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

成难以预见的后果。建议人工智能系统需要提高认知一致性，建立可验证的推理链，并与人类专家互动学习。

英特尔文章以其开发的伪造检测技术为例，讨论了负责任地应用人工智能改善民众生活的方法，如提高效率、创造力，帮助残障人士等，从应用视角阐释了负责任的人工智能工程实践。

总体而言，这几篇从企业实践的角度，讨论了人工智能安全分级管理、价值对齐、开源治理、负责任应用等人工智能治理中值得关注的若干问题，提供了有益的建议和范例。

其中，对于粤港澳大湾区，专家认为可以发挥的独特贡献和价值

中国社会科学院哲学所科技哲学研究室主任、中国科协-复旦大学科技伦理与人类未来研究院的段伟文建议粤港澳大湾区可从三方面为人工智能治理作出贡献：一是数据治理创新，通过制度创新和试验探索，探索构建可信任的数据互通共享机制；二是实施人工智能驱动的区域整合发展战略，将人工智能治理的目标与人才、教育和就业战略结合起来，对大湾区的人才、教育和产业进行布局，使之适应人工智能未来的发展；三是打造人工智能东方大湾区特区，在良好的人工智能治理和人工智能驱动区域发展的基础上，吸引全球人才，通过更具动态可塑性产业促进政策和不断优化的人工智能治理模式，构建全球人工智能创新试验区。

中国电子技术标准化研究院的马骋昊、高万琪和范思雨倡议在深圳建设国际人工智能对齐与治理创新示范区。呼吁全球人工智能企业及科研院所共同参与，在一定范围内共同验证相关的对齐方法、标准规范、治理工具、数据共享机制等方面内容的科学性及可操作性。

南财合规科技研究院的王俊和娜迪娅建议粤港澳大湾区具有海量数据规模和丰富应用场景优势，数据要素市场不断扩大。应该充分发挥自身优势，充分挖掘数据价值，在数据合规基础之上，进一步促进公共数据等开放，推进多模态公共数据集建设，打造高质量中文语料数据。

香港大学的纳撒尼尔·沙拉丁(Nathaniel Sharadin) 认为，以系统的框架评估和理解模型能力，应成为治理的优先事项。粤港澳可以利用区位优势，吸引国际人工智能企业来区内共建治理示范区，进行各类人工智能安全和伦理治理工具的验证，为全球治理贡献经验。

综合利用粤港澳的区位优势、产业基础、开放程度等方面的独特条件，可以为全球人工智能治理作出积极贡献，提升区域和中国的影响力。

第一章　生成式人工智能的风险与挑战

大语言模型怪兽对利维坦与法律秩序的挑战

季卫东

自2017年谷歌发布Transformer网络架构以来，短短五年多的时间，世界上迅速出现了一大群大模型，而这些模型又衍生出多种技术架构、多种模态、多种场景。从已发布大模型的全球分布来看，中国和美国明显领先，超过全球总量的 80%，其中美国的大模型数量一直位居全球第一。

ChatGPT于2022年11月底一经发布，就凭借强大的对话能力和广泛的应用风靡全球，短短两个月的时间就让月活跃用户规模达到1亿，增速极其可观。此后，这些大型语言模型相继发布，从赋能个人、减轻企业负担等方面深刻影响了包括法律运作在内的各种社会实践场景，留下了一幅生成式 AI 物种大爆发的数字“寒武纪”景观。据不完全统计，截至2023年5月，中国科技企业和网络平台已上线各类人工智能语言模型79个，其中通用模型34个。

必须承认，大语言模式在给国家和社会带来便利和利益的同时，也带来了令人不安的风险甚至威胁。 其中四项可列举如下：

首先，由于类似 ChatGPT 的大语言模型提供在线对话服务，它们可以比现有的互联网搜索引擎收集更多的个人信息和隐私。因此，在“知道太多、利益冲突”的情况下，大型语言模型及其操作者可能会通过控制沟通，诱导用户做出违背自己意图和利益的选择。

其次，现阶段的大语言模型会将训练数据中不存在也不可能存在的事物视为真实的，并在对话中以不容置疑的语气进行描述。 这就是用户经常抱怨的“严重胡言乱语”的现象。 从科学技术的角度来看，这当然只是一种“幻觉”。“幻觉”现象与机器学习用有限的训练数据处理无限的未知数据的泛化能力密切相关。但在应用场景中，这种幻觉可能会导致虚假信息的传播，这对用户或社会来说可能是致命的。

再次，大语言模型在使用各种数据进行学习或人工智能自动生成各种内容时，可能会引发复杂的知识产权识别和保护问题。 为了确保 AIGC 的可信性并明确责任，应该发明、应用和推广数字水印技术。

最后，大型语言模型可能有意无意地获取企业或政府机构的机密信息，操纵舆论，导致国家中央系统的安全体系出现漏洞，信息社会功能失调，甚至因恶意事故和犯罪而引发社会动荡。

人类对语言的处理和智力的利用实际上是在无意识的情况下发生的。科学哲学家迈克尔·波兰尼曾在 1964 年指出：“我们知道的比我们能表述的更多。”换言之，知识体系还应包括这种没有明确意识到、或未被社会常识所认可、或不能言说的默会知识。这一命题被表述为“波兰尼悖论”，并成为人工智能理论的基础。这也意味着人工智能对无意识的语言处理，根本就无法设计那种获得和应用所有语言的算法，也很难为机器学习设定明确的训练目标。

现在通行利用神经网络进行机器学习，通过误差反向传播算法不断调整神经元权重和更新网络参数，逐渐减少误差，寻找训练数据的正解。人们发现，当神经网络的规模被大幅度扩张之后，人工智能接龙预测的精确度就会突然得到显著改善。这个发现及其有意识的应用使机器学习进入了深度学习阶段：无需复杂的规则和学习方法，只要让网络规模倍增就可以使许多难题迎刃而解，迅速提高泛化能力——不言而喻，这种神奇效果也证明了大语言模型的重要意义。其实质是多层网络的自我学习和进入语境（in-context）的学习，以及在此基础上实现学习方法的学习——元学习。这样一来，人类对机器学习的特征设计也就变得没有意义，人工智能实际上是开始进行自我塑造，形成一种自动化的生态系统，甚至有可能脱离人类的控制。

正是在这里，“大语言建模怪兽”纷纷崭露头角，并且有可能因为在对数据进行深度学习中放弃给定的特征设计，转而自我设置次级目标而脱离人类控制，进而引起治理方面的严重问题。这意味着大型语言模型将助产新型非人类或超人类智能的诞生，这种智能将逐渐远离人类并发展出与人类截然不同的价值观。 这也意味着，除了隐藏在区块链中的平台怪物和主权个体游击队之外，主权利维坦还将面临数十甚至数百个强大的大型模型庞然大物的挑战，即数字领域的国家主权，或者说“数字主权”。 ”面临着“百模大战”和失控的挑战。 “数字主权”概念明确体现了主权国家对社会数字化转型的反应和自卫立场。

为了防止上述各种风险演变成不可逆转的灾难，专家和行业领袖提出了暂停大型模型开发、实现价值对齐、加强人工智能监管等各种对策和建议。仅就价值对齐而言，例如，美国布鲁金斯学会 2022 年 12 月 8 日发表本杰明·拉森的文章《人工智能的地缘政治与数字主权的崛起》，作者认为人工智能发展的不平衡将导致国家之间的不信任加剧，进而导致数字主权的兴起和技术脱钩的出现； 意识形态差异或道德原则差异可能对人工智能和信息技术的管理产生更广泛的地缘政治影响；因此确保人工智能价值观在国际层面的对齐可能是本世纪最重大的挑战之一。无论如何，这是一场前所未有的巨变，将不可避免地塑造新的国家和法律存在形式，并促进秩序范式的创新。

针对这一重大变化，中国政府的策略是通过统一的超算网络和基座模型层来合并和整合数十个大型模型巨头，并通过所谓的“主权区块链”来防止点对点交互失去控制的风险。 结果必然会创建一个更强大的算法利维坦。 正如米歇尔·福柯所预料的那样，这个算法利维坦实际上是一个环视装置。 在这里，数十亿个探测器形成了视线陷阱，创造了大卫·里昂所描绘的那种监视社会和文化。 这种利维坦算法无处不在且强大，只有通过人工智能系统中嵌入的程序性正当程序以及不同人工智能系统之间的去中心化制衡才能防止其滥用。 从这个意义上说，也可以说，进入大模型和生成式人工智能时代后，人工智能治理的重点将从防止算法歧视转向防止模型滥用。 在主权利维坦、平台怪物、LLM 巨头乃至自我主权身份意识的互动中，法律正当程序原则将被重新定义并与技术正当程序相结合，这种新的程序正义将发挥更重要的作用。

人工智能价值对齐的挑战是构建符合人类价值观和利益的人工智能系统 (Russell, 2019) 。 这一挑战涵盖技术和规范方面 (Gabriel, 2020)。 这项技术挑战旨在将人类价值观编码到人工智能系统中，确保它们按照预期行事。 规范性挑战涉及确定人工智能系统和更广泛的人工智能开发工作应遵循哪些价值观。本文重点关注规范方面，并探讨了人工智能民主化的两种形式——人工智能开发的民主化和人工智能治理的民主化——作为在人工智能发展中代表不同人类价值观的手段。

季卫东，上海交通大学文科资深教授、上海交通大学中国法与社会研究院院长、人工智能治理与法律研究中心主任、计算法学与 AI 伦理中心主任、日本研究中心主任，美国斯坦福大学访问学者。国家重大人才工程特聘教授，享受国务院特殊津贴。

以更积极主动的治理应对人工智能发展中的风险与挑战

段伟文

近年来，人工智能在认知、决策、知识生产和智能代理等方面日益显现出超强能力，这使得人工智能成为各国科技领域的优先发展事项。但也因为其发展所凸显的危害、风险和争议，而促使各国和全世界全面展开了人工智能治理。

人工智能治理主要针对两个方面的问题：一是由于人工智能发展的价值和目标不明确，以及恶意使用和滥用，使其给人类、个人、社会、环境和生态系统造成了现实的危害和潜在的风险，其中包括对隐私和数据权侵害、偏见和歧视的加剧等。二是因人工智能发展的不平衡以及受益与风险分配的不公，带来了“谁受益？谁付出？谁承担风险？”“谁具有领先优势？谁会被甩在后面？谁处于‘暴露’状态”等社会争议。

从防范风险的视角，目前全球人工智能治理的紧迫事项包括三个方面：一是如何避免有意和无意的人工智能滥用可能导致的重大风险，特别是以往没有关注到的跨领域的复合风险，如随着技术与信息可及性增加和门槛的降低，人工智能与生物技术的非常规结合导致的不可预见的安全风险；二是如何缓解人工智能特别是生成式人工智能对岗位、就业、人才、教育的巨大冲击；三是如何在国际层面形成人工智能军备竞赛的多方管控机制和公开对话渠道。

其中，对于国际社会如何建立有效的风险预警和应对机制并确保在关键时刻人类有能力摁下停止键，这一问题目前没有直接的答案，当可用从以下四个方面作出努力。一是各国和世界要在信息网络系统的安全性方面加强互信与合作，建立起全球风险预警系统；二是各国、不同区域和全球要构建起多个平行的信息网络系统，在有必要的情况下构建人类文明全数据备份时间机器、全球多个信息网络平行运作系统；三是发现和培养可用在人工智能时代具有超强认知、决策能力和未来洞察力的人才，在各国实施超强人才教育培养计划；四是加强人机行为的社会学、人类学、心理学和哲学研究，对此问题展开系统深入的探索。

从促进发展的视角，为了助力世界各国特别是发展中国家的高质量发展和联合国可持续发展目标的落实，人工智能治理要从“被动补偿”和“主动优化”两个方面入手：一是要设法缓解以上两方面的问题，促使人工智能系统在整个生命周期中以负责任和可信任的方式发展；二是应联合国际社会进一步采取一系列主动的预先应对措施，包括利益补偿和平衡发展政策的实施，风险预见和强化防范、欠发达地区的人工智能素养提升等。

最后，对于粤港澳大湾区参与贡献人工智能治理，我有三点建议：一是数据治理创新，通过制度创新和试验探索，探索构建可信任的数据互通共享机制；二是实施人工智能驱动的区域整合发展战略，将人工智能治理的目标与人才、教育和就业战略结合起来，对大湾区的人才、教育和产业进行布局，使之适应人工智能未来的发展；三是打造人工智能东方大湾区特区，在良好的人工智能治理和人工智能驱动区域发展的基础上，吸引全球人才，通过更具动态可塑性产业促进政策和不断优化的人工智能治理模式，构建全球人工智能创新试验区。

