·法学研究·

DOI:10.20004/j.cnki.ujn.2025.05.015

人工智能分类分级监管研究

——从风险基准到重要性导向的逻辑转变

许建峰,吕昭诗

(中国政法大学 数据法治研究院,北京 100088)

[摘　要]欧盟提出的以风险为基准的人工智能分类分级监管模式已成为多国学习和模仿的对象,我国亦基于此理念开展了人工智能分类分级治理实践。但随着人工智能技术的发展,我国面临着以要素为划分的分类分级方式滞后、现有的分类和分级监管模式交叉混乱、人工智能业态分层和分类分级存在重叠等困境,以风险为基准的分类分级方式本身亦存在预先负面判断、难以量化评估、分级机制僵化等弊端。在此背景下,应以重要性为导向构建人工智能分类分级监管新路径,该路径以价值理论为底层逻辑,通过价值的主客观结合衡量重要性的高低,具有客观性强、灵活性高、成本低等优点。在重要性导向下,我国可采取划分关键人工智能和一般人工智能的分级监管思路,对关键人工智能开发者、提供者施以更多监管要求;沿用以场景为划分的分类监管方式,明确"先分类,后分级"的流程次序;在人工智能分层业态下对基础模型层、专业模型层、服务应用层人工智能实施差异性监管。

[关键词]人工智能;分类分级;风险监管;关键人工智能

[中图分类号]D922.17　　[文献标识码]A　　[文章编号]1671-3842(2025)05-0148-13

自1956年达茅斯会议提出人工智能概念以来,人工智能技术便在半个多世纪里不断迭代发展。2022年底,Open AI公司推出人工智能聊天机器人模型Chat GPT,随即引发了新一轮的人工智能技术发展高潮。Chat GPT以其海量的数据训练和超强的逻辑推理能力实现了人工智能科技的跨越式发展,引领世界进入了生成式人工智能时代。2024年上半年,Open AI公司陆续发布具有划时代意义的人工智能文生视频大模型Sora和视听大模型GPT-4o,这两个系统分别实现了对视图分析和视听反应能力的巨大突破,预示着人工智能正在加速迈向高度智能化。

人工智能技术的日新月异给现代社会生活方式带来了翻天覆地的变化,其深入各行各业并为之赋能,给人类生产、生活带来极大便利,已然成为新质生产力的代表。但在人工智能迅猛发展的背后,其所带来的负面影响也不容忽视,诸如伦理错乱、贫富分化、技术滥用以及黑箱社会等一系列危害正在日益显现①。当前,各式各样的人工智能纷纷涌现,逐步呈现"基础模型—专业模型—服务应用"的分层业态②,不同人工智能所产生的影响也千差万别。在此背景下,对人工智能采取怎样的监管方式便成为现实的问题。如果统一对所有人工智能施以严监管,将不免阻碍人工智能产业的创新性发展;若对人工智能采取弱监管,其背后的风险危害又将难以遏制。

[基金项目]国家社会科学基金重点项目"生成式人工智能的法律定位与分层治理研究"(项目编号:23AFX009)。

[作者简介]许建峰,中国政法大学数据法治研究院教授、博士生导师;吕昭诗(通讯作者),中国政法大学数据法治研究院博士研究生,邮箱:lvzhaoshi@126.com。

①参见马长山:《人工智能的社会风险及其法律规制》,《法律科学》(西北政法大学学报),2018年第6期。

②参见张凌寒:《生成式人工智能的法律定位与分层治理》,《现代法学》,2023年第4期。

全球范围内，欧盟率先对该难题给出了创新性的解决思路。2021年4月，欧盟提出对人工智能进行分类分级监管的《人工智能法（草案）》。2024年8月1日，最终版的《人工智能法》正式生效实施，该法案以人工智能可能带来的风险为基准，将人工智能分为不同的类别，并对各类人工智能实施差异化的监管措施。这种对人工智能进行分类分级的方法有效避免了过于僵化的“一刀切”式监管，体现了对人工智能发展特性的深刻理解和顺应。2023年5月，我国国家网信办颁布《生成式人工智能服务管理暂行办法》（以下简称“《办法》”），该《办法》第3条提出，我国要对生成式人工智能服务实行分类分级监管[①]。尽管分类分级被视为当前对人工智能进行有效监管的重要模式，但以风险为基准的分类分级方法并非必然适用于我国。我国目前处于人工智能发展的初期阶段，对人工智能过早地进行严格分级可能会抑制产业的创新与发展。此外，若机械地遵循以风险为基准的监管模式，不仅无法在人工智能治理中体现中国特色，还可能使我国陷入“布鲁塞尔效应”[②]的风险之中。

由此，面对我国人工智能发展的新阶段新业态，亟需有效的人工智能分类分级监管新模式。未来我国对人工智能分类分级监管应遵循怎样的基本逻辑？不同分类分级的人工智能有哪些差异性的监管要求？对人工智能的基础模型、专业模型、服务应用各业态应实行怎样的针对性监管措施？本文将就以上问题予以一一探讨。

一、人工智能分类分级监管的实践检视

自欧盟提出以风险为基准分类分级监管人工智能的理念以来，多国在人工智能领域纷纷学习和效仿，风险监管逐渐成为目前世界范围内治理人工智能最普遍采用的模式。近年来，我国对数据、算法、平台等领域亦开展了风险监管的实践探索，但随着人工智能技术的进一步发展，现有的分类分级监管模式遇到了诸多挑战。

（一）以风险为基准的人工智能分类分级监管实践

欧盟在《人工智能法》中使用“基于风险的路径”（Risk-based approach）监管人工智能，一脉相承其在《通用数据保护条例》（GDPR）中对数据所确立的风险监管思路，其将人工智能划分为四个风险等级：不可接受风险、高风险、有限风险和最低风险[③]。针对不同风险等级的人工智能，欧盟设定了不同的规制措施。具体来说，不可接受风险的人工智能是指严重侵犯基本人权或违反欧盟价值观的人工智能，该人工智能被欧盟完全禁止；高风险的人工智能一般指可能对人类安全和生活造成严重影响的人工智能，该人工智能需遵守从事前到事后的一系列严格监管要求；有限风险人工智能是指缺乏透明度而带来一定风险的人工智能，该人工智能需符合透明度要求以确保用户意识到自身与人工智能的交互；不属于以上风险类型的都是最低风险人工智能，这些系统没有对公民权利或安全构成威胁或威胁很小，欧盟对这类人工智能没有特殊的限制，而是鼓励相关主体自发地遵守行业内规范。

除欧盟外，美国、日本、韩国等也在人工智能领域积极探索构建相应的风险监管法律体系。2023年1月，美国国家标准与技术研究院（NIST）发布了软性规则《人工智能风险监管框架》（以下简

①《生成式人工智能服务管理暂行办法》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。

②“布鲁塞尔效应”是指，欧盟通过其规章制度在全球市场和监管标准上施加的影响力，这种影响力不需要直接的强制或施压，而是凭借其市场力量吸引跨国公司主动遵守其规则以获取准入资格，使欧盟标准成为事实上全球标准的准则。这种影响力超出了欧盟成员国，深刻影响着欧盟之外国家自身的制度体系。

③Artificial Intelligence Act，http://data. europa. eu/eli/reg/2024/1689/oj，2024. 07. 12.

