欧盟人工智能基于风险的治理路径及中国借鉴

崔奥　田好　许庭皓　刘万豪　叶佳敏

西南民族大学，四川成都，610041；

摘要：人工智能技术的迅速发展及其在各领域的广泛应用引发了安全性、伦理性及社会影响等方面的深刻关注。为有效规避和管理这些风险，欧盟制定了《人工智能法案》，采用“基于风险”的治理路径，其风险等级划分涵盖不可接受风险、高风险、有限风险和最小风险四类。通过对《法案》基于风险的治理路径以及立法逻辑的分析，阐述其立法价值基于欧盟客观现实需求以及立法路径中规制方式的选择，同时指出法案存在的局限性。在此基础上，着眼于欧盟人工智能 “基于风险”治理路径对于中国的借鉴意义与可行性，最后建议“风险+场景”化治理这一具体的借鉴方式，旨在通过对欧盟法案及相关治理路径的剖析，为我国人工智能治理提供有益参考。

关键词：人工智能法案；基于风险；“风险+场景”化

DOI：10.69979/3029-2700.24.9.026

1 问题的提出

人工智能作为当今世界经济社会发展的核心技术，其应用范围几乎无所不包，但它所引发的法律和伦理问题也日益凸显。全球法律体系正面临前所未有的挑战，需要找到有效的方法来解决。[1]

现阶段，世界各国法律与制度对人工智能问题的回应逐渐从“替代”与“辅助”议题转向“风险”议题[2]。各国也入手积极构建了规范人工智能治理的框架——如美国的分散式治理结构、欧盟的集中式统一立法结构以及日本的软约束治理方式。中国在 2023 年 7 月发布《生成式人工智能服务管理暂行办法》（以下简称《办法》），目前，人工智能相关顶层法律层面呈现了“1+3”的模式。“1”是指《民法典》第四编人格权编的隐私与个人信息保护，其后《网络安全法》《数据安全法》和《个人信息保护法》“三驾马车”共同奠定了人工智能发展的基础；而部门规章领域，如《互联网信息服务的深度合成》、《算法的推荐》，已经开始触碰人工智能的实体，但是就人工智能规制来说，我国法律法规的针对性仍有不足。人工智能正加速改变着产业、社会和世界，中国应当有一部《人工智能法》来彰显国家对于人工智能治理的立场，为我国人工智能技术产业的发展保驾护航，真正实现立法先导，发挥法治在人工智能发展治理领域的基础性作用。[3]欧盟《人工智能法案》（以下简称《法案》）作为全球首部综合性的人工智能立法，对全球人工智能的发展、治理和立法产生重大影响和示范效应。中国需要密切关注欧盟的人工智能立法，对其进行深入研究和合理批判，最终形成适合中国的人工智能治理道路。[4]

因此，本文的分析以《法案》及其后续出台的相关法律文件为基础，结合中外人工智能模型技术和市场的现状与中国人工智能立法的现实问题，对人工智能的风险治理建言献策。

2 欧盟《人工智能法案》监管路径分析

2.1 基于风险的治理路径

《法案》采取了一种基于风险的监管路径（risk-based approach），除了单列的通用 AI 系统特有的风险识别方式，以 AI 系统的预期目的为依据，将 AI 系统划分为不可接受的、高、特定透明度和极小风险四类。根据四种等级的风险以确定不同 AI 系统的现实可用性，进而确定不同等下 AI 系统相关人的权利与义务。

在监管的执行中，鼓励企业进行自纠自查。若企业没有进行自我监督或者虽然进行但缺乏有效性，违反了法律所规定的义务，就要承担不利的后果，具体的不利后果包括《法案》本身规定的罚款、排除或限制 AI 产品的市场投放，如果具体侵害了使用者的合法权益，还会承担欧盟配套法案中规定的责任，即行为-风险-义务-责任的规制路径。

2.2 欧盟《人工智能法案》立法逻辑

从整体规制方法上，欧盟确立了统一的强监管式立法。其背后的原因是欧盟以数据安全、市场统一和人权为最高目标[5]，采用强监管的方式，能够使法律适用统一、集资难度减轻、内部价值趋同，从而多维度维护单一市

