认识现有人工智能监管对伦理道德、角色划分以及监督责任的需求，提出了伦理嵌入型框架[18]、责任分配型框架[19]及反馈型框架[20]等学说，对人工智能风险监管框架的分析和发展具有一定的

启发意义。在人工智能监管领域，为确保人工智能的可信赖发展，欧盟于2017年即呼吁通过进一步研究包括监管领域、监管措施以及监管主体等方面，形成AI的有效监管，打造前沿性人工智能监管框架。2018年，欧盟建立人工智能高级专家小组加快建立统一的人工智能监管框架的步伐。2021年欧盟公布《法案》草案，经数次研讨后在两年后的2023年12月通过。该《法案》试图建立面向风险的集成性风险监管框架，将一系列关于政策、技术等的人工智能监管置于专门机关与利益相关者的合作之下，既能在国家层面协调和支持各项政策，以联盟层面的统一监管减轻AI系统产生的累积性风险，满足对基本权利的高度保护，又能在微观上整合AI领域利益相关方的行动，促进人工智能监管的具体落实。

3 欧盟面向风险的人工智能监管创新

3.1 监管主体多元化

人工智能在促进社会经济高速发展与制度性变革的同时，极易产生科技发展与社会需求之间“共生演化”的波动性落差，造成监管失范[21]。为解决这一潜在危机，欧盟积极扩展监管主体范畴，促进横、纵向监管主体的多元交互。（1）纵向层面，欧盟监管机构不再限于人工智能办公室等专门机构，而是基于利益考量、权力分配等因素具体化地分布在各成员国内，包括市场监督机关、通知机关等各类机关组织中，面向人工智能风险形成“联盟－成员国－市场”的同质性利益共识。（2）横向层面，欧盟将人工智能监管的参与主体扩展到社会各层面的可能受AI系统影响的公民个体、企业代表、独立专家等利益相关者，并利用学术科研互动平台促进利益相关者理念认知、技术标准等多层面的“主体间共识”。特别是高级专家组和欧洲人工智能联盟在线论坛平台，前者在AI模型技术标准、测试实践以及基础数据领域促进各主体的研究互通。后者鼓励行业代表、中小企业、初创企业和学术界等利益相关者的咨言献策，现已成为欧洲AI领域具有极大影响力的信息交互基地。

3.2 监管形式链条化

3.2.1 形成事前、事中、事后监管层次

随着人工智能网络化、数智化和动态化的转变趋势，欧盟形成了层次化的监管策略。在事前层面，欧盟主要采取以下策略：（1）制定行为准则。欧盟根据伦理、价值保护和发展情况等因素，在人工智能生命周期体系中探讨、制定人工智能监管政策文件，为AI各主体提供理念引领和行为指导，如《法案》专门专章解决企业间及企业与政府间的信息传递、基本权利评估、市场异议等问题。（2）设立人工智能监管框架。欧盟为人工智能监管建立有效的集成化框架，预先分析和解决AI开发、部署以及使用等流程中的相关问题，加强欧盟成员国间的协同体系建设，促进特定部门或特定领域的产业创新。（3）实施市场准入机制。欧盟设置不可接受风险的AI系统市场限制令，要求高风险AI系统进入欧盟市场，必须获得授权并遵守强制性义务，如部署者将高风险AI系统投放市场之前需进行基本权利影响评估，以便用户就使用AI系统做出合理的风险预期。

相较于详细、严格的事前监管，欧盟还在事中监管层面设立透明度合规义务和测试试验制度以防控风险，包括：（1）制定严密的透明度标准。由于算法黑箱与技术保护等因素影响，AI透明度对于监管极为重要。欧盟要求通用AI系统及其模型遵守发布关于AI训练内容详细摘要等要求以满足特定的透明度合规义务。（2）创立监管沙盒和真实环境测试制度。这两种制度允许在特定保障措施下对AI系统开发、运作的全过程进行实时实地检测，为监管主体提供必要信息以进行有效的风险监督和信息反馈，为创新AI系统的开发、测试和验证建立一个受控的环境，建立公众对人工智能技术的风险度信任。

