全球人工智能治理的目标、挑战与中国方案

鲁传颖

【内容提要】全球人工智能治理是指国家、市场和技术社群行为体为实现人工智能在全球的安全发展与和平利用而共同制定实施一系列原则、规范和制度的过程。随着生成式人工智能取得技术突破，全球人工智能治理进程加速发展，各个行为体和平台围绕人工智能的伦理、规范和安全建立了诸多治理机制。但是，地缘政治因素和复杂的政治、市场、技术逻辑之间的分歧制约了全球人工智能治理进程。作为人工智能技术大国，中国及时提出《全球人工智能治理倡议》，为全球人工智能治理提供了切实可行的方案。

【关键词】人工智能治理　全球秩序　机制复合体　中国方案

2023 年 7 月，联合国秘书长古特雷斯呼吁成立实体机构负责人工智能安全治理，并召集成立“人工智能高级别咨询委员会”进行探讨，拉开了全球人工智能治理的序幕。2023 年 2 月和 11 月，荷兰与英国分别举办了全球性的人工智能峰会，发布了《军事领域负责任使用人工智能行动倡议》和《布莱切利宣言》。2023 年 10 月，中国国家主席习近平在第三届“一带一路”国际合作高峰论坛开幕式上的主旨演讲中提出《全球人工智能治理倡议》，向国际社会展示中国在人工智能治理方面的政策主张。此外，联合国互联网治理论坛、世界经济论坛等国际组织也陆续提出有关全球人工智能治理的倡议。据不完全统计，目前全球关于人工智能的倡议已有 50 多个，这标志着全球人工智能治理进入了全面发展的新阶段。

开启全球人工智能治理的新篇章

全球人工智能治理是指国家、市场和技术社群等行为体为实现人工智能在全球的安全发展与和平利用而共同制定和落实一系列原则、规范和制度的过程。人工智能作为一项战略性技术，具有改变全球格局和人类发展进程的潜力。其中，人工智能技术发展与应用所涉及的广泛行为体和行业领域远远超出了常规技术，是最复杂也是影响力最大的战略技术。正因为人工智能涉及议题领域广泛、行为体众多，如果缺乏规则和秩序，各方在伦理、规范和安全领域的很多分歧都难以解决，必将导致更多的冲突。可以说，较之于核武器与网络安全，人工智能治理对于国际体系的安全与发展具有更重要的战略性和全局性意义。当前，全球人工智能治理进程正在不断加速，并且形成了复合型的机制建构路径。

首先，安全发展与和平利用已经成为全球人工智能治理的新秩序。开展全球人工智能治理，需要在已有的治理实践之上进一步明确人工智能治理的秩序和制度体系。目前，人工智能可以参考的治理秩序主要有二：一个是在核武器领域成功构建的以“战略稳定”为基础的秩序，另一个是在网络安全领域以“负责任国家行为准则”为目标但却未能实现的秩序构建。“战略稳定”为核大国提供了制定核政策、应对危机及军备竞赛的秩序框架，而“负责任国家行为准则”则由于未能有效减少各国在网络安全上面临的挑战和冲突而举步维艰。导致这一结果的主要原因在于各方对网络安全秩序的目标和制度体系设计没有达成共识。人工智能的战略性、全局性较核武器和网络安全有过之无不及，因此对建立共同遵守的规范和秩序的需求也更高。

人工智能全球秩序的内涵就是安全发展与和平利用。就技术本身而言，构成人工智能的模型、算法、数

（AA/ABACA/澎湃影像/IC photo图片）

2023年7月18日，美国纽约，联合国安理会举行主题为“人工智能给国际和平与安全带来的机遇与风险”高级别公开会。这是安理会首次就人工智能问题举行会议。

据具有较大的安全隐患，存在着“黑箱操作”和“失控”风险。人工智能的安全风险一部分源自算法自身的“黑箱”，开发者缺乏对人工智能决策机制的理解；另一部分是由于人类对于人工智能能力的理解不足，存在滥用风险。如果不加预防，一旦出错便可能产生灾难性影响。因此，如何确保人工智能的安全使用至关重要，任何人工智能技术的发展也都应当以安全为前提。国际社会应当对此形成共识，共同推动人工智能安全发展。