称“《框架》”)[①],该《框架》给出了评估人工智能风险的方法论,认为风险评估有七点关键要素,即有效与可靠、安全、弹性、问责性与透明、可解释性、隐私保护和公平不偏见,旨在为人工智能的风险监管提供指导原则。日本采取基于风险的软法引导,其在2019年3月发布的《以人为本的人工智能社会原则》[②]中提出,要通过人工智能实现“尊重人类尊严、多元包容、可持续发展”的“AI-Ready社会”,并在医疗、交通、教育等领域根据不同使用场景对人工智能分别进行了相应监管。韩国国民议会在近几年提出了《促进人工智能产业发展和确保信任的法案》《人工智能问责法案》,这两部法案均对高风险的人工智能提出了更多责任要求,包括强制性要求事先通知用户、确保系统可靠性和安全性等,体现了韩国内部对人工智能根据风险分类分级监管的态度。

回看国内,2023年11月,我国在全球首届AI安全峰会上签署了全球第一份针对人工智能的国际性声明《布莱切利宣言》(以下简称“《宣言》”),该《宣言》亦遵循风险监管的思路展开人工智能治理。深圳和上海两市同样走在人工智能分类分级监管的前沿,2022年9月,两市分别公布《深圳经济特区人工智能产业促进条例》和《上海市促进人工智能产业发展条例》,在全国范围内率先开展了人工智能领域分类分级监管的探索[③],两市均将人工智能依据风险进行划分,采取了对高风险人工智能进行事前审查评估、对中低风险人工智能进行事前披露和事后跟踪的监管模式。

通过仔细观察以风险为基准的分类分级监管框架,不难发现,其将风险作为分类和分级的底层逻辑,设计的初衷沿着人工智能的复杂特性展开,意图通过量化人工智能可能造成的风险损害,从而对不同风险的人工智能采取差异化监管措施。该框架的底层逻辑如下:其一,从风险评估角度看,《人工智能法》第3条将风险损害程度和风险发生概率相结合以期提升评估精准度。由此可以看出,以风险为基准的评估采用了一种综合分析的方法,不单单着眼于潜在损害的严重性,也包含了风险发生的可能性[④]。通过这种较为全面的考量,风险评估将更为科学。其二,从防控成本角度看,风险监管意图实现资源投入和风险防控的最优平衡。《人工智能法》序言第65条提出,风险管理系统应根据人工智能的最新技术确定最合适的成本管理措施。欧盟意图在有限的监管资源和广泛的人工智能应用情况下,通过有效的风险分类分级系统来优化资源分配[⑤],实施相应的差异化监管,从而实现成本效益最大化。其三,从市场准入角度看,风险监管在认可人工智能可用性的同时致力于实现可控性。欧盟以风险为基准的分类分级监管模式承认大部分人工智能的可用性,但对风险较高的人工智能设置了市场准入门槛。《人工智能法》序言第46条明确:只有在符合某些强制性要求的情况下,高风险人工智能系统才能投放欧盟市场。这种方法充分体现了风险监管中的“预防原

---

①NIST: AI Risk Management Framework (AI RMF), https://nvlpubs. nist. gov/nistpubs/ai/NIST. AI. 100-1. pdf, 2025. 07. 07.

②Social Principles of Human-Centric AI, https://www. cas. go. jp/jp/seisaku/jinkouchinou/pdf/humancentricai. pdf, 2024. 12. 27.

③《深圳经济特区人工智能产业促进条例》第66条:“市人民政府及其有关部门根据人工智能应用的风险等级、应用场景、影响范围等具体情境,实施分级、分类差异化监管,完善人工智能领域监管机制。高风险的人工智能应用应当采用事前评估和风险预警的监管模式。中低风险的人工智能应用应当采用事前披露和事后跟踪的监管模式。人工智能应用分类分级监管办法由市人民政府另行制定。”《上海市促进人工智能产业发展条例》第65条:“对高风险的人工智能产品和服务实行清单式管理,遵循必要、正当、可控等原则进行合规审查。对中低风险的人工智能产品和服务采用事前披露和事后控制的治理模式,促进先行先试。具体办法由市人民政府另行制定。”

④Gellert R. , *The risk-based approach to data protection*, Oxford: Oxford University Press, 2020, pp. 27-28.

⑤Davies G. J. , Kendall G. , Soane E. , et al. , Regulators as 'agents': Power and personality in risk regulation and a role for agent-based simulation, Journal of Risk Research, Vol. 13, No. 8, 2010, p. 963.

则”[①],从而加强监管以保障人工智能技术的发展在可控范围内。

尽管以风险为基准的分类分级监管框架底层逻辑意图实现有效的人工智能治理,但其在实践中可能遇到各种各样现实的问题。风险监管框架本身亦可能包含固有缺陷,该底层的设计逻辑是否能够实现仍需要进一步讨论。

(二)我国人工智能分类分级监管的现实困境

2021年9月,国家网信办等9部门联合颁布《关于加强互联网信息服务算法综合治理的指导意见》(以下简称“《意见》”),该《意见》在我国首次提出“分类分级”概念[②]。自此之后,我国陆续在数据、算法、主体等方面构建起了分类分级监管规则。但近年来以Chat GPT为代表的生成式人工智能技术的跨越式发展,颠覆了以往的监管格局,原有的分类分级模式愈发不适应人工智能产业发展的实际需求,面临着诸多困境。

