上海国际问题研究院
SHANGHAI INSTITUTES FOR INTERNATIONAL STUDIES
清华大学战略与安全研究中心
CENTER FOR INTERNATIONAL STRATEGY AND SECURITY TSINGHUA UNIVERSITY
北京大学互联网发展研究中心

建构人工智能发展的国际规则
--趋势、领域与中国角色

作　者：鲁传颖
田　丽
封　帅
周亦奇
王玉柱
王天禅
张璐瑶

SIIS REPORT
总第29期
2023年10月

上海国际问题研究院
SHANGHAI INSTITUTES FOR INTERNATIONAL STUDIES

清华大学战略与安全研究中心
CENTER FOR
INTERNATIONAL STRATEGY AND SECURITY
TSINGHUA UNIVERSITY

北京大学
互联网发展研究中心
Internet Development Research Institution

建构人工智能发展的国际规则
——趋势、领域与中国角色

鲁传颖　田　丽　封　帅　周亦奇
王玉柱　王天禅　张璐瑶

上海国际问题研究院
国际传播中心

2023 年 10 月

作者简介

鲁传颖
上海国际问题研究院公共政策与创新研究所副所长
上研院网络空间国际治理研究中心秘书长
清华大学战略与安全研究中心特约专家

田丽
北京大学新媒体研究院副院长
北京大学互联网发展研究中心主任

封帅
上海国际问题研究院国际战略与安全研究所
所长助理

周亦奇
上海国际问题研究院公共政策与创新研究所
副研究员

王玉柱
上海国际问题研究院世界经济研究所研究员
上研院“一带一路”与上海研究中心秘书长

王天禅
上海国际问题研究院网络空间国际治理研究中心
实习生

张璐瑶
上海国际问题研究院网络空间国际治理研究中心
实习生

目录

引　言

2023 年以来，人工智能技术的突破又一次成为全球瞩目的焦点。在高速进步的生成式人工智能，以及取得突破性进展的大语言模型的带动下，人类似乎已经开始触及通用人工智能技术的门槛。围绕着人工智能技术跨越式发展而展开的关于智能化社会的想象吸引了全球社交媒体与资本市场的追捧，但关于人工智能技术可能引发的各种安全风险和社会问题也成为社会关注的焦点。为了确保人工智能技术始终保持良性发展，确保技术进步与人类价值观相向而行，为人工智能发展设定合理的国际规则逐渐成为全球范围的重要共识。

2023 年 7 月 18 日，联合国安理会举行了主题为“人工智能给国际和平与安全带来的机遇与风险”的高级别公开会议，这是联合国安理会首次举行人工智能主题会议。在本次会议上，联合国秘书长古特雷斯公开呼吁对人工智能技术风险进行有效治理，并且提议在联合国框架内成立一个新的监管机构推动人工智能领域的全球治理。国际规则具有重要的导向性作用，世界各国政府和企业都希望参与甚至主导人工智能国际规则的建构议程。可以预见，全球围绕人工智能国际规则制定问题将会展开复杂的竞合博弈，最终形成的国际规则体系也将成为影响全球人工智能技术、产业和安全的重要变量。

多年来，在有关各方的共同努力下，我国人工智能技术和产业都取得了较为丰硕的成果，人工智能已经成为推动我国经济高质量发展的重要力量，中国也已经成为全球人工智能版图上的重要一环。中国在人工智能领域已经积累了丰富的治理经验，能够为全球人工智能规则建设提供宝贵的发展中国家视角。因此，中国应积极参与人工智能国际规则制定，一方面，将中国在人工智能领域的治理经验转化为国际规则。另一方面，通过参与国际规则的讨论，加深对先进治理理念的理解，更好地促进国内人工智能的安全与发展。

陈东晓

一、人工智能国际规则建构的总体趋势

人工智能国际规则建构之所以受到世界各国的普遍重视，在于人工智能已经成为影响国家安全和发展的重要力量。美国、中国、欧盟等主要大国先后发布了人工智能的国家发展战略，围绕着人工智能而展开的国际竞争已经成为当前国际议程的重要组成部分。人工智能及其相关衍生技术本身具有强渗透性和高赋能性，能够与诸多社会生产部门建立深度链接。如果其技术潜能得到充分释放，可能对人类社会的生产生活带来颠覆式的影响。因此，在推动人工智能技术进步的过程中，必须充分重视技术发展与社会稳定之间的动态平衡。

在这种情况下，世界各国逐渐认识到通过规则建构推动人工智能治理的必要性。美欧等发达国家认为掌握人工智能规则建构的主动权是保持其在技术与产业领域优势地位的重要一环。他们积极推动相关议题的研究，并试图主导规则建构的国际议程，尝试在国内、国际两个层面着手进行机制完善和规则构建。然而，相比于更具同质性的国内环境而言，国家之间的互动关系更为复杂、不同主体之间的利益纠缠更为突出、文化和价值观更为多元，这也造成了形成人工智能国际规则的共识更加困难。