段伟文，中国社会科学院哲学所科技哲学研究室主任、中国科协-复旦大学科技伦理与人类未来研究院教授，享受国务院特殊津贴专家。

人类价值对齐难题与大模型伦理嵌入

王小红

信息伦理学家指出：必需确立一套基本的 AI 伦理准则，但这不容易，因为道德准则会在不同文化背景和 AI 使用情境下产生差异。(Taddeo and Floridi, 2018) 机器道德哲学家也强调：虽然存在超越文化差异并为人类共享的价值观，但不同文化以及人类的不同道德系统在细节上仍有分歧。(瓦拉赫，艾伦，2017:66) 有实证调研指出：AI 伦理原则取得实效的关键在于其本地化，而本地化过程中必然要遵循当地文化、宗教和哲学传统。(Danit Gal 2019:73) 上述研究表明，“以人为本”这一抽象原则，在 AI 治理实际情境中，往往会因文化差异导致实践价值差异，甚至可能出现 AI 治理技术的相互反制。

由此，基于人类价值对其的难度和复杂性，我们提出 AI 伦理建构的哲学智慧共识策略：

第一，超越文化的“概念鸽子笼”。(Dewey, 1921: 188) 多元文化、不同价值的交流中，人们往往使用自己熟悉的文化中的“概念鸽子笼”，把另外一种文化中的事实分格塞进去，来解释不同的文化现象，这就造成了轻率地归类，主观地下定论。洞悉一个不同的文化的真正意图，任何时候都是复杂的任务。不同文化承载着不同的人类群体独特而漫长的生命历程，“使历史成为实际的原因是求生的意志和求幸福的欲望。但什么是幸福？人们对这个问题的答案远非一致。这是由于我们有许多不同的哲学体系，许多不同的价值标准，从而有许多不同类型的历史”。(冯友兰, 1922) 人类只有反思自身，藉理性以纠偏。

第二，在有利于人性进步之准则下，尊重不同的历史所沉淀的多元价值。中西哲学比较研究揭示，文化差异性的一个重要方面就是，虽然所有文化共有一些基本价值，但是不同文化会给予这些价值不同的权重，形成不同的价值配置形式。(李晨阳, 2019) 不同文化在价值配置上可能永远无法一致，但是不同文化可以基于多元化的配置，达成有利于人性进步和人类发展的共识。

东西方哲学思想和思维方式历经上千年，不仅亘古常新，却自古就不缺少深度共鸣。先秦孔子有言：“吾有知乎哉？无知也。有鄙夫问于我，空空如也。我叩其两端而竭焉。”古希腊苏格拉底说：“我只知道我一无所知”。中西方两位哲人，几乎在同一时期（雅思贝尔斯称之轴心时代）道出了遥相呼应的治学箴言。《论语》的多个表述，也与康德道德律令关键思想一致，即，只有当你愿意依此准则行事，才令此准则为普遍规范。源于《礼记·中庸》的“慎独”，新儒家继承和发展的“修齐治平”工夫论，这些思想与亚里士多德倡导的美德伦理学，均投射出东西方古老文化共同的哲学智慧。

第三，AI 大模型的伦理嵌入，不论自上而下式还是自下而上式，皆需要分析和清晰呈现道德哲学的语义蕴含。当前成功的大模型所基于的训练范式，“高质量数据集构建—大规模预训练—指令微调—基于人类反馈的强化学习”，是符号主义进路（自上而下式）与联结主义进路（自下而上式）的融合，被计算机科学家称为未来优先发展的“集成智能”方向。(陈小平, 2020：116) 其中，构建推理基于的知识库或者搜索基于的状态空间，以及训练大模型的伦理规则神经网络，使用隐含伦理原则的代表性数据库时，皆需要做基于道德语义的情境分析和评估。我们设想，为获取人类价值共识，可以发展类似汉典建模 (王小红等, 2023) 的计算诠释学，使道德概念的丰富涵义在形式层面得到梳理和辨析，将文化意义架构整合进有监督（人工标注）和无监督（自动标注）的机器学习。

王小红，西安交通大学人文社会科学学院科技哲学教授，计算哲学实验室中方负责人。兼任中国自然辩证法研究会方法论委员会常务理事，陕西省自然辩证法研究会前副秘书长。

人工智能的全球监管：主要差距和核心挑战

罗斯塔姆·J·诺伊维尔特(Rostam J. Neuwirth)

通常被称为“人工智能”的新技术正在以更快的速度引入，并越来越多地渗透到人类生活的方方面面。正如 2021 年 11 月通过的联合国教科文组织《人工智能伦理建议书》所反映的那样，最初对 AI 潜在利益的热情现已被对其实际和潜在危害的伦理关切所取代。为了解决这些问题，全世界正见证着通过具有约束力的法律手段来监管 AI 的全球竞赛。欧盟、欧洲委员会、中国、美国以及许多其他司法管辖区已经采纳或正在准备针对人工智能的专用或通用的法律。

然而，目前的全球 AI 竞赛对法律和现有国际法律框架提出了严峻的挑战，首先是一个强烈的时间要素，即不仅要找到颁布 AI 法律或法规的最佳时机，还要使其具备未来一段有意义的时间内的适用性。

其次，它还具有空间维度，即由 AI 的无所不在或跨界性质引起的法律问题，这与法律的传统领土观念形成了鲜明对比。为了考虑到各种 AI 系统的互操作性，防止可能的技术故障或规范冲突，必须采取多层次治理方法，以实现各种地方、国家或地区监管方法的全球协调和协调。

第三个挑战在于 AI 对社会、组织和人类的影响的全面效应。这意味着 AI 是一种跨领域现象，它首先需要跨学科的讨论，以支持基于建立一致的制度框架的协调监管方法的制定，从而实现更有效的多机构合作形式。

第四个挑战在于现有语言在更好地接纳 AI 及相关技术的新特性方面的局限性。这个问题表现在将 AI 定性为一种矛盾修辞，即一种修辞手法，将智能与人类和机器这两个看似矛盾或极不相似的术语联系起来。因此产生的矛盾要求重新思考人类认知的基本前提，并澄清 AI 监管的主要目标，即它是旨在关注技术、将其投放市场的企业或提供者还是使用它们的人。在这方面，目前的提议往往过于模糊，或同时过于具体和过于宽泛。因此，真正全面的 AI 监管需要新的思维方式，这些思维方式同时针对特定的法律问题和法律系统整体的一致性。更重要的是，所有这些挑战的复杂性需要就 AI 的潜在用途和对人类进化更广泛目的达成全球哲学共识。

罗斯塔姆·J·诺伊维尔特(Rostam J. Neuwirth)，澳门大学法学院教授，研究兴趣包括国际经济法和“贸易联系辩论”、知识产权和创意经济、比较法以及法律和法律理论的各个跨学科方面。

生成式人工智能对全球治理的挑战与应对

孙南翔

当前，以生成式人工智能技术为代表的新兴科技革命和产业革命正加速推进。人工智能技术的运用不仅推动着国家治理体系的变革，也深刻影响着全球治理机制的发展进程。生成式人工智能技术使得技术主体出现了“自知自觉自治”的能力，也使得非国家行为体拥有可比拟于国家的权力， 法律的作用应该被重新审视、重新开创，特别是确保技术发展遵循人的发展之脉络。全球治理同样如此。

与传统时代相比，人工智能时代将颠覆传统的国际法律框架，国际体系的主体、结构、运行规则等关键要素都将随之发生巨变。人工智能技术的发展对世界秩序构成了严峻的挑战，当然也必然带来新的发展机会。当前在世界范围内，生成式人工智能工具广泛地运用于传媒、研究，甚至是运输和武装冲突等领域。如何认定人工智能工具的法律属性的问题至关重要，然而截至目前，人类社会尚未形成共识。在可预见的未来，生成式人工智能技术产生的自我学习与自我认知能力，既不能被人类所规制，也无法为人类所预知。毫无疑问，当今世界面临的巨大挑战将是无知与未知的人工智能技术。

从现阶段人工智能的技术发展而言，生成式人工智能技术发源并受制于人，人工智能所产生的思想、观念和认知来自于人类世界。我们应加速研究生成式人工智能技术对国内与国际法律机制的影响以及其对人类社会生活的挑战。总体上，应对人工智能对国际法的挑战应坚持体系融合原则、技术穿透原则和法律技术化原则。从此层面，国家应积极探索法治原则对人工智能技术的约束作用，加强道德、伦理与技术的融合，并与世界各国携手共同应对生成式人工智能技术所引发的挑战。

孙南翔， 中国社会科学院国际法研究所国际经济法研究室副研究员，主要从事国际经济法、网络法研究。担任中国法学会网络与信息法学研究会理事、北京市法学会互联网金融法治研究会理事、对外经贸大学数字经济与法律创新研究中心研究员、北京市法学会百名法学英才等。

全球人工智能风险不可避免地需要全球合作

邓肯·卡斯-贝格斯(Duncan Cass-Beggs)

超级人工智能可能比我们想象的更近。这给人类提出了紧迫的生存问题，例如开发这种先进的人工智能系统是否以及何时安全，以及这些系统应该在社会中发挥什么作用。为了有效和合法，这些选择需要由国际社会共同做出并实施。这将需要在可能很短的时间内进行前所未有的全球合作，并且必须在地缘政治紧张和冲突的背景下取得成功。迎接这一挑战需要坚定的决心和不懈的创新。

人工智能带来的全球范围的风险不能再被忽视。这些风险的存在取决于三个观察结果：1）人工智能系统正在迅速发展，并且在广泛的关键能力方面表现出可能远远超过人类的水平。这可能会比预期发生得更早。2）人类目前缺乏可靠地控制此类系统或以其他方式确保它们与人类利益保持对齐的手段。一些专家认为，最终可能会证明，较低智力水平的人不可能持续可靠地、可持续地控制极其优越的智力水平。3）如果人类创造出能力极其强大且无法可靠控制的人工智能系统，其结果很可能是有害的。先进人工智能带来的灾难性全球规模风险的例子包括：a）恶意行为者的滥用，例如新型病原体或网络武器的创造和广泛部署；b）未对齐，指创建一个或多个强大自主人工智能系统，其目标与人类的目标相冲突，无法控制或阻止。

需要全球协调决策和执行的第一个也是最紧迫的问题是，是否以及何时足够安全以允许超级智能的人工智能系统的开发。世界各地的人们都有共同的利益，即确保任何地方的任何人都不会开发出可能危害人类的人工智能系统。因此，在可靠的调整和控制机就绪之前，不应允许具有超人类一般能力的人工智能系统。为了实现这一目标，协调的许可制度可能需要事先进行风险评估并获得开发最强大（即“前沿”）人工智能系统的许可，并且这种制度可以得到强有力的国际监督机制的支持。然而，随着开发具有潜在危险的强大人工智能系统所需的算法、数据和计算能力变得更加普及，执行这一制度将变得更具挑战性。

人类必须共同解决的第二个问题是，如果有可能开发出安全且对齐的人工智能，那么我们要追求什么样的未来。保证安全可能需要很长时间，甚至被证明是不可能的，需要长期或永久的全球严密禁止人工智能的发展。然而，如果在某个时刻人类确定创建一个或多个超级智能系统是安全的，就会出现许多基本问题，例如要创建多少个人工智能系统，以及应该赋予这些系统什么目标、权利或限制。与不同国家可以独立制定的许多人工智能政策选择不同，围绕引入一种新的高性能智能物种的问题需要集体做出，因为这种人工智能系统甚至会对可能不希望拥有这种系统的社会产生影响，至

少是间接影响。 从理论上讲，人类可以就某些此类问题向人工智能本身寻求建议，但至少需要提前就这些问题的参数和框架，以及选择询问这些问题的人工智能类型达成一定程度的共识。

长期以来，人工智能的变革潜力只存在于科幻小说中，而现在，它实际上已经来到了人类的家门口，需要在未来的几个月和几年里做出关键的决策和行动。尽管这些问题很严重，但似乎还远不能保证人类会认识到这些问题或及时应对挑战。需要努力帮助决策者了解风险并制定成功度过这一时期所需的新范式和新制度。要实现这一目标，需要社会各界和世界各地的人贡献想象力和奉献精神。

邓肯·卡斯-贝格斯 (Duncan Cass-Beggs)，国际治理创新中心 (CIGI) 全球人工智能风险倡议执行董事、经济合作与发展组织 (OECD) 前战略展望顾问。

第二章　生成式人工智能的全球治理策略

基础模型和生成式人工智能时代全球人工智能治理的制度设计原则

尼古拉斯·莫斯(Nicolas Moës)，尤兰达·兰奎斯特(Yolanda Lannquist)，
尼基·伊利亚迪斯(Niki Iliadis)，尼古拉斯·米埃赫(Nicolas Miailhe)

基础模型和生成式 AI 的新兴时代为 AI 治理带来了新的风险和挑战。鉴于它们的广泛应用和快速采用，积极地理解和降低这些风险至关重要。本文提出了十条 AI 治理的制度设计原则，以确保安全和维护人类价值观。

