生成式人工智能
安全与全球治理报告

Safety and Global Governance of Generative AI Report

世界工程组织联合会创新技术专委会
深圳市科学技术协会

2024年1月
Jan 2024

目　录

编者的话

本报告是由多元的作者观点汇集而成，旨在引起公众对生成式人工智能技术发展的安全性和治理问题的关注，并激发进一步的思考。我们认识到，这一领域的发展速度迅猛，伴随着许多潜在的挑战和机遇。报告中的具体观点仅代表各个作者本人，而不代表世界工程组织联合会创新技术专委会（WFEO-CEIT）或深圳科学技术协会的立场。我们强调，对于生成式人工智能技术的探索和应用，需要行业内外的广泛合作与持续对话，以确保科技的进步能够造福全人类，并在伦理和法律框架内得到妥善管理。通过这份报告，我们希望促进更多的交流和合作，共同探索这一前沿科技的未来。

序言

人工智能治理：为了人工智能发展得更好更快，以加速实现全球可持续发展目标

龚克

首先祝贺世界工程组织联合会创新技术专委会（WFEO-CEIT）和深圳市科学技术协会共同组织编写了这份报告，做了一件很有意义的工作。在这份报告中，来自不同国家和地区、不同行业和领域的专家们，为我们带来了在不同视角下对人工智能治理的观察和思考以及有益的实践经验，他们从不同的角度提出完善人工智能治理的建议，包含了非常重要的共识：**比如，加快建立人工智能全球多方共同治理的机制和开展广泛的对话，将伦理作为人工智能治理的最重要的基础，将风险较高的领域作为加快建立全球治理规范的优先领域，等等。**

发布这个报告的时间，恰好处于《联合国 2030 年可持续发展议程》的中点。在不久前举行 2023 年联合国可持续发展目标峰会上，各国领导人一致呼吁要加倍努力，加速实现可持续发展目标（SDGs）。联合国秘书长古特雷斯指出，可持续发展目标不仅仅是一系列目标，它们承载着各国人民的希望、梦想、权利和期许。然而如今，只有15%的目标按预期进展，很多目标甚至出现了倒退。现在急需制定一项全球计划来挽救这些目标的实现。古特雷斯强调要在 6 个关键领域采取行动，其中之一就是“利用数字化转型机遇”。**可以说，我们急需人工智能成为推动 SDG 加速的实现的重要动力。**

人工智能是革命性的通用目的技术，是驱动第四次工业革命和经济社会数字化转型的先进生产力。无论是从全球的层面（如加速实现可持续发展转型），区域和国家的层面（如结合区域与国家实际的能源转型行动、促进经济增长和就业），行业和企业以及各种组织的层面（如提高行业的数字化转型、增进企业竞争力和组织效能），还是个人的层面（如提升职业能力、提升家庭生活的便捷性等等），人工智能都具有极大的潜力。**因此，人工智能的治理无论如何都不是也不应该是阻碍人工智能发展的治理，而是促使它更好更快发展的治理。**

人工智能的快速发展尤其今年以来生成式人工智能的快速发展，在给人们带来前所未有的体验和惊喜的同时，也加剧了人们对人工智能安全和伦理的关切，甚至出现

了一定程度上的社会焦虑。**这就凸显了完善人工智能治理、保证人工智能可控、向善的重要性和紧迫性。**

鉴于人工智能等新兴数字技术从本质上将是全球性的技术，这些技术不认可地缘政治边界。人工智能的发展和治理，涉及全人类的共同利益，它们产生的影响（无论是正面的或是负面的）都会产生跨越国界、跨越行业和专业的全球性、全局性影响。**因此，对于人工智能的有效治理必须是全球的、多利益相关方参与的共同治理。**

事实上，国际组织（联合国、G20、G7、OECD、欧盟等）和各国政府以及人工智能企业已经在人工智能治理上采取行动，从这个报告中可以看到这些努力和重要的治理发展以及有益的实践。然而，尽管各国、各个组织提出的这些治理原则在极大程度上是一致或相近的，但是仍然缺乏广泛的明确的全球共识，作为进一步加强全球行动的基础。**在当前已有的治理发展中，应该特别重视联合国教科文组织（UNESCO）的人工智能伦理建议书。**

世界工程组织联合会（WFEO）从工程促进可持续发展的使命出发，在高度重视促进人工智能发展和应用以加快双重转型的同时，也高度重视人工智能的治理。2020年 WFEO-CEIT 在第一个世界工程日发布了在工程中负责任应用大数据和人工智能的七项原则、2021 年 WFEO 支持联合国经济和社会事务部（UNDESA）和联合国秘书长技术事务特使办公室一起发布了《人工智能发展战略资源指南》，WFEO 还积极参与了 UNESCO 的《人工智能伦理建议书》的咨询工作。**我们认为，鉴于人工智能的发展和应用都离不开工程，而且只有工程化的人工智能才能真正在人类生产和生活中发生作用，所以，工程界应该成为人工智能共同治理中重要的、积极的一员。**

