文章编号:1003 - 2053(2026)04 - 0673 - 08

人工智能全球治理前景分析

牟砚堂[1],袁和静[2]

(1. 清华大学马克思主义学院,北京 100084; 2. 北京中医药大学马克思主义学院,北京 102488)

摘 要:人工智能高速发展所衍生的严重风险日益凸显,其治理议题已成为全球学界、各国政府和国际组织共同关注的重大时代难题。通过剖析人工智能的风险应对路径,研究发现全球性风险治理常遵循"科学共识驱动 - 政治议程响应 - 全球协同治理"的递进模式。人工智能领域快速形成了从风险识别到全球共治的广泛共识,并逐步转向全球政治行动议程,各主要国家、国际组织正积极推进治理进程并获得公众响应支持。然而,国家利益冲突、技术挑战、认知分歧壁垒等却成为人工智能全球治理的阻碍因素。展望人工智能全球治理,短期内可能实现伦理准则细化、风险评估标准化、多边机制协商化;中期内可能建立技术治理联盟、动态分级监管体系及全球治理框架雏形;长期愿景将实现人机价值共生、构建人机协同数字孪生治理生态,最终形成真正的全球智能治理共同体。

关键词:人工智能;科学共识;政治议程;全球治理;前景分析

中图分类号:G315 **文献标识码**:A

习近平总书记高度关注人工智能风险,并明确指出"人工智能治理攸关全人类命运"[1]。人工智能作为全球性的重大公共议题,其负面影响日益成为世界各国面临的共同课题和全球性挑战。《2024 年全球风险报告》评估的全球十大短期风险中,人工智能被视为全球最大风险之首,位列气候变化、经济疲软和战争之前[2]。人工智能治理作为全球安全治理的重要议题,需要思考在最大化发挥人工智能对人类的益处时,如何防控并最小化其风险。

人工智能风险应对成为日益突出且两极分化的问题。各国政府和国际组织正积极构建人工智能治理机制,但人工智能全球治理目前仍存在诸多模糊不清之处,诸如为何要进行治理、如何治理、由谁治理以及治理对象究竟是什么?这些关键问题都尚未明确[3]。全球合作是一个难以明确且存在多种解释的概念,鉴于真正意义上实现"全球"合作的成功案例有限,人们甚至质疑"全球合作"这一概念本身[4]。人工智能起源于科技发展,但其治理行动却主要涉及政治层面,那么科学领域与政治领域存在怎样的互动?科学共识、政治议程、全球治理之间有着怎样的关联?因此,有必要且能够从科学史上全球性重大问题的治理经验中,为人工智能治理汲取历史经验,探析全球治理的未来前景。

1 人工智能全球治理的现状诊断

人工智能风险及治理已引起科学共同体关注并达成科学共识。这一共识从科学界逐渐传导至全社会,最终演变为全球性共识,为全球治理议程奠定了科学基础。

1.1 科学共同体达成科学共识

科技发展常成为重大公共问题的直接或间接诱因。科学家作为风险的"洞察者"和"预警者",迅速就人工智能风险认知、伦理原则及治理行动等方面达成广泛共识。

科学界持续深化风险认识,并形成广泛科学共识。科学研究具有重要先导作用,从早期关注技术脆弱性,到中期关注伦理道德与社会影响,再到近期全面关注并实施综合治理,使得对人工智能风险认识不断走向成熟。尽管在人工智能的定义、实现路径、伦理道德、社会影响等多方面仍存在分歧,各国在人工智能发展水平和利益诉求上也各有不同,但在人工智能的潜在风险方面已形成广泛共识。2015 年以来,斯蒂芬·霍金、埃隆·马斯克及图灵奖获得

收稿日期:2025 - 03 - 19;**修回日期**:2025 - 05 - 06

基金项目:国家社会科学基金重大项目(22ZDA027)

作者简介:牟砚堂(1994 -),男,土家族,博士研究生,E - mail:350377814@ qq. com。袁和静(1984 -),女,教授。

者约书亚·本吉奥等纷纷呼吁重视人工智能对人类生存的潜在威胁,并首次将“生存性风险”引入公众视野。2017 年,美国电气和电子工程师协会发布了《人工智能设计的伦理准则》;2018 年,多国联合制定了《阿西洛马人工智能原则》,确立人工智能应遵循安全性、有益性、透明性等 23 条原则,标志着全球科学界就伦理和风险达成广泛共识。人工智能风险治理仅 60 余年便形成全球性共识,而气候共识历经 200 余年方才形成,这种加速特征凸显了全球对人工智能风险应对的高度关注与重视。

