·人工智能与社会发展·

推进全球人工智能安全治理的中国方案及战略实践

刘世强,徐惠嫄

(西南财经大学马克思主义学院,四川 成都 611130)

摘 要: 人工智能技术的广泛应用在促进经济发展和社会进步的同时也带来了一系列安全风险。当前全球人工智能安全治理陷入困境,体现在安全环境恶化、安全风险增多、安全机制缺失、安全理念滞后等多个方面。中国积极践行全球安全倡议,从共同安全、综合安全、合作安全和可持续安全四大层面深入阐释了人工智能安全治理的目标愿景、主要任务、方法手段与核心动力,形成了全球人工智能安全治理的科学方案。在此指引下,中国加快自身人工智能发展,搭建人工智能安全治理框架,推动人工智能安全国际合作,扩大人工智能安全公共产品供给,有力推动人工智能更好服务于世界和平与人类发展。

关键词: 人工智能安全;全球人工智能治理;全球安全观;数字命运共同体;中国方案

中图分类号: TP18 **文献标识码:** A **文章编号:** 1674-5094(2025)06-0012-10

近年来,随着生成式人工智能实现颠覆性突破,其应用已深度嵌入经济社会各领域,成为推动经济增长与产业变革的重要引擎,深刻重塑着人类的生产和生活方式。然而,人工智能的迅猛发展也带来了日益复杂的安全风险,影响世界和平和各国安全。在此背景下,推进全球人工智能安全治理,已成为国际社会必须直面的重大议题。2023 年 10 月,在第三届“一带一路”国际合作高峰论坛上,中国发布《全球人工智能治理倡议》,展现了中国推动人工智能健康、安全和可持续发展的大国担当。同年 11 月,首届“全球 AI 安全峰会”召开,中国等 28 国与欧盟联合签署首个全球性 AI 声明《布莱切利宣言》,进一步凸显中国在人工智能领域共治风险、共促安全的合作意愿 。2025 年 7 月,中国政府倡议成立世界人工智能合作组织并发表《人工智能全球治理行动计划》,以实际行动促进人工智能向善普惠发展。可见,面对人工智能迅猛发展及其带来的安全问题,中国始终以负责任大国姿态参与其全球治理进程,坚持理念创新与实践推进并重,为破解全球人工智能安全困境、保障全球人工智能稳定发展提供了中国方案、贡献了中国力量。

一、推进全球人工智能安全治理的战略价值

人工智能作为引领新一轮科技与产业革命的核心技术,因其强适应性、广融合性、高协同

基金项目: 国家社会科学基金重大项目“世界百年未有之大变局加速演进的理论与实践研究”(23ZDA023);国家社会科学基金项目“高校思想政治理论课引导大学生坚定历史自信研究”(22VSZ124)。

作者简介: 刘世强,西南财经大学马克思主义学院教授,国家社科重大项目首席专家,博士生导师,研究方向:全球治理与中国特色大国外交。

引文格式: 刘世强,徐惠嫄. 推进全球人工智能安全治理的中国方案及战略实践[J]. 西南石油大学学报(社会科学版),2025,27(06):12–21.

性等突出特性，在实际规模化应用中释放出了巨大的智能红利与发展潜力。与此同时，人工智能在发展过程中存在着一系列现实和潜在的安全风险。世界各国纷纷针对人工智能进行战略布局以抢占未来发展的制高点，进一步引发了激烈的国际竞争。因此，推进全球人工智能安全治理既是顺应数字技术变革及全球化趋势的应有之义，也是重塑安全环境、应对安全挑战、健全安全机制和革新安全理念的必然选择。