从工程的角度看，应该特别重视将人工智能治理的伦理原则、法律规定落实到可以检验的技术标准之中。这些标准应该是全球性的，可以互通的和具有互操作性的。而要是这些原则和标准落到实处而不是停留于纸面，应当优先发展支持治理的技术手段和工具，比如隐私计算的技术、伦理审计的技术，等等。

WFEO 还强调，人工智能的发展和治理离不开包括工程教育在内的广泛的能力建设，特别是要采取实际行动减少与人工智能相关的数字能力鸿沟，这本身也应该成为人工智能全球治理的题中应有之义。

总而言之，作为全球工程界的领导者——WFEO 愿意在人工智能全球共同治理中发挥积极的作用。我们相信，人工智能先进技术的发展和应用，是无可阻挡的，人工智能治理应该是促进性的治理，即以人工智能更好更快的发展为目标，最大限度发挥它的技术潜力为人类和地球的可持续发展服务；我们重申，人工智能的有效治理必须是全球的、多利益相关方参与的共同治理，当前应该的联合国的框架内组织广泛参与

的对话以促进明确的治理共识，并形成常效机制（如同气候协定），作为进一步推进共同治理行动的基础；我们强调，人工智能的治理应该是基于伦理的治理，UNESCO的《人工智能伦理建议书》为此提供了重要的基础；我们还注意到，已经提出的治理原则和正在进行的治理实践，都采取基于风险的差异化治理，因此我们呼吁对于人工智能的风险认识应成为全球多元对象的优先事项；我们还主张，要把人工智能发展和治理的能力建设，特别是缩小人工智能能力差距，作为人工智能治理的重要方面，并在把帮助发展中国家建设人工智能能力方面，实施有力且紧迫的行动。

龚克，WFEO 前任主席（2019-2022），WFEO-CEIT 顾问。

介绍

2023 年 10 月 18 日，习近平主席在第三届“一带一路”国际合作高峰论坛开幕式主旨演讲中宣布中方将提出《全球人工智能治理倡议》，并于同日由中央网信办正式发布，围绕人工智能的发展、安全和治理阐述立场主张，表示愿同各方就全球人工智能治理开展沟通交流、务实合作，推动人工智能技术造福全人类。10 月 26 日，联合国秘书长古特雷斯宣布，正式组建一个新的高级别人工智能咨询机构，全球 39 名专家共商人工智能治理，以探讨这项技术带来的风险和机遇，并为国际社会加强治理提供支持。11 月 1 日，首届全球人工智能安全峰会在英国布莱切利园拉开帷幕。包括中国、美国在内的 28 个国家和欧盟，共同签署了《布莱切利人工智能安全宣言》，一致认为人工智能对人类构成了潜在的灾难性风险。

在这个背景下，本报告汇集了全球 40 多位人工智能治理、科技伦理、大模型安全和对齐、通用人工智能风险等领域的政策制定者、企业家、专家学者、工程师等的 29 篇评论，旨在引起对生成式人工智能发展与安全与治理的关注和进一步的思考，呼吁开展广泛的合作。其中的具体观点并不代表任何主办和主编机构。按照讨论主题分为以下五章：

第一章，生成式人工智能的风险与挑战

从近期和长远两个时间维度来看，专家们关注的近期风险和挑战包括：一是大语言模型的隐私和安全隐患；二是大语言模型生成的虚假信息和“幻觉”问题；三是模型的价值观偏差和缺乏解释性；四是模型滥用造成的道德和伦理风险；五是人工智能应用带来的知识产权和法律监管问题。

而长远风险和挑战包括：一是人工智能可能导致经济和社会的重大变革，需要统筹应对；二是人工智能可能颠覆现有国际法律体系和世界秩序；三是人工智能存在全球共同的风险，需要建立国际规范和监管；四是不同国家和文化在人工智能价值观上存在分歧；五是强人工智能可能脱离人类控制，产生灾难性风险。

总体而言，近期的风险更多集中在个别模型和应用层面，而长期的风险和挑战更多关乎人工智能技术的整体发展方向和社会影响。但无论近远，建立国际合作和制定伦理规范对于应对人工智能风险都至关重要。

第二章，生成式人工智能的全球治理策略

专家们关注的优先事项为：一是全球合作与协调：在人工智能治理中强调国际合作的必要性，尽快启动多边协调与合作进程，促进广泛国家参与治理；二是风险识别

与管理：集中关注人工智能系统可能带来的共同大规模高风险危害；三是伦理和透明度：在人工智能的设计、开发和部署中强调伦理原则和透明度；四是技术发展与安全的平衡：在促进技术创新的同时，确保安全和合规性。