从技术应用角度看，人工智能在军事与情报领域具有巨大的发展潜力，但也带来极大的潜在风险。倘若人工智能在战场上被滥用，将有可能造成无法挽回的损失。例如，利用人工智能武器对核武器装置发起攻击很可能会打破“战略稳定”，而核武器装置采用的人工智能控制系统本身也可能存在不可控的风险。因此，禁止人工智能武器化的呼声在国际社会得到广泛支持。为此，应当明确在冲突中应用人工智能的目的是为了和平而非战争。此外，人工智能作为一项战略性技术已经成为世界主要军事大国竞相争夺的新战略制高点。一些国家在人工智能军备方面投入大量资金，用于情报、监视和侦察 (ISR)，保障网络安全，指挥和控制各种半自动和自动驾驶车辆以及提高日常工作效率，包括后勤、招聘、绩效、维护等。将和平利用人工智能作为前提可有效地避免其过度军事化，管控人工智能驱动的军备竞赛。

其次，机制复合体建构成为全球人工智能治理的实现路径。全球人工智能的治理机制对于管控分歧、寻求共识、实现治理目标至关重要。不同于原则性的秩序，基于解决某一类问题而专门设计的治理机制更加具体也更具约束性。人工智能作为通用型技术，其涉及的治理领域极为广泛，需要有一系列松散耦合的机制来共同形

成全球治理机制复合体。每一个治理机制都有明确的议题、参与行为体、互动模式。全球关于人工智能治理的倡议、规范和机制已经超过 50 个，其治理实践涉及普适性的道德和伦理、国家行为规则、技术标准、行业应用实践等多个层面，覆盖国际组织、各国政府、私营部门、技术社群和公民团体等多个层次的参与主体。

从议题角度来看，当前人工智能治理主要关注伦理、规范和安全三个领域。人工智能发展与治理始终是相伴相随的，人工智能治理是对技术和应用所带来问题和挑战的回应，随着技术进步和应用突破，治理的议题在不断拓展，能力也应不断提升。早期人工智能治理关注伦理问题，主要是机器与人之间的关系，提出了保持人对机器的控制这一理念。随着无人机等军事人工智能应用的突破，规范成为治理的焦点。负责任地使用军事人工智能不仅涉及如何将已有的国际规范落实到军事人工智能领域，也包括如何构建新的规范来进一步约束不负责任地使用军事人工智能的行为。随着大语言模型的突破，针对模型、算法和数据的安全治理成为新的重大议题。

这些不同治理机制之间虽然目标不同，但鉴于人工智能的战略性和全局性，“各自为政”的治理模式有可能带来更多冲突，不利于解决问题。例如，由于人工智能的军民两用性，在安全治理中被视为合理要求的信息共享，有可能成为军事人工智能领域的国家安全风险。统筹不同机制之间的关系有助于行为体在参与人工智能全球治理的进程中，作出相对理性、全面的决策。此外，人工智能各个治理机制之间的关联程度要比其他治理领域更高。由于人工智能技术在不同议题上都扮演了重要角色，因此，其作为一根主线贯穿在全球人工智能治理机制复合体之中，具体体现在不同议题之间的关联度以及行为体之间的互动模式上。这使得人工智能的治理理念、模式需要作出相应的调整。例如，更加突出多利益攸关方的治理作用，要解决技术上的问题就离不开企业和技术社群的参与。

全球人工智能治理面临的挑战

全球人工智能治理作为一项新兴议题面临双重挑战，即人工智能安全发展与和平利用的秩序面临地缘政治的干扰，机制复合体的治理路径面临政治、市场、技术三种治理逻辑之间的冲突。这些挑战折射出参与治理的主体对人工智能安全、发展和利用过程中不同的价值判断、路径偏好以及利益设定。

一方面，地缘政治因素制约人工智能全球秩序的构建。全球治理与地缘政治之间是一对矛盾。全球治理寻求对议题的共同理解和治理过程的共同参与，地缘政治的目标是民族国家权力的扩张，背后受到地理环境、意识形态以及军事、经济和科技实力等因素的影响。[1] 全球治理的目标是建立共同遵守的秩序，地缘政治的目标则是获得国家战略收益。[2] 因此，全球人工智能治理能否取得成功，一定程度上取决于能否克服地缘政治博弈所带来的挑战。

地缘政治冲突首先体现在由谁来主导规则制定平台。联合国作为全球人工智能治理主渠道遇到了美国主导的所谓“理念一致国家同盟”的挑战。美国和一些西方国家从地缘政治的角度出发，认为随着以中国为代表的新兴市场国家在联合国中的影响力大增，联合国已经不能反映西方的意愿和关切。因此，他们更加倾向于将传统网络治理中大力推行的所谓“理念一致国家联盟”运用到人工智能治理上，坚持以意识形态划线，塑造战略竞争对手，打造人工智能治理“小圈子”，并加强在军事人工智能领域的互操作性。[3] 例如，美国推动下的全球人工智能合作伙伴（GPAI）、七国集团（G7）“广岛人工智能进程”等机制将西式“人权”“自由”“民主”等价值理念作为“负责任的人工智能”的参考标准，美英澳三边安全伙伴关系（AUKUS）更是在加快推动人工智能军事运用的互操作性。这类排他性和安全性的联盟正在加剧全球人工智能治理安全化和军事化。