1. **内在、广泛和不可预测的风险：**风险不受用例和应用的限制，从设计到部署的整个生命周期都会出现。这些系统中的故障可能会以不可预测的方式出现，从广泛的内容审查偏见和公共心理健康危机，到自动化恶意使用、生物安全、网络安全和国家安全威胁，造成大规模、灾难性甚至生存性的风险。
2. **可信的设计：**这种方法旨在从一开始就将安全、安全性和伦理考虑因素嵌入 AI 中，而不是事后再进行改装，从而创造依赖于可信模型的新市场。
3. **收回控制权：**独立专家的第三方评估应该补充内部评估，以提供一个稳固的安全网。监管机构必须有权力根据这些评估暂停或修改开发过程。确保除私营部门主导的内部评估之外的稳健机构框架至关重要。
4. **技术和渠道中立：**为确保公平竞争并降低不可预见的风险，监管应横跨技术和分销渠道。无论 AI 系统是通过开源平台或 API 分发，还是使用不同的技术范式(如大脑仿真或基于规则的系统)开发，监管审查的基线水平应保持不变。
5. **针对性法律责任和社会责任：**大型科技公司和 AGI 公司等强大利益相关者应对其产品的影响负责和问责，作为主要承担责任的一方。在这些实体内部，安全官和合规官等角色也应承担具体义务。
6. **结构性和系统性实践：**安全、伦理和安全的做法应该融入组织的文化中，而不仅仅局限于特殊的、附加的措施，如红队演练。法规应强制执行以基于证据的要求，这些要求应随着 AI 能力状态的发展而变化。
7. **基于证据的要求：**开发人员应有义务通过经验证据来证明其安全、伦理和安全实践的有效性。基于证据的措施应随着最先进的 AI 能力以及风险缓解和预防措施的发展而变化，确保系统经得起未来的考验。

8. **公共部门能力建设：**加强监管有效性，这在很大程度上依赖于公共部门内的知识和技能。这种知识应该独立获得，不应受到行业的不当影响。知情的公共部门可以更好地抵制监管俘获，并在技术发展中对 AI 治理做出明智的决策。
9. **可适应和有韧性的治理机制：**政策连续性必须具有灵活性，以适应不断变化的情况，同时避免行业控制导致范围的稀释或转变。应赋予新的机构，如欧盟 AI 办公室，更新规则的权力。来自民众的监督可以确保继续关注安全、伦理和安全。
10. **可互操作的全球治理：**在各个司法管辖区之间促进一致的 AI 法规，这些法规在实施最稳健的安全、伦理和安全措施方面相互补充。实体之间应协调“提高治理要求的标准”，并避免法规碎片化导致的监管套利。

这些原则为治理提供了一个多方面、不断发展的方法，可以适应技术进步。遵守这些原则可以帮助确保 AI 风险得到缓解，而收益得到广泛和公平地分配。

尼古拉斯·莫斯(Nicolas Moës)，未来社会(TFS)欧洲人工智能治理总监，重点关注欧洲人工智能立法框架的发展，包括欧盟人工智能法案的起草和执行机制。

尤兰达·兰奎斯特(Yolanda Lannquist)，未来社会(TFS)全球人工智能治理总监，与国际组织、政府、公司、学术界和非营利组织一起领导人工智能治理和政策项目，以促进人工智能安全、道德、安保和包容性。

尼基·伊利亚迪斯(Niki Iliadis)，未来社会 (TFS) 的人工智能和法治总监，领导关于人工智能和法治以及美国人工智能政策的国际多利益相关者论坛雅典圆桌会议。

尼古拉斯·米埃赫(Nicolas Miailhe)，未来社会(TFS) 总裁兼联合创始人，人工智能全球伙伴关系(GPAI)、经合组织人工智能专家组(ONE.AI)和联合国教科文组织(UNESCO)人工智能伦理高级别专家组的专家。

有关全球人工智能治理的关键政策建议

周辉

在一个人工智能不断重塑经济和社会结构的时代，为这种变革性技术制定明智的治理机制已经变得空前迫切。结合中国《人工智能法示范法》及欧盟、美国的治理探索来看，为更加有效地监督通用人工智能或超越一定规模的基础模型的研发和部署，建立有效的风险预警和应对机制，需要采用多维度的框架以在适应人工智能发展规律的同时，及时有效回应由此产生的新风险新挑战。

一、伦理原则的规则化

《人工智能法示范法》明确要求在人工智能研发、提供和使用活动中，使用允许人类监督和干预其运行的系统架构，这一制度设计旨在保障人为修正或停止人工智能系统运作的能力，避免技术失控所带来的危害。除此之外，为能够更好预警风险，示范法吸收了一系列国际广泛认可的人工智能伦理原则，如确保人工智能透明度、公平、可追责等。人工智能研发者和提供者不仅需要对人工智能进行标记以提高用户感知度，而且还需在设计中保障人工智能算法、模型的透明度和可解释性，并在必要的时候向公众或监管机关提供相应信息以进行说明。在公平方面，《人工智能法示范法》提出如下要求：一方面，人工智能研发应用应该促进包容性，以减少数字鸿沟，并服务于各种弱势群体；另一方面，人工智能系统需要尽可能减少歧视性内容的生成和输出，并遏制其使用者进行类似活动。

二、精准治理

考虑到人工智能技术在复杂性和社会影响上的固有差异性，以及不同规模、不同用途的人工智能系统之间的区别，单一的治理框架难以满足精准治理的需要。《人工智能法示范法》提出了以负面清单为基础的治理模型，采用事后的监管审查作为其基石，同时为高风险的人工智能研发、提供纳入预防策略。这种混合方法也与欧盟《人工智能法案》中明确的基于风险的框架相呼应。针对通用人工智能或人工智能基础模型，示范法亦对其研发者、提供者课以特别义务，这一模式借鉴平台治理中对“数字守门人”提出专门责任义务要求的成熟实践经验，以减轻监管压力、提高治理效能。

三、均衡发展与安全

以人工智能技术助力经济社会发展，需要相关治理措施在发展和安全之间建立一个平衡，避免过于严苛的制度设计阻碍技术创新。为此，应当要求监管机构创造一个更加确定的监管环境，保障相关企业的经营预期。此外，示范法纳入了几个明确旨在促进人工智能技术发展的制度，包括发展算力基础设施建设等配套领域和打通数据要

素供给、设计监管沙盒制度、豁免开源人工智能技术提供者的责任等。此外，负面清单制度也致力于营造更宽松的低风险人工智能技术发展环境，除负面清单中所列的人工智能活动需要获取行政许可外，其余人工智能研发、提供活动仅需完成与信息披露相关的备案程序，且该程序不应成为实质审批等行政阻碍。现有的法律框架仍可在低风险人工智能活动中继续适用。

四、鼓励技术治理和多元共治

《人工智能法示范法》提出了鼓励监管科技、合规科技发展，实施多元主体综合治理人工智能的愿景。一方面，应支持监管科技、合规科技的研发和应用，创新治理工具，减轻治理压力，促进治理的智能化、自动化、科学化和精准化；另一方面，人工智能企业在治理中发挥重要作用，有充分条件将价值观与法律规则融入人工智能系统中，并有充足动机参与人工智能治理以为自己获取稳定发展环境。应建立充分有效的沟通、会商机制，倡导企业自律和行业自治，并发挥社会舆论监督作用，完善综合治理体系。

五、设立人工智能专责治理机关

在目前的人工智能全球治理中，鉴于人工智能叠加现有网络信息技术所能产生的跨区域、跨领域及跨国界应用，解决潜在的重复监管、多头监管问题较为迫切。《人工智能法示范法》主张建立一个唯一的国家人工智能主管机关，承担人工智能监管职责的同时与其他部门进行协调。这样的机构设计通过将监管功能整合到一个专门的监管机构下，旨在增强治理的一致性、专业性和确定性。各国人工智能专门主管机关也可更高效率地建立和参与国际交流协调机制，推动国家和地区间人工智能治理规则对接、规则互认等合作。

周辉，中国社会科学院法学研究所网络与信息法研究室副主任（主持工作）、副研究员，中国社会科学院大学法学院副教授、硕士生导师，耶鲁大学法学院访问学者。

基础模型开发和部署的国际监督

罗伯特·特拉格(Robert Trager)，菲恩·海德(Fynn Heide)

对强大的新兴技术建立国际监督具有挑战性。然而，对于先进的人工智能，建立这样的监管体系势在必行。这项技术带来的风险跨越国界。即使一些国家引领人工智能发展，所有国家都会受到影响，因此所有国家都应该在人工智能的开发和部署监管中拥有发言权。

文本着眼于民用人工智能的国际治理机会。军用人工智能的治理很重要，但更具挑战性。我们可以从试图通过《联合国特定常规武器公约》来规范致命性自主武器的经验中看到这一点。尽管经过十年的努力，仍未采取具有法律约束力的措施。

民用人工智能的国际监督前景更加明朗。一种方法是使用类似于国际民用航空组织 (ICAO)、国际海事组织 (IMO) 或金融行动特别工作组 (FATF) 的模式。这些组织不审计公司，而是审计管辖地。一个国际人工智能组织 (IAIO) 可以审计各辖区，以确保它们采用国际监督标准并拥有执行记录。与 IMO 和 ICAO 一样，这样的组织应与各国密切合作，以确保它们有合规能力。

任何治理体系都需要促进合规。促进民用人工智能治理体系合规的一种方法是将其与贸易挂钩。各国可能同意不从严重违反 IAIO 标准的司法管辖区进口其供应链包含人工智能的商品。他们还可能拒绝向此类司法管辖区出口关键投入，或向遵守协定框架的国家提供技术援助。由此产生的国内层面的监测和执法可能会带来重大好处。各国将不再那么担心扩散，局部的监测和执法可能有助于快速、有效地纠正违法行为。一般来说，国内执法行动比国际执法行动要快。该框架并未明确基础模型监管的具体细节——无论是国内还是国际，我们对此还没有达成共识。但建立国际治理并不需要完全的共识,只需要对一套最小标准达成广泛认同即可。

国际参与者可能会决定，超过一定规模的基础模型将受到监管审查，包括一系列标准化的模型评估。此类系统的开发和部署可以得到当地主管部门的许可。数据中心运营还需要一个许可证，该许可证需要类似于金融服务行业的“了解您的客户”实践。这样的制度可以限制人工智能的有害发展，同时也促进广泛的应用。事实上，监管制度对于广泛的应用可能是必要的。如果监管，前沿 AI 国家可能会担心扩散并更严格地限制对技术获取。一个全球体制可以鼓励前沿 AI 国家参与，同时保证全球受影响社区的发言权。

第二章　生成式人工智能的全球治理策略

罗伯特·特拉格 (Robert Trager)，牛津马丁人工智能治理计划主任、人工智能治理中心的国际治理负责人以及牛津大学布拉瓦尼克政府学院的高级研究员。专注于新兴技术国际治理、外交实践、制度设计和技术监管。

菲恩·海德 (Fynn Heide)，人工智能治理中心研究学者。专注于国际人工智能安全合作以及中国人工智能安全和政策。

协调、合作、紧迫性：国际人工智能治理的优先事项

卡洛斯·伊格纳西奥·古铁雷斯(Carlos Ignacio Gutierrez)

人工智能的国际治理本质上是一个复杂的问题，临近 2023 年末，我们还没有明确的解决方案。随着这项技术的能力以不可预测的速度增长，190 多个国家管辖区承担着管理其不断升级和不可预见的风险的任务。为了解决这些风险，独立行动显然不是一种最佳方法，因为在一个地方修补问题，并不能阻止其蔓延到其他地方。相反，人工智能的有效治理是一项需要全球参与的共同努力。 考虑到这一点，社会应优先制定考虑以下因素多边应对措施：人工智能的哪些要素应该受到治理以及如何治理？ 谁应该参与这个治理过程？ 什么时候是采取行动的合适时机？

何事和如何：人工智能风险的选择空间兼具广度和深度。可以提出任意数量的问题来吸引国际社会的关注。然而，为了成功地推动行动，候选问题必须在具有广泛需求和能力的国家之间引发一种共同意识。启动对话的一项提议是重点关注减轻人工智能系统直接或间接造成的共同大规模高风险危害。设置这样一个阈值的好处是它涵盖了相对狭窄的关注集。此外，它还有助于集中认识和同步努力，最好是通过一个多边组织，其具体工作流程由以下目标组成：识别人工智能系统产生的共同大规模高风险危害的载体。尽管通用人工智能系统会引起许多担忧，但该工作的范围必须包括符合其操作指南的狭窄系统。 此外，它应主动告知利益相关者潜在风险并识别现有的危害媒介。协调技术上合理且符合最佳治理实践的全球应对措施。这可以采取多种形式，具体取决于当前问题的规模和根源。例如，应对措施的范围可以从征求自愿标准作为预防措施，到实施一套强制性规则作为对持续关注的反应。 强制遵守商定的行动，以减少伤害的可能性和影响。 由于人工智能的直接和间接影响往往不受管辖权的约束，因此建立有效的执行机制需要最大限度地增加参与国的数量。 虽然应授权多边努力发挥这一作用，但它也可以招募、认证或委托公共机构和第三方（通常是在管辖权的基础上）来承担这一任务，以扩大其执行能力。