从实际进展情况来看，当前人工智能国际规则仍处于早期阶段，各主体在不同层次上都在尝试推动各种形式的治理方案，归纳起来包括以下几个方面的内容：

第一，在多边层面，联合国积极主导人工智能国际治理规则建设。2023 年 6 月，联合国秘书长古特雷斯公开表示，支持建立人工智能领域的“国际原子能机构（IAEA）”，对人工智能技术进行监管。这一事件意味着联合国希望在未来的人工智能国际规则制定的进程中将进行角色调整和功能增强，将其作用范围从原则层面深入到机制层面，建设更具约束力的人工智能监管机构，这可能会成为未来人工智能国际规则建构的基础架构。此外，在 2023 年 7 月联合国最新出台的《和平新议程》中，也专门提及人工智能和致命性自主武器系统对国际和平与安全的影响，建议各国紧急制定关于负责任地设计、开发和使用人工智能的国家战略；通过多边、多方进程推动人工智能军事领域的设计、开发和使用；最终就

人工智能全球范围内的监管框架达成一致。

近年，联合国下属多个机构积极参与人工智能治理议程，取得阶段性成果。

2021 年，由联合国教科文组织人工智能伦理建议特设专家组出台的《人工智能伦理问题建议书》（Recommendation on the Ethics of Artificial Intelligence）是人工智能国际规则制定的重要里程碑。该建议书从人类共同的价值观和可持续发展目标的基本立场出发，提出了十一项关于人工智能发展的指导性原则，被 193 个成员国共同接纳。2014 年起，联合国在特定常规武器公约会谈机制下召开了多次非正式专家会议，围绕着致命性自主武器系统的相关问题进行了讨论，已有近 100 个国家参与了相关活动。此外，在联合国裁军研究所、联合国毒品和犯罪问题办公室等机构的活动和议程中，也有将致命性自主武器系统的伦理和安全、深度伪造的政治风险等问题纳入讨论的范围，但尚未出台明确的原则倡议或国际规则。

第二，在发达国家阵营，美国、英国、荷兰加大筹码试图垄断人工智能国际规则主导权。美国在推动其国内规则国际化方面做了很多工作，在美国《关键与新兴技术国家战略》（National Strategy for Critical and Emerging Technology）等战略文件的引导下，美国将“推动与盟友的合作” “建设新兴技术的世界领导者地位”作为基本立场，建立了一系列人工智能的国际合作机制。例如，在军事防务领域，美国不仅在北约框架内发布《人工智能战略》，宣扬自身的安全和地缘竞争立场，还新建立了人工智能防务伙伴关系（AI PfD）、美英澳三国集团（AUKUS）新机制，进行国防和安全领域人工智能的使用规则协调；在标准协调领域，美国则依托四方（QUAD）关键新兴技术工作组、美欧贸易和技术委员会（TTC）以及美印人工智能倡议（USIAI）等机制，融合亚太、印太和西欧的关键盟友，进行技术合作并推行共同的人工智能技术标准；在价值观领域，美国则通过民主峰会、自由在线联盟人工智能与人权工作组（T-FAIR）等机制，选择性听取其“小圈子”内的伙伴对于意识形态和价值观的赞许和认同，理所当然地证实其观念的真理性，并积极推动其从区域化走向国际化。

英国、荷兰等西方国家也开始搭建起人工智能领域的“观念一致同盟”，希望把主导人工智能国际规则的能力控制在少数几个发达国家之内。2023 年 2 月，荷兰在海牙举行“军事领域负责任人工智能”峰会，该峰会邀请了全球 80 多个国家参加，会议核心目标有三：提升军事领域负责人使用人工智能的政治重要性、

动员激活更广泛的利益攸关方参与、分享经验实践。该会议虽号称寻求凝聚国际社会在人工智能在军事领域内的共识，但其实际包容性依然受到西方地缘政治议程的影响，俄罗斯未获此会议邀请。在会上中国代表详细阐释了中国在人工智能安全治理领域的主张，提出智能向善、以人为本、多边主义的治理原则。2023 年 3 月，英国政府发布题为《促进创新的人工智能监管方法》白皮书，提出了自己的人工智能监管框架。此外，英国宣布将在 2023 年内举办首届人工智能安全全球峰会。该会议虽然打着“协调各国应对人工智能风险的共同方法”的旗号，似乎能够对人工智能国际规则制定起到正向的促进作用。但事实上，英国首相苏纳克却无意透露出英国的真实意图：努力协调英美合作并巩固“在未来技术领域的共同领导地位”。英国试图在伦敦设立人工智能监管国际机构，从而掌控人工智能规则制定方面的国际话语权。

第三，发展中国家在人工智能国际规则建立上仍处弱势地位。发展中国家作为一个整体，在人工智能的国际规则制定上还处于早期阶段。一方面，这与广大发展中国家的人工智能发展还处于落后阶段有关；另一方面也与发展中国家参与国际规则的动力和意识不强有很大关系。面对西方国家不断强化人工智能国际规则体系，广大发展中国家应当快速觉醒，加大参与力度，贡献各自智慧。作为非西方国家和发展中国家的代表，中国是为数不多在人工智能国际规则领域积极发声的国家。中国不仅积极参与联合国在人工智能国际规则领域的工作，还在 G20、金砖国家和上合组织等国际组织中积极推动和加强人工智能国际规则方面的合作。不仅如此，中国也在积极推动广大发展中国家在人工智能领域规则制定方面的合作。例如，中国与东盟国家和中东国家围绕着人工智能议题举办了各种形式的会晤合作，并成功建立起了的默契和互动规范。

