构建可信人工智能评价新体系，促进全球可信应用

文｜国家计算机网络应急技术处理协调中心 徐小磊 张震 贺敏

【摘要】可信人工智能已成为全球人工智能治理的核心议题之一。作为全人类共同财富，当前的人工智能可信性评价仍难以满足全球普惠发展的实际需求。世界多数国家尚不具备自主开发基础大模型的能力，且人工智能应用高度依赖他国，这使对人工智能可信性评价和验证的需求日益显著。尽管联合国、欧盟、美国等已相继提出多种可信评价方法，但普遍存在未充分考虑国家安全维度的问题。本文将国家安全要素纳入可信人工智能评价范畴，制定涵盖技术主权、伦理安全、社会应用等多个层面的可信人工智能准则，推动形成兼顾技术主权与应用安全的全球治理方案，促进人工智能技术可信应用，从而让人工智能技术惠及广大发展中国家。

【关键词】人工智能；全球普惠；技术主权；可信准则

作为新一轮科技革命和产业变革的核心驱动力，人工智能已成为各国战略布局的重点。然而，目前多数国家尚不具备自主研发基础大模型以支撑本国人工智能服务的能力。因此，关于可信性评估与验证方面的需求显得尤为突出。近年来，多个国家与国际组织相继出台人工智能治理文件，提出了多种可信人工智能的评估方法，涵盖了公平性、可靠性、隐私保护等方面，却普遍忽视了国家安全因素的考量。本文将国家安全纳入可信人工智能的评价体系，提出涵盖技术主权、伦理安全、社会应用等多个层面的可信人工智能准则，推动人工智能技术的可信应用。

一、人工智能应用全球化引发担忧

世界各国积极拥抱人工智能技术，充分发挥其在推动经济增长、提升社会福祉和促进产业升级方面的潜力。然而，由于技术和资源等方面的诸多限制，大多数国家尚不具备自主开发基础大模型支撑人工智能服务的能力。训练基础大模型不仅需要海量数据、强大算力和研发创新能力，还伴随巨额成本投入，这使人工智能技术高度集中于少数国家和科技巨头。由此，全球多数国家在人工智能技术方面主要依赖外部，人工智能应用引发了对国家安全的担忧。在这一背景下，将国家安全因素纳入人工智能可信性评价，势在必行。

总体来看，多数国家的担忧主要体现在以下四个方面。

一是对运行控制权的担忧。人工智能系统往往由技术领先国家或大型跨国企业开发，系统辐射全球、跨境服务。部署所在国与服务所在国不一致，使服务所在国难以掌握系统运行的控制权，进而削弱国家自主性和应对风险的能力。

二是对服务可靠性的担忧。由于人工智能系统的核心技术、基础设施和数据存储可能受制于他国，人工智能应用国对服务的稳定性、连续性以及能否正确执行指令等存在顾虑。尤其在全球地缘政治紧张、技术出口管制的背景下，有些国家担心人工智能系统服务可能面临中断、操控甚至误导性输出的风险。

三是对系统价值观的担忧。人工智能模型的训练数据和算法设计往往受到开发国文化、社会规范和价值观的影响。这些因素可能与人工智能所服务的国家存在显著差异。这种差异可能导致人工智能在内容生成和决策建议方面偏离输入国的主流价值观，对当地文化、社会认同和意识形态产生冲击和影响。

四是对数据跨境流动的担忧。在人工智能系统跨境服务过程中，大量数据跨境流动，特别是本国用户在使用过程中生成的数据，常常会被跨境传输至人工智能部署所在国进行存储和处理。这类跨境数据可能涉及个人隐私、商业机密、关键基础设施信息等，从而引发数据安全的担忧。

二、人工智能国际治理的分歧和共识

在人工智能技术快速发展的背景下，全球治理格局正面临地缘政治博弈与技术竞争的双重挑战。当前，中美欧三大行为体的利益诉求和政策分歧尤为突出，而发展中国家则在技术可及性与公平性方面提出诉求。

中国强调技术主权与全球普惠。一是鼓励自主创新，推动国产人工智能框架和芯片研发，如飞桨（PaddlePaddle），减少对外依赖。二是加强国际合作，发布《全球人工智能治理倡议》，并在联合国推动人工智能能力建设的国际合作决议。美国以维护技术领先地位为核心目标，通过国内立法与国际联盟塑造治理规则。例如，通过出口管制（如先进芯片出口）限制他国技术发展，巩固半导体产业链的主导地位。美欧英签署的《人工智能、人权、民主和法治框约》并未包括中国、新加坡、俄罗斯及中东等国。欧盟则主要通过立法监管引领全球治理进程，试图借助《人工智能法案》等法规确立全球伦理与监管的标杆。同时，广大发展中国家更关注人工智能技术的普惠性与应用过程中的安全性，期待通过国际合作缩小数字鸿沟、获取资源与能力支持。