（一）秩序重塑：为营造人工智能安全发展环境提供基本条件

安全是人类社会的基本需求，营造稳定安全环境是推进全球人工智能安全治理的首要任务。随着人工智能技术的突破性发展，“技术工具演变为地缘权力载体”[1]，成为大国战略博弈的“新边疆”。西方大国将人工智能视为维护霸权和争夺权力的重要抓手，推动人工智能国际竞争全面升级。其一，人工智能战略化引发安全焦虑。作为一项新兴颠覆性技术，人工智能成为国家实力地位的现实表征和硬核支撑，其赋予的技术权力一定程度上决定着一国在国际权力结构中的位置。为此，美国就把确保人工智能领导地位作为首要战略目标，将人工智能竞争纳入权力竞争轨道[2]。例如，美国推出《关键和新兴技术国家战略》《2021 战略竞争法案》《捍卫美国 5G 未来法案》等一系列文件，旨在维护自身在人工智能领域的先发优势。然而，广大发展中国家因发展阶段、要素禀赋等因素在发展人工智能方面仍远远落后于西方发达国家。这使得二者之间的数字鸿沟进一步扩大和固化，引发发展失衡和安全焦虑。其二，人工智能阵营化破坏安全结构。随着以中国为代表的新兴国家在人工智能领域的崛起，美国出于霸权护持的战略考虑，不仅采取多种措施限制高端芯片、软件、设备出口中国，遏制中国人工智能技术和产业发展，而且通过强化“五眼联盟”“美日印澳四方安全对话”以及“美英澳三方安全伙伴关系”等盟伴体系对中国进行数字封锁与围堵。这种利用自身技术优势、建立技术联盟来肆意打压他国的霸权行径，严重冲击了现有国际安全结构的稳定性。其三，人工智能军事化加剧安全威胁。人工智能技术深度嵌入情报战、网络攻防及自动化武器系统等军事领域，固然能够显著提升国家军事能力，但也为发动基于人工智能的新型战争打开了方便之门。在俄乌冲突中，美国依托人工智能技术的高效性与隐蔽性，将俄罗斯作为网络攻击的主要目标，导致其政府、金融、科研机构部门网站一度频繁出现页面瘫痪或无法访问的情况[3]。此外，俄罗斯和乌克兰还利用无人机持续攻击对方的军事目标和民用设施，以低成本、灵活性、高隐蔽性的攻击方式极大改变了传统战争形态。总之，人工智能的发展特别是各国围绕人工智能展开的权力博弈恶化了国际安全环境。只有切实推进全球人工智能安全治理，消除“技术霸权主义”的影响，才能为各国发展与交往营造安全稳定的国际环境。

（二）现实纾困：为应对全球人工智能安全挑战提供有效保障

人工智能技术的加速迭代与深度融合应用，在驱动社会变革的同时也带来了多重安全风险与现实挑战。基于人工智能的多重属性，可将其安全风险区分为技术本体性风险与应用衍生性风险。技术本体性风险是指人工智能技术因自身属性所产生的固有风险。生成式人工智能的出现，“意味着人工智能从依赖固定规则的系统向更灵活、更复杂的数据算法驱动模型发展”[4]，但由于算法本身的复杂性、不确定性与“黑箱”性，存在较大的安全隐患。在实际操作与运用过程中，不可避免地会因其模型设计缺陷、训练冗余量不足等因素而出现数据误差或运算错误，导致技术失灵、系统事故、安全漏洞等问题。例如，人工智能模型的生成内容与客观事实不符，甚至出现完全虚构的信息内容，进而影响使用者决策的正确性与科学性。应用衍生性风险是人工智能技术融入经济社会发展而引发的外溢性风险。一是加剧失业风险。人工智能

使用的低成本和高效率不仅使得劳动密集型的传统就业岗位大量消失,也对法律、教育、金融服务等知识密集型行业带来剧烈冲击。二是催生伦理风险。在人机互动过程中,由于人工智能技术的强自主性,剥夺了人的理性认知、自控能力和自主选择[5],可能造成人主体性丧失等伦理危机。三是带来法律风险。在人工智能技术迅猛发展的过程中,也存在技术滥用现象。比如,利用人工智能技术进行数据窃取、网络诈骗、伪造和传播虚假信息等网络犯罪行为激增,严重冲击社会秩序与国家安全。在此背景下,防范和化解人工智能安全风险是所有国家面临的共同挑战。推进全球人工智能安全治理,是系统性应对人工智能技术内生缺陷及其复杂外溢风险的关键。

(三)机制供给:为应对全球人工智能安全挑战提供制度保障

在完整的治理体系中,制度安排是其核心构成要件,"表现为标准、规则以及机构的机制总和"[6]。当前,全球人工智能安全治理正处于多方博弈的早期阶段,系统性治理框架尚未形成,面临治理标准单极化、治理规则软性化、治理机构分散化等困境,导致治理制度供给不足。首先,治理标准单极化加剧治理公平赤字。人工智能安全技术标准的制定与设置正成为大国争夺治理主导权的焦点。部分技术领先国家倚仗先发优势,垄断国际技术标准制定话语权,将自身的战略意图与价值观念嵌入技术标准的制定过程。同时,他们还试图建立具有封闭性和排他性的技术标准联盟,牢牢控制人工智能发展的技术与安全标准,使广大发展中国家只能默认或接受其制定的规范。其次,治理规则软性化导致治理效用不足。当前,人工智能安全领域的治理规则多表现为"围绕特定议题或产业形成的原则共识、发展指南、倡议协议、最佳实践、标准指南等软性规则"[7]。这些规则缺乏强有力的执行机制,难以得到国际社会的普遍遵守。此外,不同主体对人工智能安全治理的领域、内容、优先序等存在认知差异,必然造成治理规则的步调不一甚至相互冲突,进而导致决策过程冗长、效率低下。最后,治理机构分散化弱化集体行动能力。目前,人工智能安全治理机构呈现"多中心"格局,主要包括联合国框架下的国际性协同平台、主权国家间形成的区域性国际组织及多方主体联合成立的非政府组织等。由于不同治理机构间相互独立,在治理理念、目标、方式等方面存在差异,加之机构间缺乏有效沟通,导致各治理机构间呈现出高度分散的状态,加剧了人工智能安全领域的集体行动困境。因此,必须大力推进全球人工智能安全治理,通过形成公正的治理标准、一致的治理规则和协同的治理机构以更好开展国际协作、推进治理进程。