第一,以要素为划分的分类分级方式滞后。近年来,我国针对数据、算法和主体等要素提出了分类分级策略,以便在各自领域展开针对性监管。在数据领域,依据数据获取和使用场景的不同,《网络安全标准实践指南—网络数据分类分级指引》将数据划分为公民个人、公共管理、信息传播等五个类别[③],并根据其潜在危害程度,将数据分为一般数据、重要数据和核心数据三个级别[④]。在算法领域,根据其应用场景,《互联网信息服务算法推荐管理规定》将算法划分为生成合成、个性化推送、排序精选等五种类型[⑤],并依据算法能力展开分类分级监管[⑥],重点关注具有舆论属性或社会动员能力的算法。在主体方面,依据平台的服务场景,《互联网平台分类分级指南(征求意见稿)》将其划分为网络销售、生活服务、社交娱乐等六类平台[⑦],并根据影响力将不同平台划分为了超级平台、大型平台和中小平台三级[⑧]。这些基于数据、算法或平台的单一要素式分类分级监管在各自领域均发挥了不可替代的作用,极大提升了监管实效。然而,随着人工智能的不断迭代,大模型参数纷纷突破万亿级别[⑨],其技术构成也愈发复杂,这些人工智能通常涉及多种要素的交互作用,数据采集、算法设计、模型训练和应用部署等要素相互影响,共同决定着人工智能的行为和性能。由此,

---

①“预防原则”主要是指在面对可能对环境或人类健康造成重大损害的风险时,即使缺乏完全的科学确证,也应采取预防措施避免或减少这些风险。

②《关于加强互联网信息服务算法综合治理的指导意见》一、总体要求(二)基本原则:……坚持风险防控,推进算法分类分级安全管理,有效识别高风险类算法,实施精准治理……;三、构建算法安全监管体系(九)有序推进算法备案工作。……健全算法分类分级体系,明确算法备案范围,有序开展备案工作……

③《网络安全标准实践指南——网络数据分类分级指引》中4.1数据分类框架指出,常见的数据分类维度,包括但不限于:公民个人维度、公共管理维度、信息传播维度、行业领域维度、组织经营维度。

④《网络安全标准实践指南——网络数据分类分级指引》中4.2数据分级框架指出,按照《中华人民共和国数据安全法》要求,根据数据一旦遭到篡改、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,将数据从低到高分成一般数据、重要数据、核心数据共三个级别。

⑤《互联网信息服务算法推荐管理规定》第2条:前款所称应用算法推荐技术,是指利用生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法技术向用户提供信息。

⑥《互联网信息服务算法推荐管理规定》第23条:网信部门会同电信、公安、市场监管等有关部门建立算法分类分级安全管理制度,根据算法推荐服务的舆论属性或者社会动员能力、内容类别、用户规模、算法推荐技术处理的数据重要程度、对用户行为的干预程度等对算法推荐服务提供者实施分类分级管理。

⑦《互联网平台分类分级指南(征求意见稿)》中2.2分类方案依据平台的连接对象和主要功能,将平台分为网络销售、生活服务、社交娱乐、信息资讯、金融服务、计算应用六类平台。

⑧《互联网平台分类分级指南(征求意见稿)》中3.2分级方案依据用户规模、业务种类以及限制能力,将平台分为超级平台、大型平台和中小平台三级。

⑨参见尚智丛,闫禹宏:《ChatGPT教育应用及其带来的变革与伦理挑战》,《东北师大学报》(哲学社会科学版),2023第5期。

原有基于单一要素的分类分级监管模式难以反映人工智能的整体特性,亟需构建一个适配人工智能整体性的分类分级监管体系。

第二,现有的分类和分级监管交叉混乱。分类分级监管包括分类和分级两个维度,前者体现横向治理思路,后者彰显纵向治理逻辑。现有的分类分级监管体系通常依据应用场景进行分类,并按照风险因素进行分级。然而,在实际的人工智能分类与分级过程中,分类和分级并无严格次序,导致两者经常交叉划分,形成了普遍存在的混乱现象。同时,对人工智能分类的基本逻辑是不同的应用场景需要不同的人工智能技术支持。然而,现实情况是,某一特定的人工智能往往不局限于单一应用场景,该系统在某一场景中的人工智能可能是高风险,但在其他应用场景中风险性便可能显著降低。例如,一个人工智能图像绘制系统,既可以应用于医疗影像合成,也可以用于美术作品创作,前者的风险级别显著高于后者。在这种情况下,不仅以场景为依据的分类难以划分,基于风险因素的分级也无法有效实施,目前这种交叉混乱的分类分级现状已然成为人工智能治理的一大阻碍。

第三,人工智能业态分层和分类分级存在重叠。不同于传统的技术支持者、服务提供者、内容生产者三种主体,当前人工智能整个产业呈现明显的纵向分层业态,形成了"基础模型层-专业模型层-服务应用层"的三级结构。其中,基础模型作为人工智能的核心,提供通用的学习能力和数据资源;专业模型则在基础模型之上进行定制化训练,以直接满足特定领域的需求;服务应用层将人工智能模型嵌入实际业务流程中,直接面向最终用户。这种分层业态是人工智能产业的自然形成,三个层级既相互联系,又各具特点。一方面,底层的基础模型可直接对服务应用层人工智能产生影响,服务应用层人工智能亦能直接对基础模型进行反馈。在这种情况下,对单一层级的人工智能进行分类分级监管必然也对其他上下游的人工智能产生影响,由此在一定程度上造成了人工智能分类分级的联动错乱;另一方面,"基础模型-专业模型-服务应用"三层模型业态各不相同,每一层人工智能都具备其独有的特点,不能等同视之,这也就意味着不同层级的人工智能需要有针对性的监管策略,以确保有效应对每一层级的独特需求和潜在风险。

二、由风险基准向重要性导向的逻辑转变

面对人工智能发展的新趋势和新业态,原有的分类分级监管模式已经遇到了越来越多的挑战。从风险监管理论的视角来看,人工智能与大数据、深度学习等能够训练、学习和预测的技术一样,其产生的风险同样难以预测和管理[①]。以风险为基准分类分级监管模式本身的不足始终无法解决,其并非是适合我国人工智能治理的最佳方案。由此,本文提出应以重要性为导向构建人工智能的分类分级监管新路径,以期实现我国人工智能领域的有效治理。

(一)以风险为基准分类分级监管模式的不足

当代对"风险"概念的理解可追溯至12世纪的意大利,当时的商人为了控制商品交易中的未来收益和损失,开始将交易风险与商品本身区分开来。从12世纪到18世纪,人们逐渐认识到,风险并非不可改变,而是可以被人类识别、计算,并在一定程度上加以控制。自18世纪起,风险逐渐成为一个成熟的概念,人们开始理性地开展风险监管[②];工业革命后,风险监管的理念逐渐扩展到金融市场、公共卫生、环境保护等领域。20世纪80年代,德国社会学家乌尔里希·贝克(Ulrich Beck)提出了

①Curzon J. ,Kosa T. A. ,Akalu R. ,et al. ,Privacy and artificial intelligence,IEEE Transactions on Artificial Intelligence,Vol. 2,No. 2,2021,p. 96.