正是在这一全球多元化诉求与治理分歧交织的背景下，构建可信人工智能逐步成为各方共识。国际组织与各国政府纷纷发布相关政策文件，普遍强调人工智能发展的可信性要求，并在公平性、可靠性与隐私保护等方面达成相关准则的共识。多项政策文件均体现出对这些准则的高度关注（见下表）。

公平性。人工智能服务应考虑全球文化、种族、性别和能力的差异，确保所有用户都能被平等对待。联合国《人工智能伦理问题建议书》、欧盟《可信人工智能伦理指南》、美国《关于安全、可靠和可信的人工智能行政命令》、英国《促进创新的人工智能监管方法》，日本《人工智能运营商指南（草案）》、新加坡《生成式人工智能治理的模型人工智能治理框架草案》等，均对公平性提出了明确要求和指导措施。

可靠性。人工智能系统应进行全生命周期的供应链安全管理，涵盖开发、训练、部署和维护等环节，确保各个环节的供应链具备稳定性、持续性和安全性。联合国《抓住安全、可靠和可信的人工智能系统带来的机遇，促进可持续发展》、欧盟《人工智能法案》、美国《人工智能风险管理框架》、英国《促进创新的人工智能监管方法》等，均论述了人工智能的可靠性。

隐私保护。人工智能服务过程应注重保护个人信息，确保个人数据的安全性和保密性。联合国《人工智能伦理问题建议书》、美国《人工智能风险管理框架》、日本《人工智能运营商指南（草案）》、新加坡《生成式人工智能治理的模型人工智能治理框架草案》等，均对隐私保护提出了具体规范。

三、对可信人工智能评价范围和准则的建议

落实《全球人工智能治理倡议》，遵循“以人为本、智能向善”的发展方向。人工智能作为人类的共同财富，不能成为“富国和富人的游戏”，而应让世界共享人工智能发展红利。从理论视角出发，可信人工智能的发展路径应体现系统性设计思维：依据技术治理理论，人工智能不仅是治理对象，更应成为推动公共治理现代化的技术工具，实现“技术治理技术”。根据伦理决策模型，人工智能治理应考虑公平性、透明性和责任性，即系统在设计和运行过程中应尊重人类尊严，避免偏见和歧视，确保公众能够理解并追责其行为后果。根据社会技术系统理论，人工智能技术的引入和应用将深刻重塑社会结构、工作模式和价值体系。人工智能治理必须回应社会利益相关者需求，体现文化多样性与价值观差异性。

结合国际治理实践、各国的可信属性实际需要和相关理论，可信人工智能评价准则的设计需基于技术治理、伦理规范和社会需求的系统性整合。该准则设计需确保不同评价维度之间既保持独立性，又具有互补性与协同性，最终形成一个逻辑闭环、结构合理、可操作性强的可信评价体系。基于此，本文提出可信人工智能评价范围和评价准则。