在事后监管层面，欧盟采取了包括问责制在内的新监管工具增加自发改革动力。（1）开展问责制。欧盟在《法案》中增加了联盟层面的执法权能，通过科层制系统的外部压力惩处AI系统的违规或侵权行为。欧盟成员国的相关部门需根据《法案》规定细化罚款规定，反馈年度罚款情况。（2）形成“互联网+”监管模式。欧盟依托互联网信息系统等监管平台，如欧洲司法系统动态数据库（Council of Europe European Commission for the Efficiency of Justice，CEPEJ），打通线上线下监管之间的枢纽渠道的同时，建立跨部门监管信息的源头追溯、信息共享等监管业务信息支撑模块，提升欧洲境内跨地区、跨部门的监管信息资源利用效能。（3）促进学科性研究与反馈。欧盟通过开展研讨会和研究项目促进官方与行业之间交流和合作，并成为吸引行业参与者与利益相关者、传递AI政策与市场创新经验的桥梁，形成良好的AI发展生态和有序市场环境，如欧洲工程院召开的“全球人工智能发展趋势”研讨会与欧盟“地平线欧洲”计划[23]支持对人工智能生命周期的风险类型、技术透明度等标准的系统性理论思考和实践总结，对欧盟监管机制的构想与落实起到重要作用。

3.2.2 构建明确的风险管理系统

高风险AI技术漏洞可能造成的风险隐患具有极强的破坏性、隐蔽性与流动性，欧盟因此规定要构建与高风险AI系统有关的风险管理系统，实现AI风险最小化。该风险管理系统主要包括风险识别、风险评估、风险应对、

风险监测四个清晰且系统性的架构设计，主要采取风险分级分类的思路，考虑不同应用场景下的核心风险因素并评估其风险的实现可能性与危害程度，实现复杂风险的有效处理（如图 1 所示）。

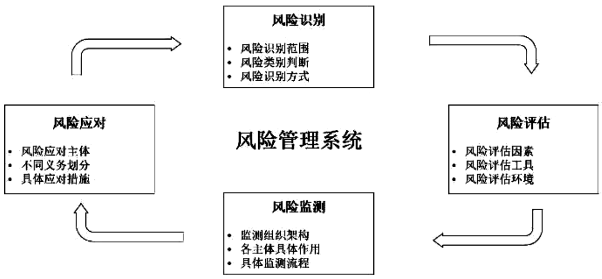


图 1　欧盟风险管理系统

（1）风险识别是用感知、判断或归类的方式对现实的和潜在的风险进行鉴别的过程。区别于美国国家标准与技术研究院以对象为标准将 AI 风险分为对人类的危害、对组织的危害和对生态系统的危害三类，欧盟对于 AI 系统的监管只涉及可以利用技术手段和组织手段减轻或消除的风险。具体而言，欧盟在包括健康、安全以及基本权利（如反歧视）等常规性个人风险的基础上，纳入了民主、法治和环境等新增的公共利益风险。AI 系统的风险从高到低被划分为不可接受的风险、高风险、非高风险（又称最小程度风险）三类等级，并有着逐渐宽松的层次性监管义务要求（如图 2 所示）。监管者一方面可以通过日常经验和事故报告预测与判断风险，另一方面可以利用后市场监测系统等特定的技术措施收集运行数据，实时分析合理预见下的健康、安全等基本权利风险。

（2）风险评估则是对 AI 风险的作用方式和危害程度予以分析，进而确定风险的类别、等级以及应对措施，包括评估因素、评估工具以及评估类型和场所等要素。欧盟规定有关主体需根据在线提供的《可信人工智能评估清单》（Assessment List for Trustworthy AI，ALTAI），考虑 AI 的预期目的、使用程度、损害或不利影响程度等因素，运用特征要素和目标数值衡量 AI 风险，以具体的检查清单和自我评估工具指导 AI 实践。在第三方评估机构、监管沙盒以及真实环境三种风险评估场所评估后，根据高风险、基本风险以及系统性风险三种等级，指定第三方合格性评估、权利影响评估或模型评估，实现各类风险的有效评估。

（3）风险应对是指对潜在的系统风险或危害结果的处理措施。由于不同主体有着对现有 AI 系统开发、应用的个性化目的与差异性激励，有必要对人工智能相关主体做出类型化区分。根据人工智能生命周期，AI 领域相关主体可被划分为提供者、部署者和使用者等类型，此外还包括市场监督机关和其他协作机关等特殊主体。不同风险等级的 AI 系统对人工智能周期的主体类型有着差异化的义务划分和流程化的风险应对措施，以实现最佳风险应对（如图 3 所示），如提供者具有提供透明度信息的义务、对违反条例行为的纠正措施以及严重事故报告制度等，否则就会受到相关机关的调查与问责。

