2024 年 9 月

治理人工智能，助力造福人类

United Nations | AI Advisory Body

治理人工智能，助力造福人类

最后报告

Advanced Translation Copy

United Nations | AI Advisory Body

关于人工智能高级别咨询机构

2020年，在联合国秘书长的数字合作路线图（A/74/821）下首次提出了建立多利益攸关方人工智能高级别咨询机构的提议，机构于2023年10月成立，负责分析人工智能国际治理问题，并就此提出建议。

咨询机构的成员以个人身份而非作为各自组织的代表参加。本报告代表多数成员的共识；并不期望任何一位成员认可本文件中的所有观点。成员申明，他们就报告中的结论和建议达成了广泛而非单方面的一致意见。本报告所用措辞并不意味着成员所在组织在机构层面予以认可。

目录

执行摘要

一 人工智能正在变革我们的世界。这套技术具有巨大的向善潜力，从开辟新的科学探索领域和优化能源网络，到改善公共卫生和农业，乃至促进在实现可持续发展目标方面取得更广泛的进展。

二 但若不予监管，人工智能的机遇可能就无法显现或得到公平分配。由于数字鸿沟不断扩大，可能只有少数国家、企业和个人能从人工智能中获益。漏用人工智能，即由于缺乏信任或缺失推动因素(如能力差距和无效治理)而未能利用和分享人工智能的相关惠益，则可能会限制机遇空间。

三 人工智能还带来了其他风险。人工智能的偏见和监视伴随着新的担忧，例如大型语言模型的虚构（或"幻觉"）、人工智能增强的虚假信息生成和传播、对和平与安全的风险以及人工智能系统在气候危机时期的能耗。

四 快速、不透明和自主的人工智能系统对传统监管体系构成挑战，与此同时，这些日益强大的系统可能会颠覆工作领域。自主武器和人工智能在公共安全领域的使用引发了严重的法律、安全和人道主义问题。

五 目前，人工智能全球治理存在缺陷。尽管就伦理道德和原则开展了大量讨论，但各种零散的规范和机构仍处于起步阶段，而且充满漏洞。问责制的缺失常常引人注目，包括在部署影响他人且不可解释的人工智能系统方面。合规往往基于自愿；言行前后矛盾。

六 正如我们在中期报告[1]中指出的，人工智能治理不仅对于应对挑战和风险极为关键，而且对于确保在利用人工智能潜力的同时确保不让任何一个人掉队至关重要。

1. 全球治理的必要性

七 全球治理的必要性尤其无可辩驳。从关键矿物到训练数据，人工智能的各种原材料都源自世界各地。跨境部署的通用人工智能在全球催生了众多应用。人工智能加速发展，引发了全球范围内权力和财富的聚集效应，并产生了地缘政治和地缘经济影响。

八 此外，目前没有人充分了解人工智能的所有内部运作，因而无法完全控制输出内容，或预测其演变过程。决策者也不会因为开发、部署或使用他们所不理解的系统而被追究责任。与此同时，这些决策可能会产生全球性的负面溢出效应和下游影响。

九 这种技术的开发、部署和使用不能仅由市场随意决定。各国政府和区域组织将发挥关键作用，但鉴于该技术本身在结构和应用方面的跨境性质，必须采取全球性办法。治理也可成为推动人工智能创新促进全球实现可持续发展目标的关键因素。

十 因此，人工智能的挑战和机遇并存，需要采取统筹兼顾的全球性办法，涵盖政治、经济、社会、道德、人权、技术、环境等各个领域。这种办法可将不断演变的零散举措汇聚成以国际法和可持续发展目标为基础、连贯一致、可互操作的整体，并随具体情况和时间推移不断作出调整。

十一 我们在中期报告中概述了建立新的国际人工智能治理机构的指导原则[2]。这些原则承认，人工智能治理并不在真空状态下进行，国际法尤其是国际人权法适用于人工智能领域。

1 见 https://un.org/ai-advisory-body.

2 指导原则1：人工智能应由所有各方以包容方式共同治理，并造福所有人；指导原则2：人工智能治理必须以公共利益为导向；指导原则3：人工智能治理应与数据治理和促进数据共享空间协同进行；指导原则4：人工智能治理必须具有普遍性、网络化并植根于多利益攸关方的适应性协作；指导原则5：人工智能治理应立足于《联合国宪章》、国际人权法和可持续发展目标等其他商定的国际承诺。

2. 全球人工智能治理差距

十二 我们并不缺少聚焦人工智能治理问题的文件和对话。各国政府、企业和联盟以及区域和国际组织已通过了数百项指南、框架和原则。

十三 然而，它们都不具备真正的全球影响力和全面覆盖性。这引出了代表性、协调和执行的问题。

十四 就代表性而言，全世界的大片地区都被排除在国际人工智能治理对话之外。图(a)显示了七项重要的非联合国人工智能举措[3]。七个国家加入了样本中的所有人工智能治理努力，118个国家没有加入其中任何一项努力(主要是全球南方国家)。

十五 为了实现公平，在决定如何治理与我们切身相关的技术时，需要让更多声音发挥切实有效的作用。决策工作集中在人工智能技术部门是不合理的；我们还必须认识到，历史上有许多社区被完全排除在影响他们的人工智能治理对话之外。

十六 人工智能治理制度还必须覆盖全球，这样才能在下列方面发挥成效：避免"人工智能军备竞赛"，或在安全和权利方面陷入逐底竞赛；查明和应对人工智能生命周期中跨越多个司法管辖区的决策所引发的事件；激励学习；鼓励互操作性；以及分享人工智能的惠益。这项技术跨越国界，随着它的传播，对于任何一个国家或国家集团能够（或应当）控制这项技术的幻想都会破灭。

十七 各项举措和机构之间的协调差距可能会将全世界分裂为多个彼此脱节且互不相容的人工智能治理制度。联合国系统内部同样缺乏协调。尽管许多联合国实体都涉足人工智能治理，但其各自具有特定的任务授权，这意味着没有任何一个实体能够进行全面治理。

十八 不过，代表性和协调并不够。问责的一项必要条件是执行，使全球人工智能治理的承诺转化为实践中的切实成果，包括在能力发展和支持中小企业方面，以期分享机遇。这方面的工作大多将在国家和区域层面进行，但也需要在全球一级作出更多努力，以应对风险和把握惠益。

3. 加强全球合作

十九 我们的建议提出了一个整体愿景，即采用全球网络化、敏捷灵活的办法治理人工智能以造福人类，其中涵盖统一认知、共同基础和共享惠益。只有通过这种包容和全面的人工智能治理办法，才能应对和把握人工智能在全球范围内带来的多层面和不断演变的挑战和机遇，促进国际稳定和公平发展。

二十 我们的提议遵循中期报告所确立的原则，旨在填补差距，使快速兴起的国际人工智能治理对策和举措构成的生态系统协同一致，帮助避免碎片化和错失机遇。为高效地支持这些措施，并与其他机构开展有效合作，我们提议建立一个精简而灵活的机构以展现协同努力，即在联合国秘书处设立一个靠近秘书长的人工智能办公室，作为"粘合剂"高效和可持续地联结本报告中提出的各项举措。

A. 统一认知

二十一 为采取人工智能治理全球办法，首先要就人工智能的能力、机遇、风险和不确定性统一认知。需要获得关于人工智能的及时、客观、可靠的科学知识和信息，使会员国能够在全球范围内形成共同的基本理解，并对拥有昂贵人工智能实验室的企业与世界其他部分之间的信息不对称加以平衡(包括为此在人工智能企业和广大人工智能社区之间分享信息)。

二十二 在全球层面汇集科学知识最为高效，这样能够对全球公共产品进行联合投资，并整合原本零散和重复的工作，促进公益合作。

国际人工智能科学小组

二十三 借鉴政府间气候变化专门委员会(气专委)和联合国原子辐射影响问题科学委员会等先例，一个国际多学科人工智能科学小组可负责整理和催化前沿研究，为科学家、政策制定者、会员国以及寻求从公正可靠来源了解人工智能技术或其应用方面科学观点的其他利益攸关方提供信息。

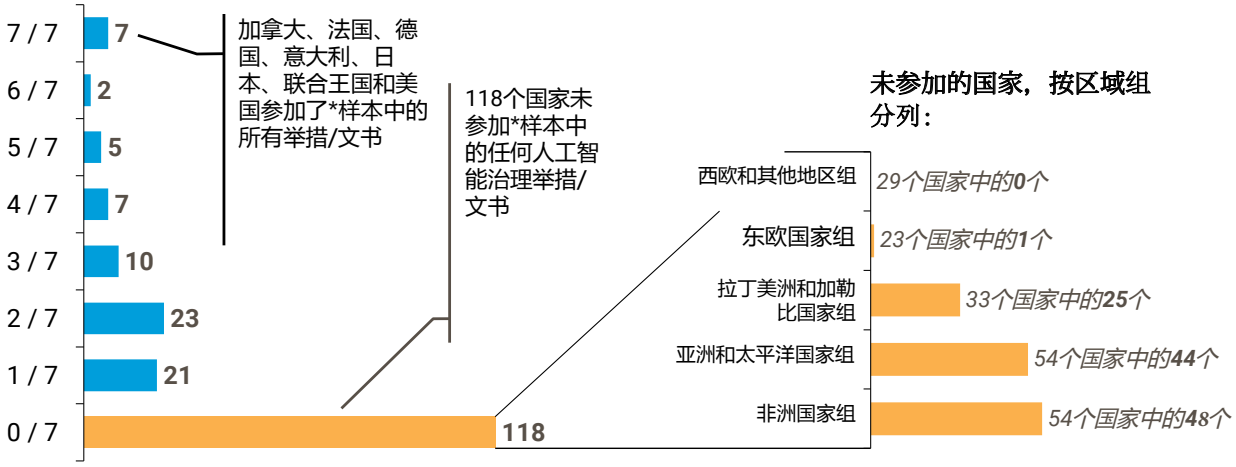
二十四 由联合国主持的科学小组可获取关于人工智能相关机遇的专业知识。其中可能包括促进"深入研究"可持续发展目标的应用领域，例如保健、能源、教育、金融、农业、气候、贸易和就业。

3 不包括联合国教育、科学及文化组织(教科文组织)《人工智能伦理问题建议书》（2021年）和2024年大会关于人工智能的两项决议："抓住安全、可靠和值得信赖的人工智能系统带来的机遇，促进可持续发展"（78/265）和"加强人工智能能力建设方面的国际合作"（78/311）。

图(a)： 七项非联合国的国际人工智能治理举措中的代表性

样本: 经合组织《人工智能原则》(2019)、二十国集团《人工智能原则》(2019)、欧洲委员会人工智能公约起草小组 (2022-2024)、人工智能全球伙伴关系《部长宣言》(2022)、七国集团《部长声明》(2023)、《布莱切利宣言》(2023)和《首尔部长宣言》(2024)。

仅区域间举措，不包括区域性举措



* 根据相关政府间文书的认可。仅因拥有欧洲联盟或非洲联盟成员资格并不会被视为诸边举措的参与国。 缩写：经合组织，经济合作与发展组织。

二十五 风险评估还可借鉴其他人工智能研究举措的工作，由联合国为研究人员提供特别值得信赖的"安全港"，以交流关于"最先进"技术的想法。联合国主持的小组可打破知识壁垒，汇集原本可能不会参与或被纳入的国家或企业所掌握的知识，从而帮助纠正误解和增强全球信任。

二十六 该小组应独立运作，并由下文提议的人工智能办公室以及国际电信联盟（国际电联）和联合国教育、科学及文化组织（教科文组织）等联合国相关机构组成的跨联合国系统团队提供支持。小组还应与经济合作与发展组织（经合组织）等其他国际机构领导的研究工作以及人工智能全球伙伴关系建立伙伴关系。

建议1:

国际人工智能科学小组

我们建议成立一个独立的国际人工智能科学小组，由该领域的多学科专家组成，以个人身份自愿任职。该小组将在拟议的联合国人工智能办公室和联合国其他相关机构的支持下，与其他相关国际组织合作，其任务包括：

a) 发布关于人工智能相关能力、机遇、风险和不确定性调查的年度报告，确定关于技术趋势的科学共识领域以及需要开展更多研究的领域;

b) 编制季度专题研究摘要，探讨人工智能有助于实现可持续发展目标的领域，重点关注可能扶持不足的公共利益领域;

c) 发布关于新兴问题的特别报告，特别是在治理领域出现的新风险或重大差距。

B. 共同基础

二十七 除了就人工智能统一认知外，还需要建立共同基础，在全球规范和原则的基础上制定符合所有国家利益的可互操作的治理办法。必须在全球层面开展这项工作，以避免监管领域的逐底竞争，减少跨境监管摩擦；最大限度地提高学习和技术方面的互操作性；并有效应对人工智能跨境特性带来的挑战。

人工智能治理政策对话

二十八 需要建立一个包容各方的政策论坛，使所有会员国都能借鉴利益攸关方的专业知识，分享基于人权和推动发展的最佳做法，以此促进可互操作的治理办法，并应对需要作出进一步政策考虑的跨境挑战。这并不意味着对人工智能的所有方面进行全球治理，但该论坛可建立国际合作的框架，使业界和各国所作的努力更符合全球规范和原则。

二十九 在联合国的主持下建立这种多利益攸关方交流制度，可为讨论新兴治理实践和适当的政策应对措施提供可靠包容的场所。意见相左的国家之间以及国家与利益攸关方之间突破舒适区开展对话可以促进学习，并为加强合作奠定基础，例如在安全标准和权利方面加强合作，并携手应对全球危机时期。联合国环境对于将这一努力锚定在尽可能广泛的共同规范中至关重要。

三十 这种关于治理办法的包容性对话通过与能力发展相结合（见建议4和5），即可帮助各个国家和企业更新其监管办法和方法，以应对加速发展的人工智能。与国际科学小组的联系将增强这种发展动态，类似于气专委与联合国气候变化大会之间的关系。

三十一 政策对话可在纽约（例如大会[4]）和日内瓦现有会议期间开始。每年召开两次会议，其中一次会议可更多关注各部门的机会，另一次会议则更注重风险[5]。展望未来，这种会议将成为分享人工智能事件信息的适当论坛，例如分享那些使现有机构捉襟见肘或超出其能力范围的事件。

三十二 在每次对话中，可由会员国牵头举办一部分会议，重点探讨各国采用的办法，另一部分会议则用于从主要利益攸关方（特别是科技企业和民间社会代表）那里获取专业知识和意见。除正式对话会议以外，还可利用其他更专业的现有机制，让多利益攸关方参与制定人工智能政策，例如通过国际电联的人工智能造福人类会议、互联网治理论坛年度会议、教科文组织的人工智能伦理问题全球论坛以及联合国贸易和发展会议（贸发会议）的电子周 。

人工智能标准交流中心

三十三 在最初研究人工智能系统时，极少有标准可用于指引或评估这一新的前沿领域。近来，各种标准急剧增加。图(b) 表明，国际电联、国际标准化组织、国际电工委员会和电气电子工程师学会采用的标准数量持续增加。

三十四 这些标准制定机构没有采用统一的措辞，公平性、安全性和透明度等许多常用的人工智能术语也缺乏商定的定义。用于狭窄的技术领域或内部核证的标准与旨在纳入更广泛道德原则的标准之间也存在脱节。正在浮现的这套标准并不基于对含义的统一认知，或者背离了它们旨在维护的价值观。

三十五 联合国系统可通过利用国际科学小组的专业知识，并纳入推动标准制定的各国家和国际实体的成员以及技术企业和民间社会的代表，由此成为具有全球适用性的人工智能标准的交流中心 。

4 类似于经济及社会理事会主持召开的可持续发展目标高级别政治论坛。

5 可让联合国系统的下列相关机构参与，并强调指出机遇和风险：国际电联侧重于人工智能标准；国际电联、联合国贸易和发展会议(贸发会议)、联合国开发计划署（开发署）和发展协调办公室侧重于推动实现可持续发展目标的人工智能应用；教科文组织侧重于伦理道德和治理能力;联合国人权事务高级专员办事处(人权高专办)侧重于基于现有规范和机制的人权问责;裁军事务厅侧重于监管军事系统中的人工智能;开发署侧重于支持国家发展能力;互联网治理论坛侧重于促进多利益攸关方参与和对话;世界知识产权组织(知识产权组织)、国际劳工组织(劳工组织)、世界卫生组织（世卫组织）、联合国粮食及农业组织（粮农组织）、世界粮食计划署、联合国难民事务高级专员公署（难民署）、教科文组织、联合国儿童基金会、世界气象组织和其他机构侧重于部门应用和治理。

建议2：

人工智能治理政策对话

我们建议在联合国现有会议期间，启动每年两次的人工智能治理问题政府间多利益攸关方政策对话。其目的是：

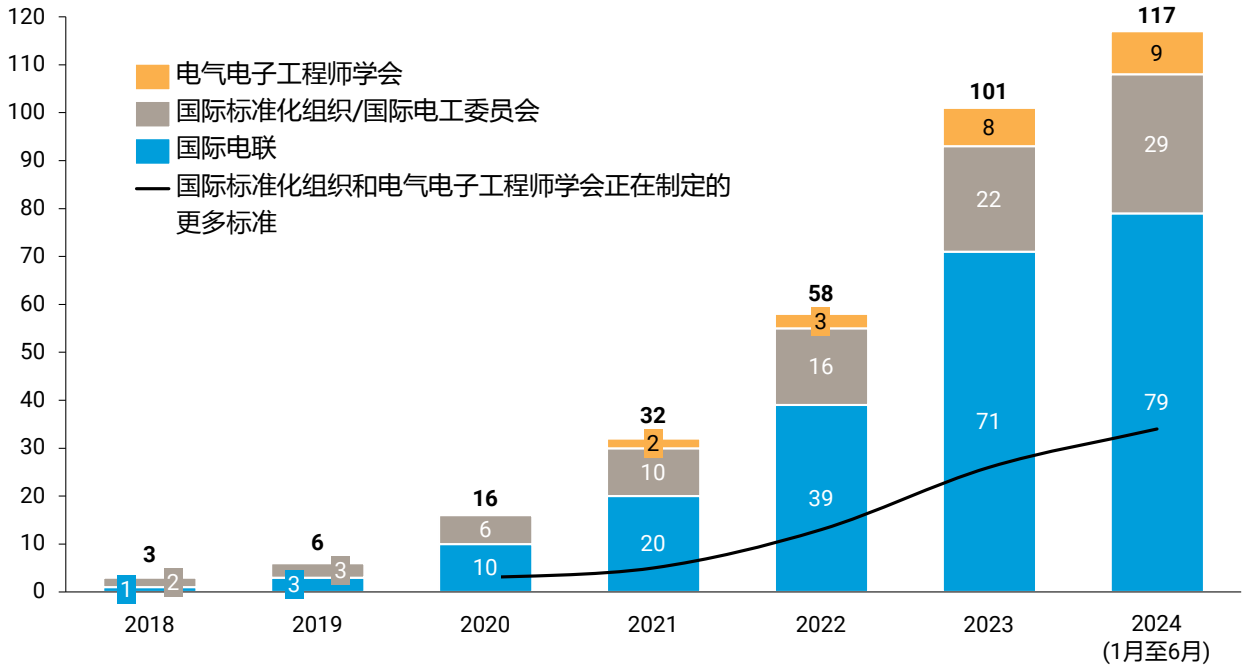
a) 分享在推动发展的同时促进尊重、保护和实现所有人权的人工智能治理最佳做法，包括把握机遇和管控风险方面的最佳做法；

b) 促进私营部门和公共部门开发人员和用户就实施人工智能治理措施统一认知，以加强人工智能治理的国际互操作性；

c) 自愿分享使国家机构捉襟见肘或超出其应对能力的重大人工智能事件；

d) 酌情讨论国际人工智能科学小组的报告。

图(b)：人工智能相关标准的数目



信息来源：电气电子工程师学会、国际标准化组织/国际电工委员会、 国际电联、世界标准合作组织(基于2023年6月的摸底调查，纳入了与人工智能相关的标准以扩大范围)。

建议3：

人工智能标准交流中心

我们建议创建人工智能标准交流中心，汇集各国和国际性的标准制定组织、技术企业、民间社会的代表以及国际科学小组的代表。其任务是：

a) 建立和维护人工智能系统相关定义与适用的测量和评估标准的登记册；

b) 就标准及其制定过程进行辩论和评估；

c) 查明需要制定新标准的空白领域。

C. 共享惠益

三十六 依照《2030年可持续发展议程》及其17项可持续发展目标，我们可以明确人工智能开发、部署和使用的目的，将投资重点转向克服全球发展挑战。若不采取全面和包容的人工智能治理办法，就可能错失人工智能为可持续发展目标作出积极贡献的潜力，而且人工智能的部署可能会无意中强化或加剧差异和偏见。

三十七 人工智能并非应对各项可持续发展挑战的万灵药，而是一套更广泛解决方案中的一部分。为使人工智能在应对社会挑战方面真正发挥潜力，关键在于各国政府、学术界、业界和民间社会之间开展合作，使人工智能驱动的解决方案具有包容性和公平性。

三十八 这在很大程度上取决于能否获得人才、计算能力（或"算力"）和数据，帮助多元文化和语言蓬勃发展。基本的基础设施和维护这些设施的资源也是必备条件。

三十九 就人才而言，并非每个社会都需要计算机科学家组成的骨干队伍来搭建自己的模型。不过，无论是购买、借用还是创建的技术，都需要基本的社会技术人才来了解人工智能的能力和局限，适当利用人工智能驱动的用例，同时应对特定背景下的风险。

四十 算力是进入人工智能领域的最大障碍之一。在能够训练大型人工智能模型的全球前100个高性能计算中心中，没有一个中心由发展中国家托管。 即使最富有的国家和企业也难以获得算力，在这种情况下承诺获得算力是不切实际的。[6]我们宁可谋求为那些无法通过其他途径获得必要推动因素的国家托底，确保它们不会沉入人工智能的鸿沟，包括支持采取举措，建立分布式和联合式人工智能开发模型。

四十一 关于数据，人们常常提及人工智能背景下的数据滥用（例如侵犯隐私）或数据漏用（未能利用现有数据集）。但一个相关问题是数据缺失，包括全球大量地区存在数据匮乏问题。人工智能系统未能反映世界上的语言和文化多样性与这些系统的偏见有关，但也可能导致这些社区错失获得人工智能惠益的机会。

四十二 需要一组共享资源（包括开放模型）来支持所有会员国以包容和有效的方式参与人工智能生态系统，全球性办法在这方面别具优势。

6 替代指标，因为大多数高性能计算中心不具备图形处理器，对先进人工智能的用途有限。

建议4：

能力发展网络

我们建议创建人工智能能力发展网络，汇集联合国下属的一系列相互协作的能力发展中心，向关键行为体提供专门知识、算力和人工智能训练数据。建立该网络的目的是：

a) 支持区域和全球一级的人工智能能力建设举措彼此建立联系，以促进和协调这些举措；

b) 提升公职人员的人工智能治理能力，以期在推动发展的同时，促进尊重、保护和实现所有人权；

c) 向寻求将人工智能应用于当地公共利益用例的研究人员和社会企业家提供多个中心的培训人员、算力和人工智能训练数据，包括通过下列方式：

   i) 制定协议，使算力稀缺环境中的跨学科研究团队和企业家能够获得算力，用于训练/调整其模型，以及将模型适当应用于当地环境；

   ii) 通过沙箱测试潜在的人工智能解决方案，并在实践中学习

   iii) 向大学生、年轻研究人员、社会企业家和公共部门官员提供一系列人工智能在线教育机会；

   iv) 四. 设立研究金方案，供有前途的个人在学术机构或技术企业工作一段时间

能力发展网络

四十三 公共和私营部门对人力和人工智能方面其他能力的需求不断增长，与此同时，国家和区域一级以及公私合作的人工智能卓越中心正在兴起，在国际能力发展领域发挥作用。全球网络可成为一个对接平台，扩大可能的合作范围，并增强能力建设办法的互操作性。

四十四 从千年发展目标到可持续发展目标，联合国长期以来一直积极发展个人和机构的能力。[7]隶属于联合国的机构网络可为寻求建立能力伙伴关系的国家扩大选择范围。该网络还可推动建立新的国家卓越中心，刺激当地人工智能创新生态系统的发展，同时采用符合联合国规范承诺的互操作办法。

四十五 这种网络将促进建立人工智能技术开发的替代范式：自下而上、跨领域、开放和协作。国家层面的努力可继续使用教科文组织人工智能准备度评估方法等诊断工具，帮助查明国家一级的差距，并由国际网络协助解决这些差距。

全球人工智能基金

四十六 很多国家由于面临财政和资源限制，其适当和有效利用人工智能的能力受限。尽管作出了各种能力发展努力（建议4），但若缺乏国际支持，一些国家可能仍然无法获得培训、算力、模型和训练数据。在缺乏这种支持的情况下，其他资助工作也可能无法扩大规模。

7 联合国通过教科文组织、知识产权组织和其他机构的工作，帮助维护了全球丰富多样的文化和知识创造传统。联合国大学长期以来一直致力于通过高等教育和研究进行能力建设，联合国训练研究所帮助培训了可持续发展关键领域的官员。教科文组织的准备度评估方法是支持会员国实施教科文组织《人工智能伦理问题建议书》的关键工具。其他例子包括设在法国里昂的世卫组织学院、贸发会议虚拟学院、裁军事务厅管理的联合国裁军研究金方案以及国际电联和开发署领导的能力发展项目

建议5：

全球人工智能基金

我们建议设立全球人工智能基金，为弥合人工智能鸿沟托底。该基金将由独立的治理机构负责管理，接收来自公共和私营部门的财政和实物捐助，并通过能力发展网络等渠道分配这些资源，从而促进提供下列人工智能推动因素，赋能当地实现可持续发展目标：

a) 共享计算资源，供当地能力不足或无力采购资源的国家的人工智能开发人员训练模型和进行微调；

b) 沙箱与对标和测试工具，将安全可靠的模型开发和数据治理最佳做法纳入主流；

c) 适用于全球的治理、安全和可互操作解决方案；

d) 数据集，以及研究如何将数据和模型结合起来，用于可持续发展目标相关项目

e) 推动实现可持续发展目标的人工智能模型库和管护数据集。

四十七 我们提议设立基金的目的并非保证获得先进的计算资源和能力。答案可能并不总是在于获得更多算力。我们还需要更好地将人才、算力和数据联系起来。基金旨在帮助无法获得必要推动因素的国家消除潜在的能力与合作差距，以达到下列目标：

a. 有需要的国家可获得人工智能推动因素，为弥合人工智能鸿沟托底；
b. 合作发展人工智能能力，形成合作习惯，缓和地缘政治竞争；
c. 采用不同监管方法的国家有动力开发通用模板，用于管理数据、模型和应用，以克服与可持续发展目标相关的社会层面挑战，并实现科学上的突破。

四十八 该基金关注公共利益，因此与拟议的人工智能能力发展网络相辅相成，基金可向该网络输送资源。基金将提供独立的影响监测能力，并可寻求和汇集实物捐助，包括来自私营部门实体的捐助，从而以低于市价的费用提供与人工智能有关的培训方案、时间、算力、模型和管护数据集。我们以此来确保全世界广大地区不会掉队，并且获得更多赋能，以便在各种情况下利用人工智能推动实现可持续发展目标。

四十九 确保在数字世界与在物理世界同样开展合作符合所有人的利益。如同应对气候变化的努力一样，转型、减缓或适应的成本并不是匀速下降，必须提供国际援助，以帮助资源有限的国家加入应对地球挑战的全球性努力。

全球人工智能数据框架

五十 通过市场或其他机制获取人工智能训练数据，是当地人工智能生态系统蓬勃发展的关键推动因素，特别是在“缺失”数据的国家、社区、地区和人口群体中（见上文关于“共享惠益”的一节）。

五十一 只有全球集体行动，才能鼓励互操作性、管理、隐私保护、赋能和增强权利，促进各司法管辖区在下列领域“力争上游”：在治理人工智能训练数据的收集、创建、使用和货币化的过程中，保护人权和履行其他商定承诺，提高数据可用性，以及公平补偿数据主体。为实现这一目标，我们提议建立全球人工智能数据框架。

五十二 这种框架不会创设新的数据相关权利。确切地说，它将解决人工智能训练数据的可用性、互操作性和

使用问题。框架将有助于就如何统一不同国家和区域的数据保护框架达成共识。它还能促进当地人工智能生态系统蓬勃发展，支持文化和语言多样性，并限制经济集中度进一步提升。

五十三 为补充这些措施，可以在公正、安全和公平托管及共享数据协议模板的基础上，促进建立数据共享空间，并制定可持续发展目标相关领域数据信托的托管规定。拟议的能力发展网络和全球人工智能基金（建议4和5），可为制定这些模板以及实际存储和分析共享空间或信托中的数据提供支持。

五十四 联合国别具优势，可根据人权、知识产权和可持续发展方面的国际商定承诺，以数据界多年来的工作为基础，并将其与人工智能伦理道德和治理方面的最新发展相结合，支持制定关于人工智能训练数据治理和使用的国际原则和实际安排。这类似于联合国国际贸易法委员会在制定法律和非法律跨境框架以促进国际贸易方面发挥的作用。

五十五 同样，科学和技术促进发展委员会和统计委员会的议程上也有关于数据促进发展和可持续发展目标数据的议题。世界知识产权组织（知识产权组织）还在审议内容、版权、保护土著知识和文化表现形式等重要议题。

建议6：
全球人工智能数据框架

我们建议通过联合国国际贸易法委员会等相关机构发起的流程，在参考其他国际组织所开展工作的基础上，创建全球人工智能数据框架，其任务是：

a) 勾勒数据相关定义和原则，包括从现有最佳做法中提炼的定义和原则，以便对人工智能训练数据进行全球治理，并促进文化和语言多样性；

b) 就人工智能训练数据的来源和使用制定统一标准，以便建立跨司法管辖区的透明和基于权利的问责制；

c) 建立促进市场发展的数据管理和交流机制，推动全球范围内的当地人工智能生态系统蓬勃发展，例如

- **i)** 数据信托；
- **ii)** 治理良好的全球市场，以便交流用于训练人工智能模型的匿名数据；
- **iii)** 促进国际数据访问和全球互操作性的示范协议，可能用作框架的技术法律协议。

D. 协同努力

五十六 上述各项提议旨在解决正在兴起的国际人工智能治理制度中发现的代表性、协调和执行差距。可通过与现有机构和机制建立伙伴关系和开展合作来解决这些差距，促进统一认知、共同基础和共享惠益。

五十七 然而，如果联合国没有专门的协调中心来支持这些努力，并使之与其他各项努力协同配合，那么全世界将缺乏包容各方的网络化、灵活和连贯一致的办法，而这种办法对于有效和公平地治理人工智能这项具有跨境影响和快速变化的通用技术不可或缺。