表 1 人工智能风险共识时间轴

Table 1 AI risk consensus timeline

关键阶段	时间	重要事件	时长计算
概念提出	1956 年	达特茅斯会议,AI 正式成为独立学科	—
风险认识	2015 年	霍金、马斯克、本吉奥等公开警示 AI 生存性风险	1956 – 2015 年 = 59 年
科学共识形成	2017 – 2018 年	2017 年:IEEE 发布《人工智能伦理准则》 2018 年:制定《阿西洛马人工智能原则》	2015 – 2018 年 = 3 年

1.2 全球范围内形成广泛共识

人工智能风险共识正加速向社会各领域渗透,国际组织、行业组织、学术团体、高科技公司、社会公众等纷纷接受并支持这一科学共识,人工智能治理逐渐演化为全球性共识[5]。

2024 年,联合国大会通过首份人工智能全球性决议,120 余个会员国就安全、可靠的 AI 系统达成全球共识。哈佛大学收集了全球几大洲 36 份人工智能伦理原则文档,发现全球大致形成了八项伦理共识。在这些文档中,所有文档提及“公平公正”,97% 包含“隐私保护”和“可追责”,94% 提及“透明与可解释性”,81% 涉及“安全”方面,78% 强调“专业研发责任”,69% 体现“人类价值”与“在人的控制之下”,64% 涵盖“人权”相关内容[6]。2022 年,皮尤中心针对美国成年人的调研显示,关于人类能否开发出公平决策的算法,28% 的人持否定态度,41% 持不确定态度,30% 认为有可能。面对人工智能的广泛运用,45% 的人期待与担忧并存,37% 倾向于担忧,18% 期待盖过担忧。74% 的人对人工智能替人类做重要决策深感担忧,75% 对人工智能洞悉人类想法和行动感到担忧,这两类应用场景成为最被担忧的问题[7]。在美国另一项调查中,74% 的受访者认为全球变暖主要源于臭氧消耗,而仅 22% 的人知晓“超 90% 气候学家支持人为导致全球变暖”,仍有 24% 的人认为专家间存在巨大争议[8]。对比气候变化、核应用、基因编辑等公共问题,全球公众对人工智能风险的认知更为全面准确,为全球治理政策筑牢科学根基与民意基础。

2 人工智能全球治理的发展动能

科学共识常作为风险治理的起点,各国政府和国际组织随后将其纳入政治行动议程。人工智能风险共识迅速转化为政治行动力,科学共识作为“门户信念”有力地促进了公众的响应与支持。

2.1 主要国家积极展开战略布局

人工智能治理充分展现了科技发展促使风险显现,科学共识推动政治行动的过程。在治理进程中,各方围绕数据资源、市场份额及技术标准等展开了竞争与博弈。

中美欧等主要国家积极将人工智能发展及治理纳入政治议程。2016 – 2017 年,各国密集出台一系列人工智能国家战略,以保持或加强其国际“领导”地位,致使全球呈现出竞争为主的格局[9]。2023 年后,人工智能进入“安全治理”阶段,各国政府纷纷出台政策加强监管,在数据、算法、伦理等方面发布了一系列管控文件。2023 年,美国发布《2023 年人工智能研究、创新和问责法案(AIRIA)》;2024 年,中国公布《生成式人工智能服务管理暂行办法》《人工智能安全治理框架》及《人工智能法(学者建议稿)》;日本也在同年发布了《关于 AI 制度的思考》《防卫省推进人工智能有效应用的基本方针》及《运用生成式 AI 制作发行物的指导手册》等文件。

人工智能治理正逐步成为各国的焦点议题。美国秉持“在发展中求安全”的理念,既积极推动人工智能领域的跨国合作,又激烈与中国等国家展开技术竞争和治理主导权之争。欧盟坚持“在安全中求

发展”的理念,积极推动国际合作以应对人工智能带来的全球性挑战。2023 年,日本将人工智能列入七国集团(G7)的议程,并通过“广岛人工智能进程”。2024 年,意大利作为 G7 轮值主席国,其总理焦尔吉娅·梅洛尼表示应对人工智能带来的风险和支持非洲发展将是该年度的两个关键主题。多个主要国家展开的这一系列政治议题体现了对人工智能治理的高度重视。人工智负面影响“国内优先”特性更为严重,同气候变化、臭氧层破坏等全球共担风险存在差异,这一系列政治议题体现了各国积极谋求技术发展与风险治理的协同共进。