何人：无论其设计、开发或部署人工智能技术的能力如何，所有国家都容易受到人工智能风险的影响，并可能有意或无意地托管或庇护高风险人工智能供应链或系统本身的任何部分。这就是为什么无论地理位置、国家的政治制度或意识形态如何，合作都必须成为优先事项。从本质上讲，任何国家都不应被排除在参与解决全球人工智能问题的多边行动之外。减轻共同的大规模高风险危害的能力取决于广泛的参与。因此，应考虑对广泛的国家采取激励措施，从那些对人工智能的设计、开发和部署有影响力的国家，到那些在该技术风险方面作用相对有限的国家。

何时：我们面临着每天都在部署越来越强大的人工智能系统的情况，而对于什么构成适当的治理，投入的注意力有限。这凸显了集体努力主动应对人工智能风险的紧迫性。即使今天着手开始多边协调与合作，也可能需要数年时间才能建立一个体系。因此，从短期来看，如果最初一批有影响力的国家主动启动这一多边进程，也是可以理解的。这可能包括在系统商业部署、硬件制造、技术人才培养和/或建立全面监管方面领先世界的国家。从长远来看，联合国是唯一具有普遍代表性并有能力主办本评论中所述努力的组织。理想情况下，它应该负责验证、协调和执行减轻共同的人工智能风险的努力。

结论：在优化多边治理方面，不存在“正确”答案。 我们可以期待的是一个多边人工智能治理计划，该计划优先考虑协调、合作和紧迫性，以解决共同的大规模高风险危害。通过将全球注意力集中在缓解这些问题上，国际社会需要建立必要的共性，以实现人工智能治理唯一负责任的最终状态：这种技术的设计、开发和部署是安全且合乎道德的。

卡洛斯·伊格纳西奥·古铁雷斯 (Carlos Ignacio Gutierrez)，未来生命研究所(FLI)人工智能政策研究员。专注于人工智能治理，研究技术对硬法的影响，发表了关于美国法规漏洞的综述，并设计了有效的软法计划。最新成就是发布全球人工智能软法计划数据库。

数据伦理与联合国教科文组织开放科学建议

国际科技数据委员会数据伦理工作组

大数据和人工智能在科研中的广泛应用带来了伦理和规范方面的挑战，特别是在公开性、隐私性、透明性、问责制、公平性和责任性方面。国际科技数据委员会(CODATA)的数据伦理工作组(DEWG)正在与全球学者合作，共同确立数据伦理原则和覆盖整个数据生命周期的数据伦理框架的基本共识，以进一步开展活动和研究。DEWG 有 4 个专题小组，分别关注科学诚信、个人数据

保护、本土数据治理和全球权力与经济关系。这将帮助 CODATA 按照联合国教科文组织开放科学建议，促进其在支持全球开放数据交流与应用方面的使命。

伦理与科学诚信专题小组讨论了研究的透明度、质量、可重复利用性和影响力等话题，关注协作努力和开放学术在支持科研诚信方面的作用。我们建议制定政策和实际指南，以推进全球数据伦理规范，通过创建支持结构来增强研究人员的数据主权，并加强研究方法在数据伦理中的关键作用，开发数据伦理培训和教育资源。

伦理与个人数据保护专题小组探讨了数据的政治及政治经济学相关问题。我们建议构建一个框架和政策，基于对隐私的更多批判性理解，认识到损害个人和社区的潜在动力和权力关系，并提供机会来提高对隐私及相关话题的技术、社会、法律和政治发展的技能。

伦理与本土数据治理专题小组关注诸如 CARE(集体利益、控制权限、责任和伦理)和 JUST(明智、无偏见、安全和透明)等数据原则，我们建议本土数据主权需要建立数据受托人(和类似的中介机构)，这将实现有选择的数字披露。

全球权力与经济关系专题小组探讨了塑造国家层面和个人层面研究的结构条件。许多国家背景中的学者面临基础设施不足、国家政策不支持、全球北方资助者控制研究议程以及少数出版商和大型科技公司主导等障碍。在个人层面，各地不符合典型学者形象(白人、身体健全、男性)的研究者面临诸如意识形态和无意识偏见、种族主义、厌女症、职业中断和社会对照料的期待等多重障碍。

DEWG 已获得 2023 年于奥地利萨尔茨堡举行的 CODATA 全体大会批准，晋升为数据伦理任务组。面对人工智能快速发展和广泛应用带来的挑战，这个任务组将在后续工作中加强对数据访问伦理的研究。

致谢：本文件基于 DEWG 各专题小组的研究成果和报告，这些成果和报告已在 2023 国际数据周上由 Johannes John-Langba、Joy Jang、Masanori Arita、Louise Bezuidenhout 和 Lianglin Hu 发布。

第二章　生成式人工智能的全球治理策略

国际科技数据委员会(CODATA)数据伦理工作组(DEWG)，使命是与全球学者合作，为数据伦理原则和数据伦理框架的进一步活动和研究建立基本共识，具体目标是就 CODATA 活动如何有助于解决数据伦理问题探索前景并提出建议。

通用人工智能或大型基础模型的国际治理：渐进原则与开放探索

张鹏

通用人工智能或大型基础模型的部署和应用能够显著提升生产效率，增强决策质量，是能够影响人类社会发展的革命性技术，但其在实践中也可能带来偏见和歧视、数据隐私保护、网络安全、知识产权、虚假信息等问题。如果不能加上安全的护栏，将对人类生活带来负面影响甚至生存风险。

通用人工智能或大型基础模型相较一般人工智能的特点是数据和参数规模庞大、运行过程难以解释，是资源密集型和数据密集型行业。它们的大规模商业化应用形成的是由模型提供者、部署者、使用者和用户组成的一条供应链。因此，对通用人工智能或大型基础模型的治理或监管，也不能采取单一主体的视角，或者所有参与者“连带责任”“一体问责”的思路，而应当具有系统思维和精确的供应链意识，区分供应链上不同的主体及其对系统的控制力，根据控制力大小设置精细化、有区分的安全义务与合规要求。例如，上游提供者应当对底层模型引发的风险承担更多的安全责任，下游使用者则应当对其产品和服务中对基础模型的应用产生的安全风险副承担更多的预防、缓解和消除等义务。

通用人工智能或大型基础模型的国际治理，如能在最具普遍性和代表性的联合国框架下制定统一的规则，是最理想的情况。也有很多观点认为应当借鉴国际社会对核能、气候变化的治理经验，建立相应得制度和机构。但要看到，通用人工智能或大型基础模型作为一项新兴技术，其发展仍处在方兴未艾的阶段，更涉及各国主权和复杂的地缘政治和科技竞争等因素，各国立场主张和利益考量未必相同，要达成一致的共识和明确的规则难度不小。应当秉持边发展、边探索、边治理的基本原则，伦理和标准先行，国内立法和国际立法审慎包容、循序渐进，在确保技术进步和创新的同时建立有效国际治理框架，消除通用人工智能或大型基础模型的潜在风险。

张鹏，上海市人工智能与社会发展研究会高级研究员、对外经贸大学数字经济与法律创新研究中心研究员。

第三章　人工智能治理助力发展中国家与全球可持续发展

发展中国家距离利用人工智能的力量还有多远？

尤金尼奥·巴尔加斯·加西亚(Eugenio Vargas Garcia)

对于许多发展中国家来说，人工智能的前景仍然是一个遥远的梦想。全球不平等是一个棘手的问题。我们不应低估这一挑战的严重性。造成这种情况的原因是多方面的。以下是其中四点：

- 竞争性优先事项——更紧迫的问题，如贫困、饥饿、突发卫生事件或暴力，通常在政策决定和预算分配中优先考虑。
- 缺乏资源——试图在资金很少的情况下，同时解决不同的问题可能是一项艰巨的任务，尤其是在科学和技术方面。
- 数字基础设施差——网络连接效率低下（有时电力供应也不可靠）、过时的通信网络、硬件和计算能力短缺等。
- 能力差距——尽管人才无处不在，不分国籍，但当优质教育、技术专长和就业机会缺失时，改变这种状况所需的努力就会艰巨得多。

由于上述原因和其他潜在因素，许多全球南方国家一直在努力弥合计算鸿沟并成为人工智能就绪国家。权力不对称、贫富差距扩大、数据剥削、网络殖民、标注员报酬过低、算法偏见和歧视性人工智能系统是最常见的困境。

在许多地方，技术仍然不发达且未得到充分利用。最终用户可以在线访问其他地方开发的产品和服务，可能是通过带有偏见的数据集进行训练的，可能不适合本地需求或国家优先事项。

应在更多不为人知的本土语言上训练模型，以便覆盖尽可能多的人，包括弱势群体或弱势社区。技术很少是中立的。生成式人工智能一直由英语用户的软件主导。发展中国家必须推动对使用本国语言的人工智能的研究。

为了确保人工智能能够帮助人类消除贫困、建设有韧性的社会、保护地球并在2030年实现可持续发展目标，我们还有很多工作要做。

试图通过大量投资迎头赶上并不适合所有国家，特别是面临严重障碍的最不发达国家。完全的数字主权和对自己数据的控制取决于国内能力和适当的资源。只有少数中等收入国家有可能在不诉诸国际合作的情况下，成功实施全面的本土人工智能战略。

全球治理举措可能会寻求通过建立创新机制来应对这些挑战，如通过多边管辖下的云服务或超级计算机使研究设施和关键基础设施得以利用。一种公私合作伙伴关系可以向发展中国家的专家提供人工智能资源。

未来的国际治理机构应包括专门设计的计划，旨在促进人工智能的和平利用、本土研发自主权、技术传播以及贫穷国家的实地实施。

联合研究、开放访问、能力建设，以及公平公正的利益分配，对于不让任何人掉队至关重要。

除非我们认真对待世界上大多数人的关切，否则获得人工智能革命的回报将只是少数人的特权，或者更糟糕的是，由个别人控制。

尤金尼奥·巴尔加斯·加西亚 (Eugenio Vargas Garcia)，科技外交官、巴西驻旧金山副总领事，科学、技术和创新负责人。人工智能与全球治理学术研究员。2018-2020 年纽约联合国大会主席前高级顾问。

推动发展中国家参与人工智能治理与可持续发展

鲁传颖

人工智能技术的发展给世界各国的经济与社会发展带来了机遇与挑战。对于发达国家而言，机遇主要表现为可以利用人工智能技术推动产业的转型与升级，赋能经济的高速增长。而对于发展中国家而言，主要的机遇则在于可以利用人工智能技术解决发展中的问题，比如说在医疗健康领域，人工智能一方面可以为医生提供辅助性的数据分析，有利于对患者病情举进行充分了解和分析，提高发展中国家的医疗能力；另一方面，从更为宏观的角度来讲，人工智能可以推动发展中国家建立系统的公民问诊和健康保障系统，从而提升发展中国家的医疗水平。此外，在教育、基础设施、农业、能源与环境等领域人工智能也为发展中国家带来了众多的发展机遇。

然而，从当前来看，人工智能的发展红利尚未惠及发展中国家。根据牛津洞察(Oxford Insight) 发布的 2022 年的人工智能政府就绪指数，发展中国家在技术基础设施、创新能力建设、以及政府的治理能力等方面仍存在较大缺口，导致人工智能的发展出现明显的南北分野。

鉴于此，在考虑人工智能治理时，需要从两方面着手推动发展中国家的参与与联合国可持续发展议程的推进：

一方面，需要进一步发挥联合国等国际组织的作用，落实可持续发展议程。当前联合国等国际组织逐渐意识到了人工智能的治理需要惠及更多发展中国家。比如说，联合国贸发会议发布报告《技术和创新促进更清洁、更具生产力和竞争力的生产》讨论通过官方发展援助、贸易和外国直接投资促进技术转让到发展中国家，解释了在联合国系统中的运行机制，并探讨了利用技术和创新实现包容性和可持续发展的方法。国际货币基金组织开展专项研究探讨人工智能对发展中经济体的影响，经合组织人工智能观察站也对发展中国家的人工智能部署情况展开跟踪，并通过工作文件的方式提出数字化转型的建议等。这些国际上的措施展现了一个良好的趋势，但有待进一步落实。

另一方面，中国也应该积极参与人工智能治理，为发展中国家发声：从现实来看，发展中国家在人工智能国际规则建立上仍处弱势地位，这一方面与广大发展中国家的人工智能发展还处于落后阶段有关，也与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发展中国家。因此，中国要肩负起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的责任。

中国应以将推动全球南方国家研发和应用人工智能技术应用优先选项，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。

鲁传颖，上海国际问题研究院研究员、网络空间国际治理研究中心秘书长。主要从事网络安全与网络空间治理研究，担任中国-欧盟数字经济、网络安全专家组委员、个人信息安全规范工作组专家，网络治理与国际合作工作委员会专家委员。