②Van der Heijden J. ,Risk as an approach to regulatory governance:An evidence synthesis and research agenda,Sage Open,Vol. 11,No. 3,2021,p. 4.

风险社会理论，该理论认为，现代社会存在复杂且相互关联的风险，需采用新的治理方式应对。在风险社会理论的影响下，风险监管的概念逐步系统化，至今成为较为成熟的监管理论。对于风险监管体系而言，需遵循以下的基本模式：第一，将危害构建为风险；第二，将风险监管视为一整套法律工具；第三，针对不同风险对象采取差异化的监管措施[①]。其核心在于：基于风险评估对不同风险进行系统的资源配置[②]。虽然风险监管理论已在多个领域适用，但从风险监管理论的基本内涵来看，该理论应用于人工智能领域的分类分级监管，却有诸多不足之处。

第一，风险划分对部分人工智能带有先天性负面判断。对人工智能进行预先的风险等级划分往往基于未经实践验证的假设和推测，这种风险判断带有高度的主观性[③]，容易导致对特定人工智能技术的先天性歧视。首先，风险的认定和分配本身就是一种社会价值负面衡量，通过将人工智能的潜在危害视为风险进行监管，无论监管者是有意抑或是无意，都是对人工智能进行先天负面判断，而这种判断与当下的社会文化、政治环境和个人偏见息息相关，反映了特定社会群体的思想和诉求。其次，技术本身具有中立性，只有在具体的应用实践中，根据技术实现的效果，才能分辨该技术是“好的”或者是“坏的”。预设技术的潜在负面效应，可能会引发不公正的风险等级划分，甚至可能会导致社会对特定人工智能技术的排斥或过度监管。某项技术在一个领域可能带来风险，但在另一个领域却可能产生正面效应，即便是引发社会普遍担忧的人脸识别技术，其应用于公共安全保障、身份核验安全等领域亦发挥了显著的积极作用[④]。因此，预设技术具有风险是不公平的，对人工智能风险的评估需要基于科学的数据、技术、实效分析，而不能仅仅基于技术的潜在负面影响来进行盲目地划分。

第二，对人工智能风险难以进行量化评估。根据风险监管理论，风险监管更适用于解决可量化的问题，对无法量化的风险则适用效果不佳。实际上，人工智能引发的风险多种多样，并非都可以直接量化，经常存在难以界定的争议[⑤]。一方面，对人工智能进行量化评估缺乏有效的数据。丰富的数据集是人工智能风险预测的基础[⑥]，但由于世界范围内人工智能产业正处于发展的早期阶段，相关应用刚刚起步，缺乏丰富的应用反馈数据，对人工智能难以进行评估量化；另一方面，全球尚未形成标准的量化评估模式。尽管欧盟《人工智能法》提出了基于损害程度和发生概率的双重评估框架，能帮助监管机构更准确地识别和划分人工智能的风险等级，但目前并没有形成成熟的评估体系，盲目地对人工智能风险进行划分不仅难以得出合理客观的评估结果，还可能浪费大量监管资源，与风险监管最初的降本增效理念背道而驰。

第三，僵化的风险分级模式难以适应人工智能的动态发展。风险分级是风险监管的关键，但由于人工智能技术和应用的多变性与复杂性，固定不变的风险分级模式将难以适应人工智能的发展。一方面，人工智能系统不断迭代，其未来发展难以预测，社会影响也无法预见[⑦]，固定的分级模式将难以涵盖人工智能技术演进带来的新型风险；另一方面，即使某一人工智能的风险等级已确

①Kaminski M. E. , Regulating the Risks of AI, Boston University Law Review, Vol. 103, 2023, p. 1389.

②参见刘鹏，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。

③Chamberlain J. , The risk-based approach of the European Union's proposed artificial intelligence regulation: Some comments from a tort law perspective, European Journal of Risk Regulation, Vol. 14, No. 1, 2023, p. 7.

④参见赵精武：《人脸识别技术应用的利益权衡与合法性认定》，《法律科学》（西北政法大学学报），2024年第1期。

⑤Willis L. E. , Performance-based consumer law, University of Chicago Law Review, Vol. 82, 2015, p. 1309.

⑥Hilb M. , Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance, Journal of Management and Governance, Vol. 24, No. 4, 2020, p. 857.

⑦De Almeida P. G. R. , dos Santos C. D. , Farias J. S. , Artificial intelligence regulation: a framework for governance, Ethics and Information Technology, Vol. 23, No. 3, 2021, p. 507.

定,其未来亦可能随技术发展而产生新风险,但固定的风险等级可能难以快速调整、改变监管框架需要经过长时间的程序流程,这就限制了监管机构对市场变化的快速响应。因此,尽管欧盟试图通过分类分级的监管框架平衡人工智能的发展与安全,但这种严格的事前审查和合规要求增加了企业的运营成本,可能对部分中小型人工智能企业的发展造成阻碍。

由此,在风险监管理论的视角下,回观前文提到的以风险为基准进行分类分级监管的底层逻辑,不难发现,该逻辑在现实中难以实现。首先,对于风险评估来说,尽管评估模式会对风险损害程度和风险发生概率进行综合考虑,但这均是在两者可以量化的前提下,由于现实中数据和有效评估方法的缺失,对损害程度和发生概率均难以衡量,更难以进行精准的风险评估。其次,对于成本控制来说,人工智能预先的风险分级本身就需要耗费大量的成本,特别在人工智能难以量化评估的情况下,其意图实现的成本效益最大化便更无从谈起,反而会造成更多不必要的成本支出。最后,对于人工智能的可控性来说,设置市场准入门槛并进行严格监管,这的确会在一定程度上加强对人工智能的管控,但这种强监管也不免限制技术的自由探索,从长远来看,可能会抑制人工智能技术的创新,不利于人工智能产业的进一步发展。