（4）风险监测目的是发现可能发生或已经发生的风险，确保组织实施风险响应措施的持续有效性以及识别影响风险态势变化的充分准确性，而对风险的准确监测离不开高效的组织架构设计[22]。为在组织层面对 AI 市场运行予以统一而全面的指导，欧盟设计了“专门机关监督－市场监督－自然人监测”这一风险监测架构。专门机关主要包含委员会内设立的欧洲人工智能办公室和专家小组。欧盟委员会行政架构下的人工智能办公室专门监测和警示通用 AI 模型和系统实施和应用中不可预知的风险。对通用 AI 模型，则特别设立由独立专家组成的科学小组专门性监测。市场监测由各成员国的市场监督机关进行，不仅负责对市场投放的 AI 系统风险予以监测，且覆盖监测 AI 真实环境测试中的运行风险。此外，由于

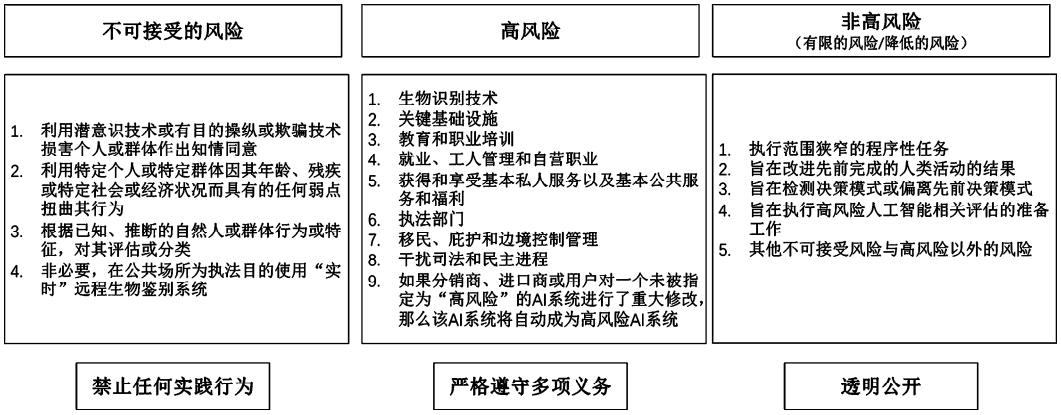


图 2　风险分类及具体内容

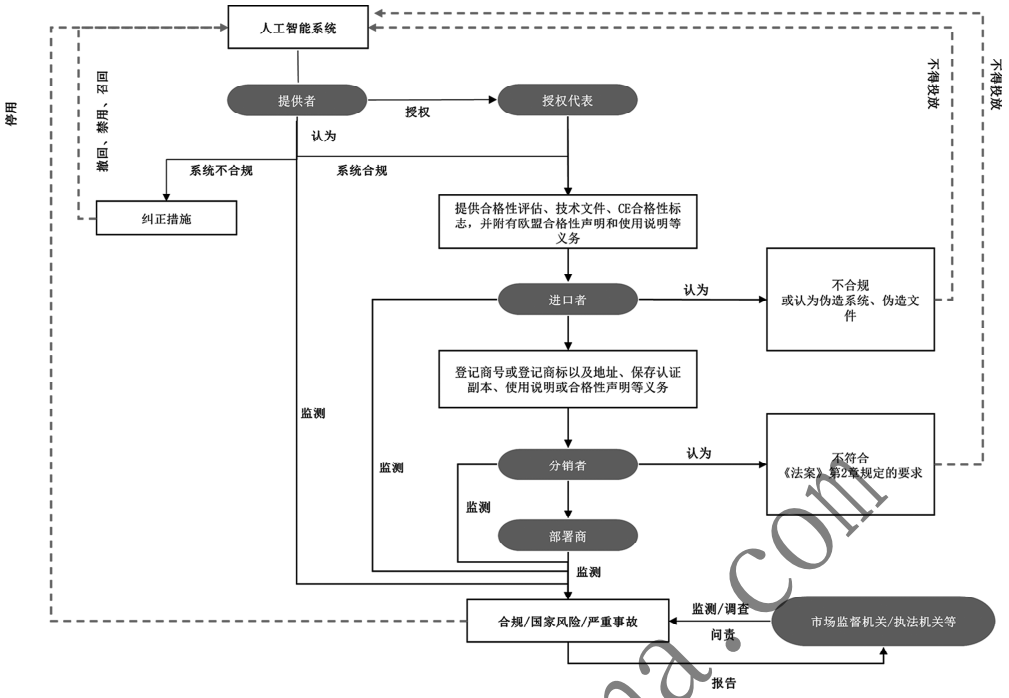


图 3　风险应对流程

国家与市场的监测活动与 AI 系统的实际运行存在着一定的时间差和空间差，欧盟规定 AI 技术人员与使用者有义务进行人为监测，必要时采取技术措施或直接行动防止 AI 系统运行中造成重大风险或严重危害。

3.3　监管内容动态化

由于投放市场后 AI 系统可能会因为使用目的、运行环境等改变造成风险变化，正如欧盟《法案》提案中本没有包含从附件三中删除高风险 AI 系统的规定。但在第三次修改研讨中，如何将系统排除在附件三规定的高风险类别之外并与主管当局建立相应的沟通渠道被重点提及。可以看出，欧盟已意识到风险的动态化并做出积极应对。具体来看，欧盟利用沙箱实验测试、市场参与以及采取灵活而实质性的风险归类方式等工具实现对动态风险的全面监管，后者即是依据 AI 系统投放市场或投入使用后风险的实质性变化而判断 AI 的风险等级，灵活把握 AI 系统的风险情况，实现对动态风险的实效性监管。