(四)价值纠偏:为革新全球人工智能发展理念提供重要支撑

当前,全球人工智能安全治理面临的现实困境,本质上反映的是治理理念的滞后。"全球安全治理受到行为主体安全观念的支配和影响,当前的全球安全乱象与西方国家过于自我的、陈旧的安全观念密切相关"[8]。一方面,个别国家固守非此即彼、你输我赢的二元对立思维,将人工智能技术视为稳固权力、排斥异己的工具,成为全球人工智能安全风险聚积的重要推手。近年来,美国政府不断制造并传播"民主与独裁""自由与专制"的虚假叙事,将不顺从的国家及其人工智能技术发展描绘为对"自由民主世界"的威胁,以意识形态和价值观为依据构建包括人工智能在内的技术联盟,不仅影响人工智能基于国际分工协作的技术和产业发展,还增加了全球人工智能安全治理的复杂性与困难度。另一方面,一些国家在发展人工智能上存在明显的工具主义倾向,表现为追求技术发展服务于国家安全竞争的单一导向,过度聚焦人工智能的地缘政治效应和军事应用场景。这一倾向造成"治理议程往往服务于效率优先、竞争优先的逻辑,而忽视对人类福祉、社会信任和伦理规范的系统回应"[9],客观上导致了技术异化风险的进

一步上升。因此，推进全球人工智能安全治理亟需构建符合人类共同利益的价值导向，以价值理性规训工具理性，实现“多边、多方主体的相互关联、相互依存”[10]。总之，推进全球人工智能安全治理需要摆脱“零和思维”和“工具主义”理念的内在缺陷，以多边、包容、公正、可持续的治理理念引导人工智能更好增进人类社会的共同福祉。

二、推进全球人工智能安全治理的中国方案

当前全球人工智能安全形势不容乐观，技术问题的跨国性、联动性、隐匿性和复杂性特质使得国际不稳定性不确定性大大增加。加之个别国家固守霸权思维，将人工智能视为权力地位的放大器，在国际上大搞“小院高墙”和封锁制裁，进一步凸显了人工智能的国际安全赤字。中国始终坚持和践行“真正的多边主义”，“倡导践行共同、综合、合作、可持续的安全观，以合作促发展、以合作促安全，构建起更为均衡、有效、可持续的安全架构”[11]，并坚持“把创新作为第一动力、把安全作为底线要求、把普惠作为价值追求”[12]，围绕目标愿景、主要任务、方法手段、核心动力等方面提出了推进全球人工智能安全治理的中国方案。

（一）倡导共同安全，以构建人工智能安全共同体为目标愿景

共同安全指向人工智能安全治理的主体维度，强调保障和实现各行为主体在人工智能领域的安全。随着全球化的深入，“安全不仅是一种现实或感知状态，更表现为一种复杂互动过程”[13]，任一地区或领域出现的风险挑战都可能发生全球性扩散。因此，安全不应被视为少数国家或国家集团的特权，而应是所有国家都应有的需求和权利。正如习近平指出的：“人类是不可分割的安全共同体。”[14]451 这无疑为推进全球人工智能安全治理锚定了正确方向。因此，中国主张通过构建人工智能安全共同体以实现共同安全。在安全利益层面，中国认为各行为主体在人工智能领域的安全利益具有相互依存性。人工智能技术及其治理具有显著的跨国特征，一国在人工智能上的安全与发展同他国乃至全球的整体安全环境密不可分。各国均拥有发展人工智能技术的权利，任何国家在人工智能领域的进步都不应建立在限制或损害他国安全利益和发展权的基础之上。正如习近平指出：“一国的安全不能建立在别国的动荡之上，他国的威胁也可能成为本国的挑战。”[15]541-542 在安全参与层面，中国认为所有行为体均享有平等参与人工智能安全治理的权利。受实力差距和国际格局的影响，当前发达国家在人工智能安全治理中占据主导地位，而发展中国家则面临着代表性缺失、话语权不足、安全空间受挤压等多重困境，导致全球人工智能安全治理格局严重失衡。改变这一局面是国际社会的普遍诉求，要求在全球人工智能安全治理框架中切实保障各方平等地位，“坚持各国无论大小、强弱、贫富，都在全球治理中平等参与、平等决策、平等受益”[16]。在安全责任层面，推进全球人工智能安全治理并非一国所能独自承担，也难以通过单边行动得以解决，亟需世界各国广泛参与和共同协作。考虑到各国发展水平与现实条件的差异，中国坚持“共同但有区别”的原则，推动各国依据自身能力与国际地位承担相应的责任，做到权责对等。总之，推动全球人工智能安全治理的目标就在于构建基于利益共生、权利共享和责任共担的安全共同体。