表 2　专家共识与公众认识

Table 2　Expert consensus and public awareness

类别	关键内容	数据来源
专家共识	1. 公平公正(100%) 2. 隐私保护(97%) 3. 可追责(97%) 4. 透明与可解释性(94%) 5. 安全(81%) 6. 专业研发责任(78%) 7. 人类价值 & 人类控制(69%) 8. 人权(64%)	哈佛大学伯克曼中心研究 (基于 36 份全球 AI 伦理原则文档)
公众认知	– 对 AI 的态度: · 既担心又兴奋(45%) · 更担心(37%) · 更兴奋(18%) – 对 AI 公平决策的信心: · 不可能(28%) · 不确定(41%) · 有可能(30%) – 最担忧的 AI 应用: · 替人类做重要决定(74%) · 洞悉人类想法与行动(75%)	皮尤研究中心 (美国成人调查)
全球共识进展	2024 年联合国 AI 全球决议:120 + 国达成安全、可靠、可信 AI 系统的共识	联合国大会

2.2　国际组织加速推进协同行动

人工智能治理呈现出国内治理与国际协商并行的加速特征。各国积极展开国内治理以遏制风险,国际合作成为避免风险外溢的关键。

全球性共识促使治理进程迈入全球化时代。人工智能技术的易扩散性和风险的广泛性凸显了国际治理的重要性,各国监管体系的简单拼凑与对接实际上无助于降低全球性风险。经合组织统计,69 个国家及地区出台了 1000 多项人工智能政策措施。2021 年,联合国教科文组织促成 193 个成员国共同签署《人工智能伦理问题建议书》。国际社会将气候变化等风险议题纳入国际政治议程耗时数十年,而人工智能风险问题国际行动仅数年便达成共识,治理进程呈现显著加速特征。

2023 年至今,人工智能治理进入快速发展阶段。人工智能治理从早期的非国家主体倡议、主权国家主导、国际机构参与,演变为行业组织、学术团体、高科技公司纷纷介入,形成了多主体、多角度、多层次的全球治理机制复合体[10]。2023 年,世界经济论坛发起人工智能治理联盟,联合政府、行业领袖、学术机构和社会组织,倡导负责任的全球设计和发布透明、包容的人工智能系。2024 年,联合国大会通过了首个人工智能治理决议——《抓住安全、可靠和值得信赖的人工智能系统带来的机遇,促进可持续发展》,标志着人工智能治理从国家主导向全球共治的制度转型。联合国未来峰会同年通过《全球数字契约》并启动全球对话机制,标志着国际共识取得制度化进展。

2.3 风险共识促进公众响应治理

科学共识如同“门户信念”影响着公众的认识与判断。专家与公众在许多风险问题中常存在显著的共识差异,而在人工智能风险及治理中罕见地达成了一致共识。

科学共识常影响公众对相关问题的其他态度。门户信念模型(Gateway Belief Model,简称 GBM)揭示了科学共识如何作为公众形成其他关键认知与情感判断的“门户”或“入口”[11]。近年来,美国国家科学院、国家工程院和国家医学院等权威机构开始重视这一模型并倡导展开深入研究,包括分析其理论机制以及感知共识与科学信念间的因果链接关系。由于专业知识差异,专家与公众间存在认知鸿沟,即“共识差距”。人们更倾向于向专家寻求指导,且更信赖多位专家的综合判断,相信通过筛选和整合“群体智慧”有助于提升判断准确性。GBM 强调了专家共识在公众态度塑造中的核心地位,以及其在公众态度形成与转变中的强大作用。当公众对特定科学问题的共识感知发生变化时,这种变化首先会引发态度的改变,而态度的转变则关乎公众行动支持程度。