其中，欧盟重点关注通用型 AI 与跨应用场景型 AI 基础模型，若有关主体对其进行了等同于实质性修改的重大修改，如更改了预期使用目的，该系统将被自动划定为高风险。相对地，若有关高风险 AI 系统不再对基本权利、健康或安全构成任何重大风险，且不会降低欧盟法律对健康、安全和基本权利的总体保护水平，该 AI 系统将被视为降低了风险而剔除出高风险 AI 名单。这一机制丰富了面向风险的欧盟人工智能监管体系，体现欧盟对 AI 动态风险的有益思考和应对举措。

4　欧盟经验对中国的启示

4.1　未来发展：人工智能监管框架与政策完善

越来越多的学者和国家开始呼吁将包含政策、技术等领域的人工智能监管置于单一机构之下，建立集中式的人工智能监管框架[24]。欧盟通过《法案》和各项政策确立了面向风险的人工智能监管，其建立的集中式人工智能监管框架在理想情况下将促进欧盟境内各大成员国在 AI 领域的集体行动。但长久来看，欧盟集成化的“一站式框架”可能会面临各成员国的规则适用差异化问题，同样带有集成性特征的 GDPR 在落地时间延长的过程中逐渐面临着监管机构裁决效率低、罚款较高等难题，跨成员国之间的合作模式因时间长、程序复杂以及各国法律之间的硬性碰撞也饱受诟病[23]。

我国《生成式人工智能服务管理暂行办法》等政策的出台标志着我国已日益重视人工智能的风险应对与日常监管。目前，国内建立统一人工智能立法的呼声日益高涨，但仍处于战略性指引、总体性意见走向具象性法律监管的过渡阶段，针对人工智能各运用场域的政策引领仍具有一定局限性，尚未建立统一的人工智能监管法律与框架体系。“他山之石，可以攻玉”，我国可基于本国国情，借鉴欧盟经验构建面向风险的集成式人工智能监管框架，在原则性指立法指引下兼顾“普适性”与