（二）倡导综合安全，以统筹人工智能多维领域安全为主要任务

综合安全指向全球人工智能安全的内容维度，强调应统筹维护各国在人工智能各领域的安全。随着技术的创新突破和快速迭代，人工智能广泛传导和渗透至经济社会生活的方方面面，从而推动人工智能安全议题的多样化。不仅如此，不同类型的安全议题往往交织叠加，采

取单一手段进行治理往往难以奏效。因此,我们需要从整体和全局视野出发,既在整体上把握人工智能风险来源、演化趋势并提出系统性的治理策略,又分类分级加强管理,避免在对象识别、风险研判和处置举措上的模糊化。具体而言,一方面,要保障各国在人工智能领域的主权安全。从主权视角看,数字主权安全构成了人工智能安全的根本前提与基础保障。所谓数字主权,是指“主权国家对其行政管辖范围内的所有数字要素拥有至高无上的排他性的政治权力”[17]。正如有学者指出,“技术不仅仅是一种工具,技术与政治统治和政治权力密不可分,它在某种意义上构成了权力的支撑系统”[18]。人工智能作为一种变革性工具,其技术权力形态集中体现了数字主权的内在属性。正因如此,中国主张应充分尊重各国数字主权,切实保障其在数字空间的平等参与权、自主决策权与正当防卫权。与此同时,国际社会需共同防范全球网络科技巨头凭借技术优势侵蚀甚至争夺主权国家的数字权益。各国只有在充分尊重和保障数字主权的基础上,才能实现人工智领域的共同安全与协调发展。另一方面,要保障人工智能功能领域安全。从功能视角看,数据、算法、算力是人工智能技术的基础要素和核心要件。具体而言,数据作为人工智能技术的关键生产资料,其规模与质量直接影响算法模型效能;算法是驱动数据转化为智能应用的核心引擎;算力则是支撑数据处理、模型训练与落地的硬件基础。三者的安全共同决定了人工智能技术的整体安全。实现综合安全,不仅要有效管控人工智能数据安全风险,对数据进行分类分级,促进数据的有序流动与平衡开发;还需“构建技术层面的人工智能算法安全体系”[19],确保算法系统的可靠性与安全性;突破高端芯片制造等关键瓶颈,构建自主可控的算力基础设施体系,夯实人工智能的技术安全基础。

(三)倡导合作安全,以推进人工智能安全领域的对话协商为方法手段

合作安全是推进全球人工智能安全的方法维度,强调以对话协商消除误解、化解矛盾,发挥“复杂开放的全球安全系统中不同安全行为体子系统的协同效应”[20]。当前,以单边主义和霸凌霸道为特征的地缘政治强势回归,大国围绕人工智能技术的争夺异常激烈,加剧了全球安全系统的分裂。对此,习近平指出:“要加强人工智能国际治理和合作,确保人工智能向善、造福全人类”[21]。换言之,实现全球人工智能安全不仅依赖各国自身努力,更需国际社会通力协作。在理念层面,中国强调扩大人工智能国际安全合作的共识。通过对话凝聚安全利益共识,是开展有效国际合作的前提。“合作安全既不认同军事手段是解决安全问题的优先方式,也不认可霸权国垄断国际和地区安全事务的霸道行径,而是欢迎各国在友好式协商和包容性参与的基础上寻求安全收益最大化”[22]。只有通过对话沟通,深入了解彼此的利益关切与战略意图,才能找到各方在安全利益上的最大公约数,进而为人工智能国际安全合作奠定互信基础。在实践层面,中国主张推进人工智能多层次、宽领域的国际安全合作。人工智能技术的快速迭代与应用场景的不断拓展,使得全球安全挑战更加复杂,治理难度持续增加。因此,传统的单一安全合作架构难以发挥治理效能,亟需构建一个合作主体多元、合作方式灵活、合作领域广泛的多维立体架构,“建立起以主要国家政府为主导、非政府组织、企业等广泛参与的全球合作网络”[23],推动多元主体在人工智能安全领域开展信息共享、技术互通、标准协商、人才交流、协同监管等多方面合作。在机制层面,中国呼吁推进全球人工智能安全治理的多边机制建设。多边主义是全球治理的基本原则,国际组织与平台是推进人工智能安全治理的重要依托。中国外交部长王毅指出:“以联合国为核心的国际体系是人类进步事业的重要保障,以协调合作为基石的多边主义理念是解决全球问题的最佳方案。”[24]推进全球人工智能安全治理,必须充分发挥联合国在维护人工智能安全方面的核心作用,并在其框架下积极参与相关国际组织