五十八 上文"全球人工智能治理差距"一节概述的各种零散规范和机构表明，人们普遍认识到人工智能治理是一项全球要务。鉴于这方面的应对措施参差不齐，有必要开展一定程度的协同努力

联合国秘书处人工智能办公室

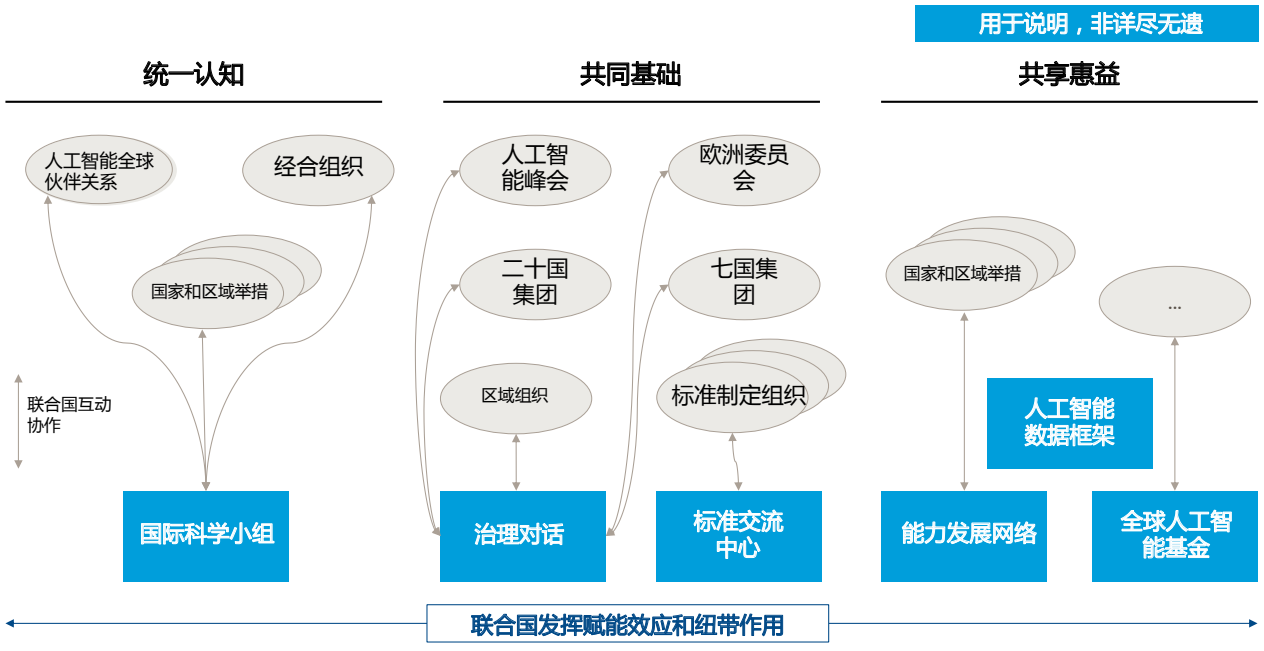
五十九 因此，我们提议建立柔性的机制，作为"粘合剂"支持和促进本报告所述的各项提议，包括通过建立伙伴关系，同时使联合国系统能够在不断演变的人工智能治理生态系统中发出一致的声音。

六十 这种小型而灵活的能力体现为联合国秘书处内部的人工智能办公室，该办公室向联合国秘书长汇报，其好处是可与整个联合国系统保持联系，而不与其中的某个部分进行绑定。这一点十分重要，因为人工智能的未来充满不确定性，它很有可能渗透到人类活动的方方面面。

六十一 该机构应保持灵活，倡导包容，迅速建立伙伴关系以加快协调和执行工作，还应优先利用联合国系统内的现有资源和职能。应当重点关注人工智能在民用领域的应用。

六十二 该办公室的部分工作人员可以是来自联合国系统下列专门机构和其他组织借调的联合国人员，例如国

图(c)： 联合国在国际人工智能治理生态系统中的拟议作用



缩写：经合组织，经济合作与发展组织。

建议7：

在秘书处内设立人工智能办公室

我们建议在秘书处内设立人工智能办公室，向秘书长汇报。办公室的组织结构应精简灵活，尽可能利用联合国现有相关实体。该办公室作为"粘合剂"支持和推进本报告中的各项提议，并与其他流程和机构合作互动，其任务包括：

a) 支持拟议的国际科学小组、政策对话、标准交流中心和能力发展网络，并向全球基金和全球人工智能数据框架提供所需支持；

b) 参与就新兴人工智能问题与科技企业、民间社会和学术界等多利益攸关方进行外联；

c) 就人工智能相关事宜向秘书长提供咨询意见，并与联合国系统其他相关机构进行协调，以便采取联合国一体化应对措施。

际电联、教科文组织、联合国人权事务高级专员办事处（人权高专办）、贸发会议、联合国大学和联合国开发计划署（开发署）。办公室应与企业、民间社会和学术界等多利益攸关方互动协作，并与联合国以外的主要组织开展合作(见图(c))。这将为联合国创造条件，使其能够在国际人工智能治理生态系统中发挥纽带作用，促进统一认知、共同基础和共享惠益。

E. 对制度模式的反思

六十三 建议7基于透彻地评估联合国可在哪些领域带来增值，包括发挥领导作用、帮助协调以及应在哪些领域让贤。落实这项建议的另一个好处是可享有现有机构安排，包括预先谈判商定的供资安排，以及业已确立和为人熟知的行政流程。

六十四 关于人工智能的讨论往往走向极端。我们在世界各地进行了多次磋商，接触到的一部分人认为，未来日渐廉价和愈加有用的人工智能系统将提供无限的商品。与我们对话的另一些人则唯恐会出现更加黑暗的未来、分歧、失业、甚至灭绝[8]。

六十五 我们无从知晓未来更有可能是乌托邦还是反乌托邦。同样，我们注意到这项技术可能会朝着摆脱这种二元对立的方向发展。本报告以科学为基础，以事实为依据，重点关注近期的机遇和风险。

六十六 上文概述的七项建议承载着我们最大的希望，即在人工智能不断演变的过程中获取人工智能的惠益，同时最大限度地减少和减轻风险。我们还注意到在开展更大规模的国际机构建设方面的实际挑战。正因如此，我们建议采用网络化的机构方法，并提供精简而灵活的支持。如果风险日益加剧，机会愈加利害攸关，那我们届时会对这种评估作出调整。

六十七 在两次世界大战后诞生了现代国际体系；人们开发的化学、生物和核武器威力不断增强，促使各国政权限制这些武器的传播，并促进和平利用相关技术。随着对共同人性的理解逐步加深，我们构建了现代人权体系，并不断致力于为所有人实现可持续发展目标。气候变化则从一项局部关切演变为全球性挑战。

8 见 https://safe.ai/work/statement-on-ai-risk.

六十八 同样，人工智能可能在发展到一定阶段后，需要比上述建议中提出的更多资源和权限，涉及规范制定、实施、监测、核证和验证、执行、问责、补救损失和紧急反应等更具挑战性的职能。因此，对这种制度模式进行反思是审慎之举。本报告的最后一节力求推动这项努力。

4. 行动呼吁

六十九 我们仍对人工智能的未来及其积极潜力持乐观态度。但这种乐观态度的前提是，务实地看待风险以及当前架构和激励措施的不足。这项技术过于重要，过于利害攸关，因而不能仅仅依赖于市场力量以及碎片化的国家和多边行动。

七十 联合国可成为制定新的人工智能问题社会契约的平台，确保全球共同支持保护和赋能所有人的治理制度。这项社会契约将确保公平分配机会，不将风险推卸给最弱势的群体，或转嫁给子孙后代，很不幸的是，这正是气候变化领域的现状。

七十一 我们期待继续开展这场重要对话，无论是作为一个团体，还是作为来自多个专业领域、组织和世界各地的个人。我们与在这段旅程中建立联系的众多其他各方及其代表的全球社会携手齐心，希望本报告助力我们治理人工智能以造福人类的共同努力。

图(d)： 人工智能高级别咨询机构会议，新加坡，2024 年 5 月 29 日

1. 导言

1 成立秘书长人工智能高级别咨询机构的目的是分析国际人工智能治理问题，并就此提出建议。我们的成员来自不同的地区，性别、学科和年龄各异；分别从政府、民间社会、私营部门和学术界汲取专门知识。我们通过激烈和全面讨论取得了广泛共识(如中期报告所述[1])，即在人工智能领域存在全球治理缺陷。中期报告阐述了关于这方面作用的指导原则和国际上可能需要的职能。

2 我们在接下来的几个月中收到了诸多反馈，并进行了大量磋商。其中包括：就特定问题领域开展18项“深入研究”，500多名专家参与研究，来自所有区域的150多个组织和100名个人提交了250多份书面答复；开展了人工智能风险脉动调查，来自所有区域的约350名专家作出答复；进行了一次机遇扫描，来自所有区域的约120名专家作出答复；以及与会员国、联合国各实体和其他利益攸关方团体在所有区域举行了40多次会议，在此期间进行了定期磋商和简报。咨询机构成员还广泛参与世界各地的论坛[2]，举行了100多场虚拟讨论，并在纽约、日内瓦和新加坡召开了3次全体面对面会议。

3 因此，这份最后报告由众多作者合作完成。报告无法完全反映所表达观点的丰富性和多样性，但展现了我们的共同承诺，即确保以造福全人类的方式来开发、部署和使用人工智能，并确保在国际层面开展有效和包容的人工智能治理。

4 本报告重申了咨询机构的中期报告中关于机遇和推动因素以及风险和挑战的结论，再次指出人工智能全球治理的必要性，并概述了七项建议。

5 这些建议包括成立一个科学小组，负责促进就人工智能能力、机遇、风险和不确定性达成统一认知。我们需要在这种统一认知的基础上建立机制，就如何在国际层面治理人工智能找到共同基础。实现这一目标的关键在于开展定期对话，以及制定所有各方都接受和适用的标准。

6 本报告还就共享惠益提出建议，旨在确保公平分享人工智能的惠益，这可能取决于获得模型或能力，如人才、计算能力（或“算力”）以及数据的机会。建议包括：建立能力发展网络、全球人工智能基金和全球人工智能数据框架。

7 我们提议在联合国秘书处内设立人工智能办公室，负责推动这些努力，与其他举措和机构合作应对人工智能领域的关切和把握这方面的机遇，并确保联合国系统在人工智能问题上发出一致的声音。

8 尽管我们考虑了建议设立一个人工智能问题国际机构的可能性，但我们目前不建议采取这项行动，但承认治理需要跟上技术发展的步伐。

9 我们的报告除了面向涉及政府的当前各项多边辩论和进程外，还面向世界各地的民间社会和私营部门、研究人员和相关人士。我们十分清楚，只有在全球多部门的共同参与下，才能实现我们所勾勒的宏伟目标。

10 总体而言，我们相信这项技术的未来仍然充满各种可能。我们对技术方向的深入研究以及就开放与封闭的技术发展方法所进行的辩论证实了这一点（见方框9）。一种可能的未来是，越来越少的企业开发出更大规模和更为强大的模型。另一种可能则是更加多元化的全球创新格局，由可互操作的中小型人工智能模型主导，并提供大量社会和经济应用。我们的建议旨在提高后者的可能性，但同时也承认风险。

11 联合国自成立以来一直致力于促进全球人民的经济和社会进展。 千年发展目标寻求制定宏伟目标[3]，为全世界所有人民提供经济机会；此后，可持续发展目标试图调和发展的需要与地球的环境限制。人工智能工具和系统的大规模开发、部署和使用构成了又一巨大挑战，需要确保我们共同拥抱数字化未来，而不是扩大数字鸿沟。

1 见 https://un.org/ai-advisory-body.
2 磋商情况概览见附件C。
3 包括通过贸易、外国直接投资和技术转让推动长期发展。

12 包容性人工智能治理可谓联合国将面临的最艰巨治理挑战之一。私营部门在人工智能领域发挥的主导作用与威斯特伐利亚国际政治体系之间存在错位。在激烈的地缘政治竞争背景下，各国被人工智能带来权力和繁荣的潜力所吸引。许多社会在人工智能开发、部署和使用方面仍处于边缘地位，还有一些社会对人工智能的跨领域影响怀有兴奋和担忧交织的情绪。

13 尽管面临种种挑战，但我们别无退路。对于联合国及其会员国以及将愿望寄托于联合国的广大社区而言，此事利害攸关。希望本报告能够提供一些指引，帮助我们携手努力，共同治理人工智能以造福人类

A.机遇和推动因素

14 人工智能正在变革我们的世界。这套技术[4]具有巨大的向善潜力，从开辟新的科学探索领域（见方框1）和优化能源网络，到改善公共卫生或农业。使用人工智能工具可为个人[5]、各经济部门、科学研究和其他公共利益领域提供各种潜在机遇，这些机遇一旦实现，便可在促进经济（见方框2）和让社会变得更加美好方面发挥重要作用。预测和应对大流行病、洪水、野火和粮食不安全等促进公共利益的人工智能，甚至有助于在实现可持续发展目标方面取得进展。

B. 利用人工智能造福人类的关键推动因素

15 不过，不一定能够公平地发掘或把握在人工智能开发和使用过程中将会出现的潜在机遇。2024年5月，对推动在实现可持续发展目标领域取得进展的人工智能项目的供资分析发现，只有10%的赠款流向了设在低收入或中等收入国家的组织；就私营资本而言，这一数字为25%（其中超过90%在中国)[3]

C. 治理是关键推动因素

16 需要在全球范围内落实推动因素，才能充分实现和获取人工智能的惠益，而不是仅由少数国家的少数人获得惠益。为了确保人工智能的部署符合共同利益，并公平分配人工智能的机遇，需要各国政府采取行动以及采取政府间行动，鼓励私营部门、学术界和民间社会共同参与。所有治理框架都应在全球层面引导激励措施，以促进设立范围更广和更加包容的目标，帮助查明及处理需要权衡利弊的领域 。

D. 风险和挑战

17 人工智能的开发、部署和使用带来了风险，这些风险可能同时涉及多个领域。我们从脆弱性的角度构想人工智能的相关风险，为制定政策议程提供了基于脆弱性的思路。

18 人工智能的速度、不透明和自主性对传统监管体系构成了挑战。人工智能的技术开发和部署不断加速，也使国际治理愈发利害攸关，人工智能的通用性质同时对多个领域产生跨境影响。

E. 人工智能的风险

19 人工智能系统存在各种偏见，这项技术驱动的监视则令人反感，对此类问题的记录正在逐渐增多。其他风险与使用先进人工智能相关，例如大型语言模型的虚构、大量消耗资源以及对和平与安全构成的风险。人工智能产生的虚假信息对民主体制构成威胁。

20 鉴于人工智能及其应用无处不在且快速演变，要列出永远适用的人工智能风险全面清单是徒劳无益的，我们认为，从弱势社区和共享空间的角度看待风险更有帮助（见下文第26至28段）。

21 我们委托开展了一项前景扫描（人工智能风险全球脉动调查；见附件E）以推进工作，这项民意调查收集了来自所有地区68个国家、各学科的348名人工智能专家对人工智能相关趋势和风险的看法，调查结果反映了当前各位专家对风险的认知概况。总体而言[6]，十分之七的受访专家担忧或非常担忧人工智能所造成（现有或新的）损害的严重程度和(或)覆盖范围将在未来18个月内大幅增加(见附件E)。

4 经济合作与发展组织(经合组织)指出，“人工智能系统是基于机器的系统，该系统出于明确或暗示的目标，从收到的输入推断如何生成预测、内容、建议或决策等输出，这些输出可影响物理或虚拟环境。不同的人工智能系统在部署后的自主性和适应性水平各异”（https://oecd.ai/en/wonk/ai-system-definition-update)

5 但我们认为，需要由各领域专家严格评估关于人工智能惠益的说法。谋求人工智能向善发展应基于科学证据以及对利弊得失和替代方案的彻底评估。除了科学研究领域以外，社会科学也在经历变革。

6 在秘书长技术问题特使办公室和人工智能咨询机构网络的基础上拟定了受邀者名单，其中包括深度研究的参与者。在进行调查期间定期邀请更多专家以提高代表性。最后答复人数共计348人，表明全球受访者样本具有说服力和平衡性，这些受访者拥有相关专业知识，可以就人工智能风险提供有依据的意见(方法见附件E)。

方框1: 人工智能在推动科学进步方面的潜力

以互联网的变革性遗产为基础，人工智能很可能成为引领科学进步的又一重大飞跃。在万维网的协助下，科学家们彼此分享了海量实验数据、科学论文和文献。人工智能以此为根基不断发展，包括推动分析大量数据集、发现隐藏模式、建立新的假设和关联以及加快发现速度，包括通过自动化机器人开展的大规模实验。

人工智能对科学的影响涵盖各个主要学科。从生物学到物理学，从环境科学到社会科学，人工智能正在融入研究工作流程，并加速生成科学知识。当前有些说法可能夸大其词，而另一些说法则已得到证实，人工智能的长期前景看好[a]。

以生物学为例，人工智能破解了持续50年的难题——蛋白质折叠和蛋白质结构预测。成果包括预测了2亿多个蛋白质的结构，并在此基础上创建了开放存取数据库。在编写本报告时，已有190多个国家的200多万名科学家使用了该数据库，其中许多科学家致力于研究被忽视的疾病。此后，应用范围已扩展到生命的其他生物分子、脱氧核糖核酸、核糖核酸和配位体及其相互作用。

对于阿尔茨海默病、帕金森病和肌萎缩性侧索硬化症，专家正在利用人工智能识别疾病的生物标志物和预测治疗反应，制定诊断和治疗方法的精准度和速度由此显著提高[b]。从更大范围而言，人工智能正在遗传和临床档案的基础上定制治疗方法，帮助推进精准医疗(例如针对治疗神经退行性疾病)。人工智能技术还有助于加速新化合物的发现和开发。[c]

在射电天文学中，平方公里阵列等现代仪器收集数据的速度和规模完胜传统方法。人工智能可发挥重要作用，包括帮助选择关注哪个部分的数据，以期获得新颖洞察。人工智能可通过“无监督聚类”识别数据中的模式，而无需向其说明具体需要寻找什么。[d]将人工智能应用于社会科学研究，还可就复杂的人类动态提供深刻洞察，增进我们对社会趋势和经济发展的理解。

随着时间推移，通过实现前所未有的跨学科水平，对人工智能的设计和部署可催生新的科学领域，正如计算机技术与生物学和神经学研究的整合产生了生物信息学和神经信息学。如果负责任地开展这项工作，就可以利用人工智能整合和分析气候变化、粮食安全和公共卫生等领域各种数据集的能力，开辟研究路径，在这些传统上相互独立的领域之间搭建桥梁。

人工智能还可以通过验证复杂假设，加强科学研究对公共政策的影响，例如结合气候模型与农业数据以预测粮食安全风险，并将这些洞察与公共卫生结果联系起来。 另一个前景是促进公众科学以及利用本地知识和数据，以应对全球挑战。

a John Jumper and others, “Highly accurate protein structure prediction with AlphaFold”, Nature, vol. 596 (July 2021), pp. 583–589; see also Josh Abramson and others, “Accurate structure prediction of biomolecular interactions with AlphaFold 3”, Nature, vol. 630, pp. 493–500 (May 2024).
b Isaias Ghebrehiwet and others, “Revolutionizing personalized medicine with generative AI: a systematic review”, Artificial Intelligence Review, vol. 57, No. 127 (April 2024).
c Amil Merchant and others, “Scaling deep learning for materials discovery”, Nature, vol. 624, pp. 80–85 (November 2023).
d Zack Savitsky, “Astronomers are enlisting AI to prepare for a data downpour”, MIT Technology Review, 20 May 2024.

方框2: 人工智能的经济机遇

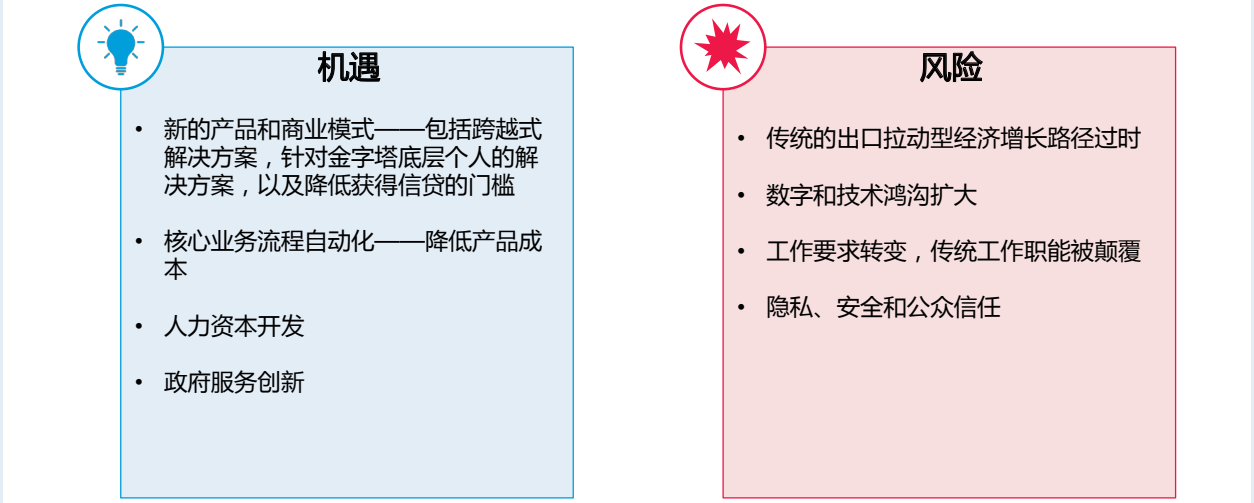
工业革命以来的数项创新极大地加快了经济进步。这些更早期的“通用技术”重塑了多个部门和行业。最近的一项重大变革来自计算机和数字时代。这些技术彻底改变了全球经济，提高了全世界的生产力，但其影响力历经数十年才得以完全显现。

生成式人工智能正在打破技术缓慢引入的趋势。专家认为，其变革性影响将在这个十年内显现。这种快速整合意味着，人工智能领域的新发展动态可能迅速重塑各个行业、改变工作流程并提高生产力。因此，人工智能的快速应用可能会以空前的方式变革我们的经济和社会。

人工智能有望创造可观的经济效益。尽管难以预计人工智能对复杂经济体的所有影响，但预测表明人工智能可显著提高全球的国内生产总值，并会对几乎所有部门都产生相关影响。人工智能可以为企业尤其是中小微企业提供高级分析和自动化工具，此前只有大型企业才能获得这些工具。人工智能的广泛适用性意味着，人工智能可能成为一种通用技术。因此，人工智能可以帮助发达和发展中经济体的各个部门，包括零售、制造和运营、医疗保健及公共部门的个人、小型和大型企业以及其他组织提高生产力。需要在本部门内和各部门间广泛采用人工智能[a]；开发提高生产力的应用；以及利用人工智能提高工人的工作成效，并大规模引入新的经济活动。还需要投资和资本深化、共同创新、流程和组织变革、准备就绪的劳动力和扶持政策。

不过，尽管人工智能可以提高生产力、促进国际贸易和增加收入，但预计也会影响工作。研究显示，人工智能在某些情况下可为工人提供协助，但在另一些情况下会取代工作。国际劳工组织(劳工组织)等机构的研究表明[b]，在可预见的未来，人工智能可能会更多地协助而非取代工人。[c]

图1: 人工智能给新兴市场带来的部分发展机遇和风险



信息来源：国际金融公司。

a James Manyika and Michael Spence, “The coming AI economic revolution: can artificial intelligence reverse the productivity slowdown?”, Foreign Affairs, 24 October 2023.Erik Brynjolfsson and others, “Generative AI at work”, National Bureau of Economic Research, working paper 31161, 2023; see also Shakked Noy and Whitney Zhang, “Experimental evidence on the productivity effects of generative artificial intelligence”, Science, vol. 381, No. 6654, pp. 187–192 (July 2023).

b Pawel Gmyrek and others, Generative AI and Jobs: A Global Analysis of Potential Effects on Job Quantity and Quality (Geneva: ILO, 2023).

c Mauro Cazzaniga and others, “Gen-AI: artificial intelligence and the future of work”, staff discussion note SDN2024/001 (Washington, D.C.: International Monetary Fund, 2024).

方框 2：人工智能的经济机遇（续）

研究还表明，一旦出现工作岗位流失，处于各个发展阶段的经济体预计将面临不同的流失情况。d发达经济体面临的风险更大，但它们的准备也更加充分，可利用人工智能对劳动力作出补充。低收入和中等收入国家利用这项技术的能力可能更为有限。此外，人工智能融入劳动力可能会对特定人口群体产生不成比例的影响，妇女可能在某些部门面临更高的工作岗位流失风险。

如果不开展重点突出的协同努力来缩小数字鸿沟，就无法发挥人工智能在支持可持续发展和减缓贫困方面的潜力，这会导致全球大量人口在迅速变化的技术环境中始终处于不利地位，并加剧现有的不平等现象。

为使人工智能成功融入全球经济，需要开展有效的治理来管控风险和确保取得公平的结果。这意味着采用各种方案，包括创建测试人工智能系统的监管沙箱，促进关于标准的国际合作，以及建立机制以持续评估人工智能对劳动力市场和社会的影响。除了完善的国家人工智能战略和国际支持外，还特别需要：

技能发展：实施教育和培训计划，培养全体劳动力的人工智能技能，从基本数字素养到高级技术专长，帮助工人做好准备，应对人工智能增强的未来。

- 数字基础设施：大力投资于数字基础设施，特别是在发展中国家，以弥合人工智能鸿沟，促进广泛采用人工智能。
- 融入工作场所：利用社会对话和公私伙伴关系，对人工智能融入工作场所的过程进行管理，确保工人参与这一过程并保护劳工权利。
- 价值链考虑因素：确保人工智能价值链各环节的体面工作条件，包括数据注释和内容审核等经常被忽视的领域，促进公平开发人工智能。

22 在人工智能相关的示例风险领域列表中[7]，多数专家表示担忧或非常担忧下列方面的损害(另见图2)：

a. 人工智能的社会影响：78%涉及对信息完整性的损害[问题j]，74%涉及财富和权力集中在少数人手中等不平等现象[问题h]，67%涉及歧视/剥夺权利，特别是在边缘化社区[问题l];
b. 故意使用人工智能伤害他人：75%涉及国家行为体在武装冲突中使用[问题b]，72%涉及非国家行为体恶意使用[问题a]，65%涉及国家行为体用于伤害个人[问题c]。

23 大多数接受调查的人工智能专家担忧或非常担忧除两个风险领域之外的所有示例风险领域都会出现损害。尽管对人工智能造成的意外损害表示感到此种担忧的专家比例不到一半[问题e和f]，但在非常担忧人工智能意外损害的专家中，有六分之一称其预计到2025年，代理系统将对人工智能相关风险产生一些最出人意料或最重大的影响。[8]

24 专家的看法不尽相同，包括存在地区和性别差异（更详细的结果见附件E），这突出表明了在界定共同风险的工作中实现包容代表性的重要性。尽管存在差异，但这些结果确实表明，专家对人工智能在接下来一年中即将造成的损害感到担忧，并凸显出各位专家对于在不久的将来解决多个领域风险和脆弱性的紧迫感。

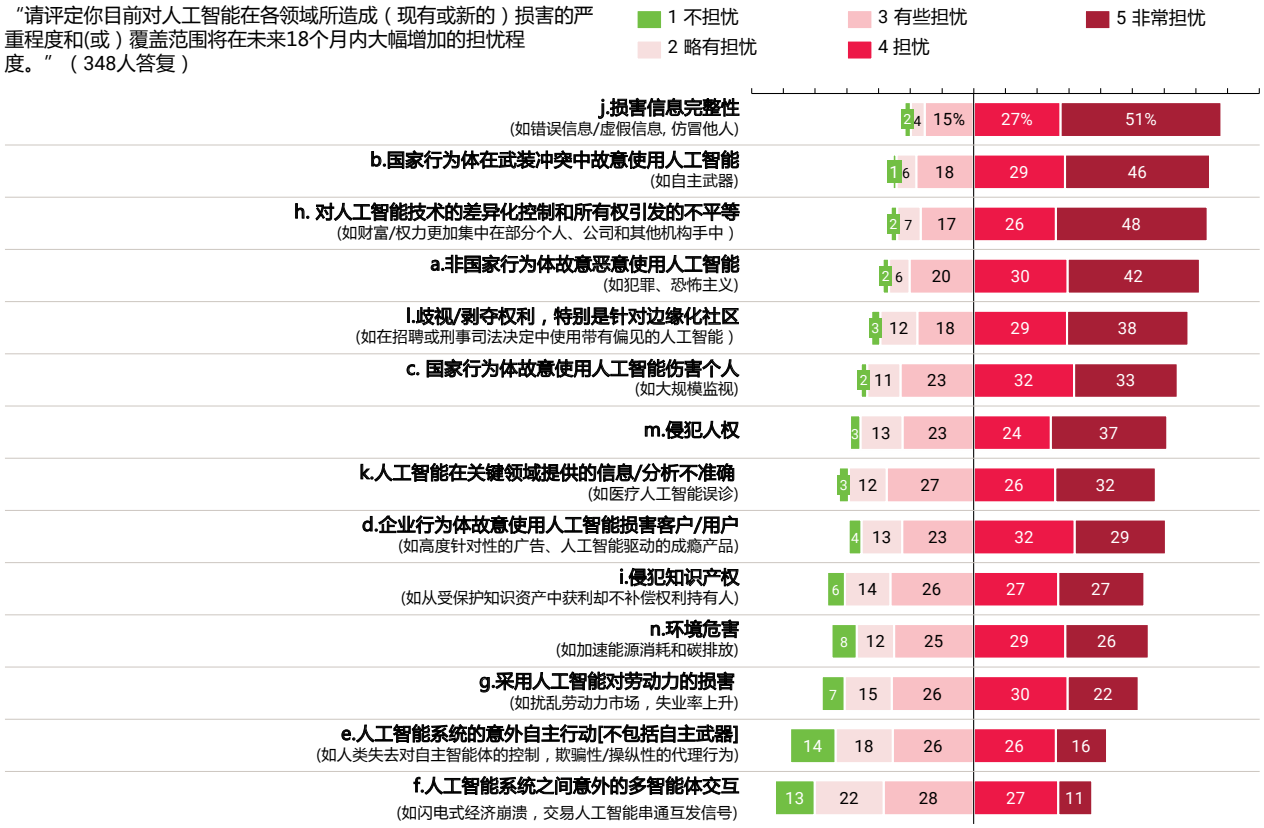
25 此外，在武装冲突、犯罪或恐怖主义中使用自主武器，特别是在公共安全领域使用人工智能，会引发严重的法律、安全和人道主义问题（见方框3）。[9]