“特殊性”，推动地方示范性人工智能监管政策建设进而破除“一站式”弊端，完善人工智能监管框架和政策体系，助推我国人工智能监管的未来发展。

4.2 综合监管：人工智能监管的模式再塑

人们普遍对人工智能单一监管模式存在体系封闭、运行迟滞的担忧。鉴于此，欧盟采取风险分级分类机制和多主体参与等形成综合性人工智能监管，使其能够广泛覆盖和应对多领域、多层次以及多形式的风险。以镜观面，我国人工智能风险发展具有跨领域、跨行业、跨主体的多风险态势，可针对性建立自发和协同监管、司法沟通和个体交互的综合监管模式，促进人工智能的风险的全面性、高效性和精准性消弭。（1）促进自发和协同监管。在统一的 AI 监管立法基础上，各省份市场监督机关等机关可设立区域性、产业性专项监管政策，适时发挥主观能动性以协同利益相关者，构建“自主”和“互助”的监管机制促进监管措施落实，成为 AI 系统有效性监管的保障者和合规性运营的积极促进者。（2）沟通司法活动。风险时代，如何根据现行立法和新法律对人工智能监管边界作出适当的解释是司法将要经历的持续性学习过程。人工智能监管机关与司法机关的互学互鉴，有利于司法界掌握有关 AI 政策、产品和服务的最新信息，正向引导与间接促进 AI 系统的创新进步和市场健康发展。（3）优化主体义务。人与 AI 之间将存在持久的复杂共存关系，目前仍需要依赖于特定领域的保障者来应对不同程度的风险隐患。我国应重视利益相关者在不同应用领域对 AI 系统的差异化义务，促进主体类型在 AI 生命周期各阶段的义务强度与风险水平相适应、成比例，避免笼统化的义务规定带来的应对失焦和权责不一致问题。

4.3 市场赋能：监管权的开放与产业创新

市场作为 AI 应用和创新的前沿阵地，在应对 AI 风险方面具有不可替代的重要作用。有效的市场监管不仅是适应 AI 发展的必然选择，也是促进 AI 创新的重要途径。欧盟特别关注人工智能监管领域的市场参与，以市场利益相关方的政策参与、实践反馈与资金支持等措施促进市场主体的监管协同和 AI 产业创新发展。我国可借鉴欧盟经验并强化如下举措：（1）赋权市场主体。面对各企业开发 AI 系统流程与内容不同，风险各异的情况，我国可通过设立咨询论坛或平台，便于利益相关方在人工智能监管领域的积极参与和辅助监管作用，针对特殊利益需求（特别是中小微型企业利益）予以反馈与保障，保障市场主体的开发、部署以及监测等必要权能，促进市场主体的监管协同性。（2）加强行业对话。国家监管机构是引领监管实践并向行业和企业传达监管要求的重要主体，为促进行业发展，监管机构必须加强与行业产品服务提供者的对话，以促进人工智能领域的标准制定和理念认同。（3）扩大市场投资激励领域。为加快人工智能领域的科学技术创新，我国可在市场应用场景中制定 AI 资助项目和产业目标、提供资金支持和设施投资等战略激励，实现人工智能产业高水平变革与高质量发展。

4.4 素质培育：基于 DigComp 开展 AI 素养培育

人工智能素养是人工智能监管领域的“防火墙”，在预防、评估和处理人工智能风险方面扮演着至关重要的角色，其不仅涉及对人工智能技术的理解和应用，还包括对其潜在风险、伦理问题、社会影响等方面的认识和应对能力。欧盟即在《法案》第 4b 条专门提出培养人工智能素养的要求，并出台《公民数字能力框架》（The Digital Competence Framework，DigComp）以发展和衡量各主体整体数字素养（包括信息、数据以及技术等素养）的能力（如图 4 所示）。我国可制定人工智能素养与能力提升策略，助力数字时代下各主体人工智能素养的评估和提升，实现人工智能风险的有效应对。（1）深入场景需求。结合相关人工智能系统的风险信息和具体内容，在人工智能的交互和协作中了解人工智能运作原理与应用场景，提高对人工智能风险的识别和应对。（2）明确培养目标。通过明确不同主体类型所需要的素养目标和内容，培养相应领域的知识、技能和能力以确保人工智能素养的持续性和有效性。（3）区分主体义务。适应人工智能生命周期不同主体的特殊义务要求，采取灵活多变的培训方式并因材施教，确保人工智能风险应对的针对性和实效性。（4）促进资源普及。在明确人工智能风险识别、应对和处理的需求和个体能力差距的前提下，提供线上线下学习渠道的综合类工具，为人工智能素养的拓展提供强有力的基础性资源保障。（5）提供正确交互。通过主体之间的沟通、报告等交互制度构建复杂场景中人工智能监管信息交流体系，为复杂场景提供清晰的风险解决思路。通过以上策略，我国不仅可以提升相关主体的人工智能素养，也对培养兼具培养技术能力和伦理责任感的高水平 AI 人才具有重要意义。