与机制的改革完善，为促进全球人工智能安全治理提供制度框架。

（四）倡导可持续安全，以夯实人工智能发展根基为核心动力

可持续安全作为全球人工智能安全的状态维度，是指“通过发展化解矛盾，消除不安全的土壤”[25]，实现各国在人工智能领域的持久安全。“发展是安全的基础，安全是发展的条件”[26]201。推进全球人工智能安全治理必须坚持安全与发展并重，通过人工智能的创新发展、多向赋能和生态完善，为全球人工智能安全治理奠定物质基础。一是强化人工智能技术创新能力，筑牢技术安全屏障。技术创新是人工智能安全发展的核心支撑。推动技术创新，必须持续提升各国在数据系统、算法模型和算力硬件等基础环节的研发与应用能力，实现技术发展能力的突破性与跨越式提升。同时，加强与各技术行为体之间的协调合作，充分发挥协同效应，共同促进人工智能技术健康发展。二是加快人工智能结构调整，完善产业安全框架。人工智能发展结构体现为技术在产业应用中的布局层次与融合程度，直接影响全球产业链供应链的安全与稳定。在产业安全方面，应进一步优化人工智能在关键行业和新兴场景中的应用布局，包括智能医疗、智能制造、智能驾驶、智能消防、智能娱乐等领域，不断深化融合层次，以高质量的产业发展筑牢产业安全基础。三是推动人工智能红利共享，弥合智能发展鸿沟。习近平指出：“世界繁荣稳定不可能建立在贫者愈贫、富者愈富的基础上。”[27] 面对全球范围内人工智能发展日益扩大的鸿沟，中国主张加强南北合作，特别是发达国家应加强对发展中国家的技术援助，帮助后者提升应对人工智能安全风险的能力，推动人工智能发展成果在全球范围内更加公平、公正和包容地共享，以可持续发展支撑全球人工智能的高水平安全。

三、中国推进全球人工智能安全治理的丰富实践

当前，人工智能安全与人类发展的前途命运息息相关。中国作为世界上最大的发展中国家，从构建人类命运共同体的高度出发，积极推进全球人工智能安全治理，从共同安全、综合安全、合作安全、可持续安全等角度提出了全球人工智能安全治理的科学方案。以此为指引，中国立足自身发展、搭建治理框架、开展交流合作、扩大公共产品供给，为人工智能的健康发展及更好造福人类作出了重要贡献。

（一）立足自身发展，夯实人工智能安全治理基础底座

以发展促安全是中国总体国家安全观的核心要义，亦是中国推进全球人工智能安全治理的基本遵循。发展与安全相互影响、相互制约，没有人工智能技术的创新突破，就难以有效解决人工智能发展过程中的安全难题。为此，中国立足自身，着力推进人工智能的创新发展，为推进人工智能安全治理提供坚实基础。第一，推动人工智能技术持续创新迭代。面对西方国家的技术封锁与战略遏制，中国充分利用自身在政策、市场、人才等方面的比较优势，集中力量突破GPU芯片、AI编译器、NLP技术等关键核心技术瓶颈，不断扩大算力规模，推动我国进入世界人工智能发展的第一梯队。据统计，2022年，我国的算力总规模达到302EFlops，位居全球第二[28]。在此基础上，涌现出以DeepSeek为代表的一大批国产大模型，凭借“高性能+低成本+开源开放”[29] 的创新模式，实现了人工智能技术的跨越式发展，为广大后发国家提供了可借鉴的技术赶超路径。第二，营造人工智能产业发展良好生态。中国高度重视人工智能技术与产业的深度融合，“推进人工智能赋能工业制造、消费、商贸流通、医疗、教育、农业、减贫等领域，推动人工智能在自动驾驶、智慧城市等场景的深度应用，构建丰富多样、健康向善的人工智

能应用生态”[30]。2025年8月,国务院关于深入实施“人工智能+”行动的意见,更进一步完善了推动人工智能赋能千行百业的顶层规划。第三,强化人工智能发展人才支撑。创新性人才是引领人工智能变革的核心因素。中国深入推进教育、科技、人才的一体化改革,完善人工智能学科布局,强化师资力量建设;积极搭建产学研合作平台,促进科研成果高效转化;完善人才培养、评价与激励的体制机制,为人工智能繁荣发展提供人才保障和制度支撑。中国在人工智能技术、产业与人才领域的高质量发展,不仅为维护自身安全提供了关键支撑,也使其在全球人工智能安全治理中扮演着日益重要的角色。