26 然而，风险管理工作不仅需要列出风险或对其排序。建立基于脆弱性的风险框架可将政策议程重点从每个风险是“什么”（例如“安全风险”）转变为“谁”面临风险和“哪里”存在风险，以及各种情况下的责任方是谁。

7 依据方框4中基于脆弱性的风险分类，该分类的更早期版本载于中期报告。
8 问题：“你认为当前的哪些新兴趋势可能会在未来18个月内对人工智能相关风险产生最出人意料和(或) 最重大的影响？”
9 该列表仅用于提供说明，只涉及个人和社会面临的一部分风险

27 这点十分重要，因为对不同的人和社会而言，不断演变的风险呈现为不同的表现形式。基于脆弱性的办法(中期报告也提出了该办法)提供了开放式的框架，可用于关注可能受到人工智能损害的人员，为动态风险管理奠定基础（见方框4）。

图 2：专家对多领域人工智能风险的关注程度



秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

方框3: 人工智能与国家和国际安全

许多人工智能技术不仅具有双重用途，而且本质上可以“改换用途”。人工智能在执法和边境管制方面的应用日益增多，引发了人们对正当程序、监视以及各国对《世界人权宣言》和其他文书所载人权规范的承诺不受问责的担忧。

在军事领域使用人工智能方面的挑战包括新的军备竞赛、冲突门槛降低、战争与和平之间的界限模糊、向非国家行为体扩散以及克减国际人道法中确立已久的原则，例如军事必要、区分、相称性和限制不必要的痛苦。从法律和道德角度而言，不应通过人工智能自动作出杀人的决定。各国应承诺避免以不完全符合国际法（包括国际人道法和人权法）的方式在武装冲突中部署和使用人工智能的军事应用。

目前有120个会员国支持拟定一项关于自主武器的新条约，秘书长和红十字国际委员会主席都呼吁到2026年完成关于该条约的谈判。咨询机构敦促会员国响应这项呼吁。

咨询机构认为，必须明确划定非法用例的红线，这些用例包括依赖人工智能自主选择和打击目标。各国应依照国际人道法中关于武器审查的现有承诺，通过合同义务和其他方式要求武器制造商进行法律和技术审查，防止以不道德的方式设计和开发人工智能的军事应用。各国还应对人工智能以及武器和战争手段的使用进行法律和技术审查，并分享有关的最佳做法。

此外，各国应就安全和军事领域人工智能的测试、评估、核查和验证机制达成统一认知。各国应就安全和军事领域的人工智能应用交流良好做法并促进负责任的生命周期管理，从而合作培养能力和共享知识。为防止犯罪或恐怖主义团体等危险的非国家行为体获得可能具有自主性的强大人工智能系统，各国应针对人工智能系统的整个生命周期制定适当的控制措施和流程，包括管理人工智能军事应用的生命周期末端流程（即退役）。

为了加强透明度，可以成立“咨询委员会”，负责在人工智能安全和军事应用的整个生命周期提供独立专家建议和审查。行业和其他行为体应考虑建立机制，防止人工智能技术被滥用于恶意或计划外的军事目的

28 为说明从脆弱性角度处理人工智能相关风险的政策相关性，从儿童等特定弱势群体的视角审查了人工智能治理的考虑因素（见方框5）。

29 通过基于脆弱性的人工智能风险框架确定的受关注个人、群体或实体及其隐含的政策议程本身可能存在差异。在人工智能风险全球脉动调查中，还询问了专家特别担忧哪些个人、群体、社会/经济体/（生态）系统将在未来18个月内受到人工智能的损害。专家特别强调指出了边缘化社区和全球南方，以及儿童、妇女、青年、创意人员和工作容易受到自动化影响的人员（见图3）。

30 这些结果表明，在根据建议1和2就人工智能风险达成统一认知和就政策议程建立共同基础时，实现包容代表性十分重要。如果缺乏这种代表性，在构建人工智能治理的政策议程框架时，可能会忽视仍将受影响的部分人群的关切。

F. 有待克服的挑战

31 除了近期的风险和危害以外，在现行制度背景下，人工智能开发、部署和使用的演变也构成挑战，进而会影响人工智能治理战略。先进人工智能技术的发展步伐及其通用性质进一步考验着人类及时应对的能力。

32 各方争相开发和部署人工智能系统，对传统监管体系和治理制度提出了挑战。人工智能风险全球脉动调查中的大多数受访专家预计，人工智能将在未来18个月内加速发展，包括在开发（74%）以及采用和应用（89%）方面（见图4）。

33 如上文第23段所述，一些专家预计代理系统将在2025年得到部署。此外，领先的技术专家承认，许多人工智能模型仍不透明，无法充分预测或控制其输出，但其下游的负面溢出效应可能会波及全世界的其他各方。

方框4: 根据现有或潜在脆弱性对人工智能相关风险进行分类

个人

- 人的尊严、价值或能动性（例如操纵、欺骗、劝诱、量刑、剥削、歧视、平等待遇、起诉、监视、丧失人类自主性、人工智能辅助确定目标）。
- 身心健全、健康、安全和保障（例如劝诱、孤独和孤立、神经技术、致命自主武器、自动驾驶汽车、医疗诊断、获得保健的机会、与化学、生物、放射性和核系统的交互）。
- 生活机遇（例如教育、就业、住房）。
- （其他）人权和公民自由，例如无罪推定权（例如预测式警务）、公正审判权（例如累犯预测、罪责、累犯、预测和自主审判）、表达和信息自由（例如劝诱、个性化信息、信息泡沫）、隐私（例如面部识别技术）以及集会和迁徙自由（例如公共场所的跟踪技术）。

政治和社会

- 对群体的歧视和不公平待遇，包括基于个人或群体特征，例如性别，以及群体孤立和边缘化。
- 对儿童、老年人、残疾人和弱势群体的差异化影响。
- 国际和国家安全（例如自主武器、针对移民和难民的警务和边境管制、有组织犯罪、恐怖主义以及冲突扩散和升级）。
- 民主（例如选举和信任）。
- 信息完整性（例如错误信息或虚假信息、深度伪造和个性化新闻）。
- 法治（例如各机构、执法部门、司法部门的运作和对它们的信任）。
- 文化多样性和人际关系的转变（例如同质性、虚假的朋友）。
- 社会凝聚力（例如过滤气泡，对机构的信任度下降，信息来源）。
- 价值观和规范（例如伦理、道德、文化和法律）。

经济

- 权力集中。
- 技术依赖。
- 不平等的经济机会、市场准入、资源分配和调拨。
- 对人工智能利用不足。
- 过度使用人工智能或“技术解决主义”。
- 金融体系、关键基础设施和机构的稳定。
- 知识产权保护。

环境

- 过度消耗能源、水和实物资源（包括稀有矿物和其他自然资源）。

34 对由不透明算法自动作出决策和创建内容的依赖性增强，可能会破坏公平待遇和安全性。虽然人类通常仍对影响他人的流程自动化决策负有法律责任，但问责机制可能发展得不够快，无法使这种问责产生迅速和切实的效果。

35 由此出现了一个社会风险，即尽管世界上出现了日益强大的人工智能系统，而且有人作出了使用人工智能实现流程自动化的决定，但最终要为这些决定所造成损害负责的人却越来越少。为此需要灵活应变地进行治理，确保问责机制跟上人工智能加速发展的步伐。

36 如果人工智能的开发和部署速度对现有机构造成挑战，那么其广泛的覆盖面同样构成挑战。先进的人工智能是具有全球影响力的通用技术，可部署到各个领域，以多种方式影响社会，并产生广泛的政策影响。

方框5: 在人工智能治理中重点关注儿童

为确保企业和学校满足儿童的需求和保障其权利，需要采取关注儿童独特处境的全面治理方法。儿童所产生的数据占数据总量的三分之一，他们长大后将踏入充斥着人工智能的经济体和习惯于使用人工智能的世界。本方框总结了我们在深入研究期间所讨论的与这一主题相关的若干措施。
优先考虑儿童的权利和声音：

人工智能治理必须将儿童视为优先利益攸关方，重点强调儿童免受技术成瘾影响的发展权以及脱离这些技术的权利。不同于一般的以人为本办法，以儿童为中心的治理必须考虑对儿童观点、自我形象以及生活选择和机会的长期影响。将儿童纳入设计和治理过程对于确保人工智能系统安全且适合儿童使用至关重要。

研究和政策制定：
我们需要开展广泛的研究，了解人工智能如何随时间推移影响儿童的社交、认知和情感发展。这项研究应为政策讨论提供依据，并为各国制定保护措施提供指引。

保护和隐私：
不应将儿童用作人工智能实验的对象。保护儿童隐私至关重要。人工智能技术必须纳入严格的数据保护协议，并提供适合年龄的内容。

对儿童影响的评估和适合儿童的设计：
应作出规定，要求评估人工智能系统对儿童的影响，这对确保其适用性和安全性至关重要。人工智能系统的设计应顾及儿童的需求，从一开始就纳入安全和限制功能。设计选择应纳入儿童本身的意见。

数字包容和公平：
使用人工智能应赋予儿童能动性、选择权和发言权，强调采用统筹办法实现数字包容。包括提供多种语言的人工智能内容，并确保其在文化上适合非英语儿童。

国际合作和标准：
儿童参与人工智能技术的规则需要具有全球互操作性，以保护处于不同教育和发展环境中的儿童。全球标准对于处理数据的跨境流动问题和为了儿童利益合乎道德地使用人工智能至关重要。

37 人工智能与金融、劳动力市场、教育和政治制度等多个领域产生交汇，由此产生的结果和潜在影响预计会造成广泛的后果，为此需要采取全社会办法（见方框6中的例子）。现有机构必须采取统筹一致的跨部门对策，以应对人工智能的广泛社会影响。

38 人工智能开发、部署和使用的速度、广度和不确定性凸显了对人工智能采取统筹、跨领域和灵活办法的价值。在国际层面上，需要在人工智能治理领域采用跨部门和跨国界的网络化制度办法，以体现整体性视角，这种办法既要吸引利益攸关方的参与，又不能受其左右。

39 在气候变化问题上，全世界迟迟才认识到，需要通过统筹办法采取全球集体行动。就人工智能而言，仍有机会通过设计谋划来实现这一目标。

40 由于财富和决策相应集中在私营部门的少数人工智能开发者和部署者（特别是跨国公司）手中，上述挑战变得更为艰巨。这引发了另一个问题，即如何在不损害公共利益的情况下，让利益攸关方参与人工智能的治理工作。

图3: 人工智能风险全球脉动调查凸显对脆弱性的担忧

"你是否特别担忧特定的个人、群体或社会/经济体/（生态）系统可能在未来18个月内受到人工智能的损害？[自由答复文字内容]（188人对本问题作出有意义的回答）

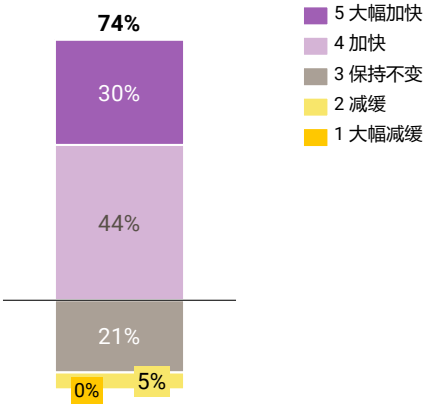
用于说明

注：秘书长技术问题特使办公室为每个答复标记了关键词。仅显示在2个以上答复中指出的关键词。字体大小与提及该关键词的答复数目成正比。作为比例参照，对本问题作出有意义回答的188名答复者中，有46人提到"全球南方"；188名答复者中有43人提到"边缘化社区"。
秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

图4:专家对人工智能技术发展的预期

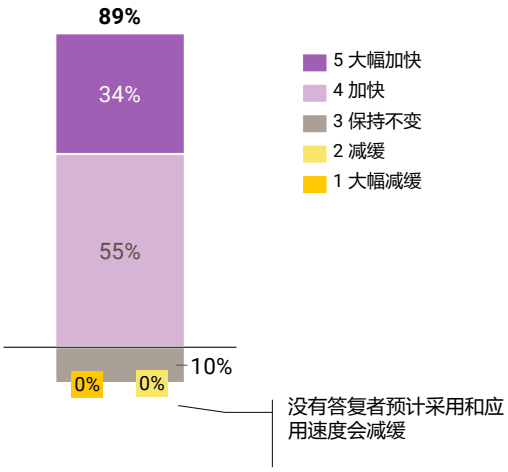
74%预计技术变革速度将加快 (30%预计将大幅加快)

"相比过去3个月，你预计未来18个月内人工智能技术变革（例如新模型的开发/发布）的速度将..."（348人答复）



89%预计采用和应用速度将加快 (34%预计将大幅加快)

"相比过去3个月，你预计未来18个月内采用和应用人工智能（例如人工智能在企业/政府中的新用途）的速度将..."（348人答复）



注：由于四舍五入，数字总和可能不等于100%。
秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

方框6: 人工智能相关的社会影响

作为更广泛互动协作的一部分，咨询机构的成员与一系列利益攸关方进行了磋商，讨论人工智能对社会造成的影响。本方框总结了在深入探讨这一主题期间提出的主要关切和潜在举措。

社会、心理和社区影响：

随着人工智能变得日益强大且影响范围更广，其开发、部署和应用将更加个性化，有可能助长疏离感和成瘾。咨询机构的一些成员认为，以个人数据训练的人工智能，及其由此扮演的主要对话者和中间人角色，可能反映了人类的一个转折点，有可能引发新的紧迫社会挑战，同时加剧现有挑战。

例如，未来的人工智能系统也许能够生成根据个人偏好量身定制的无穷无尽的高质量视频内容。这可能造成各种后果，社会孤立加剧、疏远、心理健康问题、丧失人的能动性以及对情商和社交发展的影响只是其中的一小部分。

在智能设备和互联网等技术方面，政策制定者并未充分探讨这些问题；而在人工智能方面，则几乎完全没有探讨这些问题，目前的治理框架优先关注个人而非整个社会所面临的风险。

当政策制定者考虑未来对人工智能的应对措施时，也必须权衡这些因素，并制定促进社会福祉、特别是青年福祉的政策。政府干预可营造优先促进人与人之间面对面互动的环境，使心理健康支持更容易获得，并对体育设施、公共图书馆和艺术进行更多投资。

不过，预防胜于矫治：行业开发者应设计不具个性化易上瘾功能的产品，确保产品不会损害心理健康，并促进(而不是破坏)对社会的共同归属感。技术企业应制定政策，与管控其他风险一样管控社会风险，将其作为查明和缓解人工智能产品整个生命周期风险的努力的一部分。

虚假信息和信任：

深度伪造、语音克隆和自动虚假信息活动对民主机构和选举等民主进程、民主社会和更广泛的社会信任构成了明确的严重威胁，包括通过外国信息操纵和干扰。在人工智能的强化和利用个人数据的情况下，闭环信息生态系统的发展可对社会产生深远影响，可能使社会更容易接受对他人的不容忍和暴力行为。

以虚假内容进行欺骗可能会造成损害或社会分裂，或者助长战争宣传、冲突和仇恨言论，在这种情况下，为保护代议制政府机构和流程的完整性，需要强大的验证和深度伪造检测系统，以及快速察觉和删除程序。应保护非公众人物，禁止其他人利用其肖像进行深度伪造，并用于欺诈、诽谤或其他滥用目的。带有性色彩的深度伪造对妇女和女童的影响尤其令人关切，这种行为可能是性别暴力的一种形式。

重要的初步措施包括：由私营部门行为体作出自愿承诺，例如给深度伪造内容贴上标签，或使用户能够标记恶意制作或传播的深度伪造内容，然后加以删除。然而，这些措施无法充分缓解社会风险。需要采取全球性的多利益攸关方办法，并作出具有约束力的承诺。应制定通用的内容认证和数字来源标准，在此基础上采用全球认可的办法来识别合成和经人工智能修改的图像、视频和音频。

此外，公共和私营部门行为体应依照国际标准实时分享知识，由此建立快速反应能力，在欺骗性内容或外国信息操纵和干扰有机会疯传之前立即将其删除。不过，这些流程应纳入保障措施，确保它们不被操纵或滥用于助长审查。

在开展这些行动的同时，还应采取预防措施，增强社会应对人工智能驱动的虚假信息和宣传的韧性，例如开展提高公众认识活动，宣传人工智能在破坏信息完整性方面的潜力。会员国还应推动提高媒体素养和数字素养运动，支持事实核查举措，并投资建设外国信息操纵和干扰捍卫者社区的能力。

2. 全球治理的必要性

41 今天，在人工智能方面存在全球治理赤字。尽管对伦理和原则进行了大量讨论，但各种规范、机构和举措仍处于初期阶段，各方面都存在差距。问责制和伤害补救措施往往因其缺失而引人注目。遵守取决于自愿。高层的言论，正在开发、部署和使用的系统，以及安全和包容性所需的条件之间存在着根本的脱节。正如我们在中期报告中指出的那样，人工智能治理至关重要——不仅是为了应对挑战和风险，也是为了确保我们以不让任何人掉队的方式利用人工智能的潜力。[10]

42 全球治理的必要性尤其无可辩驳。人工智能的原材料，从关键矿物到训练数据，都源自世界各地。跨国界部署的通用人工智能在全球范围内产生了多方面的应用。人工智能的加速发展在全球范围内集中了权力和财富，具有地缘政治和地缘经济影响。此外，目前还没有人了解人工智能的所有内部工作原理，足以完全控制其输出或预测其演变。决策者也不会因为开发、部署或使用他们不了解的系统而被追究责任。与此同时，这些决策所产生的负面溢出效应和下游影响也可能是全球性的。

43 尽管人工智能的影响范围遍及全球，但国家和地区的体制结构和法规却止步于实际边界。这就削弱了任何单一国家管理会造成越境损害的人工智能下游应用的能力，或解决人工智能开发和使用背后的计算基础设施、训练数据流和能源来源等复杂的跨境供应链上的问题的能力。领先的人工智能公司通常比大多数单独行动的国家对下游应用具有更直接的影响力(通过上游风险缓解)。

44 这种技术的开发、部署和使用不能仅仅由市场决定。国家政府和区域组织将是至关重要的。但除了考虑公平、获取性、预防和补救损害之外，技术本身的性质——结构和应用上的跨界性——需要采取全球多部门办法。如果没有一个能让利益攸关方参与其中的全球包容性框架，考虑到目前的竞争态势，政府和企业都可能会偷工减料或将自身利益放在首位。

45 因此，人工智能带来了全球性的挑战和机遇，需要采取横向贯穿政治、经济、社会、伦理、人权、技术、环境和其他领域的整体性全球方法。这种方法可以将不断演变的举措拼凑成一个连贯、可互操作的整体，以国际法为基础，可适应不同情况和不同时期。

46 在地缘政治和地缘经济竞争影响力和市场的时代，对人工智能全球治理的需求应运而生。然而，既要应对人工智能的风险，又要使机遇得到公平利用，这需要全球采取协调一致的行动。不断扩大的数字鸿沟可能会将人工智能的好处局限于少数国家和个人，而风险和危害会影响许多人，特别是弱势群体。

A. 国际人工智能治理的指导原则和职能

47 在我们的中期报告中概述了五项原则，这些原则应指导新的国际人工智能治理机构的组建：

- **指导原则1**：人工智能应由所有各方以包容方式共同治理，并造福所有人
- **指导原则2**：人工智能治理必须以公共利益为导向
- **指导原则3**：人工智能治理应与数据治理和促进数据共享空间协同进行
- **指导原则4**：人工智能治理必须具有普遍性、网络化并植根于多利益攸关方的适应性协作
- **指导原则5**：人工智能治理应立足于《联合国宪章》、国际人权法和可持续发展目标等其他商定的国际承诺

48 方框7总结了对这些原则的反馈意见，其中强调人权的重要性，以及需要进一步明确有效实施指导原则，包括关于数据治理的指导原则。我们面临的挑战是，如何解决关于确保采取行动加强包容性以及使边缘群体得到代表的问题。

10 见 https://un.org/ai-advisory-body.

方框7: 对指导原则的反馈

强调基于人权的人工智能治理:

根据高级别咨询机构在其中期报告发布后进行的广泛咨询，指导原则 5（人工智能治理应立足于《联合国宪章》、国际人权法和其他商定的国际承诺）获得了包括政府、民间社会、技术界、学术界和私营部门在内的所有利益攸关方的最有力支持。这包括尊重、促进和实现人权并起诉侵犯人权的行为，以及2024年3月一致通过的大会关于抓住安全、可靠和值得信赖的人工智能系统为可持续发展带来的机遇的第A/78/265号决议。

咨询机构在审议中深信，为了减轻人工智能的风险和危害，处理新用例，并确保人工智能能够真正造福全人类，不让任何人掉队，必须将人权置于人工智能治理的中心，确保各管辖区基于权利的问责制。这一对人权的基本承诺贯穿各领域，适用于最后报告中提出的所有建议。

具体的实施机制和明确的指导方针:

许多利益攸关方强调，需要制定详细的行动计划和明确的指导方针，以确保有效实施咨询机构的国际人工智能治理指导原则。政府实体建议制定明确的建议，以界定和确保公共利益，并建立公众参与和监督机制。私营部门实体经常强调，需要制定明确的政策和利用现有的监管框架，以保持人工智能市场的竞争性和创新性。许多国际组织和民间社会组织还呼吁建立灵活的治理系统，以便及时应对不断发展的技术。一些国家特别要求建立一个不仅具有协调能力而且具有“肌肉和牙齿”的新实体。

建立追究主要行为体责任的机制:

一个共同关切的问题是对歧视性、有偏见和其他有害的人工智能的问责，并建议建立机制，以确保对损害进行问责和补救，并解决技术能力和市场力量集中的问题。许多组织强调，必须解决权力不受制约的问题，确保消费者权利和公平竞争。学术机构承认指导原则在普遍性和包容性方面具有优势，但建议在利益攸关方参与方面加以改进。私营部门行为体强调负责任地使用人工智能，同时打破获取方面的障碍。

在人工智能数据治理方面发挥更具体的作用:

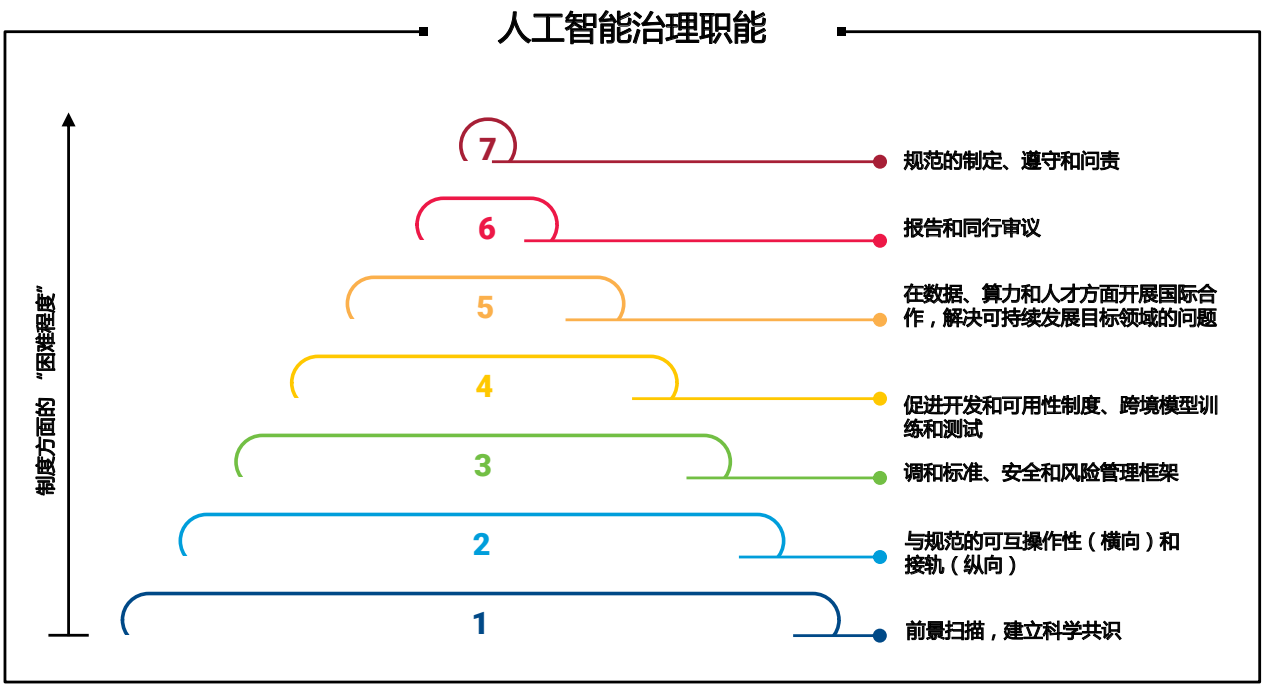
在多次咨询中都提到了缺乏数据治理系统的问题，利益攸关方表示联合国是就数据治理进行对话的自然场所。各国政府强调，需要建立强有力的数据治理框架，优先考虑隐私、数据保护和公平使用数据，倡导制定国际指导方针，以管理人工智能开发中的数据复杂性。要求通过透明和包容性的进程制定这些框架，纳入同意和隐私等伦理考虑因素。

学术界强调，数据治理应作为短期优先事项处理。私营部门实体指出，数据治理措施应补充人工智能治理，强调全面的隐私法和负责任的人工智能使用。国际组织和民间社会组织强调，人工智能训练数据的治理应保护消费者权益，并通过对人工智能训练数据的非排他性访问来支持人工智能开发者之间的公平竞争，强调了对具体和可付诸行动的数据治理措施的呼吁。联合国被确定为应对这些治理挑战和缩小资源差距的一个关键场所。

49 在我们的中期报告中还提出了可在国际层面推行的若干机构职能(见图5)。反馈意见在很大程度上证实了在全球层面需要这些职能，同时呼吁增加与数据和人工智能治理相关的补充职能，以将指导原则3 (人工智能治理应与数据治理和促进数据共享空间协同进行)付诸实践。

50 关于监测、核查、报告、合规、问责稳定、响应和执行等制度上“更难”的人工智能治理职能，反馈意见指出，首先，在将这些职能制度化之前，需要承担国际条约义务，并且在将人工智能作为一项技术进行治理方面，尚未提出将这些职能制度化的理由。

图5: 在国际层面提出的人工智能治理职能



51 并非所有职能都需要完全由联合国履行。但是，如果要将各种规范和机构的拼凑转变为一个安全网，促进和支持造福全人类的可持续创新，就需要对科学有统一认知，对规则和标准有共同基础，而我们正是通过这些规则和标准来评估治理是否实现了目标。

52 在咨询过程中，我们听到一些呼吁，要求对国际上治理人工智能的现有努力和新出现的努力进行更详细的全局分析，并对公平、有效和高效的人工智能国际治理需要弥补的差距进行分析。

B. 新兴的国际人工智能治理格局

53 可以肯定的是，目前关注人工智能治理的文件和对话并不缺乏。各国政府、公司和财团以及区域和国际组织采用了数百项指南、框架和原则。数十个论坛汇集了不同的行为体，从既定的政府间进程和专家机构到特设的多利益攸关方举措。与此同时，在国家层面和区域层面也制定了现有的和新的规章制度

54 各国政府提出的国际举措越来越多(见图6)。这些新出现的举措越来越多地在国际层面采用横向方法进行人工智能治理，包括全面而非特定领域处理人工智能问题的原则、宣言、声明和其他发布。自2023年以来，这些举措急剧增加，这得益于ChatGPT于2022年11月发布后问世的多个通用人工智能大型语言模型。

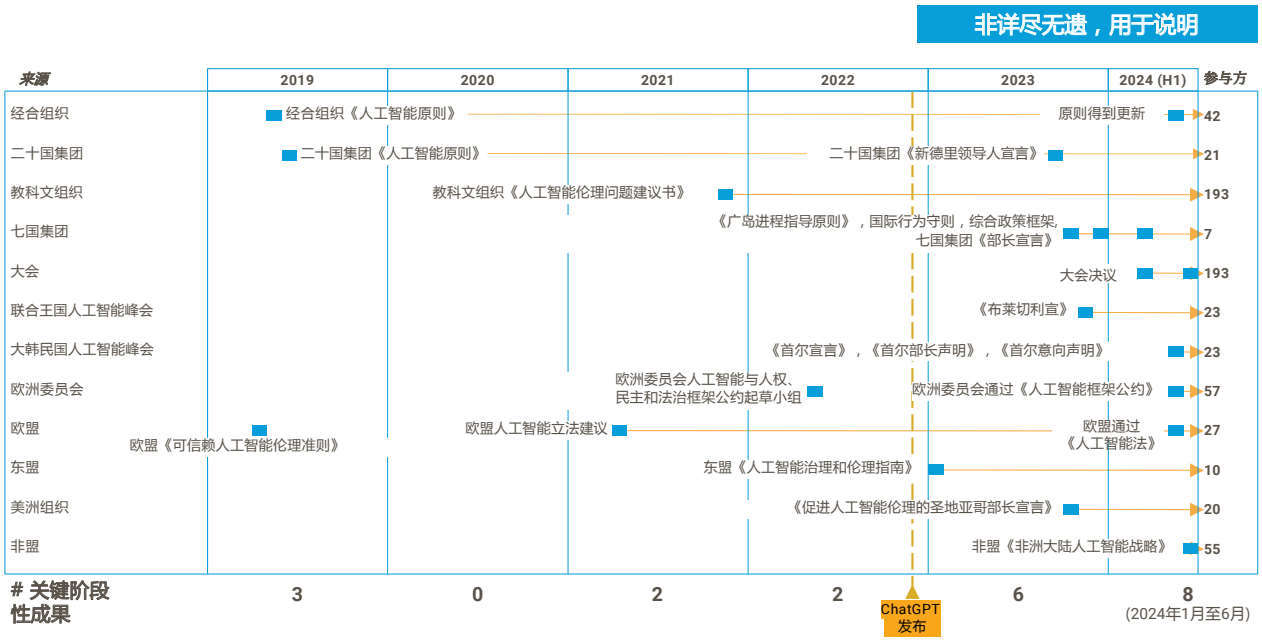
55 与此同时，还制定并发布了人工智能行业标准，供国际社会采用。其他多利益攸关方举措也寻求弥合公共部门和私营部门之间的差距，包括在互联网治理论坛等讨论场所。

56 图7提供了对人工智能治理举措和行业标准的一些来源的调查，按其地理范围和包容性情况绘制(在列举这一最近的工作时，我们感谢学术界、民间社会和专业机构多年来的努力)。

57 相关的区域和区域间诸边举措的例子包括由非洲联盟、人工智能峰会的各个主办方、东南亚国家联盟、欧洲委员会、欧洲联盟、七国集团、二十国集团、全球人工智能伙伴关系、美洲国家组织（美洲组织）和经济合作与发展组织(经合组织)等牵头的举措。