5 结论

在数字时代，AI 技术的突破性发展和大规模应用带来了巨大的经济潜力和创新机会的同时，也带来跨领域、多样态的动态风险。欧盟出台世界上第一部统一的人工智能监管规则《人工智能法案》以构建面向风险的人工智能监管，吸纳利益相关者制定集成式人工智能监管框架，为成员国和企业、社会团体等利益相关者提供了统

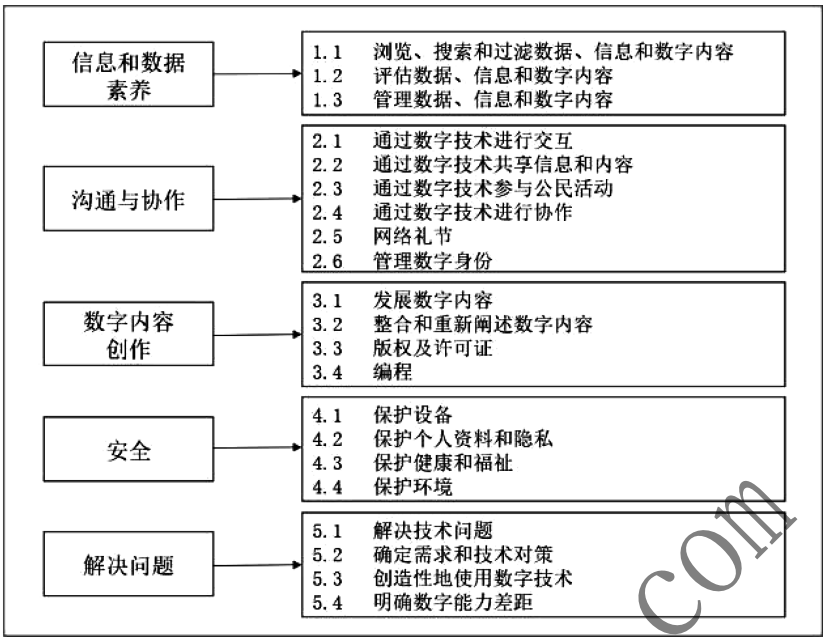


图 4　欧盟《公民数字能力框架》

一的监管体系，培养人工智能相关主体的伦理责任感和人工智能素养以应对纷繁复杂的人工智能风险，实现人工智能市场创新发展。我国可取欧盟精华并立足本国实际，建立既能促进人工智能发展又能保护个体基本权利的人工智能监管模式。在国家层面，基于本土国情制定面向风险的全国统一性、区域试点性的人工智能法律政策与框架体系，确保各领域的法律保障和人工智能监管合法性；在行业层面，构建司法机关、市场监督机关以及企业等主体的综合协同模式，促进信息共享和协同合作，并在关键产业领域，强化市场赋权、行业对话与投资激励等措施，推动 AI 技术进步与市场发展。同时，以人工智能素养的培育设置人工智能监管的风险“安全阀”，持续性、动态性和负责任地防范、应对和消除人工智能风险，实现我国公正、和谐和创新的未来人工智能发展。

参考文献

[1] UNESCO. Recommendation on the ethics of artificial intelligence [EB/OL]. [2024 - 03 - 19]. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.

[2] CARAI, MSB, CETS, et al. Applications of robotics and artificial intelligence to reduce risk and improve effectiveness: a study for the United States Army [J]. Robotics and Computer-Integrated Manufacturing, 1984, 1 (2): 191 - 222.

[3] SHARMA G D, YADAV A, CHOPRA R. Artificial intelligence and effective governance: a review, critique and research agenda [J]. Sustainable Futures, 2020, 2: 100004.

[4] FILGUEIRAS F, ALMEIDA V. The digital world and governance structures [M]. Cham: Springer International Publishing, 2021.

[5] ASADUZZAMAN M, VIRTANEN P. Governance theories and models [M]. Cham: Springer International Publishing, 2016.

[6] ERMAN E, FURENDAL M. The democratization of global AI governance and the role of tech companies [J]. Nature Machine Intelligence, 6 (3): 246 - 248.