以中国为代表的新兴市场国家和发展中国家认为，维护联合国的合法性和权威性是领导全球人工智能治理的先决条件。从联合国的角度来看，如果不能在人工智能这样的关键领域发挥重要作用，未来联合国在国际上的影响力和权威性会进一步丧失。2018 年以来，联合国相继成立数字合作高级别小组、发布“数字合作路线图”、创办人工智能高级别咨询委员会，倡导开发“可

信赖、基于人权、安全和可持续并促进和平”的人工智能。联合国下属机构也将人工智能视为实现可持续发展目标的工具，发布了各项人工智能治理倡议。联合国秘书长古特雷斯于 2023 年 7 月提出的“新和平纲领”政策报告，明确阐述了联合国引领的“基于多边”“弥合鸿沟”“平等参与”的人工智能治理原则。

此外，人工智能领域的实力地位影响了国家对待人工智能的治理态度。在规则的宽严度上，以美国为首的部分西方国家倾向于制定非约束性的规范治理，实力较弱的国家希望制定具有法律约束力的规则。这两者的差异在于，强国更加希望利用非约束性的规范空间来实现多元的战略目标。弱国目标相对单一，希望通过规则来约束各方的行为从而获取国际安全保障。特别是在军事人工智能等关键议题上，强国和弱国的诉求差异更加明显。强国希望在保持人工智能发展的同时，也获得安全方面的战略优势；弱国缺乏军事实力，更希望禁止人工智能军事化应用。

另一方面，全球人工智能治理机制复合体实现路径始终面临复杂的政治、市场、技术逻辑之间的分歧，并且已经贯穿在伦理、安全和技术治理的每一个层面。政治介入的合法性在于人工智能的安全失灵隐患需要政府来加强监管。由于人工智能是一项资金密集型高端技术，市场和技术在推动人工智能产业发展的过程中发挥着极为重要的作用。理想中的全球人工智能治理机制复合体需要国家与市场、技术等方面参与治理行为体明晰彼此的界限，并从各自的角度为治理提供解决方案。实际上，国家和市场、技术之间的界限模糊，并且存在着复杂的利益关系，使得全球人工智能治理体系难以形成统一的机制复合体。[4]

国家行为体与非国家行为体在治理中的边界模糊，对治理的关注点也不尽相同。政府部门普遍关注人工智能相关的高政治领域，市场和技术行为体则更加关注技术、标准开发过程中的治理。政治视角与市场和技术视角之间不仅关注的议题领域不同，也暗含着复杂的博弈关系。政府更加关注军事、安全风险，将自己凌驾于市场和技术之上。例如，美国发布《关于负责任地军事使用人工智能和自主技术的政治宣言》、英国召开首届人工智能安全峰会，都体现了政府对人工智能引发军事安全风险的担忧。不仅如此，国家行为体还强调自身的监管作用，通过立法和制定战略来赋予自身领导权。以欧盟为例，其发布《人工智能政策》《人工智能白皮书：通往卓越与信任的欧洲之路》《人工智能公约》等政策文件，以确保人工智能的设计、部署和使用都符合欧洲委员会关于人权、民主和法治的标准。2024 年 3 月通过的欧盟《人工智能法案》成为世界上首部全面监管人工智能的法律，致力于构建具有强约束力的人工智能监管框架，引领全球人工智能治理进程。

在市场和技术视角看来，人工智能技术复杂，迭代速度极快，政府在一定程度上离技术较远，因而其干预并不符合技术发展的逻辑，在很大程度上会阻碍技术发展。企业和市场在人工智能全球治理中应当扮演更加重要的角色。目前，微软、OpenAI、Anthropic、阿里巴巴、腾讯、百度等人工智能领域领先企业纷纷发布了全球人工智能治理相关倡议。这些倡议认为，技术开发环节的治理是确保人工智能安全的关键，强调算法和模型的透明度和负责任的研发、人工智能在价值上与人类对