公平合理的国际规则需要兼顾西方与非西方国家的立场和声音。在当前失衡的现实背景下，需要非西方国家更加团结努力，在国际规则制定中积极作为，为自身争取应有的利益。一个真正合理的国际规则体系应该是包容而非排斥的，是尊重最广泛国家利益和关切，而非仅仅关注小团体利益的。人工智能可能带来的各种安全风险是人类需要共同面对的挑战，非西方国家在人工智能领域的合理诉求不能被忽视，非西方国家在治理方面的智慧也值得借鉴。

二、人工智能国际规则覆盖的领域

人工智能已经成为数字经济发展的核心驱动力。尤其是在生成式人工智能（AIGC）和通用型人工智能（AGI）不断取得技术突破的背景下，人工智能在技术上的治理需求和主权国家治理效能之间的鸿沟有日益扩大之势。目前，全球范围内人工智能的国际规则制定仍处于起步阶段，在形式和内容上主要以国际软法和间接规制为主。但是，由于全球主要国家和地区组织都提高了对人工智能领域建章立制的重视程度，全球范围内的人工智能规制竞赛实际上已然开始。各国和地区不仅积极起草、出台规范人工智能技术的法律法规，还竞相出台与人工智能技术向善发展相关的政策指南等。这种规制竞赛有助于推动全球形成一个完善的、基于普遍共识的人工智能治理框架，也为我国争取国际规则制定的“先手”带来了机遇与挑战。

根据现阶段人工智能技术发展状况，结合各国在人工智能治理方面的经验总结，目前全球人工智能国际规则已经形成了较为稳定的覆盖领域，在相关领域的规则建构将成为国际规则的主要支柱。

（一）伦理问题

人工智能技术伦理问题一直是国际规则制定的重点领域，其目的是通过对人工智能伦理的有效规范，确保技术进步不会偏离人类基本的利益轨道和道德法则。在伦理问题方面，国际上已经出现了很多关于规则的探索，部分重要的国际组织和机构提出了人工智能的伦理原则和指导方针，也有一些国家也开始规范专门领域的技术伦理问题。

其中有代表性的成果包括下列内容：2021 年 11 月联合国教科文组织（UNESCO）发布了《人工智能伦理问题建议书》，成为首个规范性的人工智能全球伦理框架，同时赋予各国在相应层面应用该框架的责任；欧盟方面，2021 年 4 月提出的《人工智能法案》经过多轮磋商后已进入最后阶段，未来极大可能成

为全球首份人工智能监管法案，并以此形成规则示范效应；美国在人工智能伦理问题上的规范则更为细化，例如，2020 年 2 月美国国防部发布《人工智能伦理原则》，以及美国国家情报总监办公室 7 月发布的《情报界人工智能伦理原则》和《情报界人工智能伦理框架》等；2021 年 12 月，中国发布《关于规范人工智能军事应用的立场文件》，呼吁各方遵守国家或地区人工智能伦理道德准则；2022 年 11 月，中国发布《关于加强人工智能伦理治理的立场文件》，从技术监管、研发、使用及国际合作等方面提出具体主张，呼吁各方秉持共商共建共享理念，推动国际人工智能伦理治理。

这些原则和指导方针虽然在具体内容上有所差异，但都强调了人工智能应该符合人类的价值观和利益，尊重人的尊严和权利，保障公平、透明、可解释、可信赖和可问责等方面的内容，该领域的规则制定将确保人工智能发展的基本方向。

（二）国际安全

人工智能技术的发展能够产生广泛的国际影响，对于全球安全具有重要意义。从国际安全的角度来看，人工智能与军事武器的结合已经对国际安全构成了实际上的挑战。

目前国际上已经有很多组织和机构在关注和讨论人工智能对军事化、核武器等方面的影响和挑战。例如，联合国裁军研究所于 2023 年 2 月发布《迈向负责任的国防人工智能：绘制和比较各国采用的人工智能原则》报告，为“迈向负责任的国防人工智能”项目第 1 阶段成果。该项目旨在建立对负责任的人工智能系统研究、设计、开发、部署、使用的关键共识，审查负责任的人工智能在国防领域的应用情况。2021 年，红十字国际委员会在日内瓦发布了《关于自主武器系统的立场》文件，明确指出当前军事利益和投资的趋势已经充分表明，如果不确立国际公认的限制标准，未来的自主武器系统可能会日益依赖人工智能和机器学习软件，由此其在设计层面存在的不可预测性令人担忧。这些组织和机构呼吁加强对人工智能在军事领域的监督和限制，防止人工智能引发新的冲突和危机，维护国际和平与安全。针对人工智能在不同领域的安全影响仍会不断扩大这一情况，建立相应的国际规则具有重要意义。