[7] WIRTZ B, WEYERER G C, STURM B J. The dark sides of artificial intelligence: an integrated AI governance framework for public administration [J]. International Journal of Public Administration, 2022, 43 (9): 818 - 829.

[8] KOVAC M. Autonomous artificial intelligence and uncontemplated hazards: towards the optimal regulatory framework [J]. European Journal of Risk Regulation, 2022, 13 (1): 94 - 113.

[9] 贾开，蒋余浩. 人工智能治理的三个基本问题：技术逻辑、风险挑战与公共政策选择 [J]. 中国行政管理，2017 (10)：40 - 45.

[10] ALMEIDA V, MENDES L S, DONEDA D. On the development of AI governance frameworks [J]. IEEE Internet Computing, 2023, 27 (1): 70 - 74.

[11] WIRTZ B, WEYERER J, KEHL I. Governance of artificial intelligence: a risk and guideline-based integrative framework [J]. Government Information Quarterly, 2022, 39 (4): 101685.

[12] 庄友刚. 风险社会理论研究述评 [J]. 哲学动态，2005 (9)：57 - 62.

[13] JURI. European civil law rules in robotics [EB/OL]. [2024 - 01 - 15]. https://data.europa.eu/doi/10.2861/946158.

[14] FLORIDI L. Translating principles into practices of digital ethics: five risks of being unethical [J]. Philosophy & Technology, 2019, 32 (2): 185 - 193.

[15] MITTELSTADT B. Principles alone cannot guarantee ethical AI [J]. Nature Machine Intelligence, 2019, 1 (11): 501 - 507.

[16] LARSSON S. AI in the EU: ethical guidelines as a governance tool [M]. Cham: Springer International Publishing, 2021.

[17] MORLEY J, FLORIDI L, KINSEY L, et al. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices [J]. Science and Engineering Ethics, 2020, 26 (4): 2141 - 2168.

[18] AMIGONI F, SCHIAFFONATI. Ethics for robots as experimental technologies [J]. IEEE Robotics & Automation Magazine, 25 (1): 30 - 36.

[19] SCHERER M U. Regulating artificial intelligence systems: risks, challenges, competencies, and strategies [J/OL]. 2015 [2024 - 03 - 18]. https://papers.ssrn.com/abstract=2609777.

[20] RAHWAN I. Society-in-the-loop: programming the algorithmic social contract [J]. Ethics and Information Technology, 2018, 20 (1): 5 - 14.

[21] JETZKOWITZ J. Co-evolution of nature and society: foundations for interdisciplinary sustainability studies [M]. Cham: Springer International Publishing, 2019.

[22] AVEN T. Risk assessment and risk management: review of recent advances on their foundation [J]. European Journal of Operational Research, 2016, 253 (1): 1 - 13.

[23] 王彦雨，李正风，高芳. 欧美人工智能治理模式比较研究 [J]. 科学学研究，2024 (3): 460 - 468.

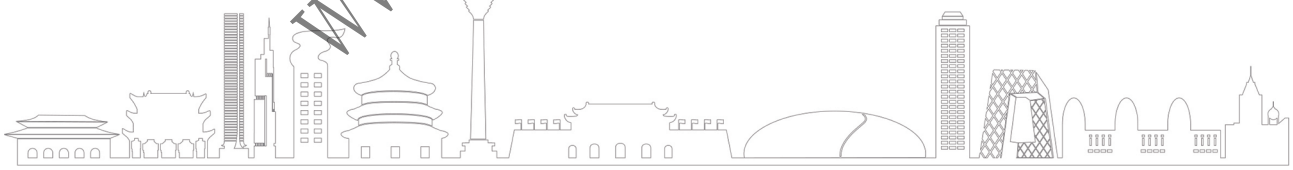
[24] ERDELYI O S, GOLDSMITH S. Regulating artificial intelligence: proposal for a global solution [J]. Government Information Quarterly, 2022, 39 (4): 1 - 25.

（收稿日期：2024 - 07 - 20）

作者简介：

王春晓（2000 - ），女，硕士研究生，主要研究方向：网络犯罪，人工智能风险规制。

李怀胜（1983 - ），男，博士，副教授，博士生导师，主要研究方向：网络犯罪、网络安全等。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部