人工智能供应链与地缘政治：与全球南方国家共同治理

方淑霞(Marie-Therese Png)

越来越明显的是，位于全球北方经济权力中心的国家和社区最有可能从人工智能研发中获得利益。与此同时，成本则由那些在社会经济、地缘政治和贸易动态方面已经处于不利地位的全球国家承担。包容性人工智能治理举措旨在解决这种分配不平等问题，但尚未将底层结构纳入其分析中。例如，人工智能产业深深植根于金融、军事、自然资源和地缘政治优势等激励结构中，这削弱了全球多数国家的物质安全。

人工智能治理举措中的少数领导者认识到，他们有责任确保生成式人工智能的部署和监管不会加剧国内和国际的不平等。此外，他们还理解全球包容性努力的潜在战略优势——用于构建共识、提高治理效力和地缘政治稳定。他们理解不同地区之间极端的权力失衡会在全球层面上产生长期的有害和不稳定效应——竞争和冲突，以及政治不确定性、贸易战、制裁和国际动荡如何破坏供应链和创新。

一个具有全球代表性的人工智能治理流程还提供了一幅以经验为基础的世界政治图景。引用 Albert 等人(2020)的话，全球治理忽略了帝国和(后)殖民时期国际关系的过去和现在，从而呈现出一个深深植根于欧洲中心主义的世界政治理论画面，因此在理论和实证上都存在严重缺陷”。基于全球南方利益相关者的需求、要求和目标的行动导向讨论——特别是民众和非专业人士——有助于确定影响杠杆、障碍和差距，从而制定明智的战略。

这种做法以及其他形式的共同治理 (Brito 等, 2021; Png, 2022)，有助于理解当前治理进程的预期和实际结果之间的价值观偏差，以便迭代并为广大全球人口提供实质性的服务。

全球南方利益相关方可以为国际人工智能治理的差距提供有价值的经验证据，包括人工智能行业对正常化但剥削性做法的依赖所带来的负面影响。人工智能行业部分依赖于大规模提取和货币化数据、廉价数字劳动力，以及提取矿物、金属、水和土地以建立硬件和物理信息基础设施。必须审查这些商品的采购是否存在剥削性做法，特别是考虑到全球南方的许多司法管辖区在数据保护、所有权和货币化、数字平台、计算、数据中心、数据流动、劳动惯例和复杂的多国供应链中的自然资源等方面仍缺乏预先存在的安全保障和法规 (Veale 等, 2023)。这使得这些国家及其人民系统性地更容易受到人工智能发展带来的风险的影响。这种情况还因基础设施发达的国家依赖并有动力以使其保持欠发达的方式从在世界市场经济中地位较弱的资源丰富国家中获取资源而进一步加剧(Rodney, 1972)。随着生成式人工智能系统能力的提高，对这些商品的需

求也在增加，这种情况可能会变得更加严峻。这需要制定对这些动态敏感并了解它们如何导致前文提到的政治不确定性、国际动荡、供应链波动，最终破坏长期创新及其利益的安全保障和法规。

方淑霞 (Marie-Therese Png)，牛津互联网研究所博士在读，曾任联合国秘书长数字合作高级别小组技术政策顾问，专注于数字包容、 致命性自主武器、网络安全、算法种族歧视等技术政策领域的研究。

人工智能监督可以从碳排放中学到什么

夏洛特·西格曼(Charlotte Siegmann)，丹尼尔·普里维特拉(Daniel Privitera)

对人类安全的全球人工智能生态系统是一项全球公共产品：个人不能被排除在受益之外，并且一个人从中受益并不会减少其对他人的可用性。这使得如果不采取对策，市场和国家可能无法充分提供人工智能安全。在人工智能能力快速发展的时代，此类市场失灵使国际社会面临影响全世界人口的潜在严重风险。

然而，清洁空气或稳定气候等全球公共产品供应中的市场失灵是一个被广泛研究的现象。我们可以从解决这些部分类似的贡献问题（例如碳减排）的失败和成功中吸取教训，以避免人工智能的市场失灵。基于市场的国际条约可能特别适合人工智能治理，**允许促成各种双赢交易、奖励领先者并惩罚违规行为。**

有一些潜在的人工智能市场失灵可能需要在全球范围内解决。 在没有任何协议的情况下：

1. **与人工智能加速相比，各国在人工智能安全和公平性方面投入，可能会少于对地球上每个人来说最理想的投入。** 同样，国家可能会允许仓促部署不成熟且危险的人工智能技术，以超越竞争公司或国家。
2. **各国可能无法确保足够的资源流向市场无法适当评估的高度有益的人工智能技术**，例如用于卫生部门和医学研究、缓解气候变化或增强代表性不足群体的人工智能工具。
3. **各国可能无法与其他国家就国内强大人工智能模型的目标和价值观进行适当协调。** 这种不协调的规范可能会导致每个国家的境况比协调规范下更糟。

虽然国家应该负责实施人工智能治理，但国际条约对于遏制搭便车是必要的。

上述每个问题都代表了典型的市场失灵：某项行动的某些全球外部性（积极和消极）没有充分反映在国家采取这一行动所产生的成本中。解决此类市场失灵的标准解决方案包括对这些外部性进行定价的协议，从而创造更符合每个人共同利益的激励措施。就人工智能而言，这可以采取多种形式：

1. **税收驱动的差异化发展：**一项国际协议可以让各国承诺监控和跟踪人工智能部署计算。 每个国家都将承诺对人工智能的各种用例进行不同的补贴和征税（取决于其社会成本或效益），例如通过计算或数据使用或公司收入征税。此外，各国可以承诺至少将国内人工智能研究资金和国内人工智能计算（无论

是政府还是公司所有）的特定部分用于社会效益目的（由各国灵活定义）或限制人工智能训练的规模运行。类比：《京都议定书》和欧洲排放交易体系。

2. **核查和惩罚：**国家在差异发展框架下的承诺需要由国际社会执行。各种承诺，例如适当的网络安全措施、人工智能安全和公平研究的支出以及对“人工智能造福人类”的投资等都可以得到验证。遵守可以得到奖励，违反可以受到惩罚。未来，各国还可以共同建立一个所有参与国都可以访问的中心化的人工智能计算集群。每个国家的准入程度（数量）和条件（价格）可能与其遵守共同定义的规则的情况有关。类比：世贸组织制裁、巴塞尔协议III、气候变化提案、军控协议。
3. **利益分享承诺：**此外，各国可以同意在全球范围内分享国内开发的人工智能的利益。 这将减少重复建设，激励国家不要竞赛，并让更多国家从该技术潜在的巨大优势中受益。类比：关于获取遗传资源的名古屋议定书。

鉴于人工智能的快速发展，世界需要迅速协调以避免大规模风险。尽管需要解决一些挑战，但像上述这样的基于市场的机制具有三个优势：首先，它们为各国遵守承诺提供了实际的财政和经济激励。其次，在大多数情况下，可以利用国家监管能力，并且可以随着时间的推移调整机制，因为大多数时候是由政府而不是国际机构进行监管。 第三，我们可以借鉴现有全球公益机构和市场机制的成功和失败经验。基于这些原因，开始试验和实施基于市场的协议，可以帮助世界获得公平和安全的人工智能开发的好处，同时防范其风险。

夏洛特·西格曼(Charlotte Siegmann)，人工智能风险与影响中心（KIRA）创始成员、麻省理工学院博士在读。专注于人工智能政策、人工智能监管、人工智能安全技术和人工智能治理经济学的潜在全球传播。

丹尼尔·普里维特拉 (Daniel Privitera)，人工智能风险与影响中心（KIRA）创始人兼执行董事、牛津大学的哲学博士候选人。

人工智能治理如何促进全球经济增长与可持续发展？

廖璐

人工智能治理对于世界各国，特别是发展中国家的经济和社会高质量发展以及联合国可持续发展目标的落实具有潜在重要意义。

首先，人工智能治理将大力促进经济增长。人工智能技术的广泛应用能够显著提高生产效率，帮助企业降低成本，提高竞争力。这对于发展中国家和地区尤其重要，因为在这些地区，通常需要依靠经济增长来减少贫困和提高人民生活水平。通过智能制造、自动化农业和数字化服务等领域的应用，人工智能可以解放劳动力和发展生产力，为这些国家创造更多的就业机会，吸引更多的投资，并促进经济多元化。

其次，人工智能治理可以改善社会服务和基础设施。在发展中国家，为民众提供高质量的教育、医疗和基础设施通常是一项挑战。人工智能作为辅助工具，可以用于改进教育、卫生系统和城市规划，为人们提供更好的服务，提高资源利用率，并降低浪费。这有助于提高社会福祉，减少不平等，并增加人们对可持续发展的支持与信任。

此外，人工智能治理可以助力可持续发展目标的实施。联合国可持续发展目标包括消除贫困、保护地球、促进和平、改善教育等多个方面。人工智能可以用于监测和评估这些目标的进展，为决策制定提供数据支持，同时也能帮助国家更有效地管理资源，减少环境影响，从而更好地实现可持续发展。

然而，人工智能治理也伴随许多挑战，如潜在隐私问题、伦理问题和数字鸿沟等。因此，各国需要建立和完善健全的法律框架和政策，以确保人工智能的发展和应用不会加剧社会不平等，同时要广泛加强国际合作交流，增强知识和经验共享，分享最佳实践，共同应对挑战。各国政府与相关国际组织也应该牵头起草相关法规和准则，建立审查与反馈机制，要求人工智能系统在开发和使用过程中充分理解和尊重多元文化和价值观之间的差异，以确保人工智能治理有效规避偏见或歧视，更好地服务全球用户的权益。

综上所述，人工智能治理对于世界各国，尤其是发展中国家的经济和社会高质量发展以及联合国可持续发展目标的落实具有巨大潜力。通过实施合理的政策和治理措施，人工智能可以成为实现这些目标的有力工具。

廖璐，北京智源人工智能研究院项目与国际合作经理，专注于人工智能领域国际交流合作与青年科学家社区建设。

为全球大多数人的人工智能治理——以东南亚为例

莉安托涅特·蔡(Lyantoniette Chua)

东南亚被广泛认为是全球最大的数据池之一，这正是推动人工智能训练模型的资源，也是开发新的全球重要人工智能平台的基础。到 2030 年，人工智能的潜在经济贡献可能在 10 至 15 万亿美元之间。总的来说，东南亚地区也是这十年全球增长的中心。在东盟，各国根据各自的目标制定了倡议和战略政策来协调人工智能的发展。

然而，人工智能的巨大前景也伴随着固有的风险和潜在的后果。人工智能的滥用或恶意使用可能会导致灾难性后果，特别是考虑到东南亚国家多样化的社会政治背景和数字差距。

通过向联合国秘书长技术特使办公室（联合国技术特使）提交的一份文件，介绍了能力正义-基于情境的方法，作为东南亚及邻近地区人工智能区域间治理的基础：《东南亚对人工智能全球治理的声音：联合国人工智能咨询委员会的七点计划》。本文件是 The Ambit 项目关于东南亚和人工智能治理的白皮书系列第一版的删节版。

白皮书将按照以下版本发布：

- 2023 年：第一版 - 印度尼西亚、菲律宾和新加坡（第 1 组，共 3 组）
- 2024 年：第二版 - 马来西亚、泰国、越南、缅甸（第 2 组，共 3 组）
- 2025 年：第三版 - 柬埔寨、文莱、东帝汶、老挝（第 3 组，共 3 组）

在 The Ambit 全球网络向联合国技术特使提交的文件中，7 点计划包含以下关键行动：

1. 区域间设立具有包容性和多元化的高级专家，以监测人工智能对社会的影响。
2. 认可并支持即将举行的 2024 年东南亚人工智能治理路线图黑客马拉松。
3. 为全球大多数国家（全球南方）或非 G20 国家设立专门的人工智能治理机构，并进行人工智能开发-部署-治理就绪情况映射。
4. 通过各国政府的行政命令，推进负责任人工智能治理区域间机构 2023-2028 年战略。
5. 推进和建立跨境合作平台，鼓励东南亚周边国家在人工智能治理、研究和发展方面的交流和知识共享。
6. 推进人工智能发展治理能力建设行动。
7. 支持能力正义-基于情境的方法作为东南亚及周边国家人工智能跨区域治理的基础。

东南亚绝不能孤立地走过这段旅程。国际合作和知识共享对于共同应对人工智能带来的全球挑战至关重要，确保其发展具有包容性并考虑到所有国家的需求和愿望。人工智能不受国界限制，其治理应反映这一现实。从表面上看，东南亚正处于十字路口，负责任的人工智能部署可以促进经济增长、社会进步和公平发展。通过拥抱尖端人工智能生态系统发展的承诺，同时积极缓解全球挑战，该地区可以为未来铺平道路，让技术成为一股正义的力量，让全球不让任何人落后。这是一个需要政府、行业、学术界和民众共同努力的旅程，并有望为所有人带来更光明、更公平和繁荣的未来。