人工智能技术在核领域的应用也逐渐受到国际社会的关注，尤其是人工智能

对核安全的影响。2020 年 9 月，第 64 届国际原子能机构大会首次讨论了人工智能在核领域的应用问题，并展示了人工智能赋能下的核技术如何造福于人类健康、水资源管理和核聚变研究。2022 年 8 月，国际原子能机构发布题为《人工智能加速核应用、核科学与核技术》的报告，介绍了人工智能在核科学和核应用、核电，以及核安全和保障核查等领域的应用。2023 年 6 月，《人工智能将如何改变核世界的信息和计算机安全》一文在《国际原子能机构简报》中刊出，专门强调了人工智能的快速发展对核安全带来的诸多风险。文章指出，恶意行为者可能会利用人工智能发起更先进、更有针对性的攻击，或利用它来破坏核设施和放射性设施中网络、系统和敏感信息的完整性。为此，国际原子能机构还制定了题为“加强核设施计算机安全事件分析”的“联合研究计划”（CRP），以支持加强计算机安全的研究。这一计划汇集了 13 个国家的代表，致力于提高核设施的计算机安全能力，包括通过人工智能技术来检测网络攻击带来的异常情况。自此，人工智能的核安全问题被正式纳入国际机制的议事日程。

（三）技术安全

从人工智能技术诞生伊始，如何确保技术本身的安全性就是各国关注的焦点。简单来说，人工智能的技术安全集中在算法、数据和场景三个层面。而通过相关标准的设定，设置技术发展的规范将构成人工智能国际规则的另一个重要组成部分。具体而言，人工智能技术的发展应秉持以下原则：其一，可扩展监督，即确保人工智能技术在开发、使用和更新过程中得到有效的监督，防止人工智能技术被用于危害人类安全的目的；其二，机制可解释性，即帮助人们理解人工智能技术的决策过程，从而更好地预防和应对人工智能技术可能产生的负面影响；其三，危险能力测试可以帮助识别人工智能技术的潜在风险，并采取措施降低这些风险。当前，国际社会已经围绕上述三项技术安全标准开展了大量工作，诸多国际组织和地区、国家都开始在标准领域发力。

在技术安全方面，国际上已经有一些标准化组织和机构在积极制定和推进人工智能的安全标准和规范。2022 年 7 月，国际标准化组织（ISO）和国际电工委员会（IEC）联合发布了两项人工智能领域新的基础标准，即《信息技术——人工智能概念和术语》（ISO/IEC 22989:2022）和《运用机器学习的人工智能

系统框架》（ISO/IEC 23053:2022）。上述技术标准和框架的提出旨在为全球数字化转型提供规范保障，并且为通用型人工智能技术的发展提供系统性框架。此外，国际电信联盟（ITU）也积极投身于人工智能技术标准与规范的建立，并通过下设的电信标准化部门（ITU-T）和无线电通信部门（ITU-R）来专门负责人工智能和机器学习相关的标准化工作。此外，由中国电信牵头在国际电信联盟制定的国际标准《AI 增强电信运营管理架构》（Framework of Artificial Intelligence Enhanced Telecom Operation and Management，AITOM）于 2021 年正式发布，将中国电信的智慧电信运营管理技术与安全实践以标准的形式推广到全球产业界，帮助行业解决电信运营管理中人工智能技术应用的问题。

在国家和地区组织层面，围绕人工智能技术规范和标准的制定也在如火如荼推进。其中，欧盟作为全球首屈一指的“规范性力量”，早在 2019 年 4 月就提出了人工智能的伦理准则，列出了 7 项评价“可信赖人工智能”的标准。2019 年 5 月 22 日，经济合作与发展组织（OECD）通过了《人工智能建议书》，这是首个人工智能领域的政府间标准，旨在通过对人工智能技术的负责任管理，确保其尊重人权和民主价值观，并促进人工智能的技术创新。2022 年，欧盟委员会提出了《人工智能责任指令》提案，进一步提出人工智能安全责任认定机制，来确定当人工智能发生故障或造成伤害时谁应该承担责任。2023 年 6 月，欧洲议会又通过欧盟《人工智能法案》草案，对人工智能系统进行风险分类，限制深度伪造，并对 ChatGPT 等生成式人工智能提出了更高透明度的要求。

此外，美国、英国等国在人工智能技术安全的基本原则和具体标准方面也展开了一定的规范实践活动。例如，英国在 2021 年 9 月 22 日发布的首份《国家人工智能战略》中指出，对内要制定跨部门标准以确保人工智能算法的透明度，对外则要参与全球人工智能标准化制定工作，并且要提高政府对人工智能技术安全的认识。美国人工智能国家安全委员会（NSCAI）在 2021 年 3 月 2 日正式发布的最终研究报告中也强调，要对人工智能系统建立合理的信心，即确保人工智能系统的坚实、鲁棒和可靠。此外，2023 年 5 月 23 日美国白宫发布了《国家人工智能研发战略计划》（The National Artificial Intelligence R&D Strategic Plan），其中的战略支柱之一就是确保人工智能系统的安全性。

值得注意的是，2020 年以来学术界和产业界对于人工智能价值对齐（AI Value Alignment）问题的讨论进入白热化阶段，并认为这是人工智能技术安全

面临的重大挑战之一。由于人工智能技术的发展产生了诸多不确定性，如何通过从设计到使用的“全链式”治理来确保人工智能与人类的预期目标、道德准则和价值观保持一致成为当前的核心关切，即价值对齐问题。目前，IBM、谷歌，以及 OpenAI 等领先的人工智能企业都发布了自己在价值对齐问题上的解决方案和路径，成为全球人工智能技术规范建立的生力军。2023 年，数百名世界领先的技术研发人员、学者和企业领袖在非政府组织人工智能安全中心（Center for AI Safety）的协调下共同签署了一份网络声明，表明“减轻人工智能带来的灭绝风险应该与流行病和核战争等其他社会规模风险一起成为全球优先事项。”