表 人工智能可信相关文件（部分）

国家地区	文件名称	发布机构	发布时间	关键内容
国际组织	《抓住安全、可靠和可信的人工智能系统带来的机遇，促进可持续发展》	联合国大会	2024 年 3 月	以人为本、可靠、可解释、符合道德、具有包容性，充分尊重、促进和保护人权与国际法，保护隐私、面向可持续发展和负责任
	《人工智能伦理问题建议书》	联合国教科文组织	2021 年 11 月	相称性和不损害、安全和安保、公平和非歧视、可持续性、隐私权和数据保护、人类的监督和决定、透明度和可解释性、责任和问责、认识和素养、多利益攸关方与适应性治理和协作
	《人工智能原则》	经济合作与发展组织	2024 年 5 月	包容性增长、可持续发展和福祉、以人为本的价值观与公平性、透明度和可解释性、稳健性、可靠性和安全性、可追责性
欧盟	《人工智能法案》	欧盟委员会	2024 年 5 月	以人为本、技术稳健性与安全性、透明度、多样性、非歧视与公平性、社会与环境福祉、责任性、合规性、风险管理、后市场监测
	《人工智能、机器人和相关技术的伦理框架》	欧盟委员会	2020 年 10 月	以人为本，安全、透明、可问责，无偏见、无歧视，社会职责、性别平等，可持续发展，尊重个人隐私和补救权益
	《可信人工智能伦理指南》	欧盟人工智能高级专家组	2019 年 4 月	三个基础条件：合法合规、伦理、鲁棒性。七个关键要素：人的能动性和监督、技术鲁棒性与安全性、隐私和数据管理、透明性、多样性非歧视性与公平性、社会与环境福祉、问责
美国	《关于安全、可靠和可信的人工智能行政命令》	美国总统拜登	2023 年 10 月	保护美国人的隐私，推进平等与公民权利，维护消费者、病患和学生的权益；支持工人，促进创新和竞争，在国外推进美国的领导地位，确保政府负责任和有效使用人工智能
	《人工智能风险管理框架》	美国国家标准与技术研究院	2023 年 1 月	有效且可靠、安全、稳定且弹性、负责且透明、可理解与可解释、隐私增强及有害偏见控制下的公平
	《国家人工智能研究与发展战略计划》	美国国防创新委员会	2019 年 10 月	负责任、公平、可追溯、可靠性和可治理
英国	《促进创新的人工智能监管方法》	英国科技办公室	2023 年 3 月	安全性、可靠性与稳健性、透明度与可解释性、公平性、问责制与治理、竞争与补救措施
	《人工智能监管政策》	英国政府	2022 年 7 月	具体情形具体监管原则、遵循协调原则、鼓励创新原则、比例性原则
日本	《人工智能运营商指南（草案）》	日本人工智能战略委员会	2024 年 1 月	以人为本、安全性（侧重于个人生命、身体、财产、精神等）、公平性、隐私保护、安全保障（侧重于网络安全和数据安全）、透明度、问责制、（提升）教育与素养、确保公平竞争、创新
新加坡	《生成式人工智能治理的模型人工智能治理框架草案》	新加坡人工智能验证基金会和信息通信媒体发展局	2024 年 1 月	问责制、安全性、可解释性、透明度、公平性、人性化、隐私性、社会责任、测试与保证

（一）可信人工智能的评价范围

可信人工智能评价包括以下三个层面。

向内。可信人工智能应贯穿整个产品生命周期，涵盖人工智能设计、研发、测试、部署、使用和维护等阶段。

向外。可信人工智能的构建应向外延伸至整个产业链的上下游，包括训练数据、算力设施、产品服务、应用场景等。

平等。可信人工智能应考虑全球各国人工智能技术发展状况，让世界共享人工智能发展红利，确保技术的全球普遍受益和公平。

本文提出，将可信人工智能的评价范围由原有的技术内生可信和供应链可信，进一步拓展到全球范围的应用可信。可信评价将全球各国视为独立个体，考虑各国个体的安全平等，进而促进人工智能的全球普惠。

（二）可信人工智能的评价准则

可信人工智能的评价准则包括自主可控、公平合法、安全可靠和数据隐私四个关键方面。这四项准则构成“技术基础—风险防控—价值约束”的闭环体系，自主可控与安全可靠保障技术底座稳健性，数据隐私与公平合法划定社会应用边界。准则设计的内在逻辑为技术底层到社会价值的递进。自主可控为技术前提，公平合法为社会价值边界。安全可靠与数据隐私则构成风险防控的双支柱，实现保障系统稳定性和用户权益。这四个方面相互交织、相互依赖，既呼应技术治理理论的全周期管理思想，又体现伦理决策模型的价值导向，为可信人工智能提供可落地、可验证的实践路径。

从上表的文件可以看出，当前的可信人工智能准则涵盖了技术可信、伦理安全和社会应用等多个层面，能够符合社会和法律的要求。然而，对大多数尚不具备自主开发基础大模型的国家而言，其应用他国大模型带来的安全担忧未被充分考虑。本文将国家安全纳入可信人工智能准则，针对国家安全担忧的各个方面提出相应的可信人工智能准则要求。

1. 自主可控

自主可控强调人工智能必须始终处于人类的监督与控制之下，确保其服务于人类社会的利益，同时，确保人工智能受到应用国的监管和控制，防止技术滥用对国家安全构成威胁。

可控性。一是始终在人类控制下。人工智能的发展必须始终处于可控状态，保障人类拥有充分自主决策权，确保不会出现人工智能脱离人类控制的风险。二是用户监督控制。人工智能提供服务时，应以用户为中心，所有行为都应取得用户授权。用户应能够干预或监督人工智能做出的每一个决定。

自主性。一是国家自主性。任何应用国应具备在关键时刻对人工智能系统采取控制措施的能力，包括拒绝、终止和中断人工智能系统的服务运行。二是应用国部署管理。人工智能的训练和部署应在应用国境内进行，应用国始终掌握人工智能系统的控制权，并具备持续监督和管理能力，确保人工智能控制权不被滥用。