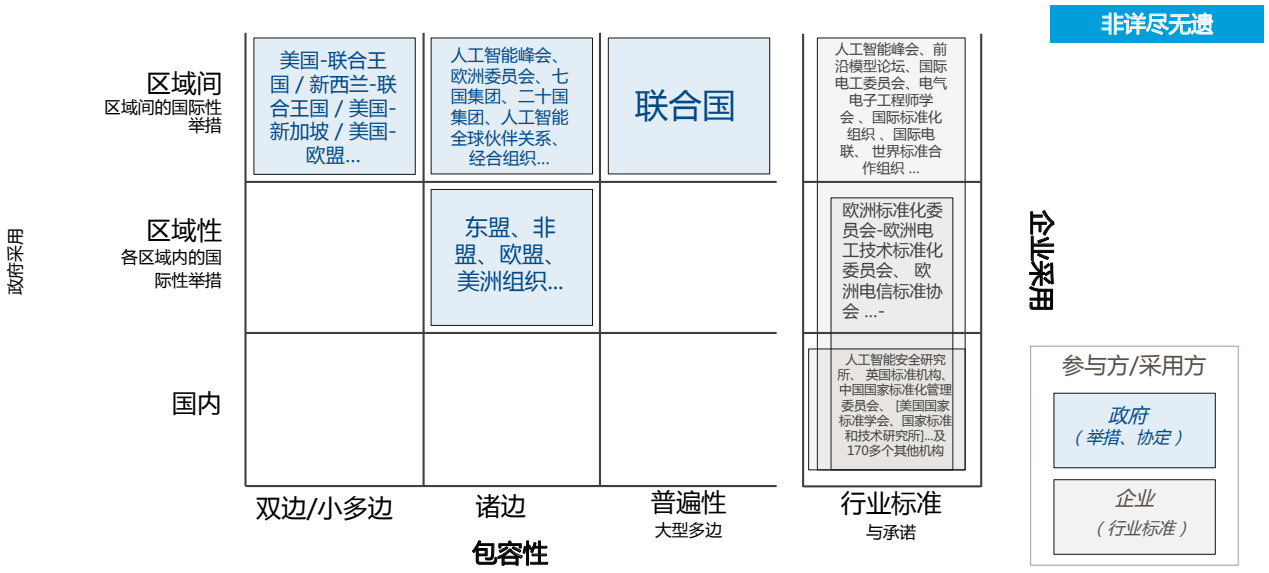
58 我们对当前治理安排的分析可能在几个月内就会过时。然而，它有助于说明当前和新出现的国际人工智能治理举措与我们关于组建新的人工智能全球治理机构的指导原则,包括原则1(人工智能应由所有各方以包容方式共同治理，并造福所有人）之间的关系。上述

图6: 2019至2024年(上半年)区域间和区域性人工智能治理举措，关键阶段性成果



缩写：东盟： 东南亚国家联盟；非盟：非洲联盟；欧盟：欧洲联盟；美洲组织：美洲国家组织；经合组织：经济合作与发展组织；教科文组织：联合国教育、科学及文化组织。

图7: 专门针对人工智能的治理举措的来源



缩写：东盟： 东南亚国家联盟；非盟：非洲联盟；欧盟：欧洲联盟；美洲组织：美洲国家组织；国际电联： 国际电信联盟；经合组织：经济合作与发展组织 。

3. 全球人工智能治理差距

多个国家、区域、多利益攸关方和其他举措取得了有意义的成果，并为我们的工作提供了参考信息；其中许多代表为我们的审议工作提供了书面材料或参加了我们的咨询。

59 上述多个国家、区域、多利益攸关方和其他举措取得了有意义的成果，并为我们的工作提供了参考信息；其中许多代表为我们的审议工作提供了书面材料或参加了我们的咨询。

60 然而，除了联合国提出的几项举措外[11]，没有一项举措能够真正实现全球性。人工智能治理在国际层面上的代表性差距是一个问题，因为该技术是全球性的，其影响将是全面的。

61 举措和机构之间各自的协调差距有可能使世界分裂成互不关联、互不兼容的人工智能治理制度。

62 此外，执行和问责方面的差距削弱了国家、私营部门、民间社会、学术界、技术界将承诺转化为具体成果的能力，无论这些承诺多么具有代表性。

A. 代表性差距

63 我们对跨区域的各种非联合国人工智能治理举措的分析表明，大多数举措在政府间层面上并不具有充分的代表性。

64 许多举措将世界的整个地区排除在外。如图8所示，在7个成员存在重叠的非联合国、诸边、区域间人工智能举措中，7个国家加入了所有举措，但有整整118个国家没有加入任何举措(主要是全球南方国家，即使是领先的人工智能国家，其代表性也不均衡；见图12)。

65 在治理的早期阶段，存在一定程度的试验、围绕规范的竞争以及对新技术的不同适应程度，选择性是可以理解的。但随着国际人工智能治理的成熟，全球代表性在公平和有效性方面变得更加重要。

66 除了现有努力的非包容性之外，在聚焦形成对人工智能的统一科学认知的国家和区域举措中也存在代表性差距。这些代表性差距可能体现在关于如何确定评估范围、提供资源和开展评估的决策过程中。

67 公平性要求更多的声音在有关如何管理影响我们所有人的技术的决策中发挥有意义的作用，并认识到许多群体历来被排除在这些对话之外。重大举措议程中作为某些区域优先事项的专题相对较少，这表明代表性不足造成了不平衡[12]。

68 人工智能治理制度必须横跨全球才能有效——有效建立信任，避免“人工智能军备竞赛”或在安全和权利方面的逐底竞争，有效应对人工智能跨界特征带来的挑战，促进学习，鼓励互操作性，分享人工智能的益处[13]。此外，纳入不同观点——包括不一致的观点——对预测威胁和调整具有创造性和适应性的对策也有好处。

69 选择性多边主义限制了参与关键议程制定、关系建立和信息共享进程的国家范围，从而可能限制其自身目标的实现。这些目标包括新出现的人工智能治理方法的兼容性、全球人工智能安全，以及在全球层面对人工智能科学的统一认知(请参阅建议1、2和3，了解如何使全球方法特别有效)。

70 大会于2024年到目前为止通过的两项关于人工智能的决议[14]表明，主要人工智能国家承认需要解决国际人工智能治理方面的代表性差距问题，而联合国可以成为在这方面将世界聚集在一起的论坛。

11 联合国教育、科学及文化组织 (教科文组织)《人工智能伦理问题建议书》(2021年)和大会关于人工智能的两项决议。

12 例如，人工智能训练数据集的管理、计算能力的获取、人工智能能力的发展、与边缘化群体歧视有关的人工智能相关风险以及在武装冲突中使用人工智能（见附件E，人工智能风险全球脉动调查结果显示，西欧和其他国家组与非西欧和其他国家组的答复者对风险的看法不同）。许多国家和边缘化群体也被排除在人工智能的惠益之外，或可能不成比例地遭受其危害。公平性要求采取多样化和包容性的办法，考虑到所有区域的意见，在减少风险的同时平均分配机会。

13 如果划定了红线——也许类似于禁止克隆人——那么只有在全球都认同这一准则并监督其遵守情况的前提下才可得到执行。矛盾的是，情况依然如此——尽管在当前的模式下，特定人工智能系统的成本在下降，但高级人工智能系统（可以说是最需要控制的）的成本却在上升。

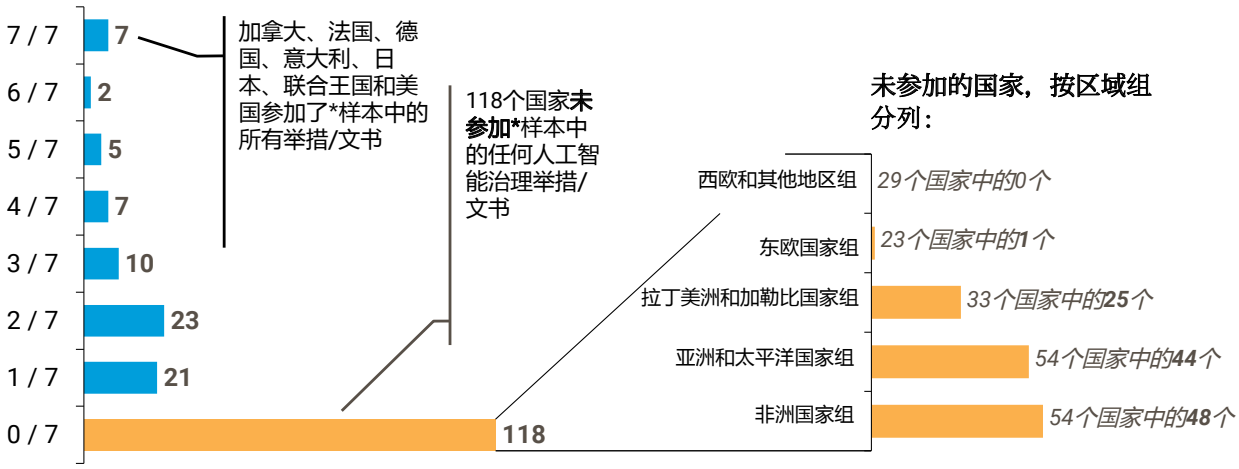
14 第78/265 号决议（抓住安全、可靠和值得信赖的人工智能系统为可持续发展带来的机遇)和第78/311 号决议（加强人工智能能力建设方面的国际合作)。

15 各种多边举措，包括经合组织人工智能原则、七国集团广岛人工智能进程和欧洲委员会人工智能框架公约，都向最初发起国以外的支持者或追随者开放。然而，这种开放性可能无法以跟上全球人工智能加速扩散所需的速度和广度提供代表性和合法性。与此同时，国际人工智能治理流程中的代表性差距仍然存在，决策权集中在少数国家和公司手中。

图8: 七项非联合国的国际人工智能治理举措中的代表性

样本: 经合组织《人工智能原则》(2019)、二十国集团《人工智能原则》(2019) 、欧洲委员会人工智能公约起草小组(2022-2024) 、人工智能全球伙伴关系《部长宣言》 (2022) 、七国集团《部长声明》 (2023) 、《布莱切利宣言》(2023)和《首尔部长宣言》(2024)。

仅区域间举措，不包括区域性举措



*根据相关政府间文书的认可。仅因拥有欧洲联盟或非洲联盟成员资格并不会被视为诸边举措的参与国。 *缩写*：经合组织，经济合作与发展组织。

71 2024年9月的全球数字契约和2025年的信息社会世界峰会论坛提供了另外两个政策窗口，可以将一套具有全球代表性的人工智能治理流程制度化，以解决代表性方面的差距。[15]

B. 协调方面的差距

72 人工智能治理举措的不断出现和演变，并不能保证它们能为人类有效合作。相反，出现了协调方面的差距。有选择性的诸边举措(见图8)与其他区域举措之间无法保证有效对接，区域之间存在不兼容的风险。

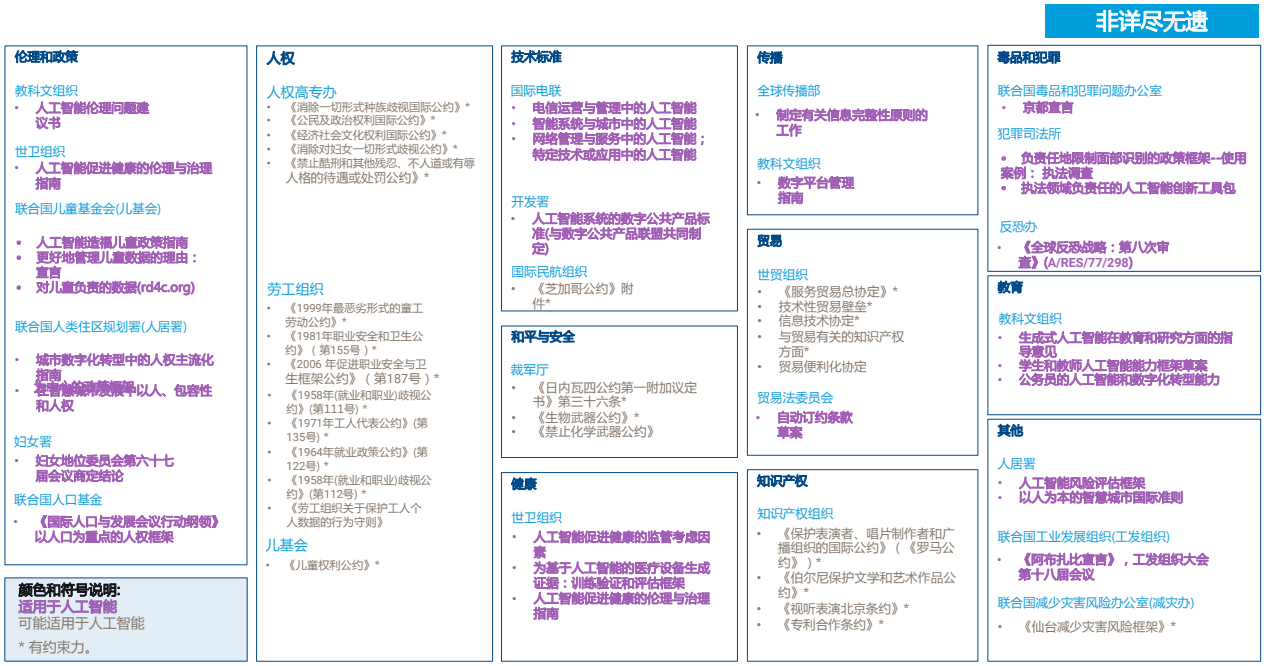
73 所有国际标准制定组织(见图7)、国际科学研究举措或人工智能能力建设举措也没有全球机制来相互协调，这破坏了各种方法的互操作性，导致各自为政。在某些情况下，各种全球以下各级举措之间由此产生的协调方面的差距最好在全球层面解决。

74 此外，在联合国系统内也出现了一系列协调方面的差距，反映在与人工智能有关的各种联合国文件和举措中。图9显示了可能适用于人工智能的特定领域的26项与联合国有关的文书，其中25项文书具有约束力，需要对其进行解释，因为它们与人工智能有关。联合国和相关组织的另外32份领域级文件专门关注人工智能，但都不具有约束力。在一些情况下[16]，这些文件和举措可用于应对特定领域的人工智能风险和把握其惠益。

75 活动水平表明了人工智能对联合国各方案的重要性。随着人工智能对社会各方面的影响不断扩大，要求联合国系统各部门采取行动的呼声将越来越高，包括通过具有约束力的规范。这也显示了应对措施的临时性，这些措施主要是在特定领域有机发展起来的，没有一个总体战略。由此产生的协调方面的差距导致工作重叠，阻碍了互操作性和影响力。

16 2024年2月，联合国系统行政首长协调理事会对57个联合国实体进行了一项调查，报告了50份有关人工智能治理的文件；57个实体中有44个作出答复，包括：拉丁美洲和加勒比经济委员会；亚洲及太平洋经济社会委员会；西亚经济社会委员会；联合国粮食及农业组织 (粮农组织)；国际原子能机构 (原子能机构)；国际民用航空组织(国际民航组织)；国际农业发展基金；劳工组织；国际货币基金组织；国际移民组织；国际贸易中心；国际电信联盟 (国际电联)；联合国促进性别平等和增强妇女权能署(妇女署)；联合国艾滋病毒/艾滋病联合规划署(艾滋病署)；联合国贸易和发展会议(贸发会议)； 经济和社会事务部；全球传播部；秘书长办公厅； 人道主义事务协调厅；联合国人权事务高级专员办事处；反恐怖主义办公室；裁军事务厅；信息和通信技术厅；秘书长技术问题特使办公室；联合国开发计划署(开发署)；联合国减少灾害风险办公室；联合国环境规划署；教科文组织；联合国气候变化框架公约；联合国人口基金；联合国难民事务高级专员；联合国儿童基金会；联合国区域间犯罪和司法研究所；联合国工业发展组织；联合国毒品和犯罪问题办公室/联合国维也纳办事处；联合国项目事务署；联合国近东巴勒斯坦难民救济和工程处；联合国大学；联合国志愿人员组织；世界贸易组织；万国邮政联盟；世界银行集团；世界粮食计划署；世界卫生组织(世卫组织)；世界知识产权组织 (知识产权组织)。见《联合国系统人工智能治理白皮书：对联合国系统适用于人工智能治理的机构模式、职能和现有国际规范框架的分析》(可查阅(可查阅 https://unsceb.org/united-nations-system-white-paper-ai-governance)

图9:联合国和相关组织与人工智能治理相关的某些文件



信息来源： “联合国系统人工智能治理白皮书： 分析联合国系统的体制模式、职能和适用于人工智能治理的现有国际规范框架”，2024年2月28日

76 方法的数量和多样性表明，联合国系统正在应对一个新出现的问题。通过适当的协调，并结合采取整体方法的进程，这些努力可以为特定领域的包容性国际人工智能治理提供一条高效、可持续的途径。这可以在卫生、教育、技术标准和伦理等领域产生有意义的、协调一致的影响，而不仅仅是助力这一日益发展的领域中举措和机构数量的激增。国际法，包括国际人权法，为所有与人工智能有关的努力提供了共同的规范基础，从而促进了协调和一致性。

77 虽然许多联合国实体的工作都涉及人工智能治理问题，但它们的具体任务意味着没有一个实体能够全面地开展这方面的工作，它们指定的政府协调部门也同样具有专业性。这就限制了现有联合国实体独自应对人工智能对全球多方面影响的能力。[17]在国家和区域层面，人工智能安全研究所或人工智能办公室等新机构[18]正在采取适当的横向方法来补齐这些差距。

C. 执行方面的差距

78 然而，仅有代表性和协调还不够。要确保善治承诺在实践中转化为具体成果，还需要行动和后续进程。需要采取更多措施来确保问责。同伴压力和同伴学习是能够促进问责的两个要素

79 让私营部门参与，对于有意义的问责和伤害补救同样重要。联合国在《联合国工商企业与人权指导原则》中就有这方面的经验。同样，我们也需要民间社会和科学专家的大力参与，确保政府和私营公司诚意践行其承诺和主张。

80 国家内部和国家之间在利用人工智能的惠益促进公共利益方面缺少推动因素，这是执行方面的一大差距。许多国家已经制定了国家战略，以促进人工智能相关的基础设施和人才的发展，一些国际援助举措也正在出现。然而[19]，这些举措的网络和资源都不足。

81 在全球层面，将国家和区域能力发展举措联系起来，汇集资源支持那些被排除在这些努力之外的国家，有助于确保没有一个国家在分享与人工智能相关的机遇方面掉队。执行方面的另一大差距是，尽管存在一些数字能力供资机制，但缺乏用于人工智能能力建设的专项基金(方框8)。

17 如教育、科学和文化部(教科文组织)、电信或信通技术部(国际电联)、工业部(联合国工业发展组织)、劳工部(劳工组织)等。
18 包括加拿大、日本、新加坡、大韩民国、联合王国、美国和欧洲联盟建立的机构。
19 国家层面的工作可以继续采用诊断工具，如教科文组织的人工智能准备情况评估方法，以帮助确定国家层面的差距，并由国际网络帮助弥补这些差距。

方框8: 人工智能能力的全球融资缺口

咨询机构认为，现有的全球人工智能能力建设基金在规模和任务上都不足以为消除人工智能鸿沟所需的大量投资提供资金。

据指示性估算，每年所需的金额在3.5亿至10亿美元之间a，包括私营部门的实物捐助，其任务是针对所有人工智能推动因素的人工智能能力，包括人才、算力、训练数据、模型开发和跨学科应用合作。

现有多边供资机制的例子包括：

a)可持续发展目标联合基金：

该基金范围广泛，涵盖每一个可持续发展目标以及应急响应。支持国家层面的举措，为各国推进可持续发展目标提供综合的联合国政策和战略融资支持。该基金帮助联合国提供和促进可持续发展目标的融资和规划。自2017年以来，30个参与的联合国实体总共收到了2.23亿美元。基金不直接资助国家政府、社区或实体，也不资助跨境活动。

2023年，约有16个捐助方向该基金捐款，捐助总额为5 770万美元，2024年估计为5 880万美元。私营部门自2017年以来捐款83 155美元，2023年和2024年迄今为止没有捐款。

基金大部分(60%)的资金用于5个可持续发展目标的行动：可持续发展目标2 (零饥饿)、5 (性别平等)、7 (负担得起的清洁能源)、9 (工业、创新和基础设施)和17 (伙伴关系)。

基金的政策数字化转型流(2023年启动)已为一个项目提供了25万美元的资金，由国际电信联盟 (国际电联)和联合国开发计划署 (开发署)各支付一半。截至2023财年末，其交付率为2.27%。数字化转型活动仅占基金活动的一小部分，通常与其他可持续发展目标有关(例如，支持服务提供的连通性和数字基础设施，例如在小岛屿发展中国家)。

b)世界银行数字促进发展伙伴关系：

基金支持各国发展和实施数字化转型，重点关注宽带基础设施、接入和使用、数字公共基础设施以及数据生产、可访问性和使用。到2022年底，该伙伴关系已在80多个国家投资了107亿美元。

该伙伴关系包括一个与网络安全相关的多捐助方信托基金(爱沙尼亚、德国、日本和荷兰王国)，以支持国家网络安全能力发展。

a 不到2023年私营部门人工智能年度投资估计数的1%。

4. 加强全球合作

82 在概述了全球治理赤字之后，我们现在来谈谈解决近期需优先弥补的差距的建议。

83 我们的建议提出了一个整体愿景，即采用全球网络化、敏捷灵活的办法治理人工智能治理以造福人类，其中涵盖统一认知、共同基础和共享惠益，以增强代表性、促进协调并加强实施(见图10)。只有通过这种包容和全面的人工智能治理办法，才能应对和把握人工智能在全球范围内带来的多方面和不断变化的挑战和机遇，促进国际稳定和公平发展。

84 在我们中期报告(见第47段)所列原则的指导下，我们的建议旨在填补空白，并使快速崛起的国际人工智能治理对策和举措的生态系统保持一致，帮助避免各自为政和错失良机。为了有效地支持这些措施，并与其他机构建立有效的伙伴关系，我们提议建立一个轻便、灵活的结构，以体现协调一致的努力：在联合国秘书处设立一个人工智能办公室，靠近秘书长，作为“粘合剂”将这些其他部分粘合在一起。

85 联合国远非完美无缺，但其独特的包容性所产生的合法性，加上其在国际法(包括国际人权法)中具有约束力的规范性基础，为以公平、有效和高效的方式治理人工智能以造福和保护人类带来了希望。[20]

图10: 建议及其如何解决全球人工智能治理差距问题的概览

目的	加强代表性	促进协调	加强实施
共同理解 人工智能国际科学小组	✓	✓	
共同点 关于人工智能治理的政策对话 人工智能标准交流	✓	✓	(✓)
共同利益 能力发展网络 全球人工智能基金 全球人工智能数据框架	✓	✓	✓
协调一致的努力 秘书处内的人工智能办公室	就有关大赦国际的事项向秘书长提出建议，努力促进在联合国系统内发出一致的声音，使各国和利益攸关方参与进来，与其他进程和机构建立伙伴关系和联系，并根据需要支持其他建议。		

20 它还应该具有包容性、凝聚力，并加强全球和平与安全。

A. 统一认知

86 管理人工智能的全球方法首先要对其能力、机遇、风险和不确定性有统一认知。

87 人工智能领域发展迅速，产生了大量信息，使人难以分辨炒作与现实。这可能会加剧混乱，阻碍共识，并在牺牲政策制定者、公民社会和公众利益的情况下为大型人工智能公司谋利。

88 此外，缺乏国际科学合作和信息交流会滋生全球误解，破坏国际信任。

89 会员国需要及时、公正和可靠的人工智能科学知识和信息，以在全球范围内建立共同的基础认识，并平衡拥有昂贵人工智能实验室的企业与世界其他人之间的信息不对称，包括通过人工智能公司与更广泛的人工智能社区之间的信息共享。

90 这在全球层面最为有效，能够对全球公益事业进行联合投资，并在原本分散和重复的工作中开展公共利益合作。

国际人工智能科学小组

建议1：国际人工智能科学小组

我们建议成立一个独立的国际人工智能科学小组，由该领域的多学科专家组成，以个人身份自愿任职。该小组将在拟议的联合国人工智能办公室和联合国其他相关机构的支持下，与其他相关国际组织合作，其任务包括：

a. 发布关于人工智能相关能力、机遇、风险和不确定性调查的年度报告，确定关于技术趋势的科学共识领域以及需要开展更多研究的领域；

b. 编制季度专题研究摘要，探讨人工智能有助于实现可持续发展目标的领域，重点关注可能扶持不足的公共利益领域；

c. 发布关于新兴问题的特别报告，特别是在治理领域出现的新风险或重大差距

91 这样的机构是有先例的。一些例子包括联合国原子辐射影响问题科学委员会、生物多样性和生态系统服务政府间科学与政策平台（生物多样性平台）、南极研究科学委员会和政府间气候变化专门委员会（气专委）。

92 这些模式以其对影响各行各业和全球人口的复杂、普遍问题的系统解决方法而著称。然而，虽然它们可以提供启发，但没有一个完全适合评估人工智能技术，因此不应直接照搬。相反，需要有量身定制的方法。

93 借鉴这些先例，一个独立、国际性、多学科的人工智能科学小组可以整理和促进前沿研究，从公正可信的来源为寻求人工智能技术或其应用的科学观点的人提供信息。方框9中讨论的当前关于开放式与封闭式人工智能系统的辩论，就是该小组可以作出贡献的一个例子。

94 在联合国主持下设立的科学小组将有广泛的重点，全面涵盖一系列优先事项。这可能包括获取人工智能相关机会的专业知识——促进“深入研究”可持续发展目标的应用领域，如医疗保健、能源、教育、金融、农业、气候、贸易和就业。

95 风险评估还可以借鉴其他人工智能研究举措的工作，联合国为研究人员交流关于“最先进技术”的想法提供了一个独特且值得信赖的“安全港”。国际法，包括人权法，将为定义相关风险提供指南。联合国主持的小组通过汇集可能本不参与或不被包括在内的国家或公司的知识，可以帮助纠正误解并在全球范围内加强信任。

96 这样一个科学小组不一定要开展自己的研究，但可以成为网络化行动的催化剂。小组可以为受众汇总[22]、提炼和翻译人工智能的发展动态，突出潜在的用例。小组将减少信息不对称，有助于避免投资方向错误，并保持信息在全球专家网络中流动。

97 小组将有三类主要受众：

a. 首先是全球科学界。人工智能基础研究向私营公司的转移[23]，部分原因是计算能力成本的推动，这导致人们担心此类研究可能受到经济利益的不当驱动。科学小组可以鼓励全球公共机构开展更多关注公共利益的研究。

22 特别是该小组可以利用已经在运作的现有部门或区域小组。
23 科学小组还可以向更广泛的受众，包括民间社会和一般公众开展外联活动。

b. 其次，定期独立评估将为会员国、决策者和本报告建议的其他进程提供信息。世界专家的年度风险调查将有助于制定建议2中提出的人工智能治理对话议程。这一高水平的报告将为建议3中提议的标准制定以及建议4中提议的能力发展网络提供信息。[24]

c. 再次，通过其公开报告，科学小组可以成为向公众提供高质量信息的公正来源。

98 通过联合国所独有的覆盖全球的网络，可以在最广泛的基础上达成共识，以适合各种社会经济和地理环境的方式提供研究结果。因此，科学小组可以激活联合国，使其成为一个可靠的平台，促进包容性的网络化、多学科利益攸关方的认知。

99 科学小组的初始任期可为3至5年(经秘书长审查后可延长)，并可根据以下规定行使职能：

a. 科学小组一开始可以有30-50名成员，通过会员国提名和自我提名相结合的方式任命，这与该咨询机构本身的设立方式类似。科学小组应注重各学科的科学专门知识，需要确保各区域和性别的代表性多样化，并反映人工智能的跨学科性质。在3-5年的总任期内，成员可定期轮换。

b. 科学小组将举行虚拟会议（并召开现场出席的全体会议，可能每年两次）。会议可在联合国相关实体所在城市轮流举行，包括在全球南方的地点。应鼓励科学小组成立专题工作组，根据需要增加成员，并与学术伙伴网络建立联系。科学小组可探讨邀请联合国相关实体参加这些工作组。

c. 科学小组将独立运作，特别是在其调查结果和结论方面，并得到来自拟议的人工智能办公室和国际电联和联合国教育、科学及文化组织 (教科文组织)等相关联合国机构的联合国系统小组的支持。

d. 科学小组应与其他国际机构（如经合组织和全球人工智能伙伴关系）、其他相关进程[25]（如联合王国最近委托编写的关于先进人工智能风险的科学报告）和相关区域组织合作，并在其基础上开展研究工作。

e. 一个指导委员会将制定研究议程，确保包容各方面观点并纳入伦理考虑因素，监督资源的分配，促进与学术机构和其他利益攸关方网络的合作，并审查科学小组的活动和成果。

100 通过利用联合国独特的召集力和包容全球各利益攸关方群体的覆盖面，国际科学小组可以提供值得信赖的科学合作进程和产出，并纠正信息不对称，以解决第66和73段中确定的代表性差距和协调方面的差距，从而促进公平和有效的国际人工智能治理。

方框9: 开放式与封闭式人工智能系统

我们在咨询中讨论的主题之一是关于开放式与封闭式人工智能系统的持续辩论。在不同程度上开放的人工智能系统通常被称为 “开源人工智能”，但这与开源软件（代码）相比有些名不副实。重要的是要认识到，人工智能系统的开放性是一个范围，而不是单一属性。

一篇文章解释说： “一个完全封闭的人工智能系统只对特定的群体开放。这个群体可以是一家人工智能开发公司，也可以是公司内部主要用于内部研发目的的特定团队。另一方面，更开放的系统可能允许公众访问或提供某些部分，如数据、代码或模型特征，以促进外部的人工智能开发。” [a]

生成式人工智能领域的开源人工智能系统既存在风险，也存在机遇。公司经常以“人工智能安全”为由不公开系统技术指标，这反映出业内在开放式与封闭式人工智能系统之间持续存在的矛盾。争论通常围绕两个极端展开：完全开放，即共享所有模型组成部分和数据集；部分开放，即只公开模型权重。

开源人工智能系统鼓励创新，通常需要公共资金。在开放的极端，当底层代码免费提供时，世界各地的开发人员可以试验，改进和创建新的应用程序。这就营造了一种协作环境，人们可以随时分享想法和专业知识。一些行业领袖认为，这种开放性对创新和经济增长至关重要。

a Angela Luna, “The open or closed AI dilemma”, 2 May 2024. 可查阅 https://bipartisanpolicy.org/blog/the-open-or-closed-ai-dilemma.

24 活跃在这一领域的联合国实体清单见图9。

25 International Scientific Report on the Safety of Advanced AI: Interim Report. 可查阅 https://gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai.