总体而言，当前国际社会中关于人工智能技术安全的标准和规范涵盖了人工智能的术语、框架、方法、理念等多个方面，旨在提升人工智能的安全性、可靠性和质量。该领域的规则制定确保人工智能技术始终在安全的标准下运动。

（四）技术创新

促进人工智能安全有效发展是人工智能国际规则的落脚点。通过规则建构保障技术要素能力的充分释放，也是规则制定的重要目的。因而可以说，技术创新发展是推动人工智能国际规则制定的充分条件，也是国际规则制定的目标与归宿。

目前，国际上较为领先的地区和国家已经着手推动和支持在人工智能技术创新发展方面的布局。欧盟委员会在 2020 年 2 月发布的《人工智能白皮书》中提出，要大幅提高人工智能研究和创新领域的投资水平。日本在 2019 年发布的《人工智能战略》中提出，要把人工智能作为国家战略性技术来发展，并推动人工智能技术发展与社会应用的结合。韩国在 2019 年 12 月发布的《人工智能国家战略》中提出扩充人工智能基础设施、确保掌握人工智能技术竞争力，以及创新规制和调整法律制度等目标。2023 年 5 月，美国白宫继 2016 年、2019 年之后发布第三版《国家人工智能研发战略计划》，在规划中将投资人工智能技术研发置于首位，并强调了人才和国际合作对于技术创新的支持作用。中国在人工智能技术发展方面也进行了规划部署，例如 2017 年国务院印发的《新一代人工智能发展规划》，是我国第一份在人工智能技术发展和应用领域进行系统部署的战略规划。2022 年 7 月，科技部等六部委联合印发《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》的通知，系统性地指导各地方和各主体加快人工智

能场景应用，以场景创新推动人工智能的技术升级、产业增长，进而实现经济高质量发展。

可以看出，当前国际社会在推动人工智能技术创新领域的着力点主要包括基础研究、应用开发、人才培养、伦理规范以及国际合作等。上述国家和国际组织通过一系列战略计划和政策措施的部署，在人工智能技术创新发展方面投入大量的资金和资源，用以培育人才和团队、建设基础设施和平台，以及促进跨领域和跨国的合作与交流。

（五）社会发展

充分利用人工智能技术，带动社会各领域的全面发展，是世界各国的共同愿望。但要保证智能技术的成果能够成功外溢到其他社会领域，相关的制度引导非常重要。从发展问题的视角来看，目前国际上已经有一些组织和机构在关注和应对人工智能对经济社会发展的影响和挑战，但总体来看相关制度建设仍有进一步完善的空间。

目前，就人工智能与社会发展问题形成的国际机制主要集中在多边领域。例如，2023 年 7 月 6 日在中国上海举行的 2023 世界人工智能大会（WAIC 2023）开幕式上，联合国工业发展组织（UNIDO）发布了成立全球工业和制造业人工智能联盟的倡议（简称“AIM-Global”联盟）。这一开创性的举措旨在联合国家政府、私营部门和国际组织，以及专注于推动人工智能技术负责的、可持续和包容性应用的行业领导者，共同为全人类创造一个共享开放、自由和安全的数字未来。此外，世界经济论坛关注到人工智能对人类工作模式可能造成的冲击。在 2023 年 6 月发布的《2023 年十大新兴技术报告》认为，生成式人工智能（AIGC）技术除了能对社会和经济带来巨大发展外，将无可避免地造成就业的流失。在教育领域，世界银行于 2021 年 9 月发布《高等教育转向面向人工智能的弹性系统》报告，介绍了世行在人工智能时代发展有效、公平、高效和有弹性的高等教育系统方面的一系列做法。

当前的国际组织和机构主要分析了人工智能对就业、教育、健康、环境等方面的影响，提出了一些应对建议和措施。但如果想实现更加全面的社会发展目标，还需要更多研讨与设计。

三、人工智能治理进程的中国价值观

中国政府高度重视人工智能技术的发展，对其可能带来的产业变革和社会变迁给予了高度关注。国务院先后出台了《新一代人工智能发展规划》等一系列政府文件，对于人工智能技术和产业发展提供了卓有成效的顶层设计。同时，中方也密切关注人工智能发展可能带来的各种风险与挑战，积极探索适用于我国国情的治理方案，确保人工智能产业的健康发展。在治理方面，中方先后发布了《新一代人工智能治理原则》等重要文件，对于人工智能治理的认知也逐渐成熟。经过一段时间的实践磨合与深入的理论探讨，具有中国特色的人工智能治理规则的独特价值观逐步形成，它也将为我国参与人工智能国际规则建构提供坚实的思想内核。长期以来，中方积极参与联合国框架内关于人工智能治理议题的讨论，曾先后在联合国框架内提交了关于规范人工智能军事应用、加强人工智能伦理治理两份立场文件。

在 2023 年 7 月 18 日联合国安理会关于人工智能的讨论中，中方又明确提出了关于人工智能治理的五条原则，即坚持伦理先行、坚持安全可控、坚持公平普惠、坚持开放包容、坚持和平利用。这五条原则是人类命运共同体理念在人工智能领域的反映，是中国关于人工智能治理价值观的充分展现，具有重要的现实意义。