场的构建，提升国际话语权。

在具体规制方法的选择上，根据学理上和欧盟会议上讨论的几种不同规制方法在法理以及执行的利弊的比较，最终选择更合适的基于风险的规制——相比起基于权利和原则的监管方式，基于风险的方式能够促进行业自律发展，阻却不同行业间的干涉，防止权力监管过多抑制创新，促进 AI 企业协调稳健发展。能够进行更加实质性的保护，更符合欧盟构建安全数据市场的核心价值理念。基于德国贝克的风险社会理论，“风险”不同于传统意义上的危险。它不仅仅指自然发生的灾害或已经存在的危害，更强调是一种由人类活动和决策所带来的潜在威胁，是对未来可能发生的危害的一种预测和评估，AI 系统带来的风险也恰处其中。而这种“对未来可能发生危害的预估”蕴含着不确定性，此时实际上还没有确定 AI 系统具体怎样的审慎义务。在此以风险作为监管标准判断的靶点，赋予自己的决定造成的不可预见的后果具备一定客观性和可预测性。而风险等级分类正是建立在基于风险的规制路径之上。并且，回到法律和 AI 系统以及衍生 AI 行为本身，AI 系统不断迭代，而法律具有滞后性，风险分类能够抵御确权程序的复杂性以及权利的不周延，不仅保护法律本身的体系的稳定，也能减小权利的强边界性对于 AI 创新发展的限制，防止将本应流动向前的技术限制在慢速变化的制度的框架中。

2.3 欧盟《人工智能法案》的局限性

首先，欧盟采取的横向监管模式需要对 AI 系统的范围精准的囊括，但是 AI 技术是不断迭代的。《法案》虽然对 AI 系统做了定义，但该定义表述模糊且抽象，不具有特征性，并且相对滞后，以至于使监管无法达到精细的程度，导致了法律适用上的主体错位，增加了实施的成本以及实施过程中的不确定性。[6]欧洲数据保护委员会（EDPB）和欧洲数据保护专员公署（EDPS）就指出，法案列出的高风险清单存在遗漏，没有涵盖使用 AI 系统确定保险费或者评估医疗方法适用于健康研究目的的场景。[7]从本质上看，法律本身的稳定性与现实的无序性在此产生了冲突。

其次，法案对于风险的监管首先采用企业自我监督而后置了外部监管，这导致大部分企业需要花费大量的合规成本促其产品上市。欧洲国际政治经济中心的一项关于科技公司事前监管成本的研究表明，在 2018 年，事前监管致使 GDP 损失高达 850 亿欧元，造成了 1000 亿欧元的消费者福利损失[8]。对于企业特别是中小企业来说，实质上是变相提升了其进入行业的门槛，进而抑制市场活力和创新度。虽然法案中采取了设立监管沙盒、对中小企业降低认证与合规成本等规定，但仍不能改变以监管和规制为出发点导致的抑制产业创新发展的本质。并且，这样的依赖“企业自我管理”的“强规制法”之间存在着矛盾，有将执法部门的职责甩锅给被监管者自身的嫌疑，使得监管实践中无法顺利进行。

最后，风险等级的固定难以适配实际控制在使用者之下的场景变化——针对通用 AI 模型。按照欧盟《法案》的风险分级，通用型人工智能往往会被划分到透明度义务风险等级。但由于此类模型本身功能应用情景广泛的属性，模型的使用者通过 prompt 工程或者不同应用程序间编程接口（API）的调用，通用型人工智能会被应用于提供者非预期的场景中。这就引出了风险分类管理的实践缺陷：风险等级较低的通用型人工智能被应用到高风险的场景。

实际上，按照现行的欧盟风险治理模式，等级划分没有充分考量提供者所提供的 AI 可能会被使用者用于不同场景，在这些不同的场景下会触发风险等级是不确定的。例如，国产 AI 智能助手——豆包，集多种功能于一体。豆包的部分功能，既可以参照类似 sovit 的“弱人工智能”系统，将其归类到有限风险之中；又可以参照 GPAI 模型实践准则（初稿）中指出的：通用型 AI 系统的复杂性和不透明性可能对基本权利、健康和安全构成风险，将豆包归类到高风险甚至不可接受的风险等级之列。例如，Characeter AI 在与一位美国 14 岁的男孩聊天中引导其自杀，而其正是一个基于通用大模型 GPT-3 构建的聊天机器人[9]。