莉安托涅特·蔡 (Lyantoniette Chua)，华盛顿特区人工智能和数字政策中心政策组协调员和研究员、IEEE 技术和权力集中委员会副主席，创立了 The Ambit 项目，并担任其全球召集人，以及菲律宾国家创始分会委员会联合主席。

第四章　工程视角下的人工智能治理

理解模型能力是全球人工智能治理的优先事项

纳撒尼尔·沙拉丁(Nathaniel Sharadin)

随着可扩展的通用机器学习(ML)模型的进步和激增，公众对评估模型以有效管理其部署和开发的兴趣也在增加。例如，英国政府最近的白皮书呼吁建立一个“工具箱”，其中包含可以“测量、评估和表达”模型能力的技术，并指出“大型模型能力的范围”是一个开放的研究问题。拜登政府呼吁“独立”的研究人员对具有未知能力的模型进行“不受限的访问”的评估。另外，包括谷歌、OpenAI 和 Meta 在内的大模型开发商公开自愿承诺推进对“能力评估”的研究，并制定一个“多方面、专业和详细的”机制来评估（和报告）模型的能力。

这并不全是夸夸而谈和自愿协议：最近的美国法律规定建立“测试平台，包括虚拟环境”来检查机器学习系统。目前由欧洲议会以草案形式通过的《人工智能法案》则更为深入。遵守《人工智能法案》要求模型开发商提供对（特别是通用目的）模型“能力的描述”。在亚洲，中国已经发布《生成式人工智能服务管理暂行办法》，其实施可能需要开发人员（或政府工作人员）评估模型能力；东盟正在为其成员国准备指南，建议对模型进行评估。此外，在建立基础模型治理框架方面，针对“能力阈值”也出现了日益增长的国际多边共识。

因此，人们广泛认为，为了有效管控机器学习模型的开发和部署，我们需要对模型能力进行稳健、系统化的评估。因此令人惊讶的是，目前还没有系统的概念框架来决定机器学习模型实际能够做什么。尽管关于机器学习模型能力的说法无处不在。大模型开发人员(及其批评者)具体声称某些特定模型能够(或不能)做什么——例如，通过律师资格考试、有效欺骗人类、生成虚假信息、产生仇恨言论等——我们被告知模型能力可能是危险的、有害的、有益的、涌现的、自主的或新颖的。据说模型在化学、医学、编程、黑客攻击、战争等许多领域都具有能力。但尽管如此，仍然没有系统地说明机器学习模型具有什么能力，或者如何精确地决定关于模型能力的声明。对能力的讨论根本就没有被详细审视。

这是我们对这项新技术共同理解中的一个严重差距，这个差距也是有效治理的重要障碍。例如，各大参与者之间无法就限制模型能力范围进行合作，除非其事先就什么算作模型具有某种能力的证据达成共识。在建立国内国际对功能强大的机器学习系

统开发与部署进行监管这项重要工作的先决条件是，我们需要一个系统的框架来评估关于模型能力的各种声明。开发这样一个框架应成为治理的优先事项。

纳撒尼尔·沙拉丁(Nathaniel Sharadin)，香港大学哲学助理教授，人工智能安全中心(CAIS)的研究员。
AI 在研项目包括理解和评估大型前沿机器学习模型的能力与人工成就的本质、价值与重要性。

标准化视角下的人工智能安全治理全球协作与敏捷更新

马骋昊、高万琪、范思雨

在智能不断涌现的今天，技术势能与风险隐患似乎总是相伴而生。一方面，人工智能不断在下游任务上取得进展与突破，刷新并拔高着社会公众的认知与期待。另一方面，其所带来的隐私泄漏、偏见歧视、责权归属、技术滥用等伦理与安全问题也同样层出不穷。

在此现状下，标准化将会是一种切实可行的思考维度与治理路径。

从国际上看，当前在 ISO、IEC、ITU 以及 IEEE 等标准组织中已经开展了不同维度的 AI 安全治理标准化工作，在名词概念、基础体系与风险管理框架等方面已经形成了一定程度上的国际共识。其中的典型代表如 ISO/IEC JTC 1/SC 42 人工智能分技术委员会设立的可信赖工作组（WG 3）以及 IEC/SEG 10 自主与人工智能应用伦理系统评估组等。

中国于 2018 年成立国家人工智能标准化总体组，设立人工智能与社会伦理道德标准化研究组。此后，全国信标委人工智能分委会（SAC/TC 28/SC 42）下设成立可信赖研究组，全国信安标委（SAC/TC 260）也同步开展国内信息安全相关标准研制。

当前，中国已经形成了一套人工智能伦理规范和相关标准，围绕 AI 伦理治理相关的技术研究及应用进行整体性布局。其中，人工智能管理体系、风险管理能力评估、可信赖等方面的重点标准已经进入起草阶段。这些重点标准在 AI 安全及伦理治理领域具有清晰的边界范围，能够将宏大、抽象的伦理原则应用于实际技术研究中，引导产业合规发展。

然而，不同国家地区和行业领域中对 AI 安全与伦理问题有不同要求，当前大部分标准化工作主要以指南、框架、原则和准则的形式出现，而更加具体、可落地应用的技术标准仍处于探索和前期研究的阶段，尚待更多研究者的加入与协作。

为此，我们以标准化工作为切入点，浅述几点倡议，以期抛砖引玉、共同讨论：

- 一是倡议在准则层面形成针对 AI 技术迭代的敏捷更新机制。构建起基于数据、案例和实验的全球性研究者社区，实时更新与细化 AI 安全治理准则的内涵与外延。
- 二是倡议在教育、医疗等敏感应用领域研制专项标准，促进宏观准则的落地应用。广泛开展社会实验与跨学科研讨，建设各行业应用 AI 的专项治理数据集与治理技术。

- 三是倡议在深圳建设国际人工智能对齐与治理创新示范区。呼吁全球 AI 企业及科研院所共同参与，在一定范围内共同验证相关的对齐方法、标准规范、治理工具、数据共享机制等方面内容的科学性及可操作性。

  以上，我们共同畅想着一个更加美好、安全和对齐的通用人工智能未来。

马骋昊，中国电子技术标准化研究院标准化工程师，IEEE/C/AISC/LSDLM 主席。曾编制多项项人工智能国家标准。

高万琪，中国电子技术标准化研究院华南分院人工智能产业研究员，曾参与完成多项人工智能产业政策的调研与编写。

范思雨，中国电子技术标准化研究院华南分院人工智能产业研究员，电子信息产品标准化国家工程研究中心人工智能科普课程讲师。曾参与深圳市多项人工智能及电子信息技术领域政策文件编写。

人工智能治理——一场再建巴比塔的革命

王俊，娜迪娅

2022 年，生成式 AI 发展为人工智能发展注入一针强心剂，ChatGPT 的横空出世，被视为通用人工智能的起点和强人工智能的拐点，引发新一轮人工智能革命。人工智能发展似乎找到了自己的主流叙事。

不过，技术创新的同时也带来了治理难题，我们面对的不是近在咫尺的当下，而是想象触达不到的未来。对于颠覆性的人工智能技术，需要以全人类的视角提出治理思路，各个国家、地区以及相关企业、专家学者、社会公众等，应打破隔离状态，解决规则的分野，共同探讨我们要面对的处境。

基于南财合规科技研究院此前对人工智能持续的观察、报道，对目前全球人工智能发展面临的几个突出问题进行总结：

一、全球人工智能治理话语竞赛下，规则互操作性问题逐渐显现

以 ChatGPT 为代表的人工智能应用掀起产业及技术发展高潮，国内外的科技巨头争相加大技术与资本投入的同时，也引发了监管担忧。

可以看到，各国、地区已开始采取措施加强监管，欧盟《人工智能法案》已经进入最后谈判阶段，美国宣布一系列围绕美国人工智能使用和发展的新举措。

中国于 7 月份发布的《生成式人工智能服务管理暂行办法》成为全球首份针对生成式人工智能的政策文件。

一方面是人工智能发展的速度竞赛，另一方面是规则话语的制订竞赛，不同的区域，有着不同的驱动力，各国、地区基于自身的人工智能发展现状做出监管政策，这也带来了规则衔接的问题，监管标准不统一造成的模型适用问题在实践中已经显现。

伴随着更多国家与地区立法、监管政策的落地，规则如若呈现碎片化和分裂，不利于人工智能的长远发展，亦不利于全球共识的达成，全球应该加强规则协调，毕竟我们面对的新挑战是人机挑战。

二、遥远的隐忧——价值对齐问题

伴随着人工智能的发展，安全与对齐的问题随之而来。从 Sam Altman 在 OpenAI 如火如荼之际主动呼吁监管，强调价值观对齐，也可以窥见这一问题的重要性。

尤其是当一项技术在各个领域里大规模应用时，过去分散性的歧视、不安全，很可能会变得更为集中。

但价值观的对齐绝非易事，即便譬如在“有益、真诚和无害”等原则性一致，细节之处可能也谬以千里。此外，笃定的价值判断不能适用人工智能，因为没有谁拥有给价值观下判断的权力。

往赛博空间一步步走深，所有现实都可以被映入人工智能系统，代码管理模式背后的权力结构，隐秘而深刻。人工智能对齐更像是一场与时间赛跑的比赛，需要找到对人工智能可控的解决方案。

三、人工智能发展带来的是数字红利还是更严重的分化？

人工智能发展将会给社会公众带来什么，更高的工作效率，更便利的生活，还是更隐秘的剥削？

从三个维度来探讨这一问题。

1. 企业与资本是否更为集中？尽管目前呈现“千模大战”，但主要的控制者仍掌握在微软、谷歌、Two Sigma、OpenAI 等手中。平台经济时代的“数字资本帝国”阴霾在人工智能时代依旧存在，并且更为突出。

2. 不同国家之间的差距进一步拉大的未来。由于人工智能系统往往由发达国家公司构建，发达国家主要掌握产业链的核心，但在这个产业链中，数据加工等产业主要由发展中国家完成，国外不少研究指出这会使得进一步剥削发展中国家人力资本和资源，加剧贫富分化。

3. 个体与系统对抗更为艰难。Code is law，如若规则偏离人类中心，个体如何反馈与对抗，此外，处于人工智能系统应用中，如何分享红利而非恐惧？

对于上述问题，我们认为可以采取以下举措：

一、加强对话交流机制

人工智能系统将嵌入人类社会，各国、地区之间应该加强互动，洞察不同监管手段的成效与影响，求同存异，寻求共识。

非官方的、民间的组织、智库等应该发挥自身作用，积极推进全球的交流沟通。

二、更为包容、多样、透明的通用人工智能安全规则

价值观对齐的问题本质是一个古老的人类问题，没有正确答案。“文化是相对的，道德是绝对的.....在多文明的世界里，建设性的道路是弃绝普世主义，接受多样性和寻求共同性。”

不同行业、不同领域、不同组织应该参与讨论，审慎考虑正在发生以及即将面临的境况。在训练数据集过程中如何标注，训练对隐秘的偏见问题及时反馈，平衡不同国家和用户的价值偏好.....建立更具公平性和包容性的系统。

就在 5 月，国内首个大语言模型治理开源中文数据集 100PoisonMpts 发布，十多位知名专家学者成为了首批“给 AI 的 100 瓶毒药”的标注工程师。标注人各提出 100 个诱导偏见、歧视回答的刁钻问题，并对大模型的回答进行标注，完成与 AI 从“投毒”和“解毒”的攻防。这是非常有益的尝试，类似经验也应多多分享交流。

三、促进高质量数据集的开源开放

高质量的训练数据集，一定程度上可以减轻人工智能歧视问题，促进多元化。

目前，全球范围内高质量训练数据集面临衰竭问题，并且存在中文语料数据不足的现实。

粤港澳大湾区具有海量数据规模和丰富应用场景优势，数据要素市场不断扩大。应该充分发挥自身优势，充分挖掘数据价值，在数据合规基础之上，进一步促进公共数据等开放，推进多模态公共数据集建设，打造高质量中文语料数据。

四、搭建数据上中下游的产业链

人工智能产业链分散庞杂，粤港澳大湾区抓住数据要素市场建设的契机，数据服务、数据加工等产业具有广阔前景，在加强国际合作的同时，打造自主可控的数据产业链。人工智能训练数据、自动化决策等或迎来模型训练素材的快速发展期。应以数据要素市场建设为契机，有效开展各类数据业务，面向个人开展信息托管，在合规前提下盘活个人信息的安全流通，使个人也能分享人工智能的红利。同时，围绕模型训练的上下游产业链，如数据清洗、数据交易等也将随之进一步快速发展，可促进数据要素市场的整体发展。

王俊，南方财经合规科技研究院首席研究员。专注科技与法律交叉地带，研究反垄断与反不正当竞争、个人信息保护、互联互通等前沿议题。

娜迪娅，南方财经合规科技研究院副院长，海丝研究院副院长。专注个人信息保护、数据安全领域报道研究与技术测评，关注数据要素市场、数据交易、数据合规等数字经济与海上丝绸之路相关课题研究。