概括而言，经过长时间的研究与磨合，中方已经初步形成了关于人工智能治理和规则制定方面的中国价值观，将为全球人工智能规则建构和治理议程提供卓越的“中国智慧”。

第一，促进技术与产业的安全有序发展是中国价值观的基本导向。

“发展才是硬道理”是改革开放时代中国社会的底层逻辑，也是所有价值观的基础支撑。中方一直将汹涌而来的人工智能技术浪潮视为重大战略机遇，希望能够借助技术进步和产业发展，带动国家竞争力整体跃升和跨越式发展。因此，中国在人工智能治理方面的一个基本导向是：人工智能治理的目标是要减少阻碍人工智能技术发展的不利因素，推动技术的广泛利用，使得更多生产部门和人口

可以享受技术带来的红利。

安全是发展的前提，如果没有安全，发展取得的成果也会得而复失。中国政府高度重视人工智能安全问题，针对人工智能技术发展可能带来的潜在安全风险积极推进相应的治理体系建构。例如，面对生成式人工智能的迅猛发展，国家网信办联合七部委于 2023 年 7 月 10 日颁布《生成式人工智能服务管理暂行办法》，成为全球最早对 AIGC 治理做出贡献的国家之一。该办法深刻地体现了我国在人工智能治理领域的基本价值理念：一方面，通过管理办法阐明我国积极支持 AIGC 技术发展，推动其创造更大的商业、社会价值的立场；另一方面，又对其生成内容与服务提出了清晰的道德伦理要求，并且明确了监管主体与责权分配原则。面对新兴技术的挑战，中国坚持以包容审慎的态度处理新技术所带来的安全问题，保持发展与安全并重的基本立场，系统建构相应的治理体系。

在人工智能治理的中国价值观中，治理本身是为了促进技术和产业的安全有序发展，其目的始终要在维持发展进程的基础上，通过相关的规则建构，确保人工智能技术安全，增进人类的共同福祉。发展与治理密不可分，治理是为了更好地发展，脱离了发展的治理就将成为无源之水无本之木。因此，在推进相关国内规则建设过程中，不能以牺牲技术发展和产业进步为代价，要始终在发展与安全、变革与稳定等多重维度之间寻找合理的平衡点。

第二，建构“负责任的人工智能”是中国价值观的核心内涵。

在国家新一代人工智能治理专业委员会颁布的《新一代人工智能治理原则》中，将“负责任的人工智能”作为治理原则的主题加以描述。发展“负责任的人工智能”作为一个重要的标准贯穿于人工智能的研发与应用过程，构成了人工智能治理的中国价值观的核心内涵。

建构“负责任的人工智能”就要求从基础研发到实践应用，始终要保持在安全可控的范围内，所有的参与主体都要以负责任的态度维持相应的伦理和安全的标准。在伦理方面，要确保人工智能技术始终符合人类的价值观和伦理道德，应以保障社会安全、尊重人类权益为前提，避免误用，禁止滥用、恶用。同时，要确保公民能够较为公平地享受技术进步的成果，推动各行各业转型升级，缩小区域差距，提升弱势群体适应性，努力消除数字鸿沟。在安全方面，则应不断提升人工智能透明性、可解释性、可靠性、可控性，逐步实现可审核、可监督、可追溯、可信赖等方面的要求。在数据获取方面则要尊重和保护个人隐私，充分保障个人

的知情权和选择权，防止信息数据滥用。

人工智能研发者、使用者及其他相关方应以建构“负责任的人工智能”为目标，以高度的社会责任感和自律意识，严格遵守法律法规、伦理道德和标准规范。建立人工智能问责机制，明确研发者、使用者和受用者等的责任。只有建立在这一价值观基础上的治理规则，才能够最大限度体现中方的意愿。

第三，坚持开放交流与多元协作是中国价值观的基本精神。

人工智能技术和产业的发展是一个系统工程，需要广泛的交流与协作。这种协作既包括多元主体之间的分工协作，又包括全球范围内的充分交流与资源分享。在长期的实践中，中国充分理解开放交流与多元协作对于人工智能技术发展的重要意义，它也构成了人工智能治理的中国价值观的基本精神。

多元协作，指的是人工智能技术的发展需要跨学科、跨领域、跨地区、跨国界的充分交流，要积极推动国际组织、政府部门、科研机构、教育机构、企业、社会组织、公众在人工智能发展与治理中的协调互动。在协作互动过程中，应以公正的态度回馈所有的贡献者，以公平的精神保障利益相关者的权益，充分释放各主体的潜能，提供均等的机会。

开放交流，则强调在人工智能领域，世界各国所处的发展阶段不同，人工智能发展存在着各种独特的应用场景，这就需要在交流沟通的基础上开展合作研究和对话，不断推动国际合作，在充分尊重各国人工智能治理原则和实践的前提下，推动形成具有广泛共识的国际人工智能治理框架和标准规范。

人工智能的发展不是一国一地之事，闭关自守和以邻为壑都不是人工智能发展的合理方式，只有在充分尊重各国人工智能治理原则和实践的前提下，推动形成具有广泛共识的国际人工智能治理框架和标准规范，才能增进人类共同福祉。