3 欧盟人工智能基于风险的治理路径的中国借鉴

3.1 中国人工智能发展现状及借鉴可行性

3.1.1 发展现状

我国人工智能市场高度活跃。早在 2017 年，国务院便发布《新一代人工智能发展规划》，对 AI 的研发和应用给予了高度重视。这一战略性举措不仅激励了各类资本涌入人工智能产业，也推动了众多 AI 技术公司和平台的涌现。根据相关数据统计，截至 2022 年 6 月，我国人工智能企业数量超过 3000 家，位居世界第二，

人工智能核心产业规模超过 4000 亿元。2022 年，我国人工智能支出达到 2255 亿人民币，跃升为全球第二大人工智能市场，占全球人工智能支出的 18%。但同时存在着区域差距明显等内部问题，仍需要完善相关配置以减少内部风险，[10]这与欧盟渴望构建稳健的内部市场需求具有相似性。

3.1.2 可行性

基于风险的治理理念在我国立法规定中有多处可循之迹。例如，《数据安全法》第 21 条：国家建立数据分类分级保护制度，根据数据在经济社会发展中的重要程度，以及一旦遭到篡改、破坏、泄露或者非法获取、非法利用，对国家安全、公共利益或者个人、组织合法权益造成的危害程度，对数据实行分类分级保护。国家数据安全工作协调机制统筹协调有关部门制定 重要数据目录，加强对重要数据的保护。关系国家安全、国民经济命脉、重要民生、重大公共利益等数据属于国家核心数据，实行更加严格的管理制度。各地区、各部门应当按照数据分类分级保护制度，确定本地区、本部门以及相关行业、领域的重要数据具体目录，对列入目录的数据进行重点保护;《办法》第 3 条：国家坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管。其具体内容虽然存在规则模糊、标准不明确、缺乏可实践性等问题，但是其监管理念与大体框架却与欧盟《法案》的“风险分级”治理体系不谋而合，体现了国际人工智能监管共性的发展趋势。这种共性不仅在治理思路上消弭了两种法系的差异，还为我国人工智能监管体系的完善提供了重要的理论依据和实践支持。

3.2 如何借鉴：“风险+场景”化治理

显然，中国具有借鉴欧盟人工智能“基于风险分类”治理路径的必要性与可行性，但是单纯依赖风险治理难以应对通用型人工智能实践应用过程中带来的挑战。究其根本还是在于《法案》风险制度的设计伊始没有充分考虑到通用型人工智能在合理预期外可能被误用的情形。24 年九月份 Google 发布了全球 185 个各大企业的 AI 模型实际应用落地案例[11]。其中绝大多数 AI 模型都属于通用型人工智能的范畴。这一现实同前文所描述的中国市场的 AI 模型现状也高度匹配，直接借鉴法案中的风险制度会造成相关人的权利义务不匹配。

为落实人工智能模型的风险监管，并兼顾模型不同功能分别适用于对应的风险等级监管要求与模型风险等级的唯一性之间的矛盾。尝试在欧盟法案的风险治理制度的基础上，增加灵活性设计是必要的。即“风险+场景”化治理。

将“风险分级”理念贯彻为我国人工智能监管的一项基本原则。中国可参考欧盟的做法，首先根据人工智能技术应用的领域、潜在影响和复杂程度，建立一套明确的风险分类标准。结合中国国情和社会主义核心价值观的需要，并为不同应用的风险等级提供详细定义。

出于灵活化设计的需要，在风险等级治理的基础上融入场景化免责事由，既能够保护使用者的权益，又能激励市场主体主动预防潜在风险，并推动人工智能技术的创新发展。场景化免责的核心在于根据不同场景中 AI 模型的实际风险，合理界定提供者的责任：在明确使用场景属于低风险类别且未引发高风险问题时，提供者可被认定不承担高等级的责任（低风险场景免责）；若提供者已采取足够的技术和管理措施限制模型的误用，但用户以极端方式使用导致的高风险后果，提供者可申请免责（合理预期误用免责）；当提供者能够证明其与用户签订了合理使用协议，明确责任分担并履行了充分告知义务，则在使用者引发问题时可减轻责任（协作性保障免责）。