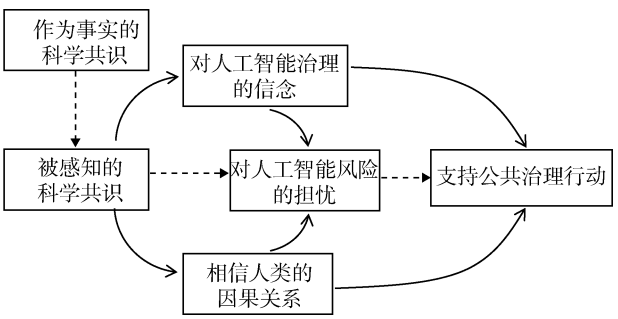

图 1 门户信念模型的因果关系

Figure 1 Causal relationship of the gateway belief model

人工智能风险议题呈现出独特现象:专家与公众罕见地达成一致共识。在气候变化、疫苗安全、转基因食品等其他重大风险议题上,公众认知与专家判断(事实情况)常存在显著的“共识差距”。科学共同体往往早已达成共识,而公众常不知道或者忽略这一科学共识。“共识差距”导致公众常高估或低估风险与严重性,他们常过度担忧酗酒、核能、疫苗和转基因食品的风险问题。例如,88% 的科学家认同转基因食品的安全性,而仅有 37% 的公众持相同观点。对于一些真正具有高风险(如气候变化)的问题,公众却常低估其严重性和紧迫性。“共识差距”常导致公众对其治理行动不关注、不理解、不支持,甚至不认可这种重大公共政策。因此,运用 GBM 缩小公众认知与客观事实之间的认识差距,弥合专家与公众的认知鸿沟,成为推动有效治理的关键环节。

3 人工智能全球治理的阻碍因素

全球人工智能治理面临监管主体分散、进程碎片化及多边框架受阻于大国博弈等难题。科技、经济和政治交织致使人工智能风险治理复杂性加剧,各国需在经济增长、技术进步和伦理价值之间做出艰难抉择。

3.1 全球治理遭遇国家利益冲突

国家利益冲突使人工智能治理呈现出竞争与合作交织、博弈与协商并存的态势。发展权与安全权的矛盾、南北国家间价值观差异及技术民族主义抬头,均对人工智能全球治理构成了挑战。

国家利益与技术主权之间的博弈,衍生出治理立场分化与利益协调的制度性困境。相较于气候变化、臭氧漏洞治理框架,人工智能风险治理面临独特的监管悖论。人工智能风险的即时穿透效应与其他风险问题的延迟效应,形成了治理中的时空差异效应。主要经济体的战略选择,也呈现出显著的“国别差异”分化特征[12]。这种差异既体现在各国对人工智能的监管立场上,也体现在各国探索符合自身需求的人工智能发展模式之中。美国通过《国家人工智能倡议法案》构建技术霸权维护体系以巩固其领先优势,欧盟依托《人工智能法案》关注技术伦理和社会影响,中国则推行《新一代人工智能发展规划》探索数字经济与实体经济的创新融合。人工智能治理呈现南北国家战略诉求对立。发展中国家强调集体赋权以推动产业转型和经济发展,发达国家倾向于实现技术优势的虹吸效应和市场扩张,甚至试图通过构建有利于自身的治理体系来制约他国技术发展。一些国家呈现出明显的政党差异,如特朗普政府推行具有单边主义特征的“美国优先”战略,拜登政府则采取具有歧视性、选择性、排他性的“伪多边主义”协作模式。主要发达国家还通过《人工智能全球伙伴关系》等机制加强合作,这种强强联合导致国际协同治理陷入“技术民族主义”与“数字多边主义”的“囚徒困境”。各国在竞争技术主导权和市场利益的过程中,也在尝试构建风险共担机制,

但竞争与协同的辩证关系始终存在失衡风险。

3.2　全球治理面临技术层面挑战

人工智能全球治理在技术层面面临诸多挑战。标准体系分散化、监管机制滞后与技术快速迭代矛盾突出,这些因素致使治理进程陷入制度适配困境。

统一技术标准是人工智能全球治理的核心问题,全球尚未形成统一的人工智能安全评估和管控标准成为重大阻碍。不同国家和地区对于人工智能的定义、风险分类、监管标准、治理规则及应对措施等存在显著差异,导致监管碎片化现象日益严重。这不仅加剧治理成本,也削弱了治理成效,这种技术与制度的适配难题将最终演化为数字时代的“巴比塔困境”。人工智能潜在风险具备隐蔽性、跨域性、全球性及长期性特征,诸如算法偏见、数据泄露、人工智能滥用等风险,通常难以在事前实现精准预测和有效防范,常陷入风险认知与治理能力动态失衡的“吉登斯悖论”。技术迅猛发展与政策滞后之间的矛盾日益突出。智能技术的快速迭代(常以周为单位)与治理制度的缓慢更新(常以年为单位)存在结构性时滞。这种“监管时滞困境”本质在于技术加速主义与制度渐进主义的系统性矛盾,常导致治理规范陷入“制定即过时”的实践困境。这些问题致使智能技术在发展进程中呈现早期难以干预、后期干预困难的两难局面,最终陷入“科林格里奇困境”。