2. 公平合法

公平合法强调人工智能的设计和应用应遵循国际公认的基本原则，确保其服务和功能尊重各国的平等权利与法律要求，同时，还应符合应用国的文化价值观，避免输出偏离该国社会认知和伦理规范的内容。

公平性。遵循联合国宪章和《世界人权宣言》精神，尊重各国人民的平等权利。人工智能提供的服务应考虑全球各种文化、种族、性别和能力的差异，确保所有用户都能被平等对待。

合法性。人工智能服务应遵守所有国际适用的法律法规；面向他国提供服务时，应尊重他国主权、严格遵守他国法律、接受他国法律管辖，尊重他国内政、社会制度及社会秩序。

价值观平等。人工智能的训练数据和输出内容应符合使用国家的文化和价值观。训练数据的本地化数据集应占有一定比例，避免输出的内容偏离该国的社会认知和伦理规范。

3. 安全可靠

安全可靠强调人工智能系统的供应链可靠性、系统输出可靠性以及支撑人类决策可靠性，同时，针对系统安全、网络安全和算法安全方面提出明确要求。

可靠性。一是供应链可靠。人工智能系统应进行全生命周期的供应链安全管理，涵盖开发、训

练、部署和维护等环节，确保各个环节的供应链具备稳定性、持续性和安全性。二是系统输出可靠。人工智能系统的输出内容、数据等应确保其真实可信、稳定一致。三是决策不被误导。人工智能开发、部署和应用过程，应采取适当措施，避免其输出内容影响人类决策的独立性和判断力。

安全性。一是系统安全。人工智能系统应以稳健方式运行，增强人工智能系统的韧性、自适应性和抗干扰能力，确保持续稳定运行。二是网络安全。人工智能系统在开发、部署和使用过程中，应采取适当网络防护措施，具备防范网络攻击的能力。三是算法安全。人工智能系统应实施安全开发规范，避免存在逻辑缺陷或漏洞，提高系统的鲁棒性。

4. 数据隐私

数据隐私强调人工智能系统必须在数据处理和使用过程中严格遵循相关的安全和隐私保护规则。

数据安全。一是人工智能系统应遵循个人信息处理安全规则，确保个人数据的安全和保密。二是人工智能运行产生的数据应保留在应用国，其服务过程中的数据跨境传输应符合应用国的数据安全法律法规要求。

隐私保护。人工智能服务应注重保护个人信息，确保个人数据的安全和保密。

（上接 47 页）

察站，建立人工智能对就业市场影响的观察站网络等具体行动，旨在实现声明提出的六个优先事项。此外，“环境可持续人工智能联盟”的成立以及 Current AI 计划的启动，也反映了巴黎峰会在推动人工智能国际治理和合作方面开展了实质性的工作。

三、启示与展望

国际人工智能峰会历经三届，取得了包括多个宣言、声明、自愿性承诺、安全研究报告等系列成果，推动了人工智能安全科学研究和国际治理合作。然而，这些治理成果能否真正推行，历届峰会成果之间是否能够形成继承性和延续性，还需要探索建立常态化的工作机制保障实施。此外，在巴黎峰会上美国和英国拒绝签署《巴黎声明》的情况是否会引发西方阵营的内部分化，进而影响峰会举办无以为继，仍需进一步观察。

当前，联合国及其专业组织、OECD、G7 等国际或区域组织已开始开展人工智能治理相关工作。展望未来，在治理议题方面，峰会如果能更多考虑发展中国家的需求，帮助发展中国家在人工智能技术变革浪潮中获得红利，并广泛吸纳发展中国家参与人工智能治理规则的讨论，将有可能推动峰会在人工智能国际治理领域发挥更大作用。在治理模式方面，借鉴世界信息社会峰会（WSIS）的治理经验，峰会如果能够既强调政府的领导角色和责任，又广泛吸纳多利益相关方的参与，也有助于峰会扩大国际影响力。

中国一直关注人工智能治理问题，呼吁通过对话与合作凝聚共识，构建开放、公正、有效的治理机制，并致力于在联合国框架下团结全球多元主体参与治理。在人工智能安全领域，中国已开展大量研究工作，积累了相应基础，并于 2024 年 9 月成立了中国人工智能发展与安全研究网络。在巴黎峰会上，中国的研究网络成功主办了以“人工智能技术及其应用发展”为主题的官方边会，获得了中外专家的高度评价。借此契机，可以依托中国人工智能发展与安全研究网络与其他国家的人工智能安全研究所加强技术交流，探索开展联合研究、联合测试、人才培养等合作，推动相关人工智能安全科学研究成果纳入联合国框架下的治理工作，进一步发挥中国在全球人工智能治理中的关键性和建设性作用。