综上所述，当前全球人工智能发展已经进入了新阶段，技术的进步正在深刻改变人类社会生活，人类正在进入到一个全新的世界。面对全新的局面，如何确保人工智能技术的发展安全、可靠、可控，且始终保持在向善的轨道上，已经成为中国与世界需要共同面对的挑战。在理论研究与实践磨合的基础上，中方逐渐形成了具有中国特色的治理思路与价值观。在它的指引下，中方将不断提升智能化技术手段，优化管理机制，完善治理体系，并以此为思想的锚点，参与到国际规则的制定中去。以此促进新一代人工智能健康有序发展，更好协调发展与治理的关系，确保人工智能推动经济、社会及生态可持续进步，共同建构人工智能领域的人类命运共同体。

四、中国参与人工智能国际规则制定的路径

为了人类经济社会生活的持续发展和人工智能技术的持续向善，中国应当以人类命运共同体理念为引导，主动参与人工智能国际规则制定，以国际社会可理解、可接受的方式来展示和推广人工智能领域的中国价值观与中国经验，并主动提出中国的人工智能国际规则建构方案，争取国际规则建构的主导权。为此，本报告提出中国参与人工智能国际规则制定的四大原则，即以联合国为主舞台开展国际治理（AI Governance by UN）、以促进全球经济社会发展作为关键目标（AI for Development）、以大国稳定机制建构作为基本保障（AI for Stability），并优先发展以私营部门的知识、经验和资源为基础的人工智能应用解决方案（AI Implication by Private Sector）。由此，形成了我国参与人工智能国际规则制定的独特路径，即“以联合国为中心实现人工智能的稳定发展以造福全人类”（Multilateral and Stable Development of AI for All People，MSDP）。

第一，以联合国为代表的多边主义平台为基本立足点。作为当前全球最具代表性和权威性的国际组织，联合国在全球数字治理和规则制定方面发挥着不可替代的主导作用。联合国秘书长古特雷斯在安理会就人工智能举行的首场公开辩论上强调，各方必须共同为发展人工智能努力，以弥合社会、数字与经济鸿沟，而不是让人们之间的距离进一步拉大。为此，中方应积极响应联合国秘书长的号召，在联合国框架内通过广泛的国际协商，在确保发展中国家合理发展利益的情况下制定相关规则。例如，在联合国秘书长科技特使办公室的引领下，积极参与 2024 年联合国未来峰会上就《全球数字契约》达成一致的进程，并在 2026 年之前与各方共同制定一个禁止使用完全自主武器系统的法律约束性协议。中国还要继续支持联合国教科文组织等机构在人工智能伦理和标准方面的工作，并促进《人工智能伦理建议》在全球范围内的实施和监督。

第二，积极参与人工智能国际规则，为弥合发展中国家参与人工智能治理的鸿沟作出更大贡献。当前，在全球人工智能治理领域存在着明显的数字鸿沟，广大发展中国家由于技术和政治的多重原因，在人工智能治理领域明显失语。他们

的内在诉求无法得到有效表达，往往只能被动接受发达国家制定的各种标准，很难深度分享人工智能发展的红利。因此，作为全球最大的发展中国家的中国，在人工智能国际规则制定过程中有着特殊的角色，既是发展中国家参与人工智能规则制定的主要代表，也是扭转发展中国家在人工智能治理方面话语权缺失的关键力量。中国需积极协调发展中国家立场，切实理解发展中国家对于相关国际规则的核心关切，并将其与中国立场有机结合，在国际规则建构过程中给予充分表达，为发展中国家参与人工智能全球治理体系建设提供有效路径。

第三，以推动全球南方的经济社会发展作为优先事项。从根本上说，想要弥平数字鸿沟，让世界各国共享人工智能技术所带来的发展红利，就要全力推进全球南方在人工智能领域的深度合作，促使其深度参与人工智能产业发展。因此，中国应以推动全球南方国家研发和应用人工智能技术为优先选项，并积极开展与重要国家和区域的技术、产业合作。尤其要重视在东南亚、中东等潜力市场，积极尝试不同的合作方案，从南方国家视角积累人工智能发展与治理方面的有益经验，积极推动南方国家参与全球人工智能产业链和价值链，在实践中增强其综合影响力和在国际规则建构过程中的话语权。此外，中国在利用人工智能促进经济社会发展方面也有着丰富的实践和案例，如智慧城市、智慧医疗、智慧教育和智慧农业等。中国应积极与全球南方国家共享在人工智能领域的知识、经验和资源，以此助力经济社会发展和治理体系建设。

第四，以建构人工智能稳定机制为契机推动大国合作。当前关于人工智能国际规则制定的形势复杂，部分国家希望将他们所主导的小圈子的共识，以国际规则的形式推广到全世界，从而获得规则制定主导权。尽管存在着激烈的竞争与博弈，但规则制定本身对于人工智能技术发展而言仍意义重大。2023 年 2 月，由荷兰与韩国联办的“在军事领域负责任使用人工智能”峰会在海牙举行，包括美国、中国在内的 60 多个国家签署联合声明，倡议负责任开发使用军事人工智能。作为发展中国家的代表，中国在坚持维护发展中国家利益的同时，亦需积极与美、欧、日等科技前沿国家开展相关对话与合作，以寻求共识和防止误判，并且据此建立多层次的对话机制以不断更新治理理念。例如，中国可以推动建立多个面向的人工智能安全稳定对话机制，在区域层面通过欧盟、亚太经合组织等建立一个协调性、互补性、协作性的人工智能合作网络，以推动人工智能的技术交流和产业协同；在双边层面，可与美国、英国等科技强国建立人工智能安全稳定对话机制，以促