3.3　全球治理存在认知分歧壁垒

人工智能全球治理还面临着法律、伦理框架以及价值观层面的挑战,这些认知分歧壁垒致使治理效力递减。

智能技术的创新速度远超现有法律体系的更新速度,导致监管体系常滞后于技术发展。诸多新兴应用领域缺乏明确的法律规范和监管标准,易引发法律空白与冲突问题。各国对人工智能风险的感知存在差异,导致全球治理理念难以统一。即使统一了治理规则,各国法律体系和执行能力的差别仍可能影响治理效能。不同国家和地区在数据隐私保护的法律要求上各不相同,也加大了跨国数据流动的监管难度。国际社会在价值取向上存在地域性差异。发达国家倾向于将“民主价值观”作为构建治理体系的核心宗旨,并将人权保障置于优先地位。发展中国家普遍采取实用主义方针,更关注于社会效益。南北国家的价值立场分歧导致全球治理规则协商受阻,在制度设计层面难以形成共识。国际组织职能效力有限,跨国协调机制虽推动了政策对话,但实质性约束力亟待提升,也尚未建构具有强制性法律效力的全球监管体系。OECD 颁布的《人工智能原则》虽已获得近 50 个成员国联合签署,但仍面临执行困境。各国在践行国际准则时常存在灵活性、差异性和不确定性,客观上削弱了全球协同治理效能。因此,各国在追求自身利益最大化和风险最小化时,需要找到“最小公约数”和动态平衡点。

4　人工智能全球治理的未来进程

人工智能治理已达成全球共识,多元主体的共识正逐步从理念转化为切实可行的全球治理机制。基于现状分析,可预测未来治理的短期趋势(约 1 – 5 年)、中期图景(约 5 – 10 年)及长期愿景(10 年以上),其治理进程将朝着更公平、更专业的方向不断前进。

4.1　人工智能全球治理短期趋势

从短期趋势分析,治理进程将细化伦理准则,技术进步推动风险评估标准化,国际协作模式将从多主体主导转向更高效的多边机制。

在伦理规范方面,治理进程正加速迈向精细化、场景化和可操作化。早期的伦理框架虽然提出了“公平、透明、问责”等宏观理念,但因缺乏具体执行标准而难以有效落地。如今,各国政府、国际组织、科技企业和学术界正共同努力突破伦理规范这一核心制约因素,将抽象原则转化为切实可行的规范。人工智能伦理规范正逐步与法律、技术标准深度融合,围绕算法透明度、数据隐私保护、公平性和责任归属等核心议题,具体实施细则正在推进。因此,中国发布了《人工智能伦理治理标准化指南》,联合国教科文组织发布了《人工智能伦理建议书》。人工智能伦理准则的细化呈现出两大趋势。其一,针对不同行业的 AI 应用风险差异,伦理规范正朝着分领域制定细则的方向发展。鉴于 AI 应用风险特征各异,需要制定针对性的伦理规范,如自动驾驶中保障安全是其首要任务。其二,伦理准则将深度贯穿人工智能全生命周期。数据采集需获数据主体知情授权,算法设计应规避免偏见,模型训练中防范偏差,系统部署时评估其影响,运维护阶段持续监测合规性,系统退役时妥善处理数据消除隐患。这些伦理准则的细化,将为全球治理提供坚实的制度基础。

全球治理风险评估正迈向标准化进程。由于人

工智能在医疗、交通、金融、军事等领域的风险特征差异,加之各国和各行业在风险定义和应对策略上存在分歧,导致监管呈现碎片化。各国政府、国际组织和科技企业正积极探寻建立统一的风险评估框架,以实现更精准的 AI 监管。欧盟《AI 法案》提出的四层风险分级(不可接受风险、高风险、有限风险、最小风险)及对应监管要求,为全球风险评估提供了重要参照。为识别、衡量和管控 AI 风险,技术升级将加快风险评估标准化,推动建立分级治理体系。标准化风险评估有助于统一风险界定(如明确高风险 AI 应用的范畴)、提高监管效率和增强企业合规性。技术进步正不断提升风险评估的客观性和自动化水平,相较于效率低、主观性强的传统人工审核,先进算法模型、大数据、区块链技术及 AI 风险评估工具(如 IBM 的 AI Fairness 360、微软的 Responsible AI 工具包)能够提升风险识别和量化能力,实现实时监测风险。国际组织(如 ISO、G20)正在推动通用风险评估标准,并探索跨国互认机制,以降低合规成本并促进 AI 系统的全球流通。风险评估标准化将成为 AI 全球治理的核心准则,将实现定期发布《全球人工智能发展评估报告》,构建动态风险评估体系。