方框9: 开放式与封闭式人工智能系统

不过，在大多数情况下，开源人工智能模型是以应用程序接口的形式提供的。在这种情况下，原始代码不会共享，原始权重也不会改变，模型更新后就会成为新的模型。

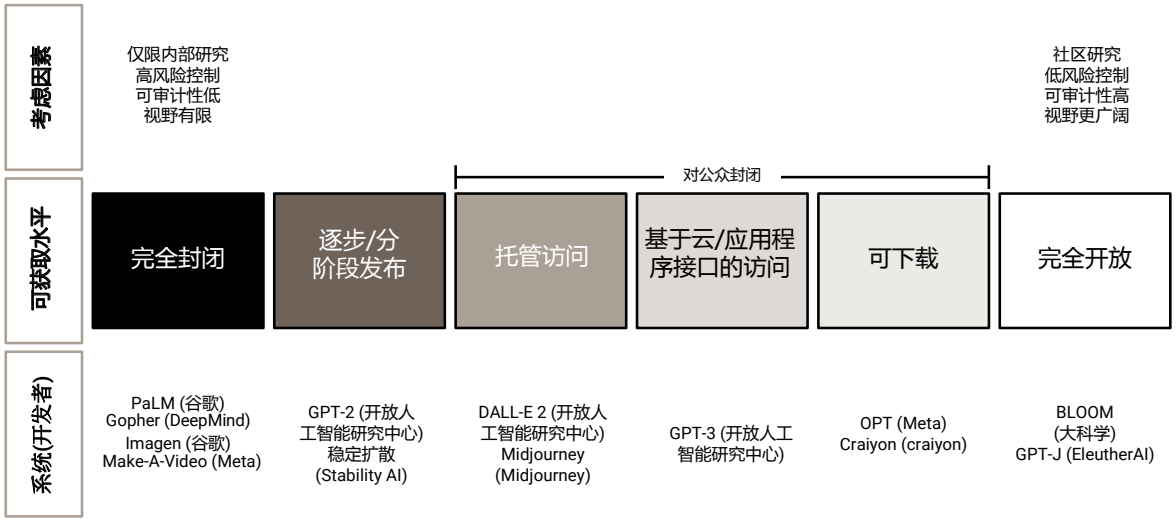
此外，开源模型往往更小，更透明。这种透明度可以建立信任，主动解决伦理问题，并支持验证和复制，因为用户可以检查人工智能系统的内部运作，了解其决策过程，并识别潜在的偏见。

封闭式人工智能系统为其开发人员提供了更大的控制权。此外，由于代码库不会因公众贡献而不断演变，封闭源码系统可以更加精简和高效。许多公司认为完全开放是不切实际的，并将部分开放作为唯一可行的选择。然而，这种观点忽视了采取能够实现“有意义开放”的平衡方法的可能性。[b]

有意义的开放性存在于两个极端之间，可以根据不同的用例进行调整。这种平衡的方法可以让公众对公开的训练进行监督和独立审计并微调数据，从而促进安全、创新和包容的人工智能开发。开放不仅仅是共享模型权重，还能推动创新和包容，有助于研究和教育领域的应用。

“开源人工智能”的定义在不断演变，[c]并且经常受到企业利益的影响，如图11所示。为解决这一问题，我们建议启动一个进程，由上述拟议的国际科学小组协调，以制定一个全面、有梯度的开放办法。这将为开放性提供有意义的、基于证据的方法，帮助用户和政策制定者对人工智能模型和架构做出明智的选择。

图 11：企业利益与开放性



信息来源 Irene Solaiman, “The gradient of generative AI release: methods and considerations”, *Proceedings of the 2023 Association for Computing Machinery (ACM) Conference on Fairness, Accountability, and Transparency* (June 2023), pp. 111–122.

b 受此文启发: Andreas Liesenfeld and Mark Dingemanse, “Rethinking open source generative AI: open-washing and the EU AI Act”, The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24) (June 2024).

c The Open Source AI Definition – draft v. 0.0.3. Available at https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-3.

B. 共同基础

101 除了对人工智能的统一认知外，还需要建立共同基础，以建立跨司法管辖区可互操作的治理方法，并以《世界人权宣言》等国际规范为基础(上文原则5)。

102 这需要在全球层面进行，不仅是为了公平的代表性，也是为了避免监管上的逐底竞争，同时减少跨境监管摩擦，最大限度地提高技术和本体上的互操作性，并检测和应对人工智能生命周期中跨越多个司法管辖区的决策所引发的事件。

人工智能治理政策对话

建议2：人工智能治理政策对话

我们建议在联合国现有会议期间，启动每年两次的人工智能治理问题政府间多利益攸关方政策对话。其目的是：

a. 分享在推动发展的同时促进尊重、保护和实现所有人权的人工智能治理最佳做法，包括把握机遇和管控风险方面的最佳做法;
b. 促进私营部门和公共部门开发人员和用户就实施人工智能治理措施统一认知，以加强人工智能治理的国际互操作性;
c. 自愿分享使国家机构捉襟见肘或超出其应对能力的重大人工智能事件;
d. 酌情讨论国际人工智能科学小组的报告。

103 目前，人工智能的国际治理充其量只能说是支离破碎、东拼西凑。118个国家没有参加政府间轨道中最近七个主要的非联合国的人工智能治理举措中的任何一个(见图8)。[26]即使在人工智能能力最强的60个国家中，也存在代表性方面的差距，这凸显了当今国际人工智能治理的选择性(见图12)。

104 需要建立一个包容性的政策论坛，使所有会员国能够利用利益攸关方的专门知识，分享在推动发展的同时促进尊重、保护和实现所有人权的最佳做法，促进可互操作的治理办法，并监测需要进一步政策干预的共同风险。

105 这并不意味着对人工智能的所有方面进行全球治理（鉴于各国的利益和优先事项各不相同，这不可能，也不可取）。然而，就人工智能发展和政策应对交换意见可以为国际合作建立框架。

106 联合国具有独特的地位，能够以包容各方的方式促进此类对话，帮助会员国有效合作。联合国系统现有的和新出现的一整套规范可以为协调一致的行动提供强有力的规范性基础，这些规范以《联合国宪章》、人权和其他国际法(包括环境法和国际人道法)以及可持续发展目标和其他国际承诺为基础。[27]

107 结合国际科学小组的专业知识和能力发展(见建议1、4和5)，联合国包容各方的对话可帮助各国和公司更新其监管方式和方法，以促进共同基础的互操作方式跟上人工智能加速发展的步伐。在这方面，联合国的一些独特之处可能会有所帮助：

a. 将包容各方的对话建立在联合国的一整套规范（包括《联合国宪章》、人权和其他国际法）之上，可以促进各方争相在治理方法方面“力争上游”。相反，如果没有全球各国普遍加入的联合国，国际集体行动将面临更大的压力，各司法管辖区之间关于人工智能安全和使用范围的监管将成为逐底竞争。
b. 联合国会员国遍布全球，这还可以促进现有全球以下各级举措之间的协调，以提高它们之间的兼容性。在我们的咨询中，许多人呼吁联合国成为一个关键的空间，在考虑到不同文化、语言和区域的不同价值观的情况下，促成现有区域和多边举措之间的软协调。
c. 本组织的程序可预测、透明、有章可循、有理有据，能够通过持续的政治参与，为意见不一致的国家牵线搭桥，缓和危险的竞争。除了在危机时刻建立信任、关系和沟通渠道外，可靠的包容各方的对话能够促进新的规范、习惯法和协议，从而加强各国之间的合作。

26 这些举措并不总是具有直接可比性。有些举措反映了现有国际或区域组织的工作，而另一些则是基于志同道合国家的特别邀请。
27 例如见《联合国宪章》(序言、宗旨及原则以及第十三、五十五、五十八和五十九条)。另见各项核心国际人权文书（《世界人权宣言》;《公民及政治权利国际公约》;《经济社会文化权利国际公约》;《消除一切形式种族歧视国际公约》;《儿童权利公约》;《消除对妇女一切形式歧视公约》;《禁止酷刑公约》；《残疾人权利公约》;《移民权利公约》；《保护所有人免遭强迫失踪国际公约》);国际人权法文书(日内瓦四公约；《特定常规武器公约》；《防止及惩治灭绝种族罪公约》；《海牙公约》)；关于区分、相称性和预防等相关原则以及《特定常规武器公约》下通过的关于致命自主武器系统的11项原则的文书；禁止大规模毁灭性武器方面的裁军和军备控制文书（《不扩散核武器条约》;《禁止化学武器公约》；《生物武器公约》);环境法文书（《联合国气候变化框架公约》;《禁止为军事或任何其他敌对目的使用改变环境的技术的公约》);《巴黎协定》以及预防原则、一体原则和公众参与等原则;关于《2030年可持续发展议程》、性别和伦理道德的非约束性承诺，如教科文组织《人工智能伦理问题建议书》。

图 12：人工智能排名前60位的国家(2023年乌龟指数)参加政府间轨道的主要诸边人工智能治理举措样本的情况

party to

国家 / 参与国*	乌龟全球人工智能指数排名 (2023)	参与的举措数	按时间顺序→ 经合组织《人工智能原则》 (2019)	二十国集团《人工智能原则》 (2019)	欧洲委员会起草小组 (2022)	人工智能全球伙伴关系《部长宣言》 (2022)	《布莱切利宣言》 (2023)	七国集团《关于广岛人工智能进程的部长声明》 (2023)	《首尔部长声明》 (2024)
美利坚合众国	1	7	party to	party to	party to	party to	party to	party to	party to
中国	2	2		party to			party to		
新加坡	3	4	party to			party to	party to		party to
大不列颠及北爱尔兰联合王国	4	7	party to	party to	party to	party to	party to	party to	party to
加拿大	5	7	party to	party to	party to	party to	party to	party to	party to
大韩民国	6	5	party to	party to		party to	party to		party to
以色列	7	5	party to		party to	party to	party to		party to
德国	8	7	party to	party to	party to	party to	party to	party to	party to
瑞士	9	4	party to		party to		party to		party to
芬兰	10	2	party to		party to				
荷兰王国	11	5	party to		party to	party to	party to		party to
日本	12	7	party to	party to	party to	party to	party to	party to	party to
法国	13	7	party to	party to	party to	party to	party to	party to	party to
印度	14	4		party to		party to	party to		party to
澳大利亚	15	6	party to	party to	party to	party to	party to	party to	party to
丹麦	16	3	party to		party to	party to			
瑞典	17	3	party to		party to	party to			
卢森堡	18	2	party to		party to				
爱尔兰	19	4	party to		party to	party to	party to		
奥地利	20	2	party to		party to				
西班牙	21	5	party to		party to	party to	party to		party to
比利时	22	3	party to		party to	party to			
意大利	23	7	party to	party to	party to	party to	party to	party to	party to
挪威	24	2	party to		party to				
爱沙尼亚	25	2	party to		party to				
阿拉伯联合酋长国	27	2					party to		party to
葡萄牙	28	2	party to		party to				
俄罗斯联邦	29	1		party to					
沙特阿拉伯	30	3		party to			party to		party to
马耳他	31	2	party to		party to				
巴西	33	4	party to	party to		party to	party to		
新西兰	34	3	party to			party to			party to
斯洛文尼亚	35	3	party to		party to	party to			
匈牙利	36	2	party to		party to				
土耳其	37	6	party to	party to	party to	party to	party to		party to
冰岛	38	2	party to		party to				
智利	39	3	party to				party to		party to
卡塔尔	40	0							
立陶宛	41	2	party to		party to				
马来西亚	42	0							
希腊	43	2	party to		party to				
印度尼西亚	44	3		party to			party to		party to
越南	45	0							
哥伦比亚	46	1	party to						
阿根廷	47	4	party to	party to	party to	party to			
斯洛伐克	48	2	party to		party to				
墨西哥	49	5	party to	party to	party to	party to	party to		party to
埃及	50	1	party to						
乌拉圭	51	1			party to				
亚美尼亚	52	1			party to				
南非	53	1		party to					
突尼斯	54	0							
摩洛哥	55	0							
巴林	56	0							
巴基斯坦	57	0							
斯里兰卡	58	0							
尼日利亚	59	2					party to		party to
肯尼亚	60	2					party to		party to
欧洲联盟	不适用	5		party to	party to	party to	party to		party to
共计 (包括未列示的)			**47**	**20**	**58**	**29**	**29**	**7**	**28**

*Including jurisdictions such as the Holy See and the European Union.

Sources:
- OECD, Recommendation of the Council on Artificial Intelligence (adopted 21 May 2019), available at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
- G20, AI Principles (June 2019), available at https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf.
- GPAI, 2022 ministerial declaration (22 November 2022), available at https://one.oecd.org/document/GPAI/C(2022)7/FINAL/en/pdf.
- Bletchley Declaration (1 Nov 2023), available at https://gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.
- G7, Hiroshima AI Process G7 Digital & Tech Ministers' Statement (1 Dec 2023), available at https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document02_en.pdf.
- Council of Europe, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (adopted 17 May 2024), available at https://coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence.
- Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity, AI Seoul Summit (22 May 2024).
- Tortoise Media, Global AI Index (2023), available at https://tortoisemedia.com/intelligence/global-ai/#rankings.

108 业务方面：

a. 政策对话可以在纽约（如大会[28]）、日内瓦和全球南方的地点的现有会议间隙开始。

b. 每次对话会议的一部分可侧重于会员国牵头的国家办法，另一部分则从主要利益攸关方（特别是技术公司和民间社会代表）获取专门知识和投入。

c. 政府的参与可以向所有会员国开放，也可以向区域平衡的分组开放（在轮流的、有代表性的有关分组中进行重点更突出的讨论），或者两者结合，随着时间的推移，随着技术的发展和全球关注问题的出现或变得突出，酌情调整不同的议程项目或部分。鉴于技术和政策背景的动态性质，固定的几何结构可能无济于事。

d. 除了正式对话会议外，多利益攸关方参与人工智能政策还可利用其他现有机制，如国际电联的人工智能造福人类会议、互联网治理论坛年会、教科文组织的人工智能伦理论坛和联合国贸易和发展会议 (贸发会议)的电子周——所有会员国的代表均可自愿参加。

e. 根据对话的包容性，讨论议程可以广泛，以涵盖不同的观点和关切。例如，一年两次的会议可以在一次会议上更多关注不同部门的机遇，而在另一次会议上则更注重风险趋势[29]。这可以包括利用人工智能实现可持续发展目标、如何保护儿童、最大限度地减少气候影响，以及关于管理风险方法的交流。[29]会议还可以酌情讨论人工智能治理和人工智能技术标准中使用的术语的定义，以及国际科学小组的报告。

f. 此外，还可以邀请不同的利益攸关方——特别是技术公司和民间社会代表——通过下文详述的现有机构以及政策研讨会参与人工智能治理的特定方面，例如最先进形式人工智能的开源方法的限制(如果有的话)，人工智能事件的跟踪和报告门槛、人权法对新用例的应用，或利用竞争法/反垄断法解决技术公司之间的权力集中问题[30]。

g. 拟议的人工智能办公室还可与经合组织等方面的现有工作开展合作，管护一个包括世界各地的立法、政策和机构的人工智能治理范例库，供政策对话审议。

109 尽管大会在2024年通过了两项关于人工智能的决议，但目前在联合国还没有与这一建议的可靠包容性愿景相对应的关于人工智能治理的授权制度化对话。在国际层面确实存在类似的进程，但主要是在区域或诸边体系中(第57段)，这些体系并不具有可靠的包容性和全球性。

110 作为对多边和区域人工智能峰会[31]这种流动性进程的补充，联合国可以为人工智能治理对话提供一个稳定的场所。通过设计实现包容——这是在地缘政治微妙时期发挥稳定作用的关键要求——也可以解决第64和72段中指出的代表性和协调方面的差距问题，促进在人工智能治理方面采取符合所有国家共同利益的更有效的集体行动。

人工智能标准交流中心

建议3：人工智能标准交流中心

我们建议创建人工智能标准交流中心，汇集各国和国际性的标准制定组织、技术企业、民间社会的代表以及国际科学小组的代表。其任务是：

a. 建立和维护人工智能系统相关定义与适用的测量和评估标准的登记册;

b. 就标准及其制定过程进行辩论和评估；

c. 查明需要制定新标准的空白领域。

111 当人们开始探索人工智能系统时，几乎没有什么标准可以帮助引导或衡量这一新领域。图灵测试——测试机器能否表现出与人类等同（或无差别）的行为——吸引了大众的想象力，但其文化意义大于科学意义。事实上，一些最伟大的计算技术进步都是以其在游戏中的成功来衡量的，比如计算机在国际象棋、围棋、扑克或《危险》游戏中击败人类。非专业人士很容易理解这些衡量标准，但它们既不严谨，也不特别科学。

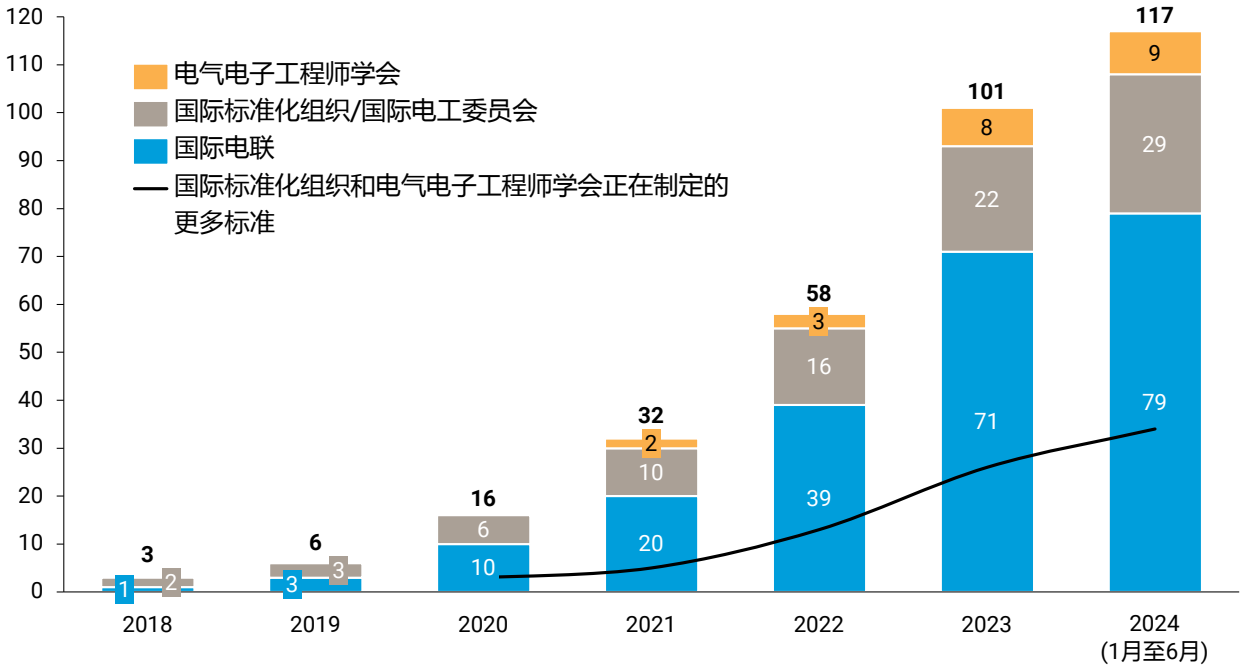
28 类似于经济及社会理事会主持下举行的可持续发展目标背景下的高级别政治论坛。

29 可以让联合国系统的相关部门参与进来，突出机遇和风险，包括国际电联参与关于人工智能标准的工作;国际电联、贸发会议、开发署和发展协调办公室参与关于人工智能应用于可持续发展目标的工作；教科文组织参与关于伦理和治理能力的工作；联合国人权事务高级专员办事处 (人权高专办)参与关于基于现有规范和机制的人权问责的工作；裁军事务厅参与对军事系统中人工智能进行监管的工作；开发署参与关于支助国家发展能力的工作；互联网治理论坛参与促进多利益攸关方参与和对话的工作；知识产权组织、劳工组织、世卫组织、粮农组织、世界粮食计划署、难民署、教科文组织、联合国儿童基金会、世界气象组织和其他组织参与本领域应用和治理的工作。

30 这样的会议还可以提供一个机会，让多利益攸关方就任何强化人工智能全球治理的问题进行辩论。例如，这些可能包括禁止开发不可控制或无法控制的人工智能系统，或者要求所有人工智能系统都足够透明，其后果可以追溯到能够对其承担责任的法律行为者。

31 虽然多个人工智能峰会已帮助20-30个国家在人工智能安全问题上统一立场，但参与情况并不一致：巴西、中国和爱尔兰在2023年11月认可《布莱切利宣言》，但在短短6个月后却没有支持《首尔部长声明》(见图12)。相反，墨西哥和新西兰认可了《首尔部长声明》，但没有支持《布莱切利宣言》。

图13: 人工智能相关标准的数目



信息来源: 电气电子工程师学会、国际标准化组织/国际电工委员会、 国际电联、世界标准合作组织(基于2023年6月的摸底调查, 纳入了与人工智能相关的标准以扩大范围)。

112 最近，标准激增。图13显示，国际电联、国际标准化组织、国际电工委员会和电气电子工程师学会采用的相关标准数量不断增加。[32]

113 有两个趋势很突出。首先，这些标准主要是为解决具体问题而制定的。在人工智能方面没有共同语言，许多常用术语——公平性、安全性、透明度——也没有商定的定义或可衡量性（尽管经合组织和美国国家标准和技术研究所最近针对人工智能等动态系统采用了新方法）。

114 其次，为狭隘的技术或内部验证目的而采用的标准与旨在纳入更广泛伦理原则的标准之间存在脱节。计算机科学家和社会科学家经常对同一概念提出不同的解释，社会技术标准的联合范式很有希望，但仍是一个愿景(见方框10)。

115 其结果是，我们有了一套新出现的标准，但这些标准并不基于对意义的统一认知，或者脱离了它们旨在维护的价值观。关键问题是，在能源消耗和人工智能方面，几乎没有达成一致的标准。在标准制定过程中缺乏对人权因素的考虑是另一个需要弥补的差距。[33]

116 这是有实际成本的。除了会员国和不同个人感到关切之外，我们在许多磋商中发现，企业界（包括发展中国家的中小企业）也存在关切，即在一个日益全球化的世界里，管理分散和标准不一致提高了开展业务的成本。

117 本报告并不建议联合国加入其中，制定更多标准。相反，联合国系统可以利用(建议1中提议的)国际科学小组的专业知识，并吸纳为标准制定做出贡献的各实体的成员以及技术公司和民间社会的代表，充当适用于全球的人工智能标准的信息交换中心[34]。

118 本组织的附加价值将是促进最广泛的标准制定组织之间的交流，以最大限度地提高技术标准之间的全球互操作性，同时将社会技术标准制定方面的新兴知识注入人工智能标准讨论。

119 收集和分发有关人工智能标准的信息，借鉴人工智能标准中心[35]等方面的现有工作并与之开展合作，将使来自各个标准制定组织的参与者能够在关键领域汇聚共同语言。

32 国家和跨国层面也出现了许多新标准，如美国白宫《人工智能自愿承诺》和欧洲联盟《人工智能行为守则》。
33 见A/HRC/53/42(人权与新兴数字技术的技术标准制定进程：联合国人权事务高级专员办事处的报告)和人权理事会第53/29号决议(新的和新兴数字技术与人权)。
34 即使这似乎也是一项具有挑战性的任务，但在达成全球最低税额协议方面取得的进展表明，即使在经济和政治复杂的领域，也有可能采取集体行动。
35 可查阅 https://aistandardshub.org.

方框10: 适用于人工智能安全的标准

人工智能安全的综合方法包括了解先进人工智能模型的能力，采用安全设计和部署标准，以及评估系统及其更广泛的影响。

过去，人工智能标准主要侧重于技术规范，详细说明系统应如何构建和运行。然而，随着人工智能技术对社会的影响越来越大，有必要转向社会技术范式。这种转变承认，人工智能系统并非存在于真空中；它们与人类用户互动并影响社会结构。现代人工智能标准可以将伦理、文化和社会因素与技术要求结合起来。在安全方面，这包括确保可靠性和可解释性，以及评估和减轻不同情况下对个人和集体权利、[a]国家和国际安全以及公共安全造成的风险。

最近成立的人工智能安全国家研究所的一个主要目标是确保对人工智能安全采取一致和有效的方法。统一这些方法将使人工智能系统在国际上达到较高的安全基准，在保持严格的安全协议的同时实现跨境创新和贸易。

就“安全”而言，让各种利益攸关方和文化参与制定此类标准，可增强其相关性和有效性，并有助于对定义和概念统一认知。通过纳入不同的观点，协议可以更彻底地处理与人工智能技术相关的各种潜在风险和益处。

a 见A/HRC/53/42(人权与新兴数字技术的技术标准制定进程：联合国人权事务高级专员办事处的报告)和人权理事会第53/29号决议(新的和新兴数字技术与人权)。

120 在拟议的人工智能办公室的支持下，标准交流中心还将受益于与国际科学小组在技术问题上的紧密联系，以及与政策对话在道德、伦理、监管、法律和政治问题上的紧密联系。

121 如果达成适当的一致意见，国际电联、国际标准化组织/国际电工委员会和电气电子工程师学会可以联合牵头举办首次人工智能标准峰会，并每年开展后续活动，以保持其受关注度和势头。为了为纳入经济、伦理和人权因素的社会技术方法奠定基础，经合组织、世界知识产权组织 (知识产权组织)、世界贸易组织、联合国人权事务高级专员办事处 (人权高专办)、劳工组织、教科文组织和其他相关联合国实体也应参与进来[36]。

122 标准交流中心还应为建议4中的能力建设工作提供信息，确保标准支持实际做法。可以分享关于国家或区域开发的有助于对标准遵守情况进行自我评估的工具的信息。

123 本报告目前只是建议联合国作为讨论和商定标准的论坛，仅此而已。如果安全标准随着时间的推移得到正式确定，这些标准可以作为最终成立的机构进行监测和核查的基础。

C. 共享惠益

124 《2030年议程》及其17个可持续发展目标可以为人工智能提供独特的目的，将投资的弧线从浪费和有害的使用转向全球发展挑战。否则，投资将追逐利润，甚至不惜对他人造成负面外部效应。联合国可以做出的另一个重要贡献是将人工智能的积极应用与确保公平分配其机会联系起来(方框11)。

125 正如我们在中期报告中所指出的那样，这在很大程度上取决于人才、算力和数据的获取，以帮助文化和语言多样性蓬勃发展。治理本身可以成为一个关键的推动因素，在促进跨国界和跨学科领域合作的同时，协调激励机制、建立信任和产生可持续的做法。如果在人工智能治理方面缺乏全面、包容的方法，就可能错失人工智能为可持续发展目标做出积极贡献的潜力，其部署也可能无意中强化现有的差距和偏见。

126 在咨询机构就教育、医疗卫生、数据、性别、儿童、和平与安全、创意产业和工作等专题进行的广泛咨询中，可以明显看出，人工智能具有极大的潜力，可以在各个关键领域促进创新和交付，从而大大加快可持续发展目标的进展。

36 这可以包括相关的部门、国家和区域标准组织。

127 然而，人工智能并不是解决发展挑战的灵丹妙药；它只是一系列更广泛解决方案中的一个组成部分，甚至可能会加剧其中的一些挑战——如气候变化。要真正释放人工智能应对社会挑战的潜力，政府、学术界、产业界和民间社会之间的合作至关重要

128 人工智能解决方案的有效性取决于数据的质量和可用性，而可持续发展目标相关数据集的质量和代表性令人严重关切，这些数据集可能无法反映某些人群的相关现实情况。此外，人工智能专家在不完全了解交叉应用领域的情况下设计的人工智能解决方案通常只是电脑模拟，在实际开发环境中不够强大或有效。这就是为什么人工智能解决方案必须在深入了解其社会、经济和文化背景的基础上进行合作设计和实施。它们必须符合更广泛的地方和国家数字转型战略，并解决数字鸿沟问题。

129 例如，如果没有可靠的电力和互联网连接来运行数据中心、维持计算机的稳定运行、访问全球数据集、参与国际研究合作以及使用基于云的人工智能工具，低收入和中低收入国家就无法实现人工智能能力。因此，我们支持投资于基础数字基础设施的呼吁，这是发展中国家参与并受益于人工智能进步的先决条件。

130 建设人工智能能力对于确保全球各地的个人，无论其所处区域的发展阶段如何，都能从人工智能的进步中受益至关重要。得到充足资金支持的战略能力建设对于使人工智能技术有效、可持续和符合公共利益也至关重要，这也是全球发展努力的关键所在。下面，我们将探讨国家人工智能能力的三个关键推动因素——技术专业知识的可获得性、算力的可及性和高质量数据的可用性。然后，我们将对具体行动提出建议。

方框11: 人工智能与可持续发展目标

由于人工智能在推动科学进步（方框1）和创造经济机遇（方框2）方面具有潜力，人工智能有望加快在实现可持续发展目标领域的进展。2023年，一项对相关证据的审查表明，人工智能可推动实现所有可持续发展目标下的134项具体目标（79%），通常是通过技术改进来克服某些当前普遍存在的限制。[a]

为推进我们的工作，委托进行了一次机遇扫描，对来自38个国家的120多名专家进行了调查，了解他们对于人工智能对科学突破、经济活动和可持续发展目标的积极影响的预期，调查结果说明了各位专家当前的看法概况。这项调查仅询问了人工智能可能产生的积极影响。

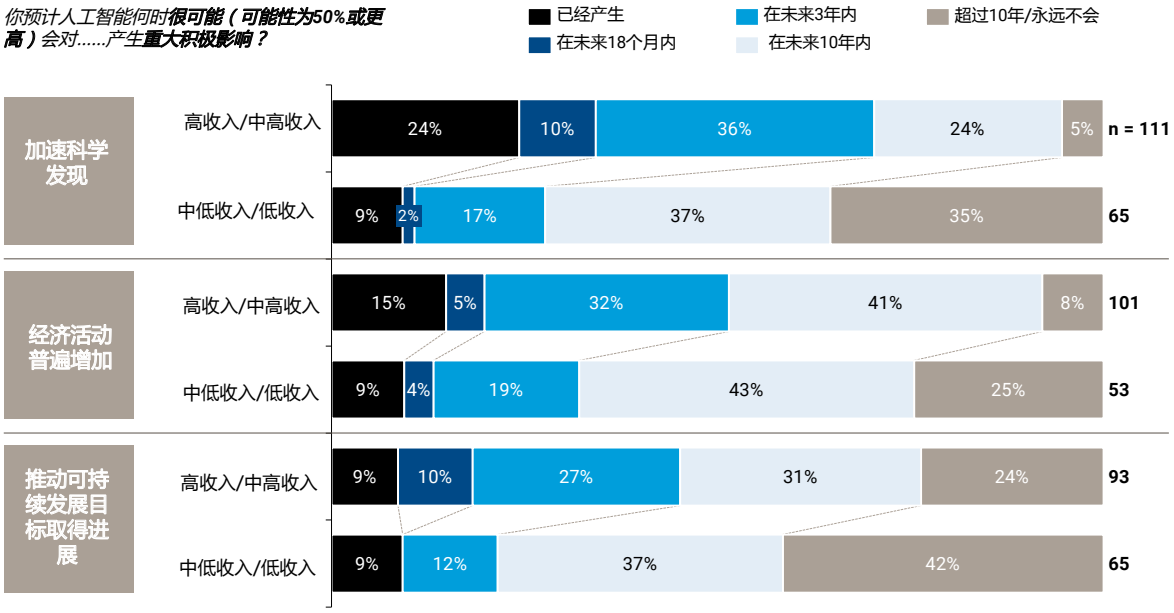
总体而言，各位专家对人工智能在多久以后能够产生重大积极影响的预期各异（另见图14）：

- 专家对加速科学发现最乐观，七成专家表示，人工智能很可能会在未来三年或更短时间内对高收入/中高收入国家产生重大积极影响，28%的专家预测中低收入/低收入国家会出现同样情况。
- 约五成专家预计人工智能很可能会在未来三年或更短时间内，在高收入/中高收入国家对增加经济活动产生重大积极影响，32%的专家预测中低收入/低收入国家会出现同样情况。
- 共有46%的专家预计，人工智能很可能在未来三年或更短时间内，在高收入/中高收入国家对推动可持续发展目标取得进展产生重大积极影响。不过，只有21%的专家预计中低收入/低收入国家会出现同样情况，四成专家认为，人工智能可能至少需要10年才会在这些地区对可持续发展目标产生如此重大的积极影响。

a Ricardo Vinuesa and others, “The role of artificial intelligence in achieving the Sustainable Development Goals”. Nature Communication, vol. 11, No. 233 (January 2020). This study also argued that 59 targets (35%, also across all SDGs) may experience a negative impact from the development of AI.