进双方在人工智能研发和应用方面的交流合作。

第五，优先发展以私营部门的知识、经验和资源为基础的人工智能应用解决方案。作为全球人工智能发展的重要力量，中国国内与人工智能技术和产业发展相关的各类主体数量庞大，包括众多技术与产业主管部门、行业协会、互联网企业、大学及科研机构、社会科学相关研究部门等主体都为中国的人工智能发展做出了不同程度的贡献。大型科技企业在推动人工智能技术发展方面发挥着其他主体难以替代的关键作用，他们是智能技术发展诸要素的主要所有者，他们也是前沿智能技术发展的主要推动者，当然也是人工智能产业的主要受益者。所以，大型互联网企业对于规则体系的建构具有更高的积极性，他们对于人工智能的潜在风险也最为敏感和熟悉。故此，在人工智能国际规则制定过程中，应充分重视中国人工智能企业的意见，关注其核心立场和诉求，积极鼓励其参与跨国企业层面的国际对话，成为国际规则制定的主力。

关于上海国际问题研究院

上海国际问题研究院成立于 1960 年，是隶属于上海市人民政府的高级研究机构和知名智库。我院的主要任务是：以服务党和政府决策为宗旨，以政策咨询为方向，通过对当代国际政治、经济、外交、安全的全方位研究，为党和政府决策提供有力的智力支持；通过与国内外研究机构和专家学者的合作交流，增强我国的国际影响力和国际话语权，提升国家的软实力。多年来，我院一直被国内外权威机构评为中国最重要的国际问题和中国外交智库之一。

上海国际问题研究院下设六个研究所和六个研究中心，分别是：全球治理研究所、外交政策研究所、世界经济研究所、国际战略研究所、比较政治和公共政策研究所、台港澳研究所；美洲研究中心、亚太研究中心、俄罗斯中亚研究中心、西亚非洲研究中心、欧洲研究中心、海洋与极地研究中心。此外，我院还是上海国际战略研究会和上海国际关系学会的机构会员。

上海国际问题研究院编辑出版的中文刊物《国际展望》双月刊，《上研院报告》（中英双版）和英文刊物《China Quarterly of International Strategic Studies》季刊已经成为国际问题研究领域的重要学术论坛。

关于清华大学战略与安全研究中心

清华大学战略与安全研究中心是清华大学校级研究机构，成立于 2018 年 11 月 7 日，旨在打造国际战略和安全领域的国际化和专业化高端智库。

中心有两大目标：一是就国际秩序、国际关系以及战略与安 全等问题开展研究，跟踪形势变化并做出判断，为决策提供参考意见和建议；二是通过开展多种形式的国际交流与合作，宣介、阐释和传播中国的理念和政策主张，增进国际社会对中国的了解，提升清华大学在战略与国关界的国际影响力。

研究中心以战略与安全、外交与国际关系等问题为主要研究领域，以全球秩序、国际安全治理、人工智能与国际安全等重大战略与安全问题作为主要研究方向。研究中心实行管委会领导下的主任负责制。外交部前副部长傅莹大使为中心的创始主任。中心设立学术委员会作为学术指导和顾问机构，下设美欧研究项目、全球治理研究项目、欧亚研究项目、人工智能与国际安全研究项目、“战略青年”研究交流项目以及中国论坛秘书处。

关于北京大学互联网发展研究中心

北京大学互联网发展研究中心是北京大学为适应时代发展，服务国家网络强国战略，推动互联网领域的知识创新和文化交流而成立的研究机构。中心致力于以互联网发展为中心，面向问题、面向社会、面向未来的研究，依托北京大学的学术优势和社会影响力，联合国内外一流研究机构和专家，开展学术性、政策性、实践性相结合的跨学科融合研究，同时也为政府机构、企事业单位、社会组织提供相关调研、咨询和培训等服务，旨在推动研究创新，从而促进互联网更好造福社会，增进人类福祉。

研究领域包括：互联网产业发展、互联网治理、互联网伦理、互联网与社会、互联网与传播、未成年人数字保护与发展等。

主编 / 出品人

陈东晓 上海国际问题研究院院长、国际传播中心主任

执行主编

李　忻 上海国际问题研究院国际传播中心执行主任

编　　辑

陈　雪　上海国际问题研究院国际传播中心助理研究员

联系方式:

地　　址：上海市徐汇区田林路 195 弄 15 号

邮政编码：200233

联系电话：+86-21 54614900

传　　真：+86-21 64850100

http://www.siis.org.cn

上海国際问题研究院
SHANGHAI INSTITUTES FOR INTERNATIONAL STUDIES

Shanghai Institutes for International Studies
195-15 Tianlin Road, Xuhui District
Shanghai 200233, PRC
Tel/Fax: +86 21 64850100
www.siis.org.cn