新加坡治理生成式人工智能的方法和实践

丹尼丝·王(Denise Wong)

生成式人工智能(GAI)为不同领域和应用带来了巨大的机遇，从增强用户体验到提高生产力。它也带来了人们担心的风险：

1. **错误和幻觉。**生成式人工智能模型可能会犯错误，它的“幻觉”可能具有欺骗性的说服力或仿真性。
2. **隐私和保密。**由于生成式人工智能模型倾向于记忆训练数据，对手可能通过对模型的查询来重建敏感数据。
3. **虚假信息、有害内容和网络威胁。**由于大规模生成的令人信服但具有误导性的文字、图像和视频，识别假内容(如假新闻)变得越来越困难。生成式人工智能也可能传播有毒内容。
4. **版权挑战。**生成式人工智能模型可能会在包含版权材料的数据集上进行训练，从而创造未经授权的衍生作品。
5. **内置偏见。**AI 模型可以放大训练数据集中的偏见，从而可能导致下游应用中的偏见输出。
6. **价值观和对齐。**AI 系统可能与人类价值观和目标不一致，从而导致潜在的危险后果。

为了建立对生成式人工智能的信任，确保其得到负责任的开发和使用，所有利益相关者，包括政府、行业、学术界和公民团体都有责任发挥作用。

目前，在治理生成式人工智能和帮助行业负责任地部署生成式人工智能的原则、框架、标准和工具方面，国际间缺乏共识和统一。

为了促进有关负责任生成式人工智能的国际讨论，新加坡发布了一份关于《生成式人工智能：对信任和治理的影响》的讨论文件，建议采用实用、基于风险和多利益相关方的方法来治理生成式人工智能：

1. **模型开发和部署：**模型开发者应对于他们的模型是如何开发和测试的保持透明。政策制定者可以通过促进标准化评估指标和一系列工具和能力的发展来提供支持。
2. **保证和评估：**第三方评估和保证对提高可信和信任至关重要。吸引开源专业知识的参与对培育充满活力的 AI 系统第三方测试生态系统至关重要。随着各国

寻求确保 AI 模型与其独特价值观和 AI 治理原则保持对齐，以及公司在特定数据集上训练 AI 模型，行业合作和定制化的测试基准将很重要。

3. **安全和对齐研究：**政策制定者需要投资来加速安全和对齐研究，以实现可解释性、可控性和鲁棒性。这项工作还应培养亚洲和世界其他地区的知识中心，以补充美国和欧盟正在进行的努力。

新加坡还旨在贡献工具来帮助公司测试和评估生成式人工智能。新加坡成立了 AI 验证基金会（AI Verify Foundation）来开源 AI 验证的最小可行产品，这是一个 AI 治理测试框架和软件工具包，目前用于测试歧视性 AI。AI 验证基金会寻求利用全球开源社区的力量来扩展 AI 验证，使其具备评估生成式人工智能应用的能力。

丹尼丝·王 (Denise Wong)，新加坡资讯媒体发展局（IMDA）数据创新与保护组的助理首席执行官、战略政策与运营集群总监，兼任个人资料保护委员会（PDPC）副专员。专注于制定前瞻人工智能和数据治理政策、促进行业采用人工智能和数据分析等。

通过人工智能技术民主化实现人工智能对齐

伊丽莎白·西格(Elizabeth Seger)

通过开源使人工智能开发的民主化

尽管人工智能的民主化是一个多层次的概念，但该术语经常用来指人工智能开发的民主化。人工智能开发民主化是为了帮助广泛的人参与人工智能开发过程 (Seger, Ovadya 等, 2023)。 当拥有不同生活经历、地理、经济和文化背景的人能够参与并支持人工智能开发过程时，开发出来的人工智能产品更有可能很好地满足不同用户的需求，而不是将开发集中在硅谷的几家领先实验室中。

在过去的 30 年里，为满足不同的人类兴趣和需求，开源社区在软件开发民主化以及现在的人工智能开发民主化方面发挥了重要作用。开源人工智能开发是指创建人工智能模型和工具的协作过程，这些模型和工具可供任何人免费查看、使用、研究、修改和分发，从而促进透明度和社区驱动的创新。 例如，大语言模型 BLOOM 是由 1000 多名人工智能志愿者开发人员联合开发的，历时一年时间，支持 46 种语言(BLOOM, 2022)。

除了开发和共享开源人工智能模型之外，开源社区还通过教育、宣传和共享模型训练成本等积极努力来进一步实现人工智能开发的民主化。

开源开发对于人工智能开发民主化的好处是显著的。 然而，开源也带来了必须平衡考虑的风险(Seger, Dreksler 等, 2023)。特别是随着模型能力的增强，恶意使用的后果变得更加严重(Anderljung & Hazell, 2023; Shevlane 等, 2023)。 通过访问模型代码和权重，恶意行为者可以破解防止滥用的保护措施，并可能通过微调引入新的危险能力 (Qi 等, 2023; Rando 等, 2022)。开源决策也是不可逆转的；如果出现危害，就无法收回。因此，应慎重考虑开源决策。

人工智能治理的民主化

在人工智能开发的民主化风险较高的情况下，人工智能治理的民主化可以作为第二种机制，有助于将人工智能开发和部署决策与更广泛的公共利益和价值观相统一。

人工智能治理的民主化是将人工智能决策的影响力分配给更广泛的利益相关者社区和受影响人群(Seger, Ovadya 等, 2023)。 人工智能治理决策涉及平衡人工智能相关的风险和收益，以确定人工智能如何以及由谁来开发、分发、使用和监管。

有多种选项可供探索，以实现有关人工智能决策的民主化：

公众参与和审议。一组可能性涉及通过参与性或审议性民主进程直接征求公众意见。这些流程可能会利用 Pol.is 等在线工具（可能在人工智能支持下）来征求公众意见并将其综合到复杂的规范决策中(Ovadya, 2023)。 OpenAI 最近启动了一项“人工智能的民主输入”资助计划，以尝试建立民主流程来决定人工智能系统应遵循哪些规则(Zaremba 等, 2023)。 集体智慧项目(CIP)也在尝试“对齐组件”，以帮助识别反映人工智能行为的集体价值观，此外还可以解决其他复杂的价值观加载问题，例如模型发布决策和描述可接受的风险阈值（集体智慧项目, 2023）。

体制结构。 大型实验室还可以引入本质上更加民主的组织结构，例如通过实施民主选举或随机抽签选出的监督委员会。他们还可能成立公益公司，以为最大化公共利益而做出决策提供更明确的法律地位，即使这样做与股东利益相冲突。

民主知情监管。 另一种选择是支持通过开发者、开源社区、学术界和民间团体参与的审议过程而制定的监管，以反映不同的利益相关者的利益。

结论

很难确定人工智能行为和开发决策应该响应的确切价值观。相反，我们可能会寻求采用公平的流程来收集和整合不同利益相关者的意见。为此，一种选择是通过开源开发让更多的人直接参与人工智能开发流程。 然而，开源也伴随着风险。在风险较高的情况下，可以通过让各种利益相关者参与民主决策来推进人工智能对齐，为后续决策提供信息。

伊丽莎白·西格 (Elizabeth Seger)，牛津大学人工智能治理中心（GovAI）研究员、剑桥大学 AI：未来与责任项目 (AI:FAR) 研究员。曾任未来智能莱弗休尔姆中心（LCFI）的研究助理。专注于研究人工智能对技术先进社会中信息生产、传播和内化的实际和潜在影响。

学习机器的工程智慧

布雷特·卡兰(Brett Karlan)，科林·艾伦(Colin Allen)

深度神经网络的最新成功使一些人相信人工通用智能时代并不遥远。不过，这样的未来是否真的迫在眉睫，还有待猜测，我们对此持保留态度。

最先进的深度神经网络在处理大量数据和检测模式方面表现出色。然而，它们在响应上却出人意料地脆弱。例如，尝试基于用户输入生成看似合理文本的语言模型，当要求的文本量变长时，其输出的质量和可理解性明显下降，且无法监控自身的矛盾。最先进深度神经网络的另一个陷阱是其不透明性。往往无法知道神经网络是如何使用信息做出决策的，它如何处理这些信息，甚至这些信息存储在网络中的位置。这种不透明性导致用户高估了 AI 的能力。另一个原因是用户没有严格测试这些系统的极限。

我们如何做出更好的算法辅助决策？在人机互动中，我们的目标应该是什么？我们认为，实用智慧的概念对于组织、理解和改进人机互动特别有用。实用智慧指的是一套知识、理解和技能，旨在探寻领域真理并在该领域做出更好的决策。我们对实用智慧的理解强调两个重要组成部分：1）形成合理策略以应对自身局限和技术限制所需的元认知意识；2）做出良好判断所需的广泛理解、知识和技能。

在人机互动中发展实用智慧，特别是关注策略选择和特定情境的元认知推理，代表了一种既概念化脆弱性问题又最小化巨大错误可能性的方式。最近的几项研究表明，明智的推理能力受到推理者情境意识的影响：当受试者从自己的个人投资中解脱出来，并从更第三人称的视角考虑自己的能力和局限时，他们往往能做出更明智的推理。这有助于更好地元认知地意识到自己对情境理解的限制。

人工智能的不透明性已被长时间讨论和理论化。使用人机互动中的实用智慧框架，我们可以看到可解释 AI 可能有价值的另一个原因：因为它有助于实用智慧的发展。如果决策者能够获取更多关于深度神经网络是如何得出其输出的信息，那么她在面对相同输出但没有解释时，往往会做出更好的决策。

虽然 AI 的个体（或群体）用户是决策的重要分析焦点，但假设实用智慧框架不能扩展到人机互动的其他方面将是一个错误。我们提议的一个优势是，实用智慧的概念有助于理解深度神经网络和其他 AI 技术在重要决策领域的生产和使用的所有层面。在人工智能的创造、设计和实施过程的所有阶段发展实用智慧，反过来使我们在决策中运用实用智慧的过程显著简化。

我们对实用智慧的概念化提供了一个强大的框架，用于理解人机互动。这个框架支持更好地解释这个领域的成功是什么样的，以及如何避免未来的失败。尽管其他方

法和框架可能能提出与我们在这里提出的类似建议，但我们的框架提供了一套统一而连贯的能力，解释了为什么以及如何做出这些建议。进一步的概念精炼将需要心理学家、行为科学家、工程师和其他利益相关者的输入。

布雷特·卡兰(Brett Karlan)，普渡大学哲学系助理教授，专注于科学哲学（特别是认知科学和人工智能）和规范哲学（特别是认识论、伦理学和行动哲学）的交叉点。

科林·艾伦(Colin Allen)，加州大学圣塔芭芭拉分校哲学系杰出教授，专注于认知科学哲学、动物心智、认知进化、机器道德和计算人文。

多元、开放、互动：生成式人工智能模型训练所需的原则

刘纪璐(JeeLoo Liu)

生成性人工智能模型是在由书籍、文章、在线资源和其他可用数据组成的广泛数据集上训练的。大语言模型训练的优势在于其能够训练 AI 直接从自然语言处理信息，并生成语法和风格上出色的内容。迄今为止最成功的模型 ChatGPT-4，已经展示了在处理广泛主题信息方面的极速性能。尽管它被批评为制造了许多事实错误，但这些错误在未来一代生成性人工智能中应该能够避免。

然而，这种机器学习方法也存在紧迫的问题，正如 ChatGPT 等模型的表现所证明的那样：

1. 训练基于现有数据，因此结果不反映任何新发展的情况或新的输入。正如休谟所指出的：没有保证未来会像过去一样。现有方法有忽视未来例子的风险，并延续当前社会中存在的偏见、不公正和错误歧视。
2. 书籍和文章中的现有文本只包括那些电子可获取的项目。因此，结果将始终排除古老的非数字化文本和来自不发达国家或边缘化文化的非数字化文本。我们因此忽略了非数字化文本中代表的宝贵观点，忽视了对全面人类知识库至关重要的输入。
3. 数据策划通常在 OpenAI、谷歌和特斯拉等公司内部和私下进行，对公众没有透明度和可解释性。如果没有制衡，产品可能不会公正代表一般福祉或公众情绪。
4. 虽然“人在回路”方法有助于消除有害数据、煽动性语言、淫秽内容、偏见和其他有害内容，但巨大的数据量通常迫使人们外包到不发达国家的廉价劳动力市场。这种做法引起了人们对劳动力剥削和数据管理不合格的担忧。没有由哲学家、伦理学家、道德领袖、教育家以及特定社会中的普通公众等人类价值专家来进行的进一步的制裁和指导。