(二)基于价值理论的重要性导向分类分级监管模式构建

虽然以风险为基准的人工智能监管模式已经成为世界的主流,我国也开展了类似的实践,但并不意味着该模式是唯一适用于我国的人工智能分类分级监管思路。面对风险监管的种种弊端,世界范围内,人们亦在探索不同于风险监管的其他路径。2022年7月,英国发布了人工智能监管新文件——《建立一种支持创新的人工智能监管方法》[①],该文件提出了一种"以结果为导向的监管方法",其认为,人工智能产生的风险主要取决于实际应用的场景。文件明确提出,英国将根据人工智能在特定应用中可能产生的结果进行监管,并非预先将风险级别与技术进行绑定。此外,美国国会参议院两党议员也于2023年11月联合提出了《2023人工智能研究、创新和问责法案》[②],该文件提出了以"影响力"为划分的分类分级监管模式,重点关注关键影响和高影响力的人工智能,其要求关键影响人工智能责任人在部署前需要每两年进行一次风险评估,高影响力人工智能责任人需要对实施情况进行年度报告。

我国目前处于人工智能产业发展的关键赶超阶段,未来的人工智能发展路径尚不明朗,现阶段不宜采取风险分级的方式对处于萌芽状态的人工智能产业过度预防,而应坚持审慎的原则为人工智能产业发展划定红线[③],营造鼓励创新的监管氛围,采用更加柔性的方式来应对治理过程中的不确定性[④],促使我国人工智能产业赶超世界领先水平。

在此背景下,我国应基于国内人工智能产业发展实际,采取"以重要性为导向"的分类分级监管模式。具体而言,该模式即是在人工智能的监管中,一切以重要性作为逻辑起点与划分依据对人工智能进行分类分级监管,对重要的人工智能投入更多的监管资源,而对其他人工智能则放宽限制。从理论基础来看,价值衡量是重要性导向的底层逻辑。任何事物的存在均有其特定的价值,价

①A pro-innovation approach to AI regulation, https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#section322, 2025.03.29.

②The AI Research, Innovation and Accountability Act of 2023, https://www.thune.senate.gov/public/_cache/files/7dea8daa-f6d1-4881-ad21-2381fcbe0785/6362CE1D0A17743166BC170A593B5CDA.ccaskfall23a15.pdf, 2024.11.15.

③参见张凌寒:《中国需要一部怎样的〈人工智能法〉?——中国人工智能立法的基本逻辑与制度架构》,《法律科学》(西北政法大学学报),2024年第3期。

④参见贾开,赵静,傅宏宇:《应对不确定性挑战:算法敏捷治理的理论界定》,《图书情报知识》,2023年第1期。

值是事物对人类主体的对象性关系，是客观性与主观性的辩证统一[①]，其可通过肯定事物自身价值、人之自觉价值存在等方式体现[②]。由此，通过对事物价值性高低的衡量来判断其重要性的大小，便成为“以重要性为导向”监管模式实施的基础。

在该模式下，人工智能价值高低的判断是关键问题。党的二十大报告曾提出“世界各国弘扬和平、发展、公平、正义、民主、自由的全人类共同价值”这一倡议，为我们提供了一个以人为本的多维价值性衡量体系，结合社会主义核心价值观，可以从客观与主观两方面结合以判断人工智能的价值高低。

首先，在客观方面，事物在现实世界经济结构中的意义是其客观价值的生动体现[③]，衡量人工智能在社会经济活动中展现出的自身价值是判断其价值高低的重要因素。目前人工智能给各行各业带来了巨大变革。例如，人工智能赋能制造业生产，使其在提高效率的同时也降低了成本；人工智能助力医疗诊断，显著提升了疑难病症治疗效果。由此，人工智能已在经济活动中发挥了重要作用，可以通过效率提升、成本降低、病患治愈率等具体指数反映人工智能在其中发挥的价值。

其次，在主观方面，人对事物不可量化的感知也是价值性体现的关键，基于经验和既有发展基础对人工智能价值进行衡量是主观判断的重要方式。人工智能不仅是一个技术工具，其未来发展的潜在影响性和不可知性使得在评估价值时，必须考虑人的主观体验和社会预期。在这一背景下，评估人工智能的价值应关注其与人类发展的契合度。在推动生产力发展的同时，遵循以人为本、实现共同富裕的目标导向是其价值的重要体现。

客观数据提供了价值判断的现实基础，主观感知则引导人们关注技术进步带来的伦理和社会问题。将客观数据与主观感知相结合，可以形成一个多维的价值衡量体系，以此综合判定人工智能重要性的大小，进而进行分类分级监管，确保人工智能在促进社会经济发展的同时，能够真正服务于全人类的共同利益。

与此同时，一个重要的问题便是“重要性”与“风险”两词似有相同之意，尤其在人工智能领域，类似“越重要的人工智能意味着更高的风险”这样的理解容易产生。然而，这样的理解存在偏颇。虽然两者之间存在联系，但其基本内涵截然不同。从词语的基本语义来看，“风险”一词通常指的是消极结果的可能性和严重性。在人工智能领域，风险通常与负面事件相关，对国家安全、网络安全、数据安全等重要领域的潜在威胁均属于典型的风险表现[④]。而“重要性”一词则关注事物的整体价值，不仅包括可能的消极效应，也包含可能带来的积极效益。在这种情况下，对人工智能进行“风险”判断，将会对人工智能造成或大或小的负面评价，不免抑制人工智能产业的创新发展，而对人工智能进行“重要性”判断则不会有此担忧。由此，相比于“风险”的单一维度，“重要性”从正面和负面两重维度进行衡量，有助于更好地平衡监管策略。

（三）以重要性为导向分类分级监管模式的优势

以重要性为导向的分类分级监管模式不采取禁止或限制性措施来抑制人工智能技术的发展，而是通过明确的监管红线，来确保这些重要人工智能的安全运行和技术创新。这种以重要性为导向的监管策略，相比于风险监管具有不可替代的优势，其反映了对未来技术发展趋势的前瞻性思考和高度的责任感，为全球人工智能的监管提供了新的思路和框架。

第一，该模式对人工智能没有预先的歧视性判断。在传统的风险分级监管模式中，对不同的人工智能进行分类分级往往涉及对其潜在风险的主观判断，相比之下，以重要性为导向的动态监管