(二)搭建治理框架,完善人工智能安全治理制度规范

在数智时代,人工智能安全已成为全球技术治理的重要议题。各国在治理实践中逐渐形成不同的理念与模式,导致全球人工智能安全治理框架呈现明显的差异化特征。在此背景下,中国从标准、规则、伦理与立法四个层面推动构建包容性治理框架,为全球人工智能安全治理提供制度性规范。在标准层面,中国积极推进人工智能技术标准的国际化,鼓励科技企业、科研机构等主体深度参与国际标准制定,以提升在人工智能安全治理中的话语权。例如,2023年,中国电子技术标准化研究院联合阿里巴巴集团、阿里云智能集团、达摩院共同发布《AIGC治理与实践白皮书》,系统梳理了中国在人工智能安全治理方面的实践经验。2025年,鹏城实验室牵头制定的IEEE Std 3404™-2025算力网国际标准正式发布,为我国后续算力各项工作的标准化和国际化打下坚实基础[31]。在规则层面,中国秉持敏捷治理原则,强调“软法”与“硬法”相结合,通过在硬性约束与软性监管之间寻求动态平衡,不断提升人工智能的透明性、可控性与可靠性。同时,中国积极推动国内安全规则与国际机制的有效衔接,支持在联合国框架下开展人工智能安全规则的平等协商,尤其注重保障发展中国家的安全利益与参与权利。在伦理层面,中国始终坚持“以人为本、智能向善”的理念宗旨。2020年发布的《国家新一代人工智能标准体系建设指南》明确了伦理安全标准的重要地位。2021年出台的《网络安全标准实践指南——人工智能伦理安全风险防范指引》,系统总结了伦理问题与安全风险的具体类型,并明晰了各方的治理责任。此后,《中国关于规范人工智能军事应用的立场文件》《新一代人工智能伦理规范》等文件陆续发布,也进一步强调了“智能向善”和“增进人类福祉”的伦理导向。在立法层面,中国逐步构建起层次分明、覆盖全面的人工智能法律体系。依托《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》等法规,确立了包容审慎、分类分级的监管方式[32],并推动建立“风险等级测试评估体系”[33],以系统应对技术应用中的各类安全风险。同时,中国“在数据跨境流动、算法透明度等争议领域,主张分类施策而非一刀切禁令,为技术迭代保留弹性空间”[34]。这一动态平衡的监管思路为国际社会提供了可借鉴的法治模式。

(三)开展对话协商,深化人工智能国际安全治理合作

习近平在中共中央政治局第二十次集体学习时指出,“人工智能可以是造福人类的国际公共产品”“要广泛展开人工智能国际合作”[35]。中国深知推动人工智能健康发展需要多元主体集体行动、协同共治。一方面,中国积极搭建高层次对话平台,推动广泛深入的国际交流。秉持平等、包容、互信的原则,中国不断拓展多边对话机制,致力于构建系统性、立体化的全球人工智能安全合作网络。自2014年起,中国已连续十年在浙江乌镇成功举办世界互联网大会(WIC)。在这一平台的持续推动下,2025年6月,世界互联网大会人工智能专业委员会主办“全球人工智能安全与治理框架设计”研讨会,汇集国际专家围绕敏捷治理、机制可持续性、风

险分级等关键议题展开深入讨论，探讨应对全球人工智能安全风险的可行路径。2025 年 7 月，在上海召开的世界人工智能大会（WAIC）“人工智能治理国际多方合作论坛”上，中国倡议成立世界人工智能合作组织，共同推进包括安全议题在内的人工智能全球治理。另一方面，中国持续深化双边和多边合作机制，推动国际治理框架的对接协调。在双边层面，中法两国以建交 60 周年为契机，共同发表《关于人工智能和全球治理的联合声明》，强调“双方还将依托联合国层面开展的工作，致力于加强人工智能治理的国际合作以及各人工智能治理框架和倡议之间的互操作性”[36] 。同时，《中阿数据安全合作倡议》的发布，以及《关于中德数据跨境流动合作的谅解备忘录》的签署，也体现了中国在跨境数据安全合作方面的积极努力。在多边层面，中国坚定支持以联合国为核心推进人工智能安全治理多边合作，积极响应联合国教科文组织、世界知识产权组织、国际电信联盟等国际组织发布的相关倡议与行动指南，不断加强与欧盟、东盟、金砖国家等多边机制的务实合作。中国在全球人工智能安全治理中的广泛参与和实质贡献，充分展现了其推动人工智能安全、可靠、可控发展的良好意愿和坚定决心。