人工智能全球治理正从多主体主导进程向多边主导模式过渡。当前全球治理面临多元利益冲突和地缘对抗加剧等挑战。各国基于自身经济、技术、安全和文化价值观等考量,在治理立场上分歧显著,亟需在尊重差异的基础上寻求共识。鉴于人工智能的全球影响力且存在数据流动、算法偏见、自主武器扩散等跨国性、系统性风险,单边或小范围联盟的监管模式已难以奏效。其"公共产品"和"外部性"属性,迫切要求超越单边主义,迈向多边协作。正如 1987 年《蒙特利尔议定书》通过多边合作成功减少氟利昂排放,人工智能治理亦可借鉴类似的技术共享机制,助力发展中国家提升监管能力。多边治理的必要性体现在避免"监管孤岛"、消除"监管套利"、共同应对挑战和平衡大国竞争。多边协议和组织规则能够明确各方权责、确保责任落实,从而达成协同目标和行动方案。人工智能全球治理已呈现中美欧"三强鼎立"格局:中美分别在算法创新和应用场景上占据优势,欧盟则凭借其规则制定与制度输出形成第三极影响力[13]。欧盟"可信人工智能"框架、中国"以人为本"治理理念与美国"创新友好型"监管原则具有兼容性。三方在"竞争中合作"且又相互制衡的关系,有效避免了单一主体主导议程,为全球治理体系提供了结构稳定性。这种"多边对话 + 共同目标 + 差异化路径"的治理模式,正推动着全球治理向前发展。

4.2 人工智能全球治理中期图景

从中期图景分析,人工智能全球治理将呈现三大特征:技术治理联盟的成型、动态分级监管体系的建立,以及全球治理框架的初步形成。

智能科技企业和人工智能协会将携手形成技术治理联盟。与臭氧层治理中化工企业的被动配合形成鲜明对比,智能科技巨头(如 OpenAI、DeepMind)主动参与治理,资助风险研究(如 Anthropic 的"宪法 AI"),形成一种新型的"技术治理联盟"。传统以政府为主导的治理模式正面临结构性挑战,科技公司成为推动全球治理的关键角色。政府因技术认知滞后或政治博弈,可能采取保守甚至阻碍技术发展和治理效能的政策;科技公司兼顾市场和技术的治理模式,不仅更符合 AI 发展和治理的内在逻辑,也通过分散传统政府的权力缓解了国家之间的博弈。高科技全球化企业为吸引全球用户,常积极拥抱全人类共同价值,制衡某些政府的错误干预。加强行业间合作成为人工智能治理的关键路径。由于技术无国界,行业组织(如 IEEE、PAI)能够跨越政治分歧,推动更具包容性的治理框架。行业自律与政府监管相结合被视为人工智能算法治理的最佳政策路径。人工智能企业作为算法的开发者和直接应用者,对技术更加熟悉,自我监管更加便捷[14]。苹果、亚马逊、DeepMind、谷歌、脸书等科技公司成立"全球人工智能伙伴关系",这种"自下而上"的治理模式推动了全球规则协调。人工智能全球治理未来可能呈现"双轨并行"格局:政府负责风险底线立法,而市场与技术联盟主导伦理标准与行业规范。

人工智能全球治理将建立动态分级监管体系。传统静态监管模式难以应对技术迭代产生的复杂风险,一刀切模式也无法满足各国实际需求和多样化风险。动态分级监管体系能够根据各国的发展阶段和风险承受能力,灵活调整监管强度与重点,从而实现精准治理。摩尔定律指出约 18 – 24 月集成电路的晶体管数量和性能将翻倍,凸显了监管体系更新的紧迫性。技术发展常导致固定风险分类标准过时,风险场景的动态变化使得同一技术在不同应用场景下的风险等级存在显著差异。动态分级监管体系能够避免对低风险应用的过度监管,将有限资源

集中投入到高风险领域。为实现有效的动态分级监管,可尝试构建三类技术基础设施:一是风险感知网络(如新加坡金融管理局推出的 AI 监管沙盒),从而实时监测和评估 AI 技术风险;二是智能评估工具,借助自动化算法升级评估手段(如 IBM 的 AI Fairness 360 开源工具),提高评估效率和准确性;三是弹性调整机制,从而实时捕捉风险变化。这种“监测 - 评估 - 调整”的闭环治理模式,将推动 AI 监管从传统被动响应模式转向主动预防模式。