①参见傅其林：《中华优秀传统文化蕴含全人类共同价值的理论、历史与现实》，《江西社会科学》，2023年第7期。

②参见李景林：《中国哲学“本义的价值”观念及其结构层次》，《中国社会科学》，2024年第5期。

③参见宋希仁：《黑格尔论“价值”》，《高校理论战线》，2002年第2期。

④参见刘艳红：《生成式人工智能的三大安全风险及法律规制——以ChatGPT为例》，《东方法学》，2023年第4期。

模式则采取了一种更为客观的视角。根据人工智能在具体应用中的重要性来划分,这种模式更多地考虑了人工智能的实际功能和作用,而非所谓的“预期目的”,即开发该系统的最初目的,因为最初开发目的可能并非实际使用效果[①]。这种动态监管模式通过针对具体应用场景对人工智能进行价值衡量,使得监管措施能够更加精准地对接技术的实际用途,也能有效预防和控制潜在的风险。

第二,该模式侧重事中的动态监管提升了监管的灵活性。风险分类分级监管受制于固有的风险划分模式,使得监管机制日益僵化。以风险为基准的分类分级模式要求相关主体在部署人工智能之前,需满足必要的安全和道德标准[②],而以重要性为导向的监管体系强调在产品进入市场后实施动态追踪和管控,通过这种方式,监管不再是一个静态的、一次性的门槛,而是一个持续的、动态的过程。而且重要性的划分也将随着技术的更新和社会的进步而变化,为应对未来难以预测的变革留出了必要的调整空间。这种监管模式的动态性和响应性不仅提升了监管的实效,也通过降低合规成本减少了市场进入障碍。

第三,该模式避免了风险分级和控制的高成本。复杂的风险分级通常基于对潜在风险的预测,定量评价智能产品与服务的风险时,需要从人工智能的性能、效率、公平性、透明度及可解释性等多个角度进行综合分析[③],这要求监管机构投入大量时间和资金来收集数据、进行复杂风险评估和制定分类分级策略,而且这种方法经常依赖于未经验证的信息和假设,容易造成监管资源的不合理投入,从而耗费大量监管成本。而以重要性为导向的分类分级监管则恰恰相反,其侧重于实际应用的影响,直接将监管重点放在人工智能的实施效果上,使监管机构专注于真正重要的人工智能,避免了分级的繁琐流程和支出成本,能够更有效地投入监管资源。

三、人工智能分类分级监管的制度革新

未来人工智能的发展将颠覆各行各业[④],而目前的分类分级现状已难以满足我国对人工智能的治理需求,亟需以重要性为导向的监管模式革新人工智能的分类分级治理路径。具体来说,在人工智能的分级中,可借鉴《人工智能法(学者建议稿)》的思路,将人工智能划分为关键人工智能和一般人工智能,对关键人工智能施以更多的监管义务;在人工智能的分类中,沿用以场景为划分的分类思想,一般情况下遵循“先分类,后分级”的监管思路,关注医疗、教育、金融等重点领域的分类治理措施;面对人工智能的分层业态,制定针对性的治理策略,基础模型层需提升数据质量水平,专业模型层应注重确保模型的可信可控,服务应用层则侧重保护用户合法权益。

(一)创设以重要性为导向的分级方法:划分关键、一般人工智能

在原有监管模式愈发无法满足我国人工智能治理需求的背景下,2024年4月,由中国政法大学等单位联合起草的《人工智能法(学者建议稿)》突破了以风险为基准的分类分级监管框架,提出对关键人工智能实施重点监管。在以重要性为导向的分级监管中,可借鉴该建议稿的思路,将重要性较高的人工智能定义为关键人工智能,其他为一般人工智能。在这一框架下,如何认定关键人工智能便成为一个至关重要的问题。根据人工智能的发展特性,可以从社会影响、人身安全、技术指标

①参见张涛:《人工智能治理中“基于风险的方法”:理论、实践与反思》,《华中科技大学学报》(社会科学版),2024年第2期。

②Malgieri G. , Pasquale F. , Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology, *Computer Law & Security Review*, Vol. 52, 2024, p. 2.

③参见陈吉栋:《以风险为基础的人工智能治理》,《法治研究》,2023年第5期。

④Peter Diamandis, A. I. Will Lead to Massive Disruption Across Industries, https://www. inc. com/tess-townsend/diamandis-artificial-intelligence. html, 2024. 09. 24.

三个维度进行分析，以期为人工智能的政策制定和技术开发提供方向性指引。

在社会影响方面，关键人工智能所涉及的领域通常具有广泛的社会效应和潜在风险。凡是在军事国防、财产安全、社会稳定等方面产生重大影响，或涉及关键基础设施的技术，均可归类为关键人工智能。具体而言，应用于军队、情报机构、国防工业等国家安全领域的人工智能，直接关系国家的战略利益与防御能力，是国家安全的关键支撑；在金融、保险等领域广泛应用的人工智能技术可能对财产安全造成重大影响，一旦出现漏洞，将引发大规模的经济损失，对经济稳定造成严重冲击；应用于社会治理、公共管理等领域的人工智能技术在维护社会秩序和防范风险方面具有显著作用，甚至逐渐形成算法权力与行政权力二元共治的新秩序[①]，其包括用于预测犯罪、舆情监测、社会风险评估等的人工智能系统，而这些人工智能技术运用不当会给社会带来巨大的威胁；电力、通信、交通、能源等领域的基础设施是社会运转的基石，其智能化水平的提高将有效促进创新资源和生产要素的优化配置[②]，这些基础设施所依赖的人工智能一旦出现中断或失效将导致严重的社会后果。由此可见，在社会影响方面具有广泛社会效应和潜在风险的人工智能，无疑是关键人工智能，需要高度重视并加强管理和监督。

在人身安全方面，关键人工智能所带来的影响直接关乎人的基本权利。凡是可能对人的健康、自由、尊严和隐私等人身基本权利产生重大影响的人工智能，均应视为关键人工智能。例如，涉及生物识别技术的人工智能可能关系到性别、宗教、种族和身体残疾等敏感信息，一旦被滥用，可能会导致严重的人身危害，应被视为关键人工智能进行监管。同时，某些人工智能可能具备操控个人意识和行为的能力，这类技术一旦被不当使用，将造成不可逆转的危害后果，亦应被重点关注。此外，深度嵌入教育、就业、医疗、司法等影响人身基本权利领域的人工智能也应纳入关键人工智能的范畴，因为这些领域的决策与执行一旦使用歧视性的算法，可能会导致对特定群体的不公正待遇[③]。