（四）承担大国责任，扩大人工智能安全公共产品供给

基于人工智能安全议题的复杂性与特殊性，现有全球治理体系缺乏公正、有效且可持续的国际公共产品。在此背景下，中国立足于自身人工智能安全治理实践，积极扩大人工智能安全公共产品的全球供给，致力于构建开放包容、普惠共享的全球人工智能安全治理体系。一是积极发布人工智能安全治理倡议和指南，彰显中国担当。2025 年 7 月，中国在世界人工智能大会暨人工智能全球治理高级别会议发表《人工智能全球治理行动计划》，从风险防治、数据管理、技术开发与服务、国际合作等多维度，为各国开展人工智能安全治理提供了系统性规范[30] 。此外，在同期举办的第四届 AI 安全国际对话会上达成并发布的《AI 安全国际对话上海共识》，聚焦“人工智能超越人类智能后的失控风险”，进一步深化了“确保高级人工智能系统的对齐与人类控制，保障人类福祉”的安全伦理规范。2025 年 9 月，《人工智能安全治理框架》（2.0 版）正式发布，在 1.0 版基础上突出了对人工智能失控风险的前瞻应对与风险动态分级的精细化治理方案，并强化了国际合作与应用安全指引，为国际社会人工智能的安全开发与应用提供了可参考的最新政策规范。二是深入推进人工智能包容普惠发展。中国积极响应联合国对人工智能的普惠性发展要求，持续为发展中国家提供支持与援助。例如，由中国援助的津巴布韦高性能超级计算机中心已交付津方使用，不仅显著提升了该国的高性能计算能力，还为其人工智能技术研发与应用提供了关键基础设施支撑，助力津方增强本土人工智能安全与数字化转型[37] 。在数据安全方面，深信服的“XDR + 安全 GPT”AI 安全运营中心在东南亚多国落地，不仅大幅提升区域威胁检测与响应效率，还提升了当地人工智能应用的安全防护水平。此外，上海人工智能实验室在探索 AI 安全与性能平衡的方案中提出的“45°平衡律”，也为发展中国家优化技术路径提供了重要参考[38] 。中国不断扩大人工智能安全公共产品的覆盖范围，并持续提高其供给质量，切实助力发展中国家加强人工智能安全能力建设，为构建均衡、普惠的全球人工智能安全治理体系作出了实质性贡献，充分彰显了中国在全球人工智能安全治理中的建设性角色。

四、结 语

在数字全球化纵深发展的时代背景下，全球人工智能安全治理已成为全球治理体系的重要组成部分。构建公平、合理、包容的全球人工智能安全治理框架，实现全球人工智能领域的

善治,不仅关乎技术本身的可持续发展,更对维护国际和平稳定、促进全球共同发展具有深远战略意义。当前,面对全球人工智能安全治理的诸多挑战,中国作为最大的发展中国家和全球治理体系的重要参与方,始终秉持负责任态度,积极履行大国义务,不断探索并实践符合多边利益、统筹发展与安全的人工智能安全治理新范式。面向未来,中国将与国际社会携手合作,共同推动构建更加开放包容、公正有序的人工智能安全架构和发展生态,为构建人类命运共同体贡献智慧与力量。

参考文献

[1] 姚旭,李琛峣.全球人工智能安全治理:面向全球南方需求的转型路径[J].中国信息安全,2025(04):38-41,46.
[2] 朱荣生,陈琪.美国对华人工智能政策:权力博弈还是安全驱动[J].和平与发展,2022(06):47-70.
[3] 马述忠,李折周.美式"数字霸权":特征、动因与影响[J].宏观经济研究,2023(10):106-115.
[4] 臧雷振,陈浩.生成式人工智能算法风险及社会治理挑战[J].中共中央党校(国家行政学院)学报,2025(01):43-53.
[5] 李猛.人工智能时代的社会公正风险:何种社会?哪些风险?[J].治理研究,2023(03):118-129.
[6] 俎文天.人工智能全球治理合作:问题、进路与中国参与[J].国际经济评论,2025(04):153-176.
[7] 薛澜,赵静.人工智能国际治理:基于技术特性与议题属性的分析[J].国际经济评论,2024(03):52-69.
[8] 刘世强.全球安全倡议的时代价值、内在逻辑与实践路径[J].当代世界,2024(01):35-40.
[9] 蔡翠红.推动构建人工智能全球善治新范式[J].国家治理,2025(14):57-65.
[10] 贾开,俞晗之,薛澜.人工智能全球治理新阶段的特征、赤字与改革方向[J].国际论坛,2024(03):62-78.
[11] 习近平.弘扬和平共处五项原则 携手构建人类命运共同体———在和平共处五项原则发表70周年纪念大会上的讲话[N].人民日报,2024-06-29(02).
[12] 习近平向2024年世界互联网大会乌镇峰会开幕视频致贺[N].人民日报,2024-11-21(01).
[13] 祁昊天.从整合、多元到协商:安全共同体理念演化与全球安全治理[J].亚太安全与海洋研究,2024(04):16-36.
[14] 习近平.习近平谈治国理政:第4卷[M].北京:外文出版社,2022.
[15] 习近平.习近平谈治国理政:第2卷[M].北京:外文出版社,2017.
[16] 习近平主持"上海合作组织+"会议并发表重要讲话[N].人民日报,2025-09-02(01).
[17] 保建云.百年变局下的全球数字治理变革及数字风险治理[J].人民论坛,2023(12):42-47.
[18] 罗有成.数字时代主权的嬗变与国际安全秩序重塑[J].国际展望,2024(06):89-112.
[19] 王秉,王渊洁.综合人工智能安全:人工智能与安全的共舞[J].湖南师范大学社会科学学报,2025(03):37-49.
[20] 黄大慧,朱榆雯.构建人类安全共同体:理论逻辑与实践路径[J].国家安全论坛,2024(01):59-74.
[21] 习近平.携手构建公正合理的全球治理体系[N].人民日报,2024-11-20(02).
[22] 毕海东.全球安全观:形成过程、丰富内涵与践行价值[J].国家安全研究,2024(01):38-58.
[23] 巩辰.人工智能时代的全球治理:一般路径与中国方案[J].人文杂志,2019(08):38-46.
[24] 王毅主持联合国安理会"践行多边主义,改革完善全球治理"高级别会议[N].人民日报,2025-02-20(03).
[25] 刘胜湘,唐探奇.安全不可分割:理论内涵与实现路径——兼论全球安全倡议[J].国际安全研究,2023(05):3-28.
[26] 习近平.习近平谈治国理政:第1卷[M].北京:外文出版社,2018.
[27] 习近平.把握时代大势 共促世界繁荣——在亚太经合组织工商领导人峰会上的书面演讲[N].人民日报,2024-11-17(02).
[28] 中国信息通信研究院.中国算力发展指数白皮书(2023年)[EB/OL].(2023-09)[2025-08-24]. http://www.caict.ac.cn/kxyj/qwfb/bps/202309/P020240326630458153765.pdf.