2023年10月30日，美国华盛顿，美国总统拜登签署一项关于人工智能的行政命令。

（视觉中国影像/IC photo图片）

2023年7月8日，上海，2023世界人工智能大会举行。图为华为展台展示的AI大模型。

（签约摄影师/IC photo图片）

齐等。不仅如此，市场和技术代表还组成联盟共同参与全球治理进程。如由苹果、亚马逊、DeepMind、谷歌、脸书等科技公司成立的人工智能伙伴关系（PAI），积极支持与人工智能伦理和治理相关的紧迫问题研究，促进教育项目和实用工具的开发，建立人工智能事件数据库等。市场和技术在全球治理层面的积极投入是为了在技术发展过程中追求更为宽松的监管环境，通过自愿承诺或标准制定等进行自我监督。市场和技术的主动作为是企业责任的一部分，但客观上也分散了政府的权力。总体而言，随着技术与市场在人工智能全球治理中的作用与影响不断增加，政治、市场与技术三者之间的博弈也会愈加激烈。

全球人工智能治理的中国方案

2023 年 10 月，随着《全球人工智能治理倡议》的提出，中国向世界系统阐述了自身关于全球人工智能治理的立场、主张和建议。这份来自中国的治理方案是对当前全球人工智能治理中的问题、难点的系统性回应，展现了中国在推动全球人工智能发展和治理合作方面的积极态度与务实行动。

首先，中国的治理方案顺应了当前人工智能技术发展的趋势，及时回应了治理需求。从议程内容来看，当今的主流人工智能治理进程主要涉及安全发展与和平利用两大主题。中国的《全球人工智能治理倡议》同样把握住了这两大主题，提出“坚持安全和发展并重的原则，促进人工智能技术造福于人类”，并进一步从技术伦理、科技向善、风险评估、隐私和数据安全、算法治理、发展赋能等具体议题角度对这两个主题进行充分诠释，重申全球人工智能治理中的重点问题。中国的主张回应了联合国《特定常规武器公约》框架下的“致命性自主武器系统”（LAWS）人工智能军事化治理进程。

（zdmphoto/IC photo图片）

2023年11月9日，浙江乌镇，世界互联网大会乌镇峰会发布大会成果之一——《发展负责任的生成式人工智能》研究报告及共识文件。

中国方案在充分参考全球各类倡议和治理进程的基础上，发出了人工智能领域加强治理合作的呼吁。具体而言，中国的治理方案从核心架构上看，需要各国秉持共同、综合、合作、可持续的安全观，坚持发展和安全并重的原则，通过对话与合作凝聚共识，构建开放、公正、有效的治理机制；从治理主体上看，需要各国政府、国际组织、企业、科研院校、民间机构和公民个人等各类主体秉持共商共建共享的理念，协同推进；从治理手段上看，倡导建立人工智能伦理准则、建立风险等级评估体系、建立健全法律规章、开发风险防范技术以及对发展中国家的援助合作机制，从而推动人工智能治理倡议的落实；从治理目标上看，希望人工智能技术进一步造福人类，推动构建人类命运共同体。

从全球趋势上看，中国的治理方案与其他人工智能双多边治理进程共同体现了当前人工智能治理“喜忧参半”的特征。一方面，各类进程在人工智能治理的大原则和大方案上形成了一定共识，基本都将“科技向善”“风险防控”“和平利用”等大原则纳入议程设计的核心，在原则实施的具体方案上认可必须加强合作、凝聚共识，并充分发挥多利益攸关方的作用等。这些宏观性的原则和方案确立了全球人工智能治理的准绳和方向。但另一方面，各类机制在原则细化和方案落实方面存在困难。从联合国教科文组织（UNESCO）发布的《人工智能伦理问题建议书》到英国人工智能安全峰会提出的《布莱切利宣言》，多数进程仍然难以在具体的议程领域取得突破性进展，遑论设立明确的、具有约束力的标准框架或行动指南。这既是人工智能技术不确定性带来的客观挑战，也是由各国不同的价值伦理和战略需求造成的主观困难，主客观因素叠加阻碍了全球人工智能治理进程的有序推进。

其次，中国的治理方案进一步完善了全球人工智能治理原则和理念。中国的治理方案不仅遵循全球人工智能治理中较为普遍的原则和实践策略，还对当前全球人工智能治理中出现的一些片面、狭隘的提法作出了一定的修正和完善，推动全球人工智能治理进程朝着更加安全、公平的方向发展。

一是完善人工智能伦理中的人权观。中国在《全球人工智能治理倡议》中强调，人工智能发展必须坚持“以人为本”理念。这一原则虽然在各种治理进程中也很常见，但是中国对其作出了更加丰富的诠释，即人工智能