为此提出的相关政策建议包括：一是建立多边组织和国际社会的共同努力：为了识别和缓解人工智能风险，需要全球性的参与和合作，推动建立国际人工智能组织来确保国际监督标准的实施；二是制定风险预警和应对机制：包括事后监管审查和预防策略，确保系统的安全性和可靠性；三是建立第三方评估机制：独立专家的第三方评估补充内部评估，以提供一个稳固的安全网；四是构建可互操作的合规体系：推动不同国家治理规则标准化和对接；五是制定国际公约：在全球范围内分享人工智能成果与利益。

总体而言，当前亟需统一全球视野，就人工智能治理原则和政策达成共识，并采取统筹协调的国际合作，以应对快速发展带来的挑战。

第三章，人工智能治理助力发展中国家与全球可持续发展

专家们认为，人工智能治理可以为发展中国家提供的助力包括：一是解决紧迫问题：人工智能可以帮助发展中国家应对贫困、饥饿、卫生事件等迫切问题，通过提供精确的数据分析和解决方案；二是弥补资源缺乏：利用人工智能，可以在资源有限的情况下有效地管理和分配资源，特别是在科技和教育领域；三是改善数字基础设施：通过人工智能推动网络和通信技术的发展，提高互联网接入和计算能力，缩小数字鸿沟；四是缩小能力差距：提供优质的教育和技术培训，提升本地人才的技术专长，增强就业机会；五是本土化人工智能应用：培养适应本地需求和文化的人工智能应用，特别是在语言和文化多样性方面；六是国际合作：推动发展中国家参与国际人工智能治理，确保它们在全球人工智能产业链和价值链中有话语权。

对全球可持续发展的助力则包括：一是促进经济增长：人工智能可提升生产效率，降低成本，增强全球竞争力，尤其对发展中国家而言，这是推动经济多元化的关键；二是改进社会服务和基础设施：在教育、医疗、城市规划等领域，人工智能能提供更高效、更精确的服务，提升资源利用率，减少浪费；三是实现联合国可持续发展目标：人工智能可用于监测和评估可持续发展目标的进展，为政策制定提供数据支持，帮助更有效地管理资源，减少环境影响；四是推动包容性增长：通过包容性人工智能治理，考虑到所有国家的需求和愿望，确保技术发展惠及全球大多数人；五是国际治理与合作：建立国际治理机构和合作平台，促进知识和资源的共享，提供经济激励促进遵守规范，共同应对全球挑战；六是敏感性和透明度：监督人工智能行业的实践，确保其符合伦理标准，尊重数据隐私和安全，减少剥削性做法。

总体而言，人工智能治理在帮助发展中国家加速发展和实现全球可持续发展目标方面发挥着重要作用，但同时也需要注意其潜在的挑战和风险，特别兼顾不同发展目标方面。

第四章，工程视角下的人工智能治理

支持治理的技术手段和工具可能包括：

了解和评估模型能力的重要性。目前还缺乏系统的概念框架来决定模型的具体能力，这阻碍了人工智能的有效治理。建议制定评估模型能力的标准化方法应成为治理的优先事项，并强调了在人机互动中形成合理策略和广泛的理解、知识和技能的重要性，可解释的人工智能也有助于发展实用智慧。

加强人工智能安全治理的标准化工作。如建立准则更新机制、研制应用领域专项标准、建设试验区等。这些标准化工作有助于引导人工智能的可控发展。

基于风险和多方参与的方法治理生成式人工智能，开发评估框架和工具，并寻求国际合作。这为负责任地应用人工智能提供了参考。

通过开源和治理的民主化使人工智能开发与部署更符合公共利益，建议构建跨文化跨语言的伦理数据库，保持开放的公众参与，这有助于提高人工智能系统的安全性和价值对齐。

此外，还需要加强国际间的对话交流、建立包容的安全规则、开源高质量数据集等，以应对当前人工智能发展中的规则分散、价值对齐难度、加剧贫富分化等问题。

总体而言，工程技术对人工智能治理起关键支撑作用。需要深化对关键问题的理解，并将之转化为模型设计、训练与验证等具体实践。

第五章，企业视角下的人工智能治理

各家企业的讨论各有侧重：

迈克尔·塞利托 (Michael Sellitto)介绍了 Anthropic 的人工智能安全级别(ASL)的概念，用于管理人工智能潜在的灾难性风险。该方法借鉴了处理危险生物材料的生物安全级别(BSL)标准，根据人工智能能力定义了风险等级，并要求不同等级采取不同的安全措施。

司晓和曹建峰讨论了人类反馈强化学习在提高大模型价值对齐中的应用，以及其他技术和治理手段如数据处理、可解释性、对抗测试等在模型价值对齐中的作用，从工程层面保障人工智能系统价值观安全和对齐的方法。

韦韬指出了近年来大语言模型在快速进步的同时，也面临缺乏认知对齐、原则性和可解释性等问题。这会导致人工智能系统产生严重的错误决策并快速扩散执行，造