[29] 周文，张奕涵．人工智能创新发展的中国范式：来自 DeepSeek 的启示[J]．社会科学辑刊，2025(05)：107-121.
[30] 新华社．人工智能全球治理行动计划(全文)[EB/OL]．(2025-07-26)[2025-08-20].https://www.gov.cn/yaowen/liebiao/202507/content_7033929.htm.
[31] 鹏城实验室牵头制定IEEE算力网国际标准正式发布[EB/OL]．(2025-08-18)[2025-08-20].https://www.pcl.ac.cn/html/943/2025-08-18/content-4622.html.
[32] 国家网信办，国家发展改革委，教育部．生成式人工智能服务管理暂行办法[EB/OL]．(2023-07-10)[2025-08-20].https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.
[33] 外交部．全球人工智能治理倡议[EB/OL]．(2023-10-20)[2025-08-20]. https://www.mfa.gov.cn/wjb_673085/zzjg_673183/jks_674633/fywj_674643/202310/t20231020_11164831.shtml.
[34] 韩娜．DeepSeek 推动 AI 全球治理新转向[EB/OL]．(2025-02-19)[2025-08-20].https://www.cssn.cn/gjaqx/202502/t20250219_5848243.shtml.
[35] 习近平在中共中央政治局第二十次集体学习时强调 坚持自立自强 突出应用导向 推动人工智能健康有序发展[N]．人民日报，2025-04-27(01).
[36] 中华人民共和国和法兰西共和国关于人工智能和全球治理的联合声明[EB/OL]．(2024-05-07)[2025-08-20]. https://www.gov.cn/yaowen/liebiao/202405/content_6949586.htm.
[37] 新华网．中国援助的超算中心助力津巴布韦数字化转型[EB/OL]．(2025-08-19)[2025-08-20]. http://www.news.cn/20250819/afc3d0142b814918a606e7ca24934f35/c.html.
[38] 《中国智·惠世界(2025)》案例集发布 展现AI国际合作多元成果[EB/OL]．(2025-07-26)[2025-08-20]. https://www.ndrc.gov.cn/fzggw/wld/zsj/zyhd/202507/t20250726_1399428_ext.html.

责任编辑：李 洁
编辑部网址：http://sk.swpuxb.com

China's Approach and Strategic Practices in Advancing Global AI Security Governance

LIU Shiqiang, XU Huiyuan
(School of Marxism, Southwest University of Finance and Economics, Chengdu Sichuan, 611130, China)

Abstract: The widespread application of artificial intelligence technology promotes economic development and social progress, but it also introduces a series of security risks. Currently, global AI security governance faces multiple challenges, including a deteriorating security environment, increasing security risks, inadequate security mechanisms, and outdated security concepts. In response, China actively implements the Global Security Initiative (GSI) and elaborates on the goals, main components, methodologies, and core drives of AI security governance from four dimensions: common security, comprehensive security, cooperative security, and sustainable security. This has led to the formulation of China's proposal for global AI security governance. Guided by this approach, China is accelerating its own AI development, establishing a framework for AI security governance, promoting international cooperation in AI security, and expanding the provision of public goods to safeguard AI security. These efforts are strong contributions ensuring that AI better serves world peace and human development.

Keywords: AI security; global AI governance; global security concept;community with a shared digital future; China's solution