技术的发展不仅需要尊重和维护个人权利，更应该以增进人类共同福祉为目标，以保障社会安全、尊重人类权益为前提，使人工智能始终朝着有利于人类文明进步的方向发展。从这个意义上来讲，中国的方案不仅有别于西方狭隘的人权思想，为人工智能技术的道德规范和发展方向提供了框架，更提出了深刻的伦理立意，倡导构建一个平衡各方利益、共同应对人类挑战的全球人工智能治理新模式。

二是强调人工智能发展中的平等观。面对发展中国家在人工智能全球治理中普遍“失声”的问题，中国作为发展中大国，通过充分发挥自身影响力，将发展中国家普遍关注的问题带入更加广泛的国际讨论中，提出各国人工智能发展与治理的“权利平等”“机会平等”“规则平等”原则，并倡导需要开展面向发展中国家的国际合作与援助，不断弥合智能鸿沟和治理能力差距。这将从能力、意愿与平台开放度等方面破解发展中国家在全球人工智能治理中的边缘化难题，[5] 通过进一步兼顾南北国家立场和声音构建公平合理的国际规则。

三是更新人工智能治理合作观。美西方国家虽然积极开展人工智能国际合作，但不论是军事防务领域的人工智能防务伙伴关系，还是产业与发展领域的美欧贸易和技术委员会（TTC），都有着极强的排他性和竞争性，从长远来看并不利于推动全球性的人工智能治理进程。中国方案明确提出人工智能发展需要跨越“小圈子”，反对特定国家将全球治理机制工具化、私物化。[6] 中国提出各国无论大小、强弱，无论社会制度如何，都有平等发展和利用人工智能的权利；强调要建立更加广泛的人工智能合作机制，共同推动人工智能治理进程。

最后，中国的治理方案为统筹当前全球范围内松散的治理机制提供了一个综合性框架。当前，全球人工智能治理进程虽然在加速推进，但是在各个功能领域，治理机制的耦合程度仍然较弱，导致在执行力和协调性方面存在明显不足。这一方面是由于各类机制的代表性不足，导致其治理标准不能为更广泛的国家所接受，另一方面则是由于议题的涵盖度不广，导致对于人工智能领域综合性较强的军事化问题、系统安全问题与发展合作问题的回应较弱。

中国的《全球人工智能治理倡议》则对以上问题进行了针对性补充。中国方案将不同意识形态、不同发展阶段的国家纳入同一个治理框架中，赋予各国平等参与和讨论的机会，强调多边主义和共同利益。通过这一方式，中国方案试图解决现有治理机制代表性不足的问题，促进全球人工智能治理标准的普遍接受和应用。同时，中国方案也扩大了治理议题的范围，将人工智能安全、发展、和平、合作等主题下多个治理议题纳入讨论范围，旨在构建一个更加全面和综合的全球人工智能治理体系，为全球人工智能治理的有效实施和未来发展奠定坚实基础。

因此，中国的人工智能治理方案具有广阔的实践前景。中国可携手国际社会利用联合国未来峰会等机遇，推动《全球人工智能治理倡议》中的关键主张在联合国层面获得认可。中国还可以积极加强同其他治理机制的多层级合作，通过原则协商、资源共享等方式，进一步加强国际技术交流和政策对接。此外，中国还可以围绕《全球人工智能治理倡议》搭建更多具体而有实践性的治理机制，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，进而为解决当前全球人工智能治理中的代表性和广泛性难题提供良策。

作者系上海国际问题研究院公共政策与创新研究所副所长、网络空间国际治理研究中心秘书长，研究员

---

[1] 蔡翠红：《全球科技发展：迭代加速与地缘政治化》，载《人民论坛》2023 年第 24 期，第 24-29 页。

[2] 鲁传颖：《全球数字地缘政治的战略态势及其影响》，载《当代世界》2023 年第 5 期，第 37-43 页。

[3] 郎平：《网络空间安全治理的全球性困境与中国对策》，载《国家治理》2022 年第 22 期，第 31-35 页。

[4] Lewin Schmitt, “Mapping Global AI Governance: A Nascent Regime in A Fragmented Landscape,” *AI and Ethics*, Vol.2, No.2, 2022, pp.303-314.

[5] 门洪华：《中国三大全球倡议的战略逻辑》，载《现代国际关系》2023 年第 7 期，第 5-21 页。

[6] 鲁传颖：《〈全球人工智能治理倡议〉中的中国智慧》，中国社会科学网，2024 年 1 月 23 日，https://www.cssn.cn/gjgc/lbt/202401/t20240124_5730523.shtml。