在技术指标方面，关键人工智能可基于其算力、参数等技术特征进行界定。通常情况下，凡是在算力、模型参数等方面达到较大规模的人工智能，均可归为关键人工智能。从实际情况来看，我国目前可以将关键人工智能浮点算力阈值设置为10^{25}（FLOPs），将参数阈值设置为万亿，凡是超过这些阈值的人工智能，便被认定为关键人工智能。通过这一标准为技术监管提供量化依据，同时推动各行业共同构建一个更为安全、透明的人工智能生态系统。

满足上述三个维度之一，便可认为该人工智能足够重要，应纳入关键人工智能的范围。此外，对于关键人工智能的认定标准并非一成不变，三个衡量维度指标将会随着社会的发展而与时俱进、动态更新，这种动态框架将会使规则制定摆脱传统约束和惯性思维，以满足分类分级的即时需求。以技术指标为例，随着未来硬件的优化和算法的更新，10^{25}（FLOPs）的浮点算力和万亿级的参数阈值将明显低于实际需求，其判定阈值必将逐步提高。由此，可成立相关的评估委员会，负责动态调整人工智能重要性判断标准以满足现实发展需要。

在具体监管措施上，监管部门负责对关键人工智能进行认定，并承担对关键人工智能开发者、提供者的告知义务，接到监管部门通知的关键人工智能开发者、提供者才需履行对该人工智能的额外主体义务。更重要的是，以重要性为导向的监管模式不对人工智能设置市场准入门槛，对一般人工智能而言，其仅需满足基本的备案义务，能够让监管机构了解算法模型的基本情况并把握其

①参见齐延平，曹立军：《智能行政中基本人权的保护路径》，《学术交流》，2024年第12期。

②参见李政，周希禛：《新时代增强国有经济创新力：理论内涵、现实意义、独特优势与实现路径》，《马克思主义与现实》，2022年第5期。

③参见罗洪洋，李相龙：《智能司法中的伦理问题及其应对》，《政法论丛》，2021年第1期。

潜在的风险因素即可[①],不必对其施加过多的监管措施,避免对一般人工智能的过度干预。但需对关键人工智能开发者、提供者提出更高的监管要求,构建多维的人工智能监管体系。

在关键人工智能部署前,要求开发者和提供者履行备案义务是提高透明度、保障安全的重要手段。人工智能备案机制旨在确保开发者和提供者对人工智能的开发和应用进行必要的信息披露,这些信息包括开发目的、技术部署背景、使用的数据类型和来源、算法模型的选择与设计、预防风险的机制等,在一定程度上确保人工智能的可解释[②],一旦系统出现问题,备案信息可以作为事后追责的基础,确保责任归属明确。同时,关键人工智能的开发者和提供者还应建立完善的应急处置机制,一旦发生安全事件,相关的应急机制需能迅速启动,采取相应的补救措施。

在关键人工智能应用过程中,可以对该人工智能施以检测和追踪措施,通过技术检测手段追踪系统的运行状态和数据变化,确保其符合预期标准,以保障系统的安全性和稳定性。为进一步提升监管效果,关键人工智能开发者和提供者还应每年至少进行一次人工智能安全评估,定期报告人工智能的应用情况和风险监管措施,以便监管部门能够了解和掌握人工智能的运行状态,及时作出相应的政策调整。

一旦关键人工智能发生事故,便需对相关责任人进行严肃追责,确保相关系统开发者、提供者对其人工智能的安全性负有明确的责任义务,形成完善的监管体系和市场秩序,从而有效倒逼人工智能开发者和提供者在系统开发与运行中更加谨慎。同时,还应对关键人工智能的实施效果展开评测,收集来自系统用户、监管机构、技术专家等多方的反馈,及时发现关键人工智能系统中的问题,以改进和优化关键人工智能的部署应用。

(二)优化基于场景的分类措施:着力关注重要领域

以场景为划分的分类措施在数据、算法、平台治理中均发挥了积极的作用。人工智能的分类虽不再遵循单一要素的分类模式,但在现阶段,人工智能治理的主要任务之一,便是实现人工智能技术的设计、开发和部署的价值协调[③],对人工智能按照应用场景进行分类监管仍然是一种合理且有效的选择。通过将人工智能按场景进行分类,监管部门能够更好地识别哪些技术会带来更大的影响,哪些领域需要更严格的监管措施,尤其是在复杂应用中可以有效解决监管难题、提升人工智能治理的效率,但也应对该模式进一步优化以适应人工智能的迭代发展。

面对人工智能分类和分级过程中可能存在的相互交叉和混乱问题,在以重要性为导向的监管模式下,可遵循以下的分类分级解决思路:一般情况下按照"先分类,后分级"的流程对人工智能进行监管。也就是说,首先应当根据人工智能所涉及的具体应用场景对其进行分类,然后在分类的基础上进一步对其进行分级。对于同一人工智能涉及多个场景难以分类的情况,分类的优先级可以依照其最重要的应用场景进行判断,在此基础上,监管部门再根据人工智能的重要性对其进行进一步的分级处理。此种分类分级模式为监管部门提供了清晰的框架,使其能够更准确地评估和把握人工智能的潜在影响,有利于将监管资源合理分配到最需要的领域。值得注意的是,人工智能不仅仅是特定领域的技术工具,其已然成为一种"通用技术"[④],在各个领域中发挥着越来越重要的作

①参见刘博涵:《智能投顾算法透明的价值辩证、规制导向与实现路径》,《深圳大学学报》(人文社会科学版),2024年第1期。

②参见周辉:《人工智能可解释的制度建构》,《山东师范大学学报》(社会科学版),2025年第1期。

③Novelli C., Taddeo M., Floridi L., Accountability in artificial intelligence: what it is and how it works, *Ai & Society*, Vol. 39, No. 4, 2024, p. 1880.

④Crafts N., Artificial intelligence as a general-purpose technology: an historical perspective, Oxford Review of Economic Policy, Vol. 37, No. 3, 2021, pp. 523-524.