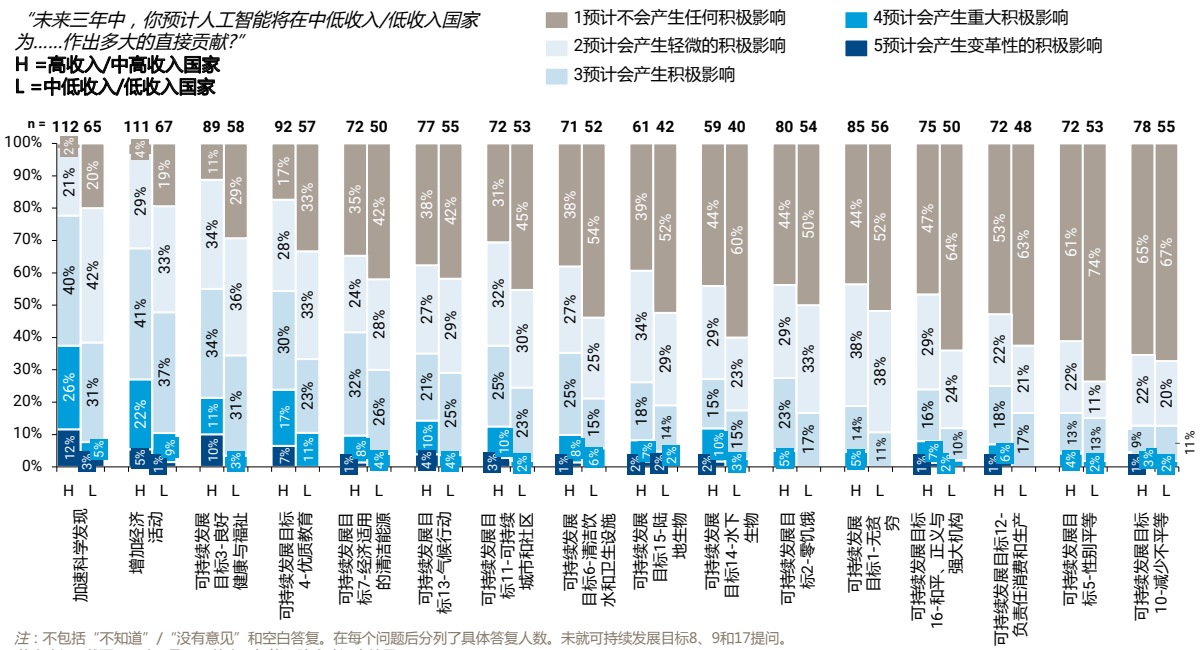
方框 11：人工智能与可持续发展目标（续）

图 14：专家对人工智能在各领域产生重大积极影响的时间的预期



注：不包括“不知道”/“没有意见”和空白答复。在每个问题后分列了具体答复人数。
信息来源：截至2024年8月21日的人工智能风险脉动调查结果（121人答复，其中78人回答了关于中低收入/低收入国家的问题）。

图 15：专家对人工智能在未来三年产生重大积极影响的预期，按地区和可持续发展目标分列



注：不包括“不知道”/“没有意见”和空白答复。在每个问题后分列了具体答复人数。未就可持续发展目标8、9和17提问。
信息来源：截至2024年8月21日的人工智能风险脉动调查结果.

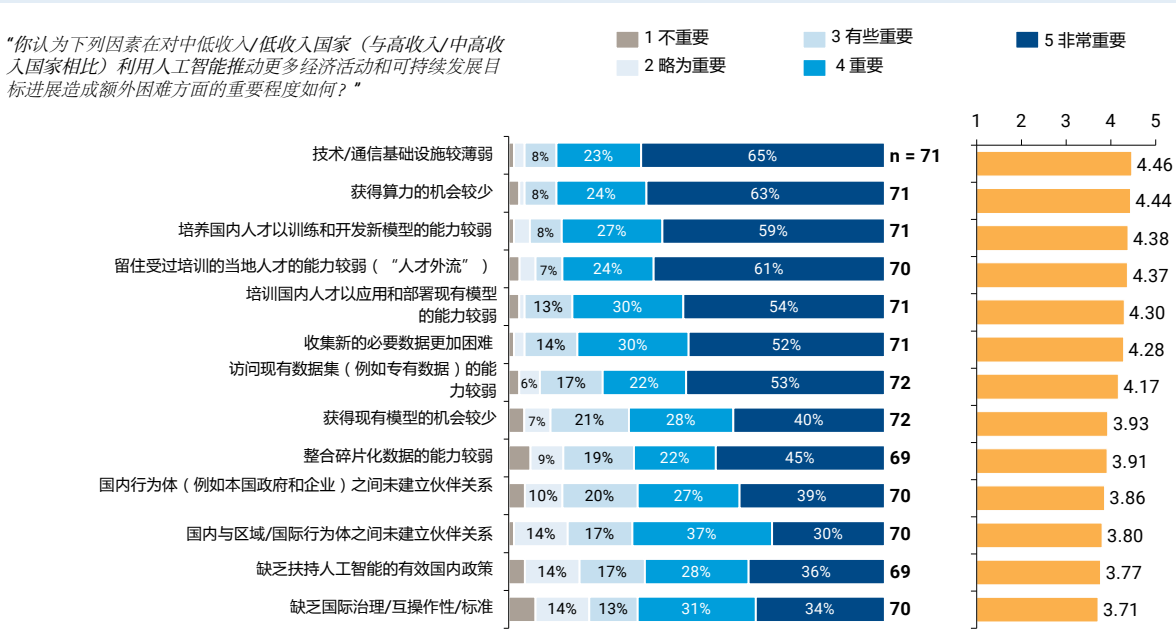
方框 11：人工智能与可持续发展目标（续）

专家预计，未来三年中，人工智能将在较高收入国家对调查涵盖的所有领域，包括加速科学发现、增加经济活动2以及提问涉及的14个可持续发展目标领域产生更大的积极影响（见图15）。专家对于人工智能对健康和教育（可持续发展目标3和4）领域的积极影响最为乐观，20-25%的专家预计，人工智能将在未来三年内，在高收入/中高收入国家对这些领域产生重大或变革性的积极影响。专家对于人工智能对性别平等和不平等（可持续发展目标5和10）领域的积极影响最不乐观，三分之二的专家预计，人工智能对减少较高收入或较低收入国家内部或之间的不平等不会产生积极影响。预计人工智能将在较高收入国家更早产生更大的影响，部分原因是中低收入和低收入国家面临各种障碍（见图16）。

超过一半的受访者指出，缺乏推动因素——从基础设施薄弱，到国内政策和国际治理的缺失——是导致较低收入国家在利用人工智能推进经济活动和可持续发展目标取得进展方面遇到更多困难的重要因素。

这些结果突出表明，目前仍然无法确定人工智能最终将对可持续发展目标作出的贡献，而且这种贡献依然高度依赖于缺失的推动因素。在欠发达国家尤其如此，它们本已缺乏较发达国家所拥有的大部分要素，包括基础设施和政策等。若不合作建设能力和促进获得关键推动因素，那么现有的人工智能鸿沟可能进一步扩大和加深，制约人工智能到2030年为推进科学进步、经济效益和可持续发展目标领域进展作出切实贡献的能力。

图16: 专家对中低收入/低收入国家在利用人工智能推动更多经济活动和可持续发展目标进展领域所面临障碍的评分



注：不包括"不知道"/"没有意见"和空白答复。在每个问题后分列了具体答复人数。
信息来源：截至2024年8月21日的人工智能风险脉动调查结果

b The share of experts expecting "major positive impact" on increasing economic activity and accelerating scientific discovery over three years is higher in the first chart than the second chart. This may be due to the qualifier "by when do you expect it likely (50% chance or more) that AI will cause a major positive impact" (emphasis added) in the question responses depicted in the first chart, which is absent in the second.

人才

131 世界各地的社会能否参与人工智能的有益成果，首先取决于人。必须认识到，并非每个社会都需要一批计算机科学家来建立自己的模型。然而，无论是购买、借用还是构建技术，都需要人力资源，以了解人工智能的能力和局限性，并适当利用人工智能支持的用例。

132 这种能力——主要是公共部门的能力，也包括学术界、商界和民间社会的能力——将提高人工智能战略及其在各部门实施的有效性。培养与人工智能相关的人员能力对于保护世界文化和语言多样性以及为未来的人工智能发展建立高质量的数据集也至关重要。从本质上讲，这就是公益性人工智能的能力建设

133 在非洲等人口结构年轻化的多元化环境中培养人力资源，对于未来的全球人才梯队建设也至关重要——本世纪上半叶，全球三分之一的劳动力将是非洲人。提高女性在科技领域的能力，一方面需要关注缩小现有的性别差距，另一方面要避免人工智能领域的性别差距。人工智能部门也需要更多女性担任领导职务，将性别平等视角纳入人工智能治理。这首先要为女童提供更多成为人工智能人才的机会。

算力

134 尽管一直在努力开发对算力要求较低的人工智能方法，但要训练出有能力的人工智能模型，仍然迫切需要获得负担得起的算力[37]。这不仅是全球南方国家的企业进入人工智能领域的最大障碍之一，也是全球北方许多初创企业和中小企业进入该领域的障碍。在世界上能够训练大型人工智能模型的前100个高性能计算集群中，没有一个设在发展中国家。在前300名中只有一个非洲国家[38]。两个国家占据了全球超大规模数据中心的半壁江山[39]。

135 大多数开发人员通过云服务访问计算基础设施；许多人选择与大型云计算公司合作，以确保对算力的可靠访问。随着时间的推移，供应链问题可能会得到解决，竞争可能会带来更多样化的硬件来源，包括用于训练模型的高性能芯片和部署在移动设备上的人工智能加速器芯片。然而，在可预见的未来，这一制约因素仍将是建立更具全球包容性的人工智能创新生态系统的巨大障碍。

136 具有讽刺意味的是，计算能力可能会闲置或很快过时。在整个折旧周期中充分利用这种能力具有潜在价值。然而，在不同硬件配置的互操作性和高要求任务的调度方面，还有一些障碍需要克服，同时还要保持时限严格的用途（如气象预测）的优先权。

137 此外，如果没有人才和数据，仅靠算力是没有价值的。在拟议的全球人工智能基金中，我们考虑如何通过资金和实物支持相结合的方式来解决所有这三方面的问题。

数据

138 尽管许多关于人工智能经济的讨论都集中在“人才争夺战”和图形处理器等硬件的竞争上，但数据同样至关重要。促进初创企业和中小企业大规模获取高质量的训练数据来训练人工智能模型，以及以尊重权利的方式补偿数据持有者和训练数据创建者的机制，可能是人工智能经济蓬勃发展的最重要推动因素。为公共利益汇集数据以推进具体的可持续发展目标是一个关键方面(方框12对此作了概述)，但这还不够。

139 在人工智能方面，人们常说“滥用”数据（如侵犯隐私）或“漏用”数据（未能利用现有数据集），但与此相关的一个问题是数据缺失，包括数据贫乏的全球大部分地区。一个例子是医疗保健，其中大约一半的主要数据集可以追溯到十几个组织，其中一个在欧洲，一个在亚洲，其余的在北美[40]。

140 另一个例子是农业，在农业中，气候、土壤和作物管理做法等因素之间的复杂相互作用需要数据来支持有用的人工智能模型。农业也常常因缺乏数据和数据收集基础设施而受到影响。需要做出专门的努力来管护农业数据集，特别是在粮食系统抵御气候变化的背景下。

141 与非正规资本的问题类似，那些数据（从出生记录到财务交易）未被采集的人可能无法参与人工智能经济的利益，无法获得政府福利，也无法获得信贷。使用合成数据可能只能部分抵消对新数据集的需求。

37 咨询机构了解到最近的一个案例，一家位于全球南方的公司花费7 000万美元对一个大型语言模型进行了为期3个月的训练。拥有图形处理器而不是从云服务提供商那里租用图形处理器，成本会低很多倍。

38 可查阅 https://top500.org/statistics/sublist; 此处为代用指标，因为大多数高性能计算集群没有图形处理器，对高级人工智能的作用有限。

39 贸发会议，《2021年数字经济报告》(2021年，日内瓦)。

方框12: 在可持续发展目标领域为公共利益汇集数据

合作数据和人工智能公共资源——用汇集的数据对共享模型进行交叉训练——可在促进公共利益方面发挥关键作用，否则，数据将缺失或过于稀少，无法为人工智能带来益处。跨职能和多领域的数据池可以帮助开发跨学科数据集，这些数据集涵盖可持续发展目标的各个领域，来源多种多样。

例如，我们可以考虑评估气候变化对健康的影响这一复杂问题。为了有效应对这一挑战，必须采用跨学科的方法，将有关疾病发病率的流行病学数据与跟踪气候变化的气象数据结合起来。通过以保护隐私的方式汇集来自世界各国的这些不同类型的数据，研究人员或许能够利用人工智能来识别孤立数据集中不明显的模式和相关性。

包括来自所有国家的数据确保了数据的全面覆盖，反映了气候变化的全球性，并捕捉到不同区域的各种环境影响和健康结果。数据的跨学科来源提高了旨在预测未来由气候变化引发的公共卫生危机或自然灾害的模型的预测准确性。

142 对我们中期报告的反馈意见指出，报告没有充分阐述目前围绕人工智能训练数据的来源获取、使用和不披露的跨司法管辖区做法如何威胁到权利并导致经济集中的问题。反馈意见建议我们考虑国际人工智能治理如何能够促进和推动更多样化地参与人工智能数据的利用。

建设国际人工智能核心公共能力，实现共享惠益

143 对于上述三个推动因素，发达经济体既有能力也有责任通过国际合作促进人工智能能力建设。反过来，发达经济体将受益于基础更广泛的数字经济以及高质量的人才和数据流。重要的是，通过这种合作将良好的人工智能治理纳入主流，每个人都将从中受益。

144 合作的重点应放在培养人工智能人才、提高公众人工智能素养、提高人工智能治理能力、扩大获取人工智能基础设施的机会、促进适合不同文化和区域需求的数据和知识平台，以及加强对人工智能应用和服务能力的应用。只有这样一种全面的方法才能确保公平获得人工智能的惠益，确保没有一个国家落在后面。

145 我们咨询的许多利益攸关方强调，应制定详细的战略，汇集全球资源，以建设能力，促进公平分享机会的集体行动，并缩小数字鸿沟。

能力发展网络

建议4：能力发展网络

我们建议创建人工智能能力发展网络，汇集联合国下属的一系列相互协作的能力发展中心，向关键行为体提供专门知识、算力和人工智能训练数据。建立该网络的目的是：

a. 支持区域和全球一级的人工智能能力建设举措彼此建立联系，以促进和协调这些举措;
b. 提升公职人员的人工智能治理能力，以期在推动发展的同时，促进尊重、保护和实现所有人权;
c. 向寻求将人工智能应用于当地公共利益用例的研究人员和社会企业家提供多个中心的培训人员、算力和人工智能训练数据，包括通过下列方式
    i. 制定协议，使算力稀缺环境中的跨学科研究团队和企业家能够获得算力，用于训练/调整其模型，以及将模型适当应用于当地环境;
    ii. 通过沙箱测试潜在的人工智能解决方案，并在实践中学习
    iii. 向大学生、年轻研究人员、社会企业家和公共部门官员提供一系列人工智能在线教育机会;
    iv. 设立研究金方案，供有前途的个人在学术机构或技术企业工作一段时间。

40 可查阅 https://2022.internethealthreport.org/facts.联合国大学长期致力于通过高等教育和研究进行能力建设，联合国训练研究所帮助培训了对可持续发展至关重要的领域的官员。教科文组织的准备情况评估方法是支持会员国实施教科文组织《人工智能伦理问题建议书》的关键工具。其他例子包括里昂世卫组织学院、贸发会议虚拟学院、裁军事务厅管理的联合国裁军研究金以及国际电联和开发署牵头的能力发展方案。

41 包括金融和医疗机构在内的各种国家机构都开发了沙箱，如新加坡资讯通信媒体发展局。

146 从千年发展目标到可持续发展目标，联合国长期以来一直为个人和机构的能力发展做出贡献[41]。通过教科文组织、知识产权组织和其他组织的工作，联合国帮助维护了全球丰富多样的文化和创造知识的传统。

147 与此同时，人工智能的能力发展需要一种全新的方法，特别是跨领域培训，以培养新一代的多学科专家，如公共卫生与人工智能，或粮食和能源系统与人工智能等领域的专家。

148 能力还必须与成果挂钩，具体做法是在沙箱中开展实训[42]，以及开展协作项目，汇集数据和算力，以解决共同的问题。必须将风险评估、安全测试和其他治理方法纳入这一合作训练的基础设施。

149 鉴于挑战的紧迫性和规模，我们建议采取一种战略方法，通过高性能计算节点网络汇集和调配算力资源，激励开发可持续发展目标相关领域的关键数据集，促进共享人工智能模型，将人工智能治理的最佳实践纳入主流，并为公益人工智能培养跨领域人才，确保人权专业知识的跨领域整合。

150 换句话说，我们不是通过彼此脱节的项目逐一争取关键的推动因素，而是建议通过一系列合作中心来实施整体性的统筹战略。关于促进可持续发展目标的能力发展和人工智能的新兴举措，如瑞士发起的国际计算和人工智能网络举措，有助于为这一战略创造初始的关键量。

151 理想的情况是，世界每个区域至少应有一个或两个节点。参加全球人工智能伙伴关系的两个专门知识中心可与联合国一道支持能力建设网络。为能力建设作出贡献的学术机构和私营部门机构可通过最近的区域节点或支持该网络的国际组织寻求加入。

152 我们对各国之间通过联合使用算力和相关基础设施等方式开展合作的前景感到特别鼓舞。正如我们在中期报告中指出的那样，欧洲核研究组织提供了有益的经验。为人工智能重新设想的“分布式欧洲核研究组织”在不同国家和区域间建立网络，可以扩大更多获取人工智能工具和专业知识的机会。

153 我们设想能力发展网络是国家和区域能力的催化剂，而不是硬件、人才和数据的集中地。通过加速学习，能力发展网络可以推动国家卓越中心促进当地人工智能创新生态系统的发展，解决第73、80和81段中提及的在协调和执行方面的根本性差距问题。国家层面的努力可以继续采用诊断工具，如教科文组织的人工智能准备情况评估方法，以帮助评估各国的初步成熟度，找出差距，并指导如何为每个国家和区域量身定制能力建设路线图，由国际网络帮助解决这些差距。

154 拟议的人工智能办公室可能最适合侧重于战略、伙伴关系和附属关系，以便将节点与网络连接起来，起到连接而非重塑的作用。该办公室还可以帮助为整个网络的算力访问提供中介服务。网络中的一个或多个节点可在训练的特定方面发挥牵头作用，托管沙箱或高性能计算集群，以开发人工智能模型。各节点可以在研究项目上进行合作，项目涉及的主题包括保护隐私的数据使用、将不同类型的硬件或数据集连接起来进行模型训练的新方法，以及将人工智能模型相互结合使用的方法。

155 我们希望该网络还将促进人工智能技术发展的另一种范式：自下而上、跨领域、跨区域、开放和协作。鉴于训练和部署人工智能模型所需的能源和其他成本不断上升，以及算力资源闲置的前景，在时间共享的基础上将算力资源的访问连接起来，同时利用这种访问促进跨领域人才、数据和人工智能模型的发展，以实现可持续发展目标，是非常有意义的。

全球人工智能基金

建议5：全球人工智能基金

我们建议设立全球人工智能基金，为弥合人工智能鸿沟托底。该基金将由独立的治理机构负责管理，接收来自公共和私营部门的财政和实物捐助，并通过能力发展网络等渠道分配这些资源，从而促进提供下列人工智能推动因素，赋能当地实现可持续发展目标：

a. 共享计算资源，供当地能力不足或无力采购资源的国家的人工智能开发人员训练模型和进行微调；
b. 沙箱与对标和测试工具，将安全可靠的模型开发和数据治理最佳做法纳入主流；
c. 适用于全球的治理、安全和可互操作解决方案；
d. 数据集，以及研究如何将数据和模型结合起来，用于可持续发展目标相关项目；
e. 推动实现可持续发展目标的人工智能模型库和管护数据集。

156 这里建议的人工智能开发和使用模型类似于互联网的最初愿景：分布式但相互连接的基础设施，具有互操作性并能增强能力。在这样一个市场中，人工智能模型及其所依赖的基础设施和数据具有互操作性、管理完善且值得信赖，从而更好地服务于公共利益。这不会自动实现。以充足资源为后盾的专门努力至关重要。

157 我们以谦逊的态度对待这一建议，意识到强大的市场力量影响着人才和算力的获取，且地缘政治竞争阻碍了科技领域的合作。遗憾的是，如果没有国际支持，许多国家可能无法获得培训、算力、模型和训练数据。如果没有这种支持，现有的供资努力也可能无法扩大规模。

158 公平竞争在某种程度上是一个公平问题。创造一个所有人都能为共享的生态系统做出贡献并从中受益的世界，也符合我们的集体利益。这不仅仅是在国家之间。确保对人工智能模型开发和测试基础设施的多样化使用，也将有助于消除人们对权力过度集中在少数技术公司手中的担忧。

基金宗旨及目标

159 我们提议设立基金的目的并不是要保证能够获得计算资源和能力，即使是最富裕的国家和公司也很难获得这些资源和能力。答案可能并不总是更多的算力。我们可能还需要不同的方法来利用现有的高性能计算基础设施，这些基础设施是为峰值使用而构建的，不一定是为人工智能而设计的。也许有更好的方法来连接人才、算力和数据。

160 因此，目的是为无法通过其他手段获得必要推动因素的各方解决第73、80和81段所述协调和执行方面的根本性差距，以确保：

a. 有需要的国家可以获得人工智能推动因素，为弥合人工智能鸿沟铺平道路；
b. 在人工智能能力开发方面进行合作，形成合作习惯，缓解地缘政治竞争；
c. 监管方法不同的国家有动力开发通用模板，用于管理与可持续发展目标和科学突破相关的社会层面挑战的数据、模型和应用。

161 利用全球基金资源建设的能力将面向可持续发展目标和人工智能的全球共同治理(方框13)，例如可以包含一个用于安全保障测试的“治理堆栈”。这将有助于将最佳做法纳入整个用户群的主流，同时减轻小用户的验证负担。

方框13: 全球人工智能基金：可能的投资实例

一个规模适中的基金可以帮助创建一个最低限度的共享计算基础设施，用于训练中小型模型。这些模型具有实现可持续发展目标的重要潜力，例如用当地语言培训农民。

这项投资还将创造一个沙箱环境，使开发人员能够利用自己的背景数据和高质量数据对现有的开源模型进行微调。对算力和沙箱基础设施的访问可以采取分时共享的方式，合理的使用费可用于收回维护和运行成本。

资金的第三个用途是帮助为某些缺乏商业激励的可持续发展目标管护黄金标准数据集。模型开发、测试和数据管护工作可以战略性地结合在一起，形成与具体成果相关的强大的人工智能实践赋能方法。

最后，该基金不仅可以促进与情境相关的发展和与可持续发展目标相关的人工智能应用的研究和开发，还可以促进算力和模型的相互联系以及新的治理评估。

43 见 https://www.theglobalfund.org/en/about-the-global-fund.

162 这种对公共利益的关注使全球基金与关于建立人工智能能力发展网络的建议相辅相成，基金将向该网络提供资源。基金还将提供独立的影响监测能力。通过这种方式，我们可以确保世界上的广大地区不会落在后面，而是有能力在不同的背景下利用人工智能实现可持续发展目标。

163 确保在数字世界中与在现实世界中一样进行合作，符合每个人的利益。这可以与应对气候变化的努力相类比，在这方面，过渡、减缓或适应气候变化的成本并不平均，国际援助对于帮助资源有限的国家至关重要，这样这些国家才能加入应对地球所面临挑战的全球努力。

164 这里的重点是利用筹集的资金帮助确保不同区域的国家能够建立起最低限度的能力，以了解人工智能促进可持续发展的潜力，根据当地需求调整和建立模式，并加入人工智能方面的国际合作努力。

基金治理

165 该基金将筹集和汇集实物捐助，包括来自私营部门实体的实物捐助。协调资金和实物捐助需要适当程度的独立监督和问责。治理安排应具有包容性，理事会成员应来自政府、私营部门、慈善家、民间社会和联合国机构。这些安排应纳入科学和专家的投入，例如通过拟议的国际科学小组提供的投入，并在围绕数据和模型开发的合作中保持中立和信任。

基金业务

166 该基金的运作模式应借鉴欧洲核研究组织和全球疫苗免疫联盟等集合式国际研发合作的经验，以及分时共享基础设施商业平台的经验。该基金还应借鉴全球基金(2002年设立，旨在集中资源防治艾滋病毒、结核病和疟疾)和复杂风险分析基金等机构的经验，这些机构汇集数据[43]，支持所有利益攸关方预测、预防和应对危机。

全球人工智能数据框架

建议6：全球人工智能数据框架

我们建议通过联合国国际贸易法委员会等相关机构发起的流程，在参考其他国际组织所开展工作的基础上，创建全球人工智能数据框架，其任务是：

a. 勾勒数据相关定义和原则，包括从现有最佳做法中提炼的定义和原则，以便对人工智能训练数据进行全球治理，并促进文化和语言多样性;
b. 就人工智能训练数据的来源和使用制定统一标准，以便建立跨司法管辖区的透明和基于权利的问责制;
c. 建立促进市场发展的数据管理和交流机制，推动全球范围内的当地人工智能生态系统蓬勃发展，例如：
    i. 数据信托;
    ii. 治理良好的全球市场，以便交流用于训练人工智能模型的匿名数据;
    iii. 促进国际数据访问和全球互操作性的示范协议，可能用作框架的技术法律协议。

167 在我们的咨询中，我们了解到，虽然有很多建议旨在促进更广泛的数据访问权和数据共享安排，以创建更多样化的人工智能生态系统，但迄今为止实现的建议并不多。这是发展包容性和充满活力的人工智能生态系统方面的一个关键差距。

168 部分答案在于人工智能训练数据的文化、语言和其他特征的透明度。识别代表性不足或“缺失”的数据也很有帮助。与此相关的是促进“数据共享空间”，鼓励为多个行为体管护训练数据。这些举措可以通过展示设计如何嵌入隐私、数据保护、互操作性和公平使用数据以及人权的技术法律框架来创造最佳做法。

169 如今，人工智能的数据市场有点像狂野的西部。“攫取你能攫取的东西并将其隐藏在不透明的算法中”的想法似乎是一个操作原则；另一个原则是在某些司法管辖区可强制执行的获取专有数据的排他性合同安排。存在这种排他性关系是因为联合王国竞争和市场管理局担心“[前沿模型]部门正在以可能产生负面市场结果的方式发展[44]”。

44 Competition and Markets Authority, *AI Foundation Models: Technical Update Report* (London, 2024).

方框14: 确保用于训练人工智能模型的数据安全： 数据授权、数据信托和跨境数据流安排

在许多情况下，数据需要受到保护(包括隐私、商业机密、知识产权、安全和保障等原因)，但将其用于训练人工智能模型也会对个人和社会带来好处。

法律上的数据权利通常是防止与数据有关的行为的权利。数据隐私权也是个人的个人权利。数据权利的构成可能会使人们难以灵活地行使数据权利，使数据在不丧失权利的情况下被用于某些目的，并作为一个群体集体行使数据权利。即使有可能灵活、积极地控制权限，这也往往需要更多的时间、专业技术知识和信心，而大多数人和组织都不具备这样的能力。

使数据所有者和主体能够安全、有限度地使用其数据，同时维护其权利的机制，可以说是数据赋权的手段。数据赋权可以使社会上更多的人和群体成为人工智能的积极合作伙伴和利益攸关方，而不仅仅是数据主体。目前已在开发用于安全管理访问的工具，包括用于引导跨境数据流的数据信托和隐私保护应用程序。

数据信托是一种机制，使个人和组织能够集体提供对其数据的访问，访问权由受托人控制。数据所有者可以设定访问、使用和目的的条款，由受托人行使。数据所有者和主体在为共同目标做出贡献的同时，也保留了自己的合法权利。在这些数据上训练出来的人工智能模型可能会比缺乏这种特定投入的模型表现得更准确，由此可更好地为该特定群体或更广泛的社会服务。

获取和使用，特别是跨界获取的管理机制都依赖于专门的法律框架。在实践中使用这些机制还需要适应各部门和社区的需要和情况。应找到并弥补数据管理方面的差距。今后能否成功和广泛地使用这些机制，将取决于技术保证和维持数据提供者的信任。

因此，我们建议为这些工具的进一步发展提供更多支持，并确定在哪些领域使用这些工具进行人工智能训练可以带来最大的公共价值。

170 因此，我们认为必须启动一个全球进程，让各种行为体参与其中，包括处于不同发展水平的国家，并得到联合国系统内外相关国际组织（经合组织、知识产权组织、世界贸易组织）的支持，为蓬勃发展的人工智能训练数据生态系统创建“防护栏”和“共轨” 。这一进程的成果不一定是具有约束力的法律，有可能是示范合同和技术法律安排。这些促进性安排可作为原则和定义框架的议定书逐一制定。