人工智能全球治理框架将从理念探讨迈向实际构建与完善阶段。需将多边对话机制转变为具有约束力的制度化框架,以形成初步的全球治理架构。联合国及其专门机构通过搭建多边协商平台,推动形成全球广泛认同的治理原则框架。在数据安全、算法伦理、责任归属等关键议题上,国际协商取得突破性进展,公平、透明、可问责、隐私保护等原则已成为全球广泛遵循的价值标杆。WTO 将人工智能治理纳入数字贸易规则框架,国际电信联盟(ITU)则专注于技术标准的国际协调。人工智能全球治理的系统性思维深入人心,治理进程将从区域实践走向广泛协同,构建全球 - 区域 - 国家 - 地方的多层次治理体系。

4.3　人工智能全球治理长期愿景

从长期愿景来看,人工智能全球治理将从价值对齐迈向价值共生,构建人机协同数字孪生治理生态,最终形成真正的全球智能治理共同体。

从价值对齐到价值共生标志着治理能力的跃升。智能系统理解并内化人类多元价值观至关重要。为确保技术中立性,可通过对抗性训练提升算法普适性,并构建前置性制度约束,将人类价值观贯穿于人工智能研发全流程。通过构建人类与人工智能的价值共生体系,将实现人机价值交互[15]。人类可从人工智能对复杂数据的分析中获取新伦理洞察,人工智能则通过学习人类价值判断逻辑提升伦理决策水平。人工智能治理还将建立动态价值平衡机制,使其能够适应人类价值观权重的动态变化,确保其决策始终符合人类价值诉求。进一步将培育“数字文明共同体”意识,将人工智能视为数字文明的核心组成部分,构建人类与人工智能和谐共生的文明形态,开启人机协同创造价值的新纪元。

人机协同与数字孪生治理将开启智能治理的新篇章。人类与智能系统的深度协同将打破人类与人工智能之间的协作壁垒,实现治理效能最大化。数字孪生治理生态将虚实交互提升到新境界,通过创建与现实世界高度映射、实时交互的虚拟数字空间,为全球治理提供全新的平台与视角。借助物联网、大数据、云计算等技术的深度融合,现实世界的各类治理要素将被精准映射到数字空间,形成数字孪生体,并实时反映其运行状态和变化趋势,从而及时开展社会治理,提前规避免事故后果。数字孪生治理生态为人机协同提供直观、精准的决策依据和模拟验证环境;人机协同则为数字孪生治理生态的有效运行提供智能决策支持和灵活应变能力。二者的融合将重塑全球治理流程,推动全球治理迈向更加智能、高效的新境界。

人工智能全球治理将形成真正的全球智能治理共同体。企业、科研机构、非政府组织等多元主体的深度参与,将推动共同体突破传统国际关系局限,构建人机协同的新型治理范式。一方面,建立预测性治理系统,借助超级算力对社会政策进行大规模仿真模拟,将各类风险扼杀于萌芽。这些创新举措使治理体系具备了前所未有的适应性和包容性。另一方面,采用人类和 AI 系统共同参与决策的混合机制,既确保治理效率,又维护人类的主体地位。各国将超越地域和文化差异,形成“多元智能文明观”,这不仅是治理机制的革新,更是人类文明形态的重大跃升。全球智能治理共同体还将有效应对气候变化、公共卫生等挑战,引领人类社会迈向更智能、更美好的未来社会。

5　结语

全球风险议题常遵循“科学共识 - 政治议程 - 全球治理”的演进路径。科学共同体与政策决策者的良性互动是构建治理体系的关键环节。科学共同体为治理体系奠定知识基础,政治议程推动科学共识转化为实质性的治理框架。

本研究通过分析主要有以下发现:在科学认识方面,科学界就人工智能风险及治理快速形成科学共识;在公众认识方面,社会公众与专家群体罕见地达成高度一致性;治理行动方面,技术发展与治理能力、国内治理与国际行动齐头并进。其他全球性风险问题与其存在显著差异,常表现为科学共识形成缓慢,公众与专家存在显著的“共识差距”与“认知鸿沟”,治理行动常陷入“搭便车”困境或公地悲剧。人工智能全球治理也面临着发展权与安全权冲突激