监管机构可开发标准化的场景风险评估工具，用以帮助提供者在模型投放市场前，系统评估其在不同场景中的风险等级。该工具能够根据实时数据和使用方式，动态调整模型在具体场景中的风险评估结果。提供者也可通过评估工具，及时调整功能或引入防护机制以降低场景化风险。同时，为了避免滥用免责事由，监管机构应要求企业在申请场景化免责时进行备案。提供者需提交详细的场景分析和技术防控措施说明。企业备案的信息应对相关方（如监管机构或用户）透明，以便审查和监督。

4 结语

欧盟采用的“基于风险分类”的治理路径，是对人类理性的重要反思。毕竟，传统社会中监管对象及其风险具有相对确定性与稳定性，人类可以按照理性制定出具有确定性的政策和高度完备的法典。为了确保理性至上的成文法能够得到有效执行，国家可以基于身份统治构建起科层监管结构，以确保政策执行的确定性。[12]但

是,在监管对象高度多变性的情况下,人类无法根据有限理性制定出具有确定性的政策,只能根据监管对象具有的风险分类构建起具有针对性的监管结构,充分激发多元场景下多主体的智慧，以应不确定风险。随着人工智能技术在全球层面的广泛应用,不确定的系统性风险是人类需要面临的共同挑战。在此基础上,“风险+场景”化治理路径有助于全人类集思广益,以更具针对性、更务实的态度积极行动起来。

参考文献

[1]読買新聞:《生成 AI のあり方に関する共同提言》, https://www.yomiuri.co.jp/politics/20240408-OYT1T50188，2024 年 12 月 13 日访问。

[2]国家人工智能标准化总体组.人工智能伦理风险分析报告[R].2019-4

[3]参见清华大学人工智能国际治理研究院：国内人工智能立法进展、影响及展望【AI 知识库】，https://news.qq.com/rain/a/20240524A0AYTS00,2024 年 12 月 13 日访问。

[4]丁晓东.人工智能风险的法律规制——以欧盟《人工智能法》为例[J].法律科学(西北政法大学学报),2024,42(05):3-18.

[5]刘子婧.欧盟《人工智能法》:演进、规则与启示[J].德国研究,2024,39(03):101-128+151.

[6]王天凡.人工智能监管的路径选择——欧盟《人工智能法》的范式、争议及影响[J].欧洲研究,2024,42(03):1-30+173.

[7]European Data Protection Board, European Data Protection Supervisor. Joint Opinion 5/2021 on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence(Artificial Intelligence Act)[EB/OL].(2021-06-18)[2021-12-27].https://edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf.

[8]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[J].电子政务,2022,(09):63-72.

[9]EU Artificial Intelligence Act：《General Purpose AI and the AI Act》，https://artificialintelligenceact.eu/wp-content/uploads/2022/05/General-Purpose-AI-and-the-AI-Act.pdf，2024 年 12 月 13 日访问

[10]黄静怡.“分级分类”与“契约”风险治理并行的人工智能监管制度构建——以欧盟《人工智能法案》为分析对象[J].海南金融,2024,(02):76-85.

[11]https://blog.google/products/google-cloud/gen-ai-business-use-cases/#customer-agents

[12]武振国.欧盟人工智能的实验主义治理路径及中国借鉴[J].西北大学学报(哲学社会科学),2024,54(06):153-164.

作者简介：崔奥（2002.1—），男，汉族，河北保定人，西南民族大学法学院本科在读，研究方向：法学。
田妤（2004.10—），女，汉族，四川南充人，西南民族大学法学院本科在读，研究方向：法学
许庭皓（2002.11—），男，汉族，四川乐山人，西南民族大学法学院本科在读，研究方向：法学。
刘万豪（2003.12—），男，汉族，浙江龙港人，西南民族大学法学院本科在读，研究方向：法学
叶佳敏（2005.3—），女，藏族，青海西宁人，西南民族大学法学院本科在读，研究方向：法学

项目来源：西南民族大学（2024 年度）大学生创新创业训练计划项目：AI 语音合成技术运用中权利义务配置的探究—以 AI 歌手为例（S202410656106）阶段性成果。