用。未来，随着人工智能技术的进一步发展，它将广泛应用于社会的各个方面，包括医疗、教育、金融等重要领域。在以重要性为导向的监管模式下，不同场景的分类监管需要特别关注这些重点领域的特殊要求，以确保人工智能技术的合理应用和规范发展。

在医疗领域，人工智能通过深度学习和数据分析，能够协助医生进行病情诊断、患者监护、药物研发等工作。当前，人工智能技术已在辅助医疗诊断、个人健康监测、慢性病治疗和传染病防控等方面展现了显著的优势[①]。相关人工智能必须遵守严格的医疗标准和伦理准则，以确保系统的可靠性和安全性。同时，由于医疗领域涉及患者隐私、健康安全等敏感问题，监管机构应要求医疗人工智能系统设置严格的保密机制，以防止算法偏差和信息泄露对患者造成负面伤害。

在教育领域，人工智能技术正迅速渗透到个性化学习系统、智能题库、在线教育平台等场景中。教育领域的人工智能应用必须特别注意公平性问题，技术的应用应确保每位学生都能平等地获得学习机会，警惕算法自动化决策而带入人的主观取向和价值判断[②]。此外，教育领域的人工智能还应强调正确价值观的传递，系统设计应符合教育部门的相关规定和指导方针，实现人工智能对正义、诚实、责任、关怀及坚毅等美德的潜移默化引导[③]。

在金融领域，人工智能技术应用涵盖了证券交易、风险评估、合规管理等多个重要业务领域[④]。借助人工智能，金融机构能够处理和分析大量的金融数据，并结合客户的资产情况和投资目标，制定更加精准的投资策略。由于金融市场高度复杂且敏感，人工智能系统可能因数据泄露、算法不透明或系统漏洞而引发市场混乱或投资者损失。为此，金融监管机构应加强对金融人工智能系统的透明性要求，确保系统的算法和决策过程具有足够的可解释性。

（三）实施人工智能多元化监管策略：适配各分层业态重要特点

在对人工智能开展以重要性为导向的分类分级整体监管后，虽然人工智能不同业态之间的相互影响仍然存在，但重要性导向的监管思路将不同业态的人工智能分类分级问题进一步简化。对于各层级人工智能相互影响导致分类分级困难这一问题来说，在重要性导向下，某一人工智能无论处于哪一层级，均应从其自身出发，将不同层级之间的联系也作为重要性判断的因素之一，利用重要性衡量的三个维度判断该系统的重要性，从而实施分类分级监管。同时，人工智能产业呈现的“基础模型——专业模型——服务应用”分层业态各层均有其独特之处，为更好地回应分层业态的治理需求，我国应遵循重要性导向，以各层的重要特点为抓手，实施多元化的监管策略。

第一，在基础模型层注重提高数据质量。基础大模型为多种专业模型和服务应用提供底层支持，实现了从“一个应用一个算法”到“一个模型多用途”的转变，而数据是基础模型的根基，基础模型的好坏很大程度上取决于数据质量的高低，结果的准确性是由其输入数据的准确性直接决定的[⑤]，当人工智能模型吸收了低质量的数据，该模型将会接受有偏差的训练，从而导致错误推理结果的输出。由此，在人工智能基础模型层，监管重点应转向大模型的训练数据集来源、设计和训练过程。面对具有广泛应用潜力的基础模型，应提高其透明度义务，要求基础模型服务提供者在进行备案时提供模型设计、数据来源的基本信息，确保训练数据遵守数据保护法律和伦理标准。

---

①参见石佳友，徐靖仪：《医疗人工智能应用的法律挑战及其治理》，《西北大学学报》（哲学社会科学版），2024年第2期。

②参见马长山：《数字时代的法学教育转型》，《上海政法学院学报》（法治论丛），2023年第1期。

③参见白钧溢，于伟：《教育人工智能伦理原则的有限性及其补正——兼论“原则-美德-法律”框架》，《中国电化教育》，2024年第2期。

④参见赵炳昊：《金融领域中人工智能应用的监管与法律规制》，《郑州大学学报》（哲学社会科学版），2024年第3期。

⑤参见商建刚：《生成式人工智能风险监管元规则研究》，《东方法学》，2023年第3期。

第二,在专业模型层注重模型的可信、可控。有效治理人工智能,保障人工智能的可信、可控是世界共同面临的课题[①],专业模型往往基于基础模型的调整和优化而来,其一般针对特定的场景和行业。在这些具体的领域,某些特定的场景和行业可能会对人工智能提出特殊的监管要求,以便于人工智能在特定场合下的负责任使用。专业模型层的细分服务领域也对人工智能大模型输出的精准度提出了更高的要求,特别是在模型应用于医疗、金融、交通、教育等重要领域时,其系统本身的可信、可控显得尤为重要。由此,在该模型层,一方面应重点关注由基础模型转为专业模型过程中调整和优化的安全性和适应性,保证专业模型设计的可信性;另一方面,专业模型在处理和输出各项运算任务时,应进行一定的监督,以确保专业模型的可控性。

第三,在服务应用层关注用户的合法权益。经历了基础模型的数据质量优化和专业模型的可信可控调整,服务应用层人工智能将被用户直接使用,其直接关系着用户的切身利益。由此,服务应用层的监管重点应放在用户的合法权益保护上,可以从数据传输、数据存储和数据处理全流程出发,通过全流程监督人工智能保护用户隐私、确保数据安全,确保服务应用层的人工智能能够充分保护用户权益。同时,为了方便辨识和追溯,应对该层人工智能生成内容添加标识,进一步提高生成内容的准确性和可靠性。监管机构亦可采取定期审查与用户反馈的方式,掌握服务应用层人工智能的运行效果,督促其持续改进并提高服务质量。

四、结 语

2025年4月26日,习近平总书记在主持中共中央政治局第二十次集体学习时指出:“要加紧制定完善相关法律法规、政策制度、应用规范、伦理准则,构建技术监测、风险预警、应急响应体系,确保人工智能安全、可靠、可控。”人工智能的分类分级监管将是完善人工智能治理规则的重要一环,但如何科学合理地对人工智能进行分类分级监管,是推动人工智能技术健康发展的关键。以风险为基准对人工智能进行分类分级监管存在诸多弊端,并非适用于我国。我国应立足于目前国内人工智能产业发展实际,设计一种既促发展又保安全的人工智能治理路径,而以重要性为导向的分类分级监管模式恰好符合这一特性,其从人工智能的价值导向出发,提供了一种避免预先负面判断、更加灵活和更低成本的动态监管模式。以重要性为底层逻辑区分场景应用的同时划定关键人工智能实施差异性治理,不仅适应我国人工智能产业的发展现状,也为未来的技术迭代预留了足够的灵活空间,其不失为我国未来对人工智能分类分级监管的有效方式。

[责任编辑:郑 毅]

①Radu R.,Steering the governance of artificial intelligence: national strategies in perspective,Policy and society,Vol. 40,No. 2,2021,pp. 180-181.