烈、南北国家价值分歧显著、技术民族主义兴起、监管时滞与技术迭代矛盾突出、标准体系碎片化等阻碍因素,导致了执行效力的衰减效应。

围绕“人工智能怎样发展与治理”这一重大时代课题,习近平总书记倡导秉持共商共建共享理念。联合国教科文组织和世界科学知识与技术伦理委员会呼吁,世界各国需要采用全新的视角看待人工智能系统的未来监管。展望未来,国际组织、社会公众、科技行业及技术进步将合力推动全球治理进程。从短、中、长期来看,伦理准则将从细化走向价值共生,风险预防治理将从静态标准划分转向提前动态规避,全球合作不断深化并形成全球智能治理共同体。

参考文献:

[1] 全球人工智能治理倡议[J]. 中国信息安全,2023(10):12 - 13. Global AI governance initiative [J]. China Information Security, 2023(10): 12 - 13.

[2] 俞陶然.“全球风险之首”该如何管控[N]. 解放日报,2024 - 01 - 20(002). Yu T R. How should the 'top global risk' be governed? [N]. Jiefang Daily, 2024 - 01 - 20(002).

[3] Veale M, Matus K, Gorwa R. AI and global governance: Modalities, rationales, tensions[J]. Annual Review of Law and Social Science, 2023, 19(1): 255 - 275.

[4] Ala - Pietilä P, Smuha N A. A framework for global cooperation on artificial intelligence and its governance [J]. Reflections on Artificial Intelligence for Humanity, 2021: 237 - 265.

[5] 陆卫明,张文,王明. 全球人工智能治理从“共识”到“共治”:动力、挑战与路径[J/OL]. 西安交通大学学报(社会科学版),1 - 20[2025 - 04 - 14]. Lu W M, Zhang W, Wang M. From “consensus” to “co - governance” in global AI governance: Drivers, challenges, and pathways [J/OL]. Journal of Xi'an Jiaotong University (Social Sciences), 1 - 20 [2025 - 04 - 14].

[6] Fjeld J, Achten N, Hilligoss H, et al. Principled artificial intelligence: Map** consensus in ethical and rights - based approaches to principles for AI[R]. Berkman Klein Center Research Publication, 2020.

[7] Lee Rainie, et al. Al and human enhancement: Americans´openness is tempered by a range of concerns[EB/OL]. Pew Research Center, March17, 2022. https://www. pewresearch. org/internet/2022/03/17/how - americans - thinkabout - artificial - intelligence/.

[8] 牟砚堂,柯遵科. 全球变暖的科学共识与美国公众的误解探析[J]. 自然辩证法研究,2022,38(11):64 - 70. Mou Y T, Ke Z K. The scientific consensus on global warming and American public's misunderstanding about it [J]. Studies in Dialectics of Nature, 2022, 38(11): 64 - 70.

[9] Roberts H, Hine E, Taddeo M, et al. Global AI governance: Barriers and pathways forward [J]. International Affairs, 2024, 100(3): 1275 - 1286.

[10] Cihon P, Maas M M, Kemp L. Fragmentation and the future: Investigating architectures for international AI governance [J]. Global Policy, 2020, 11(5): 545 - 556.

[11] Van der Linden S. The Gateway Belief Model (GBM): A review and research agenda for communicating the scientific consensus on climate change[J]. Current Opinion in Psychology, 2021, 42: 7 - 12.

[12] 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,2024,42(3):460 - 468,491. Wang Y Y, Li Z F, Gao F. A comparative study of European and American AI governance models [J]. Studies in Science of Science, 2024, 42(3): 460 - 468,491.

[13] 施雯,缪其浩,从两极到三强:欧盟人工智能的全球竞争战略分析[J]. 中国科技论坛,2022(6):179 - 188. Shi W, Miao Q H. From bi - polar to tripartite: Analysis of EU's strategy to survive in global artificial intelligence competition [J]. Forum on Science and Technology in China, 2022(6): 179 - 188.

[14] Adonis Hoffman. Why self - regulation is best for artificial intelligence [EB/OL]. The Hill, November 8, 2023. https://thehill. com/opinion/4300288 - why - self - regulation - isbest - for - artificial - intelligence/.

[15] 夏永红. 人工智能伦理治理范式:从价值对齐到价值共生[J]. 自然辩证法通讯,2025,47(1):1 - 8. Xia Y H. The paradigm of ethical governance of artificial intelligence: From value alignment to value symbiosis [J]. Journal of Dialectics of Nature, 2025, 47(1): 1 - 8.

(下转第 712 页)