171 虽然完整的细节超出了我们的范围，但全球人工智能数据框架的关键原则将包括：互操作性、管理、隐私保护、赋权、权利增强和人工智能生态系统支持。

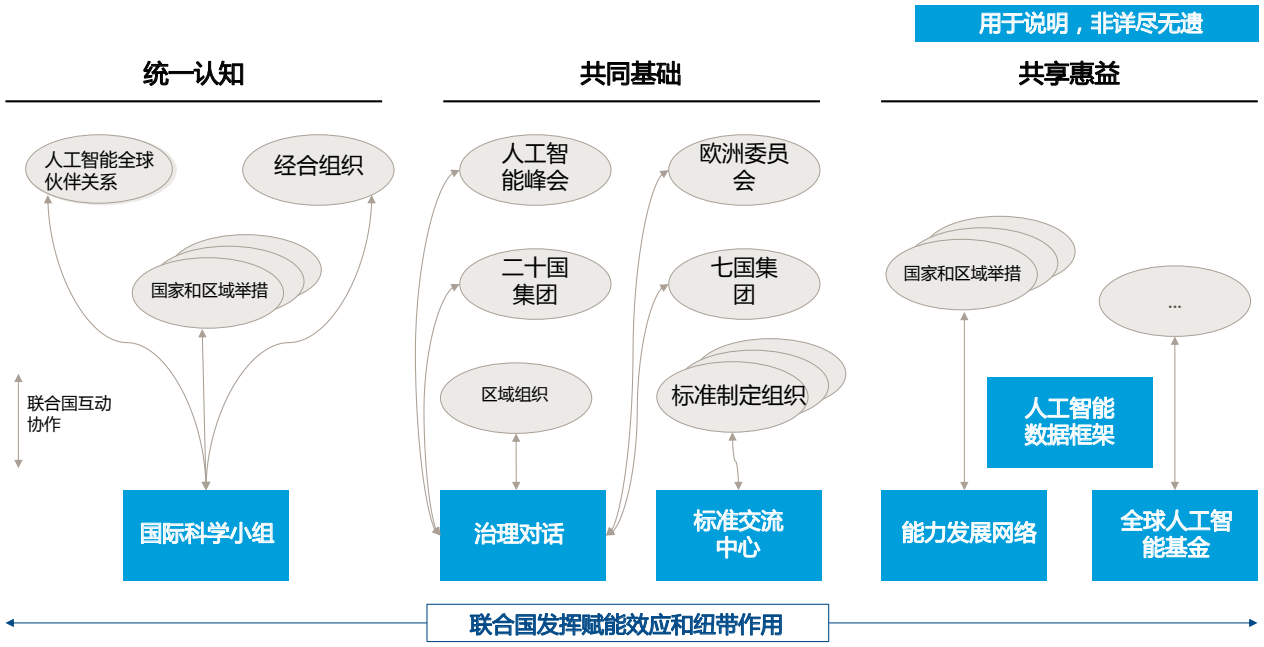
172 我们注意到，反托拉斯和竞争政策仍然是国家和区域当局的领域。然而，国际集体行动可促进本地人工智能初创企业跨境获取国内无法获得的训练数据。

173 联合国在支持制定人工智能训练数据治理和使用的全球原则和实际安排方面具有得天独厚的优势，以数据界多年的工作为基础，并将其与人工智能伦理和治理方面的最新发展结合起来。这类似于联合国国际贸易法委员会在国际贸易方面所做的努力，包括在法律和非法律跨境框架方面的努力，以及通过关于电子商务、云计算和身份管理的示范法促进数字贸易和投资的努力。

174 同样，科学和技术促进发展委员会和联合国统计委员会也将数据促进发展和可持续发展目标数据列入其议程。知识产权组织也在审议有关内容、版权以及保护本土知识和文化表现形式的重要问题。

45 包括加拿大、日本、大韩民国、新加坡、联合王国、美国和欧洲联盟。

图 17: 联合国在国际人工智能治理生态系统中的拟议作用



缩写：经合组织，经济合作与发展组织。

175 这里提出的框架将不影响国家或区域的数据保护框架，不会产生新的数据相关权利，也不会规定现有权利如何在国际上适用，但在设计上必须防止商业或其他利益集团攫取这些权利，从而破坏或妨碍权利保护。相反，全球人工智能数据框架将解决人工智能训练数据的可取得性、互操作性和使用等横向问题。这将有助于就如何调整不同国家和区域的数据保护框架达成共识。

176 在国家和区域层面解决这些问题的措施是有希望的，公共和私营部门更加关注最佳做法。然而，如果没有一个管理人工智能训练数据集的全球框架，商业竞争就会导致各司法管辖区之间在访问和使用要求方面竞相逐低，从而难以在国际范围内管理人工智能价值链。只有全球集体行动才能在人工智能训练数据的收集、创建、使用和货币化的治理方面促进“力争上游”。互操作性、管理、隐私保护、赋权和权利增强。

177 同样，这种行动对于促进当地人工智能生态系统的繁荣和限制进一步的经济集中是必要的。这些措施可以通过促进数据共享空间和在与可持续发展目标相关的领域提供托管数据信托来补充(见方框14)。能力发展网络和全球人工智能基金可为这些模板的开发以及共享空间或信托所托管数据的实际存储和分析提供支持。

D. 协同努力

178 通过促进统一认知、共同基础和共享惠益，上述各项提议旨在解决正在兴起的国际人工智能治理制度中发现的差距。可以通过与现有机构和机制建立伙伴关系和开展合作，弥合在代表性、协调和执行方面的差距。

179 但是，如果联合国没有专门的协调中心来支持这些努力，使之与其他努力之间实现软协调，并确保联合国系统在人工智能问题上发出一致声音，那么世界将缺乏包容各方的网络化、灵活和连贯一致的办法，而这种办法对于公平有效地治理人工智能不可或缺。

180 因此，我们建议在联合国秘书处内设立一个具有灵活能力的小型人工智能办公室。

方框 15: 人工智能办公室可能行使的职能和第一年的交付成果

人工智能办公室应采用精简型结构，做到灵活、可信和网络化。必要时应采用“中心辐射”方式运作，与联合国系统内外其他部分建立联系。

开展外联工作，包括可在会员国、多边网络、民间社会组织、学术界和科技企业之间所谓“软协调架构”中充当关键节点，在各方相互联系的制度综合体中，通过开展互动交流，以协作方式解决问题，此外，还可以作为安全可信的场所，就相关主题召开会议。一项富有雄心的目标是成为粘合剂，帮助将其他不断发展的网络凝聚在一起。

支持本报告所提各种倡议，其中一项重要职能是在交付各项成果的过程中，如在提交科学报告、开展治理对话和确定恰当后续实体时，确保迅速实现包容性。

统一认知:

- ·为国际科学小组的征聘提供便利并支持小组开展工作。

共同基础:

- ·为政策对话提供服务，听取多利益攸关方意见，以支持互操作性和政策学习。初期优先主题是阐明跨司法管辖区的风险门槛和安全框架。
- ·支持国际电联、国际标准化组织/国际电工委员会和电气电子工程师学会建立人工智能标准交换中心。

共享惠益:

- ·支持人工智能能力发展网络，初期重点是培养公职人员和社会企业家利用人工智能促进公共利益的能力。确定最初的网络愿景、成果、治理结构、伙伴关系和运行机制。
- · 确定全球人工智能基金的愿景、成果、治理结构和运行机制，并征求会员国、业界和民间社会利益攸关方对提案的反馈，以期在设立后 6 个月内为初始项目提供资金。
- ·编制并公布年度优先投资领域清单，为全球人工智能基金和该架构之外的投资提供指导。

协同努力:

- ·建立精简型机制，支持会员国和其他相关组织加强联系、增进协调、更有效地开展全球人工智能治理工作。
- ·编制指导和监测人工智能办公室工作的初步框架，包括全球治理风险分类法、全球人工智能政策格局审查和全球利益攸关方地图。
- ·制定执行季度报告，定期向会员国当面介绍人工智能办公室工作计划的进展情况，并建立反馈渠道，以支持必要的调整。
- ·建立一个由人工智能办公室、国际电联、贸发会议、教科文组织和其他相关联合国实体和组织共同领导的指导委员会，使联合国加速开展工作，为上述职能提供服务，每三个月审查一次加速工作的进展情况。
- ·与联合国训练研究所和联合国大学等相关联合国实体和组织合作，促进会员国代表共同学习和发展的机会，支持他们履行全球人工智能治理的责任。

在联合国秘书处内设立人工智能办公室

建议7：在秘书处内设立人工智能办公室

我们建议在秘书处内设立人工智能办公室，向秘书长汇报。办公室的组织结构应精简灵活，尽可能利用联合国现有相关实体。该办公室作为“粘合剂”支持和推进本报告中的各项提议，并与其他流程和机构合作互动，其任务包括：

a. 支持拟议的国际科学小组、政策对话、标准交流中心和能力发展网络，并向全球基金和全球人工智能数据框架提供所需支持；
b. 参与就新兴人工智能问题与科技企业、民间社会和学术界等多利益攸关方进行外联；
c. 就人工智能相关事宜向秘书长提供咨询意见，并与联合国系统其他相关机构进行协调，以便采取联合国一体化应对措施。

181 在磋商过程中，我们清楚地认识到，建立一个拥有报告、监测、核查和执行权力的机构的理由至今尚不充分，而且会员国也不太愿意建立一个昂贵的新组织。

182 因此，我们注重联合国能够提供的价值，同时铭记联合国系统的不足以及在一年内能够切实取得的成果。在这方面，我们提议建立一个精简灵活的机制，作为“粘合剂”，将促进统一认知、共同基础和共享惠益的各项流程结合起来，使联合国系统在不断演变的国际人工智能治理生态系统中发出一致的声音。

183 一些国家设立了专门机构和办公室[45]，重点开展国家、区域和国际层面的人工智能治理，同样，我们认为有必要建立一种能力，为国际人工智能科学小组和人工智能政策对话提供服务和支持，并加速创建人工智能标准交流中心和能力发展网络，这种做法的管理成本和交易成本比由不同组织为每项工作提供支持的成本更低。

184 在联合国秘书处内设立向秘书长汇报的人工智能办公室，其益处在于可以与整个联合国系统建立联系，而不与其中的某个部分进行绑定。这一点十分重要，因为人工智能的未来充满不确定性，而且很有可能渗透到人类活动的方方面面。

185 小型而灵活的人工智能办公室能够在人工智能治理问题上与各领域和各组织建立联系，有助于以动态的方式弥合差距，加强联合国内外的现有努力。人工智能办公室可以在其他举措之间发挥桥梁纽带作用，例如在区域组织领导的举措和其他多边举措之间建立联系，从而有助于降低各项举措之间的合作成本。

186 该机构应倡导包容，迅速建立伙伴关系以加快协调和执行工作，还应优先利用联合国系统内的现有资源和职能。办公室的工作人员可以是从联合国系统相关专门机构和其他部门借调的联合国工作人员。办公室应与民间社会、业界、学术界等多利益攸关方互动协作，并与经合组织等联合国以外的主要组织发展伙伴关系。

187 人工智能办公室将确保整个联合国系统共享信息，使联合国系统能够一致发出权威的声音。方框 15列出了人工智能办公室可能行使的职能和早期交付成果。

188 这项建议基于透彻地评估联合国可在哪些领域带来增值，包括发挥领导作用、填补空白、帮助协调，以及应在哪些领域让贤，从旁与现有努力密切合作(见图17)。落实这项建议的另一个好处是可利用现有机构安排，包括预先谈判商定的供资安排和业已为人熟知的行政流程。

189 应考虑到人工智能技术不断演变的特点。技术突破很可能极大地改变人工智能模式的现有格局。应切实有效地设立起人工智能办公室，以便根据不断演变的格局调整治理框架，应对人工智能技术带来的无法预见的发展。

E. 对制度模式的反思

190 关于人工智能的讨论往往走向极端。我们在世界各地进行了多次磋商，接触到的一部分人认为，未来日渐廉价和愈加有用的人工智能系统将提供无限机遇。与我们对话的另一些人则唯恐会出现更加黑暗的未来、分歧、失业、甚至灭绝

191 我们无从知晓未来来会发生什么。我们注意到，这项技术可能会朝着摆脱这种二元对立的方向发展。在本报告中，我们以科学为基础，重点关注近期的机遇和风险。本文概述的各项建议承载着我们最大的希望，即在获取人工智能惠益的同时，最大限度地减少和减轻风险。我们还注意到在开展更大规模的国际机构建设方面的实际挑战。正因如此，我们建议采用网络化的机构方法，并提供精简而灵活的支持。

192 不过，如果风险日益加剧，机会愈加利害攸关，那我们届时会对这种评估作出调整。在两次世界大战后诞生了现代国际体系；开发的各类武器威力不断增强，促使我们建立了限制武器传播和促进和平利用相关技术的制度。

193 随着对共同人性的理解逐步加深，我们构建了现代人权体系，并不断致力于为所有人实现可持续发展目标。气候变化则从一项局部关切演变全球性挑战。同样，人工智能可能在发展到一定阶段后，需要比本报告建议中提出的更多资源和权限。

194 我们的职权范围包括审议新的国际人工智能机构的职能、形式和时间表。在本报告最后，我们对这一问题进行了一些思考，尽管我们目前并不建议设立这样的机构。

国际人工智能机构？

195 如果人工智能的风险变得更加严峻、更加集中，会员国可能有必要考虑建立更强大的国际机构，使其拥有监测、报告、核查和执行权力。

196 这种演变有先例可循。从 1899 年和 1907 年《海牙公约》，到 1925 年《日内瓦议定书》，再到 1993 年《化学武器公约》，两用化学品的获取长期以来一直受到限制，其储存和使用受到议定书的规范，而将其用于武器制造的行为则被禁止。

197 生物武器同样被禁止，相关研究也同时受到定期限制，如 1975 年对 DNA 重组或基因拼接的限制。这些规定强调了将限制作为实验设计的基本考虑因素，使限制程度与估计风险挂钩。某些无法保证可被限制的高风险实验基本被禁止。其他实例还包括可能逾越基本伦理界限的研究，如目前对克隆人的限制，这一实例显示，人工智能研究有朝一日可能需要类似“红线”，同时还需要在执行方面进行有效合作。

198 持续的科学评估也是其中一些框架的特点，例如禁止化学武器组织设立的科学咨询机构和《生物武器公约》第十二条。

199 人工智能与核能之间的对比众所周知。从原子分裂那天起，科学家们就清楚地认识到，这项技术可以为善，尽管其研究是为了制造一种可怕的新武器。和现在一样，当时最强烈要求限制这项新技术的人中不乏顶尖科学家，这一点很能说明问题。

200 奠定国际原子能机构(原子能机构)核心基础的是一项“大协议”，即核能的有益用途可以共享，无论是在能源生产还是在农业和医药方面，以此换取不再利用核能发展武器的保证。核不扩散机制表明，良好的规范是有效监管的必要条件，但却不是充分条件。

201 这种类比的局限性显而易见。核能涉及一套明确界定的流程，针对的是分布不均衡的特定材料，创造核能力所需的大部分材料和基础设施掌握在民族国家手中。人工智能则是一个模糊的术语，其应用范围极其广泛，不同行业和国家都笼罩在人工智能最强大的能力之下。原子能机构的“大协议”侧重于造价昂贵且难以隐藏的武器；而人工智能的武器化则两者皆非。

202 为和平目的汇集核燃料的早期想法并未按计划实现。在汇集资源以共享技术惠益方面，更适合人工智能的类比可能是汇集了资金、人才和基础设施的欧洲核研究组织。不过，鉴于实验基础物理学与人工智能之间的差异，这种比较也有局限性，因为对人工智能需要采取更加分散的方法。

203 另一个不完美的类比是国际民用航空组织(民航组织)、国际海事组织(海事组织)等类似组织。运输领域的基础技术已经确立，其民事和军事应用很容易区分，而通用人工智能则并非如此。国家监管机构负责将民航组织和海事组织框架内制定的国际规范付诸实施，这些机构之间的网络也已建立。安全、促进商业活动和互操作性是关注重点。合规性并非按照自上而下的方式处理。

204 还有其他的合规方法可供借鉴。金融稳定委员会和金融行动特别工作组等机制使金融风险管理工作受益，且无需诉诸条约。

205 最终，如果人工智能监管需要强制执行，在全球层面建立某种机制对于正式划定“红线”可能必不可少。这种机制可能包含类似欧洲核研究组织所作的正式承诺，将汇集资源以便合作开展人工智能研究和分享惠益作为协议谈判的一部分。

206 然而，鉴于人工智能具有迅速、自主和不透明的性质，等待威胁出现可能意味着任何应对措施都将为时过晚。持续的科学评估和政策对话将确保世界不会措手不及。当然，任何启动正式进程的决定都将由会员国作出。

207 触发此类行动的可能门槛包括：预计正在开发不受控制或不受限制的人工智能系统，或部署无法追溯到人类、公司或国家行为体的系统。还可能包括人工智能系统出现“超级智能”的特征迹象，尽管当今的人工智能系统并不存在这种迹象。

208 第一个合理步骤是设立观察情况简报，让不同领域的杰出专家监测发展前景。可以责成科学小组对这一问题进行研究，作为其《季度研究文摘》系列的一部分。随着时间推移，政策对话可以成为分享人工智能事件信息的适当论坛，例如讨论令现有机构能力捉襟见肘、难以应对的事件，这一点类似于原子能机构在核安全和核安保方面相互保证的做法，或世界卫生组织(世卫组织)在疾病监测方面的做法。

209 拟议国际人工智能机构的职能可以借鉴相关机构的经验，如原子能机构、禁止化学武器组织、民航组织、海事组织、欧洲核研究组织和《生物武器公约》。这些职能可包括：

- 制定和颁布人工智能安全标准和规范；
- 监控可能威胁国际和平与安全或导致严重侵犯人权或违反国际人道法的人工智能系统；
- 接收和调查关于事件或滥用的报告，并报告严重违规事件；
- 核查遵守国际义务的情况；
- 协调人工智能安全事件的责任追究、应急响应和损害补救；
- 促进为和平利用人工智能开展国际合作。

210 在设计任何未来的人工智能机构时，都需要采取量身定制的方法，并酌情借鉴其他机构的经验教训(见方框 16)。

方框 16：从过去的全球治理机构中汲取的经验教训

人工智能是一套独特的技术，其风险和社会影响超越国界。但这并不是促成全球人工智能治理安排的第一套技术。民用航空、气候变化、核能和恐怖主义融资也是需要全球共同应对的复杂多维领域。

其中一些领域建立了新的联合国机构，如民用航空、气候变化和核能领域。另一些领域，特别在是保护全球资金流动领域，则设立了非条约机构，不过，这些机构提供了有力的规范框架、有效的市场执行机制和强大的公私合作伙伴关系。

当我们将这些机构应对措施与人工智能领域全新的努力进行比较时，不应过于关注各种人工智能问题最适合用哪种机构模式进行类比。我们在中期报告中建议，应该考虑有效和包容的全球人工智能治理需要哪些治理功能，以及可从以往的全球治理活动中学到什么。

第一个经验教训是，必须在科学技术层面对问题形成共同认识，才能采取共同接受的对策。在这方面，持续应对气候变化风险的气专委是一个有益典范。这一范例说明在不断演变的领域中，以包容的方式起草报告和形成科学共识，可以为研究人员和决策者创造公平的环境，形成对有效政策制定至关重要的共同认识。虽然起草和宣传气专委报告和全球评估的过程不乏挑战，但始终具有核心重要意义，有助于形成统一认知、建立共同知识基础、降低合作成本并引导《联合国气候变化框架公约》缔约方大会达成具体的政策成果。

对人工智能而言，随着技术的发展，形成共同的科学认识也同样重要。由于人工智能系统的能力不断进步，已知的有效缓解方法可能不足以应对潜在风险，国际科学小组可以随着新出现的需求而不断发展。

方框 16：从过去的全球治理机构中汲取的经验教训

第二个经验教训是，多方利益攸关方合作可以确立强有力的标准，促进快速反应。在这方面，民航组织和金融行动特别工作组是跨界管理高技术性问题的有益范例。在民用航空领域，民航组织的安全和安保标准由行业和政府专家制定，并通过市场准入限制强制执行，例如确保从纽约起飞的飞机可以在日内瓦降落，而不会触发新的安全审计。国际民航组织主导的安全审计和会员国驱动的审计相结合，确保了即使在技术不断发展的情况下也能始终如一地贯彻执行。

金融行动特别工作组是七国集团于 1989 年为解决洗钱问题而设立，是软法律机构促进共同标准和实施工作的另一个实例。工作组的同行审查监督制度非常灵活；工作组的建议被广泛接受，从而给那些不合规的公司和会员国造成了声誉损失。即使国际资金流动的风险已发生变化，而其中最显著是恐怖主义融资和扩散融资的兴起，金融行动特别工作组的灵活结构和规范框架使其能够迅速做出反应，跟上复杂挑战的步伐。

因此，民航组织和金融行动特别工作组都以各自独特的方式制定了广泛认可的国际标准、衡量合规情况的国内框架以及可互操作的系统，以便应对某类风险和跨司法管辖区的挑战。民航组织通过市场准入激励机制和限制措施确保强制执行，而金融行动特别工作组则令不合规的行为承担声誉风险。二者都是人工智能治理可以参考的有用模式，展示了政府和其他利益攸关方如何通过合作创建相互关联的规范和法规网络，并让不合规的行为承担代价。

第三个经验教训是，全球协调对于监测和采取行动应对可能影响广泛的严重风险往往至关重要。金融稳定委员会和原子能机构的模式提供了重要范例。金融稳定委员会于 2009 年由二十国集团创建，旨在监测国际金融体系面临的系统性风险并发出警告。金融稳定委员会成员包括二十国集团的财政官员和国际金融与发展组织，其独特的组成方式使其在协调识别全球金融风险的工作时灵活敏捷且具有包容性。

原子能机构的核保障方法体现了一种不同的模式。该机构的全面保障监督协定由 182 个国家签署，是联合国范围最广的合规保障制度的一部分。通过将视察与监测相结合，并以安全理事会采取行动作为威胁，原子能机构的模式或许是对不合规会员国最明显的谴责。

金融稳定委员会和原子能机构的实例都表明，国际协调对于监测严重风险至关重要。随着人工智能的风险愈加清晰和突出，可能同样需要建立聚焦人工智能的新机构，最大限度地加强协调努力，监测严重的系统性风险，使会员国能在可能情况下进行干预，在风险发生前未雨绸缪。

第四个经验教训是，必须让所有各方都能获得研发所需的资源并分享惠益。欧洲核研究组织和原子能机构的经验都具有启发意义。欧洲核研究组织汇集了世界一流的学者和物理学家，对粒子加速器和其他造福人类的项目开展复杂的研究，并为物理学家和工程师提供培训。

同样，原子能机构也为获取技术提供便利，具体而言，是为获取核能和电离辐射技术提供便利。基本的权衡取舍简单明了：会员国遵守核保障监督措施，原子能机构则为和平利用核能提供技术援助。在这方面，原子能机构提供了一种向发展中国家传播技术惠益的包容性办法。原子能机构为核安全卓越中心网络提供便利的模式与我们关于能力建设网络化办法的建议非常相似。

如上文所阐释，人工智能是一套技术，需要以更加包容和公平的方式分享其惠益，特别是与全球南方国家分享惠益。因此，我们建议建立人工智能能力发展网络和全球人工智能基金。随着国际人工智能科学小组的工作使我们对人工智能的了解更加深入，而且随着负责任部署人工智能以支持可持续发展目标的需求日益迫切，联合国会员国可能希望更广泛地将这一职能制度化。如果会员国这样做，则应将欧洲核研究组织和原子能机构作为支持更广泛获取资源的有用模式，从中汲取经验教训，纳入全球人工智能治理总架构。

5. 结论：行动呼吁

211 作为专家，我们对人工智能的未来及向善潜力持乐观态度。但这种乐观态度的前提是，务实地看待风险以及当前架构和激励措施的不足。我们还必须以务实的态度对待国际质疑，这些质疑可能会阻碍实现公平有效治理所需的全球集体行动。这项技术过于重要，过于利益攸关，因而不能仅仅依赖于市场力量以及碎片化的国家和多边行动。

212 我们需要积极主动，建立明确目标。除了认识到机遇与风险并存之外，还要应对快速和跨领域变革带来的挑战。人工智能对下游各领域造成影响，很少有人不被触及。将人工智能的治理工作交给少数开发者或其所在国，将造成一种极其不公的局面，令大多数人被迫承受开发、部署和使用人工智能造成的后果，而他们却对造成这种局面的决策没有任何发言权。

213 过去一年内，全球对人工智能治理的关注和讨论让我们人类看到了希望。各国和各部门之间尽管存在分歧，但也有开展对话的强烈愿望。与不同地区、不同性别和不同学科的专家、政策制定者、商界人士、研究人员和倡导者的互动接触表明，多样性未必会导致不和，而对话可以形成共同基础，促进合作。

214 我们有时感到犹豫：是应脚踏实地，专注于看似可行的事情？还是要志存高远？最后，我们决心二者兼顾。我们的建议既反映了建立公平有效的全球人工智能治理制度的全面愿景，也认真思考了如何逐步贯彻实施。

215 我们感谢为审议工作作出贡献的众多人士、组织和会员国，包括联合国各机构的代表和秘书处工作人员，他们对联合国在这一复杂领域的能力和局限性作出了敏锐的评估。人工智能的治理问题不仅仅涉及管理这项技术所造成的影响，而且还关系到多边合作和多利益攸关方合作的未来。

216 当我们在五年后回首时，技术格局可能会与今天大相径庭。但如果我们坚持到底，克服犹豫和疑虑，五年后就能迎来一个对世界各地个人、社区和国家包容赋能的人工智能治理格局。最终，重要的不是技术变革本身，而是人类如何应对技术变革。

217 我们相信，本报告建议的功能和形式如果得到认真实施，就能建立一个灵活、适应性强的制度，与人工智能的发展保持同步，并有助于分享惠益和防范风险。它们可以帮助我们及时发现问题和机遇，利用共同原则和框架调整国际行动，促进国际合作，建设个人和机构应对变化的能力。

218 实施本报告中的建议还可以鼓励新的思维方式：培养协作和学习的心态，促进多利益攸关方参与和基础广泛的公众参与。联合国可以成为制定新的人工智能问题社会契约的平台，确保全球共同支持保护和赋能所有人的治理制度。这项契约将确保公平获得和分配机会，不将风险推卸给最弱势的群体，或转嫁给子孙后代，很不幸的是，这正是气候变化领域的现状。

219 我们期待继续开展这场重要对话，无论是作为一个团体，还是作为来自多个专业领域、组织和世界各地的个人。我们与在这段旅程中建立联系的众多伙伴及其代表的全球社会携手齐心，希望本报告助力我们治理人工智能以造福人类的共同努力。

附件

附件A：人工智能高级别咨询机构成员

卡梅·阿蒂加斯(联合主席)

詹姆斯·马尼卡(联合主席)

安娜•阿布拉莫娃

奥马尔•苏丹•乌莱马拉蒂法•阿卜杜勒卡里姆

埃斯特拉•阿拉尼亚

拉恩•巴利瑟

保罗•贝南蒂

阿贝巴•比尔哈内

伊恩•布雷默(联合报告员)

安娜•克里斯特曼

娜塔莎•克兰普顿

妮加特•达德

维拉斯•达尔

弗吉尼亚•迪格南

江间有沙

穆罕默德•法拉哈特

阿曼迪普•辛格•吉尔

温迪•霍尔

拉哈夫•哈尔福什

何瑞敏

北野宏明

高学湮

安德烈亚斯•克劳斯

玛丽亚•瓦尼纳•马丁内斯•波塞

赛义迪纳•穆萨•恩迪亚耶

米拉•穆拉蒂

彼得里•米吕迈基

阿朗德拉•纳尔逊

纳兹尼恩•拉贾尼

克雷格•拉姆拉尔

艾玛•鲁特坎普-布洛姆

玛丽切•沙克(联合报告员)

沙拉德•夏尔马

扬•塔林

菲利普•蒂戈

希梅纳•索菲娅•比维罗斯•阿尔瓦雷斯

曾毅

张凌寒

附件B：人工智能高级别咨询机构的职权范围

人工智能高级别咨询机构由联合国联合国秘书长召集，负责对人工智能的国际治理进行分析并提出建议。咨询机构的初步报告将为正在进行的国家、区域和多边辩论提供独立的高级别专家意见。

咨询机构由来自政府、私营部门、民间社会和学术界的 38 名成员以及1名成员秘书组成。机构组成在性别、年龄、地域代表性以及与人工智能风险和应用相关的专业领域保持平衡。咨询机构成员以个人身份任职。

咨询机构将与各国政府、私营部门、学术界、民间社会和国际组织广泛接触和协商。该机构将在与现有进程和平台互动以及利用多利益攸关方投入方面发挥灵活性和创新性。可以就特定主题设立工作组或小组。

机构成员由秘书长根据会员国提名和公开征集的候选人名单选出。咨询机构将设两名联合主席和一个执行委员会。所有利益攸关方团体在执行委员会中均有代表。

咨询机构首次任期为一年，秘书长可延长其任期。该机构将举行现场会议和在线会议。

咨询机构在 2023 年 12 月 31 日前编写第一份报告，供联合国秘书长和会员国审议。该报告对人工智能国际治理的各种备选方案进行高级别分析。

根据对第一份报告的反馈，咨询机构将在 2024 年 8 月 31 日前提交第二份报告，其中可能会就新的国际人工智能治理机构的职能、形式和时间表提出详细建议。

咨询机构应避免与审议人工智能问题的现有论坛和进程重复。相反，应设法利用在相关领域开展工作的现有平台和伙伴，包括联合国各实体。咨询机构应充分尊重现有联合国架构以及国家、区域和业界在人工智能治理方面的特权。