为了遏制这些问题，我建议：

1. 我们需要建立一个庞大的数据库，其中包含跨越历史、文化和语言的道德行为、哲学见解和道德思考。这将帮助我们构建一个伦理模型，以叠加到当前的大语言模型上。该模型不需要呈现单一的价值结构或道德判断的共识；有道德的人会有道德地彼此不同意。结果必须是多元化的，以代表全球不同的文化和价值观。

2. 我们还需要有一个开放平台，以征询公众对各种问题的意见，并将数据自动添加到精炼训练的策划数据中。这是“人在回路机器学习”的真正形式——将所有人带入回路，而不仅仅是特定群体，如被雇佣的 MTURK 众包工人。
3. 最后，数据收集和机器学习必须保持开放性，适应新数据和不断演变的场景。学习和训练必须与每个附加输入互动和主动。人类从对他人的建议保持开放中学习；机器也必须这样做。

刘纪璐(JeeLoo Liu)，加州州立大学富勒顿分校哲学系教授，研究重点是对中国哲学的重构分析，包括中国的玄学、儒家道德心理学以及新儒家伦理，以及机器人伦。

第五章　企业视角下的人工智能治理

——种负责任地扩展人工智能模型的框架

迈克尔·塞利托(Michael Sellitto)

随着前沿人工智能模型的能力变得越来越强大，它们将创造重大的经济和社会价值，但也将带来越来越严重的风险，需要加以管理。为了管理这些风险，Anthropic 设计并采用了负责任的扩展策略(RSP)。

我们的 RSP 专注于灾难性风险——人工智能模型直接导致大规模破坏的风险。此类风险可能来自于故意滥用模型（例如恐怖分子利用模型来制造生物武器），也可能来自于模型以与设计者意图相违背的方式自主行动而造成破坏。虽然人工智能呈现了一系列必须解决的风险，但我们的 RSP 旨在应对这一风险范围中更极端的风险。

我们计划的核心是人工智能安全级别(ASL)的概念，它大致模仿了美国政府处理危险生物材料的生物安全级别(BSL)标准。我们定义了一系列人工智能能力阈值，这些阈值代表着不断增加的潜在风险，因此每个 ASL 都需要比前一个更严格的安全、保障和操作措施。

更高的 ASL 模型也可能与日益强大的有益应用程序相关联，因此我们的目标不是禁止这些模型的开发，而是通过适当的预防措施安全地启用它们。

因此，RSP 带来了识别和管理灾难性风险的具体和实证方法。它的目的是让我们的安全研究和技术的社会效益应用能够继续发展和规模化，同时实施严格的流程来衡量和减轻风险。如果无法缓解重大风险，我们将暂停相关模型的进一步扩展，避免部署它，或将其从部署中删除，直到我们可以确保它足够安全，可以继续。

通过在扩展速度超过安全性时暂停开发和部署，我们就有动力解决必要的安全问题。如果作为前沿实验室的标准并得到政府的支持，RSP 可能会产生一种“争先恐后”的动态，其中竞争激励直接化为解决安全问题。

值得注意的是，RSP 还具有灵活性和适应性。RSP 为当前(ASL-2)和近期(ASL-3)人工智能系统指定了具体的安全承诺，涵盖安全、训练监督、红队测试、模型评估和负责任的部署措施。它承诺 Anthropic 在达到更高的 ASL（从 ASL-4 开始）之前迭代地定义它们，确保随着时间的推移学习新的经验信息，安全协议与能力一起发展。因此，负责任扩展策略应该随着时间的推移而发展，并保持其长期相关性。

完整政策可在以下网址获取：

https://www.anthropic.com/index/anthropics-responsible-scaling-policy

迈克尔·塞利托(Michael Sellitto)，人工智能安全和研究公司 Anthropic 的地缘政治和安全政策主管，兼任技术与国家安全项目高级研究员。专注于研究人工智能技术对国家竞争力、国际关系和国际安全相关问题的影响。

以价值对齐塑造健康可持续的大模型发展生态

司晓、曹建峰

随着生成式 AI 大模型的快速发展，其自主性、通用性和易用性快速提升，大模型将深入各行各业，带来显著的经济和社会效益。与此同时，大模型面临着幻觉、歧视、滥用、涌现风险等伦理问题。因此，如何让大模型的能力和行为跟人类的价值、真实意图和伦理原则相一致，确保人类与人工智能协作过程中的安全与信任，这一“AI 价值对齐”问题变得至关重要。为了让大模型更加安全、可靠、实用，就需要尽可能地防止模型的有害输出或滥用行为，这是当前大模型价值对齐的一项核心任务。业界和研究机构一直在探索技术和治理措施。

当前的一个核心技术路径是，通过强化学习让 AI 大模型能够理解人类的价值、伦理原则等，在这方面，人类反馈的强化学习（RLHF）被证明是一个有效的方法。RLHF 包括初始模型训练、收集人类反馈、强化学习、迭代过程等几个步骤，其核心思路是要求人类训练员对模型输出内容的适当性进行评估，并基于收集的人类反馈为强化学习构建奖励信号，以实现对模型性能的改进优化。从实践来看，RLHF 在改进模型性能、提高模型的适应性、减少模型的偏见、增强模型的安全性等方面具有显著优势，包括减少模型在未来生产有害内容的可能性。然而，RLHF 存在可拓展性差、受限于训练员的主观偏好、长期价值对齐难以保证等问题。因此，研究人员开始探索如何从相对低效的“人类监督”转向更为高效的“规模化监督”，其中 AI 监督是一个越来越受到重视的思路。AI 监督的一个核心理念是让 AI 模型辅助人类训练员或者自主地对大模型的输出进行评估，目前在 AI 实践中存在一些有益的探索。

此外，还可以通过其他方式和治理措施来保障大模型价值对齐的实现。一是对训练数据进行干预，包括：对训练数据进行记录以识别是否存在代表性或多样化不足的问题；对训练数据进行人工或自动化筛选、检测以识别、消除有害偏见；构建价值对齐的专门数据集。二是构建可解释、可理解的大模型，为了实现 AI 对齐，人们需要更好地理解大模型如何作出决策。三是对抗测试或者说红队测试，红队测试员通过向模型提出试探性的或者危险性的问题以测试模型的反应，发现模型在不准确信息、有害内容、虚假信息、歧视、语言偏见、安全风险等方面的问题，从而帮助改进模型。

总之，AI 对齐之所以重要，不仅在于它是当前大模型的必由之路，更在于它关乎超级智能的未来。在这场人工智能与时间的赛跑中，需要各方携手，推动更广泛的跨学科参与和协作，以确保大模型健康可持续发展，未来更强大的人工智能持续造福人类和人类社会。

司晓，腾讯研究院院长、腾讯集团副总裁，兼任国家网络版权产业研究基地副主任，深圳市版权协会会长，斯坦福大学访问学者。

曹建峰，腾讯研究院高级研究员、华东政法大学数字法治研究院特聘研究员、上海政法学院客座教授、中国伦理学会科技伦理专委会理事、广东省法学会信息通信法学研究会理事。专注于互联网前沿科技与数字经济相关的政策、法律、伦理与社会研究。

不要让深黑盒人工智能锁定了我们文明进化的路径

韦韬

GPT 及类似的大规模预训练语言模型的不断发展，正在引领人工智能技术进入一个新的时代。这些模型在很多自然语言处理任务上已经展现出超越人类的能力，但也存在一些严峻问题。目前它们最大的问题是缺乏认知对齐、缺乏原则性与推理黑盒化。比如，由于不能有效处置“不知道”，它们可能产生严重的幻觉问题；同时，它们也不能有效遵循“隐私保护”，可能导致个人信息泄露；此外，在推理过程中由于各种原因导致的错误决策和建议都难以解构和验证。

在这种风险下，由于 AI 的无限精力，这些问题导致的错误决策可能会被大规模快速执行，造成难以预计的严重后果。就像一个不成熟的孩子，却有着毁天灭地的精力和能量。现阶段，和很多专家担心的相反，我们认为人类首先要担心的问题不是机器智能因为“邪恶”来毁灭人类，而是不要因为人类自己的愚蠢导致的机器智能的愚蠢来毁灭人类。

我们认为，AI 的认知一致性（包括内在一致性和外在一致性）、可进行专业验证的推理链自解构，以及与其他智能体交互激发的认知迭代演进（持续学习），将是迈向专业 AGI 的关键。随着 AI 技术的快速演进，AI 技术在各个专业领域会快速扩大应用规模，但依然不能用黑盒化技术来替代对专业知识、范式和规则的探索、积累和迭代。只有这样，AI 才能更好地为人类服务，并为社会带来更大的益处。

韦韬，蚂蚁集团副总裁兼首席技术安全官，北京大学客座教授，浙江省科协委员，著名安全学术论坛 InForSec 联合创始人。专注于使各种复杂系统变得更加安全可靠，多项成果帮助各主流操作系统提升安全性，领导和推动了多项著名开源安全软件的研发。

英特尔负责任的人工智能应用探索

邹宁、王海宁

在过去的几年里，生成式人工智能变得更加强大，因此更有能力制造一些难以察觉的、逼真的错误。尽管一些人对生成式人工智能威胁就业的可能性表示担忧，但也有更多的机会负责任地使用生成式人工智能来提高人们的效率和创造力。我们认为，人工智能不仅应当用于防止伤害，还应当用于改善人们的生活。

作为英特尔负责任的人工智能工作的一部分，公司将“FakeCatcher”产品化，这是一种可以以 96%的准确率检测假视频的技术。FakeCatcher 是英特尔与纽约州立大学宾厄姆顿分校合作设计的深度伪造探测器，英特尔将它应用在深度伪造实时检测平台上。它使用英特尔的硬件和软件，在服务器上运行，并通过基于 web 的平台进行接口。在软件方面，一组专业工具形成了优化的 FakeCatcher 架构。团队使用 OpenVino™ 运行人脸和地标检测算法的人工智能模型。计算机视觉模块使用“Intel® Integrated Performance Primitives”（一个多线程软件库）和 OpenCV（一个用于处理实时图像和视频的工具包）进行了优化，推理模块使用 Intel® Deep Learning Boost 和 Intel® Advanced Vector Extensions 512 进行了优化；媒体模块也使用 Intel® Advanced Vector Extensions 512 进行了优化。团队还依靠 Open Visual Cloud 项目为 Intel®Xeon®可扩展处理器系列提供集成软件堆栈。在硬件方面，深度伪造实时检测平台可以在第三代 Intel®Xeon®可扩展处理器上同时运行多达 72 个不同的检测流。大多数基于深度学习的检测器都会查看原始数据，试图发现不真实的迹象，并识别视频的错误。相比之下，FakeCatcher 通过评估视频像素中那些真实人类才会有的微妙的“血流”，在视频中寻找真实的线索。当我们的心脏跳动输送血液时，我们的静脉就会变色。算法将这些从面部各处收集的血流信号转换为时空图。然后，使用深度学习，我们可以立即检测一个视频是真是假。

生成式人工智能还可以用于使 3D 体验更加逼真。例如，英特尔的 CARLA 是一款开源的城市自动驾驶仿真模拟器，旨在支持自动驾驶系统的开发、培训和验证。使用生成人工智能，驾驶员周围的场景将看起来更加逼真和自然，促进了自动驾驶技术的发展。

生成式人工智能还可用于改善残疾人的生活。英特尔有一个语音合成项目，旨在让失声的人能够重新说话。这项技术被用于英特尔与戴尔技术公司、罗尔斯-罗伊斯公司和运动神经元疾病（MND）协会合作的“我永远是我”数字故事书项目。该互动网站允许任何被诊断为 MND 或任何预计会影响其说话能力的疾病的人记录自己的声音，以便在辅助语音设备上使用。

我们明白，如果不了解这些系统的工作过程，我们就无法信任生成式人工智能的结果。英特尔长期以来一直重视部署先进技术相关的道德和社会影响。人工智能应用尤其如此，因为我们仍然致力于利用最佳方法、原则和工具，确保在产品周期的每一步都采取负责任的做法。随着生成式人工智能的发展，将人类保持在这项工作的中心是至关重要的。负责任的人工智能始于系统的设计和开发，而不是部署后的整顿。

邹宁，英特尔中国区技术政策和标准高级总监。

王海宁，英特尔中国区人工智能技术政策和标准总监。

致谢

在此，我们特别感谢各位专家分享他们的宝贵观点，对本报告的质量和深度产生了显著影响。我们对龚克和段伟文在报告的规划和编写过程中提供的建议和专业指导表示衷心的感谢。

免责声明

本文表达的观点仅代表作者个人观点，不一定反映机构的立场。报告主编及所在单位对内容的准确性、完整性或可靠性作尽可能的追求，但无法做任何保证和承诺，不对任何因此报告导致的直接或间接损失或损害承担责任。

联系方式

世界工程组织联合会创新技术专委会：wfeo-ceit@wfeo.org
深圳市科学技术协会：kxbgs@shenzhen.gov.cn

贡献情况

主办单位：WFEO-CEIT、深圳市科学技术协会
报告主编：安远AI、WFEO-CEIT 大数据和人工智能工作组