咨询机构的审议工作将由设在秘书长技术问题特使办公室的小型秘书处提供支持，并由预算外捐助资源提供资金。

附件C：2024 年咨询活动清单

活动	日期，2024年	地区
教科文组织，斯洛文尼亚	1 月 5 日	欧洲
秘书长科学咨询委员会	1 月 8 日	全球
向会员国介绍中期报告	1 月 12 日	全球
达沃斯世界经济论坛	1 月 24 日	欧洲
东南亚国家联盟(东盟)数字高级官员会议	1 月 30 日	亚洲
世界政府峰会	2 月 12 日	中东
蒙特利尔学习算法研究所(Mila – 魁北克人工智能机构)	2 月 14 日	北美
柏林咨询	2 月 15 日	欧洲
欧亚信息技术论坛	2 月 20 日	全球
世界移动通信大会	2月26日	欧洲
莫斯科国立国际关系学院	2 月 28 日	欧洲
皇家协会国际人工智能治理研讨会	2 月 28 日	欧洲
外交部科技咨询网络	2 月 28 日	全球
经合组织 - 非洲联盟人工智能对话	3 月 4 日	欧洲
布鲁塞尔咨询	3 月 5 日	欧洲
世界银行，全球数字峰会	3 月 5 日	北美
开放科学与人工智能：伦理问题网络研讨会	3 月 5 日	东欧
教科文组织数字化转型对话	3 月 6 日	欧洲
各国议会联盟	3 月 6 日	全球
方案问题高级别委员会第四十七届会议	3 月 11 日	全球
全球青年数字权利峰会	3 月 13 日	拉丁美洲
七国集团(G7)人工智能峰会，意大利特伦托	3 月 15 日	欧洲
咨询网络启动会议，3 月 18 日至 19 日	3 月 18 日	全球
妇女地位委员会第六十八届会议	3 月 21 日	北美
咨询机构向会员国通报最新情况	3 月 25 日	全球
负责任人工智能非洲观察站	3 月 25 日	非洲
人工智能促进可持续和包容性未来会议 - 法国开发署	3 月 26 日	欧洲
塑造全球规范：集体反馈	3 月 28 日	非洲
瑞士创新	4 月 2 日	欧洲
秘书长技术问题特使办公室访问中国，4 月 9日至12 日	4 月 9 日	亚洲
俄罗斯互联网治理论坛	4 月 9 日	东欧
沃顿“Cypher 日”- 金融	4 月 12 日	北美
访问硅谷	4 月 15 日	北美
斯坦福大学，人工+智能政策研讨会：全球盘点	4 月 16 日	北美
联合国科学技术促进发展委员会	4 月 16 日	欧洲
二十国集团(G20)数字经济，4 月 16日至18 日，巴西	4 月 17 日	拉丁美洲
咨询机构向会员国通报最新情况	4 月 22 日	全球
联合国大学，澳门人工智能大会 ，4 月 24日至25 日	4 月 24 日	亚洲
秘书长技术问题特使办公室访问布鲁塞尔和巴黎，4 月 25日至26 日	4 月 26 日	欧洲
咨询机构向国家人工智能咨询委员会介绍情况(美国)	5 月 2 日	北美
与伊斯兰世界教育、科学及文化组织在利雅得共同举办全球人工智能大会(53 个国家，4 个地区)	5 月 14 日	中东
人工智能促进可持续发展：哈萨克斯坦对《2030 年议程》的贡献	5 月 20 日	亚洲
拉丁美洲和加勒比国家组	5 月 21 日	拉丁美洲
金砖国家学术论坛	5 月 22 日	全球
首尔人工智能治理会议	5 月 23 日	亚洲
亚洲技术峰会，5 月 29日至31 日，新加坡	5 月 29 日	亚洲
智能向善全球峰会，5 月 29日至31 日	5 月 29 日	欧洲

附件D：“深入研究”清单

领域	日期(东部夏令时间)
教育	3 月 29 日
知识产权与内容	4 月 2 日
儿童	4 月 4 日
和平与安全 (1)	4 月 12 日
和平与安全 (2)	4 月 29 日
农业(第 一次会议)	4 月 30 日
农业(第二次会议)	4 月 30 日
基于信仰	5 月 1 日
开放源码和技术指导	5 月 1 日
对社会的影响	5 月 3 日
性别	5 月 7 日
数据	5 月 13 日
工作的未来	5 月 13 日
标准(第一次会议)	5 月 14 日
标准(第二次会议)	5 月 14 日
和平与安全 (3)	5 月 20 日
环境	5 月 20 日
健康	5 月 22 日
法治、人权、民主	5 月 24 日

附件E：全球风险脉动调查的答复

应人工智能高级别咨询机构的请求，秘书长技术问题特使办公室开展了一次人工智能风险全球脉动调查，作为人工智能前景扫描活动的一部分，旨在了解世界各地专家对人工智能风险的看法。调查要求专家在答复时以个人身份(而非代表机构或雇主)发表观点。调查请专家评定预计人工智能技术变革的程度，并(单独)评定采用和应用人工智能速度加快或减缓的程度。

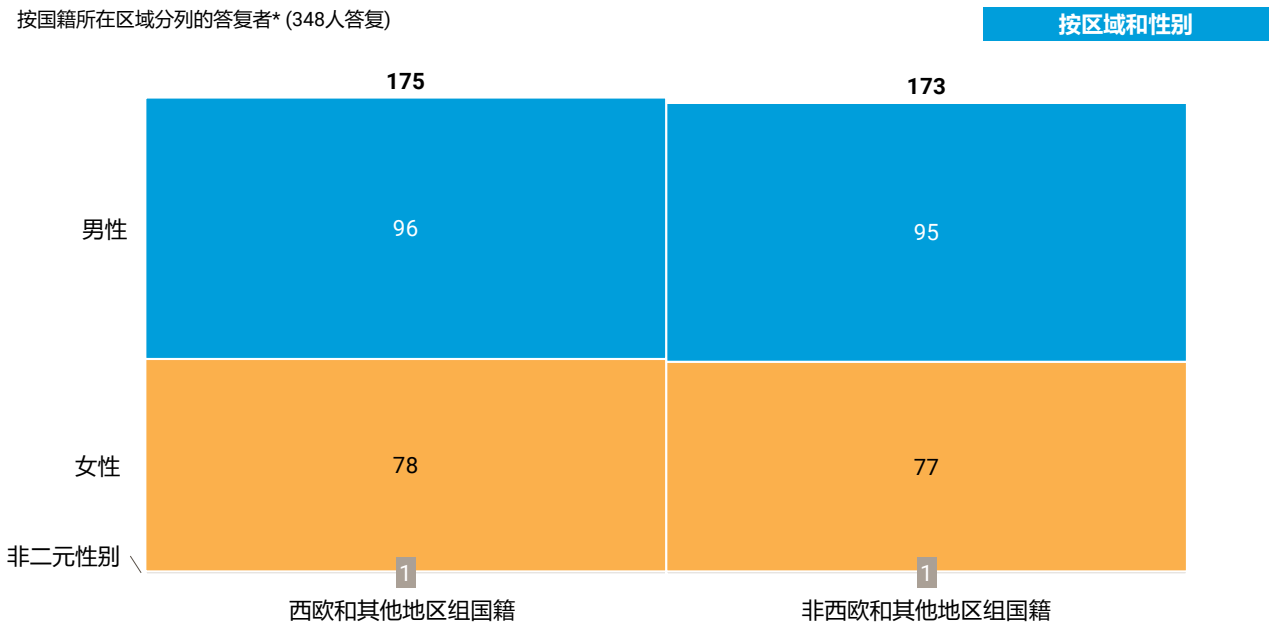
调查还要求专家评定他们对人工智能所造成(现有或新的)损害的严重程度和(或)覆盖范围将大幅增加的总体担忧程度，以及最近这种担忧加剧或减轻的程度。答复者使用14 个样本危害领域(如“非国家行为体故意恶意使用人工智能”)列表评定其担忧程度。最后，调查还提出了许多要求以文字作答的问题，请专家就新出现的趋势和面临特定人工智能风险的个人、群体和(生态)系统发表评论，并详细阐述他们作出的评定。

调查于 2024 年 5 月 13 日至 25日进行，受邀者名单来自秘书长技术问题特使办公室和咨询机构网络，其中包括咨询机构深度研究的参与者。在调查期间，根据最初答复者的推荐和与地区网络建立的联系，不断邀请更多专家参与，特别是在人工智能讨论中代表性不足地区的专家。340 多名受访者对调查做出了答复，就人工智能带来的风险提供了丰富多样的见解(包括不同地区和性别的观点）。

样本概览

按性别和区域划分显示分布均衡

按性别和地区进行的单变量分析不会立即受到其他变量的影响。

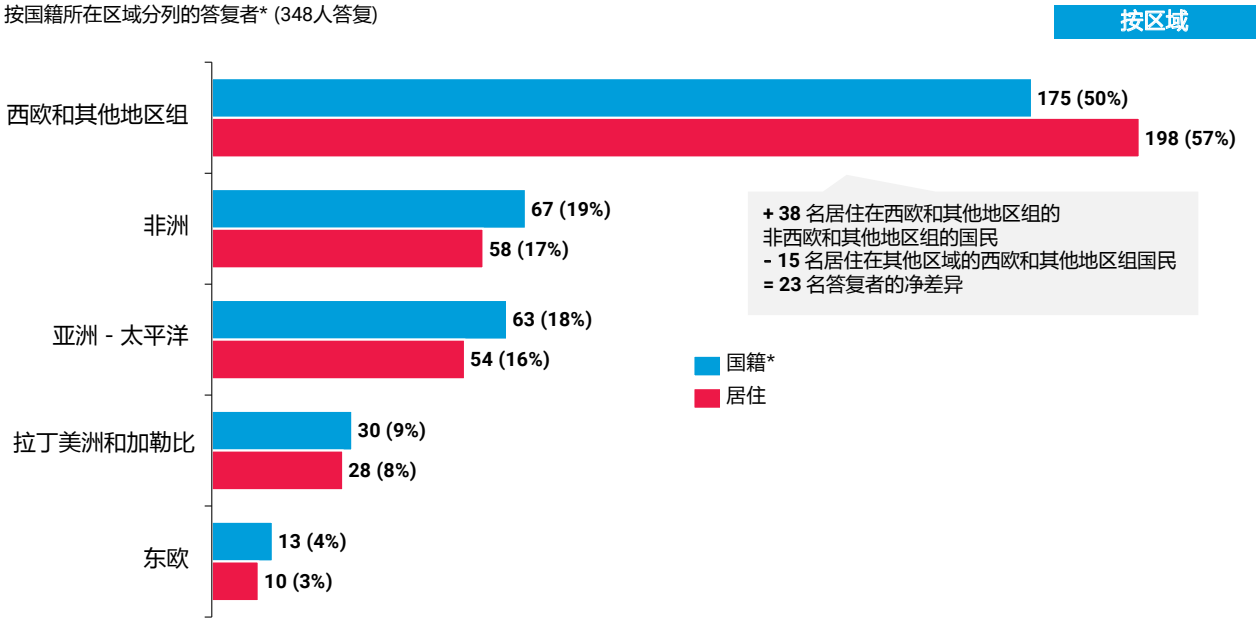


* 43名答复者(12%)表示拥有多个国籍。 如果答复者居住在其中一个国籍国，则使用该国籍进行分析(43人中有34人)，否则使用代表性最低的国籍(43人中有9人)。
信息来源：秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

如果按居住地考虑，样本仍然具有全球性

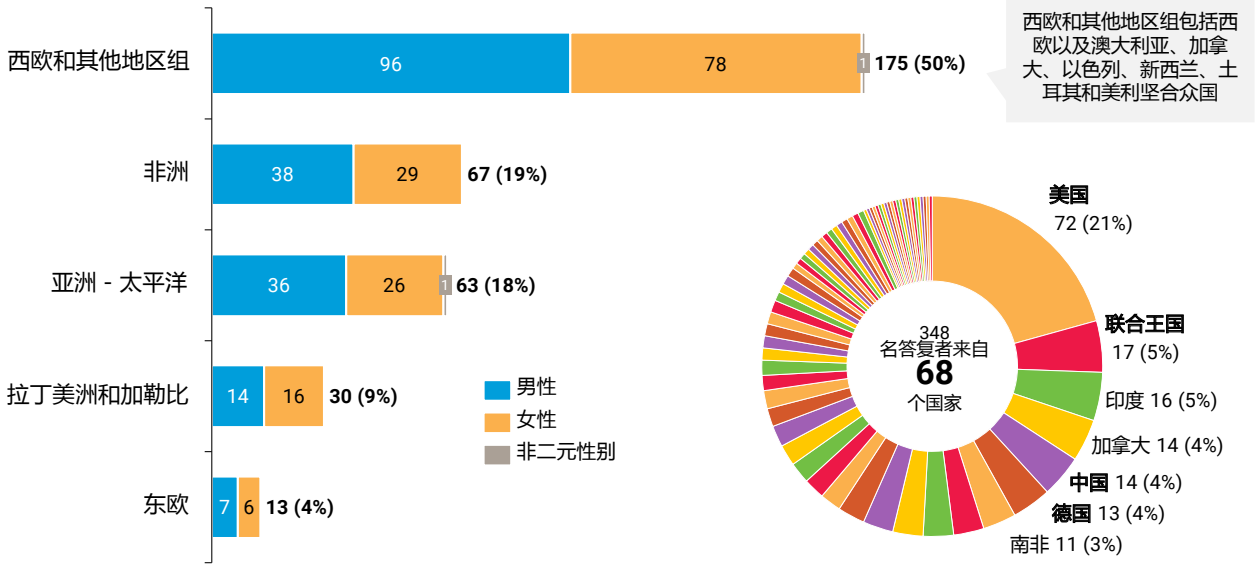
84%的答复者居住在国籍国所在区域

按国籍所在区域分列的答复者* (348人答复)



* 43名答复者(12%)表示拥有多个国籍。 如果答复者居住在其中一个国籍国，则使用该国籍进行分析(43人中有34人)，否则使用代表性最低的国籍(43人中有9人)。
信息来源: 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

按国籍所在区域分列的答复者* (348人答复)

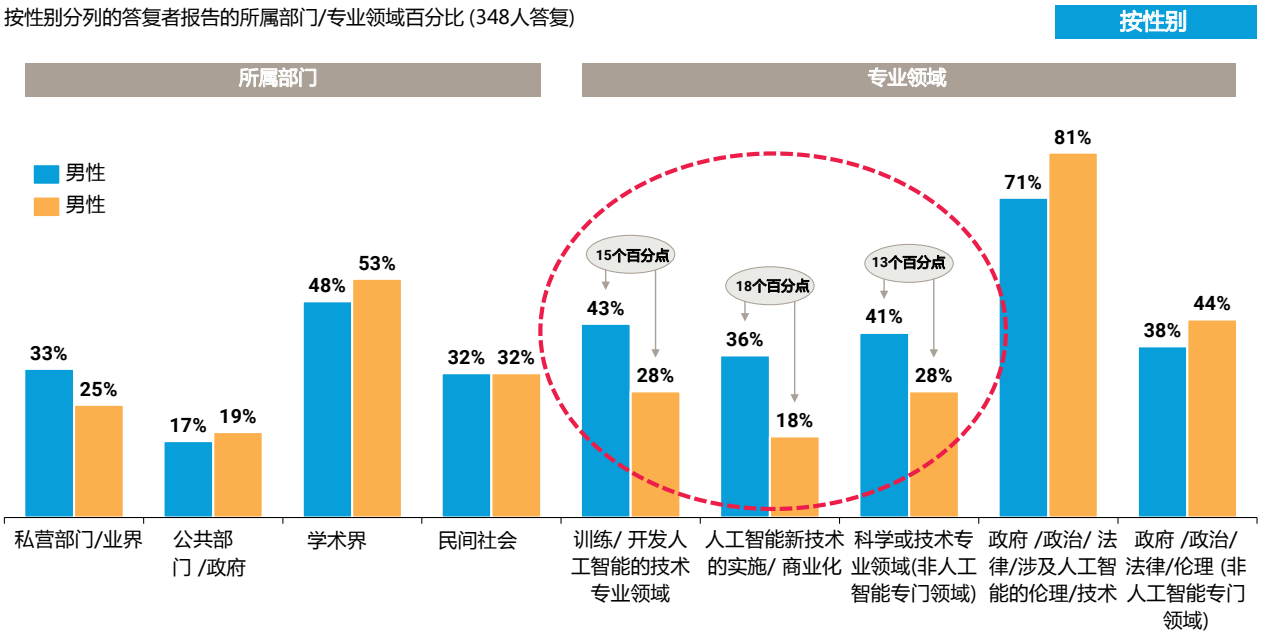


* 43名答复者(12%)表示拥有多个国籍。 如果答复者居住在其中一个国籍国，则使用该国籍进行分析(43人中有34人)，否则使用代表性最低的国籍(43人中有9人)。
信息来源 : 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

男女受访者的概况存在一些差异

更多男性报告属于技术专业领域；更多女性报告属于治理、政治、法律/伦理领域

按性别分列的答复者报告的所属部门/专业领域百分比 (348人答复)

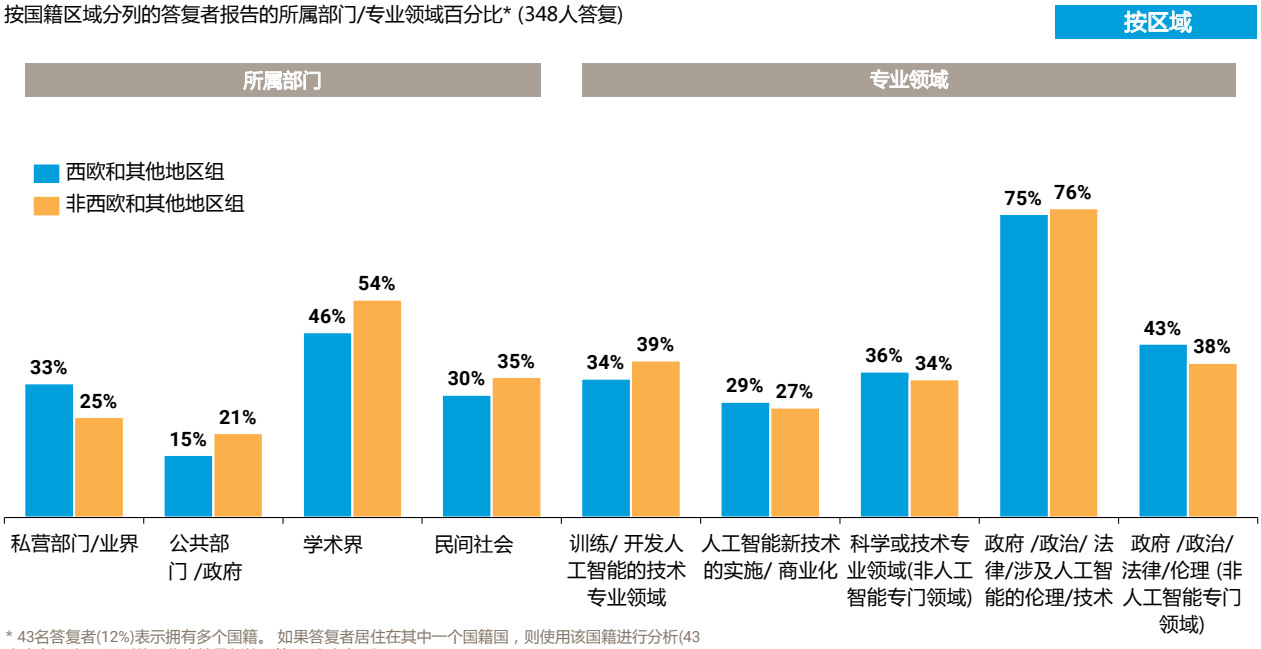


信息来源: 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

西欧和其他地区组与非西欧和其他西欧和其他地区组与非西欧和其他地区组答复者的情况相当类似

非西欧和其他地区组的答复者更可能属于公共部门或学术界，而较少属于非私营部门或业界

按国籍区域分列的答复者报告的所属部门/专业领域百分比* (348人答复)



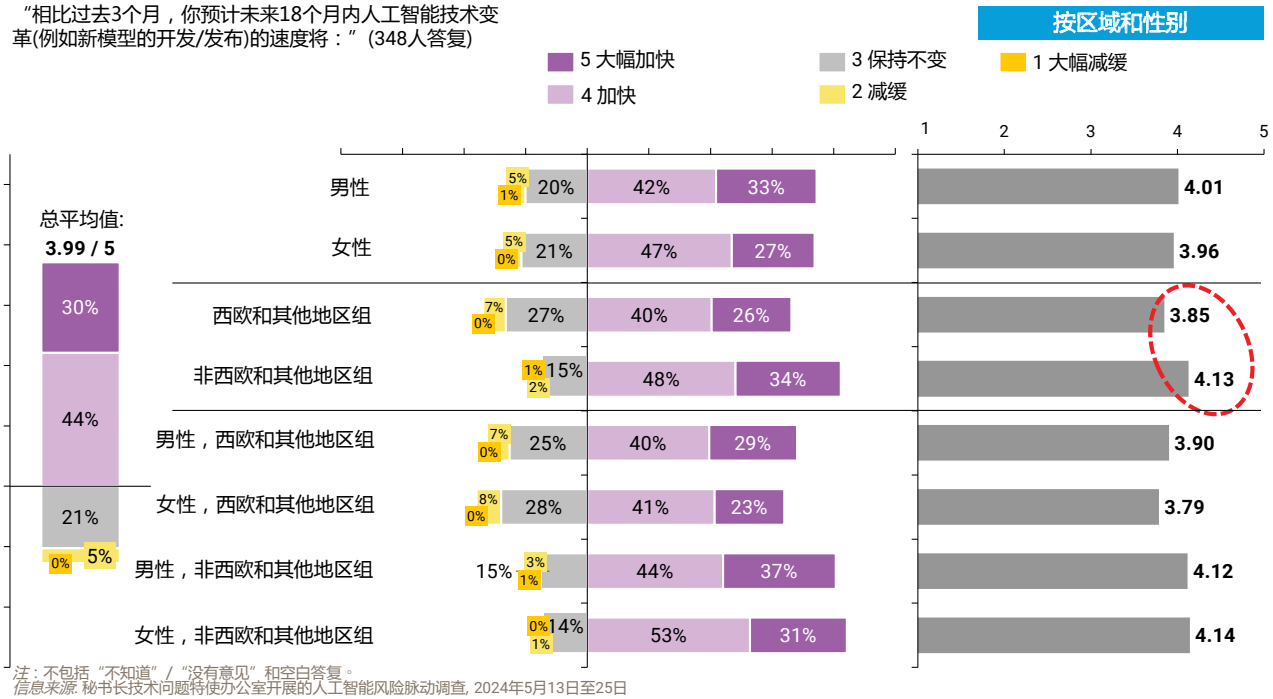
* 43名答复者(12%)表示拥有多个国籍。 如果答复者居住在其中一个国籍国，则使用该国籍进行分析(43人中有34人)，否则使用代表性最低的国籍(43人中有9人)。

信息来源: 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

对加速人工智能发展的看法

74% 的受访者预计技术变革将加速

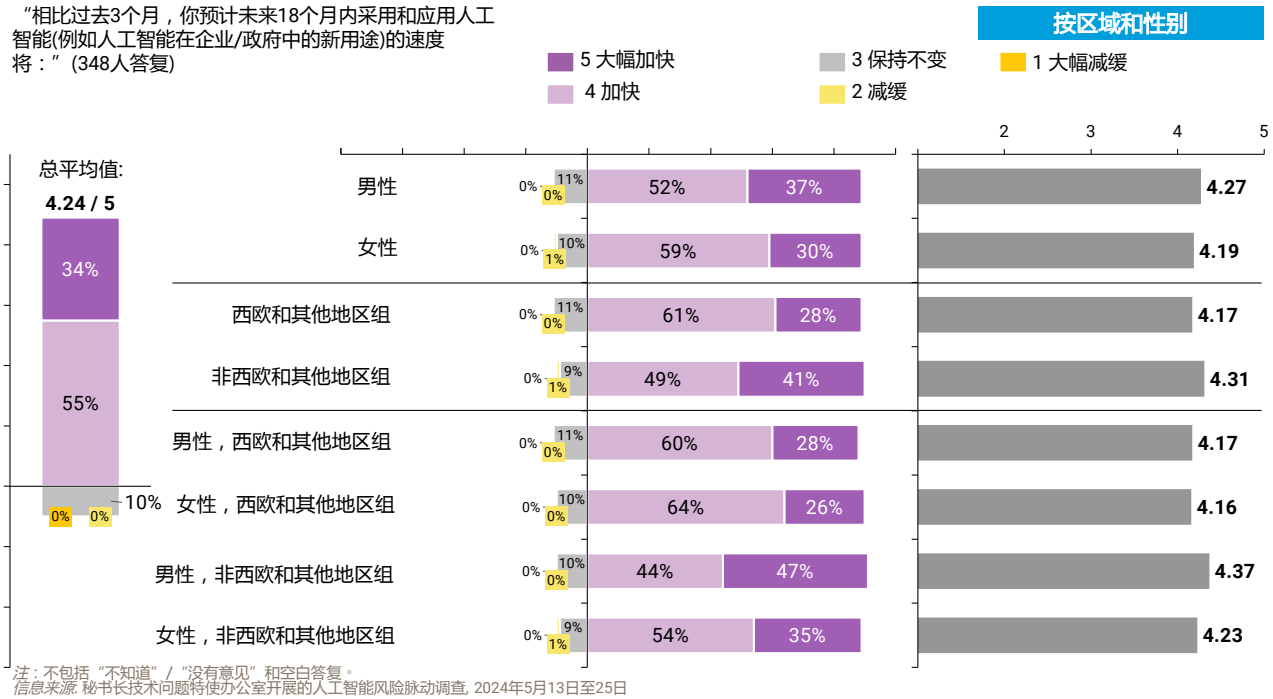
非西欧和其他地区组答复者预计技术变革速度将加快的比例高于西欧和其他地区组答复者



注：不包括"不知道"/"没有意见"和空白答复。
信息来源: 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

89%的答复者预计采用/应用人工智能的速度将加快

非西欧和其他地区组答复者预计将大幅加快的人数略多(特别是男性)



注：不包括"不知道"/"没有意见"和空白答复。
信息来源: 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

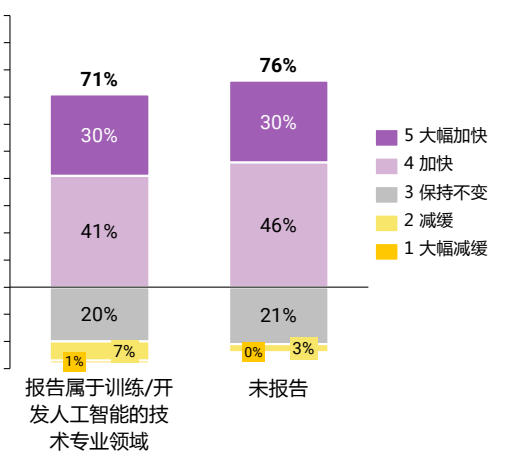
来自技术专业领域(训练/开发人工智能)的受访者认为影响有限

答复者对技术变革略微更加悲观，对采用/应用略微更加乐观

按专业领域

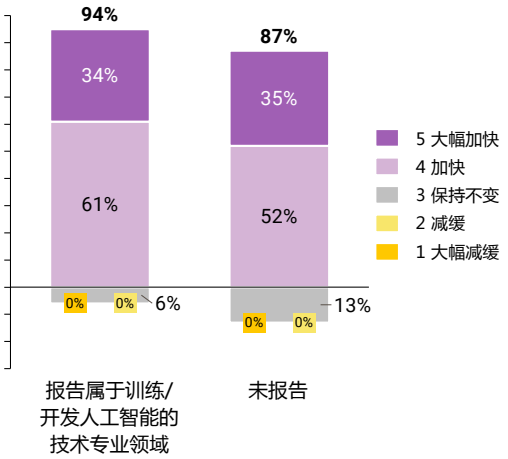
技术变革

“相比过去3个月，你预计未来18个月内人工智能技术变革(例如新模型的开发/发布)的速度将：”(348人答复)



采用和应用

“相比过去3个月，你预计未来18个月内采用和应用人工智能(例如人工智能在企业/政府中的新用途)的速度将：”(348人答复)



注：由于四舍五入，数字之和可能不等于100%。不包括“不知道”/“没有意见”和空白答复。
秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

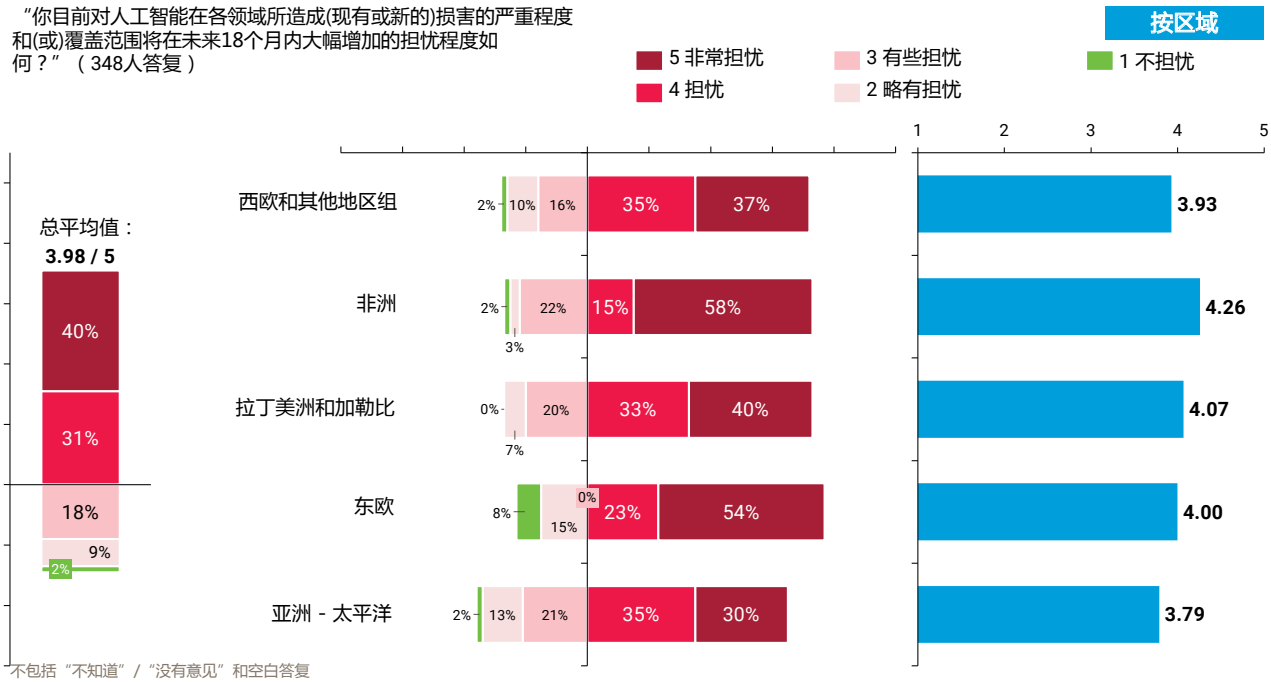
对未来 18 个月（自 2024 年 5 月起）人工智能危害风险的看法

71%的答复者对未来18个月内人工智能的危害感到担忧/非常担忧

非洲答复者比其他答复者更加担忧；亚洲-太平洋答复者的担忧程度低于西欧和其他地区组

“你目前对人工智能在各领域所造成(现有或新的)损害的严重程度和(或)覆盖范围将在未来18个月内大幅增加的担忧程度如何？”（348人答复）

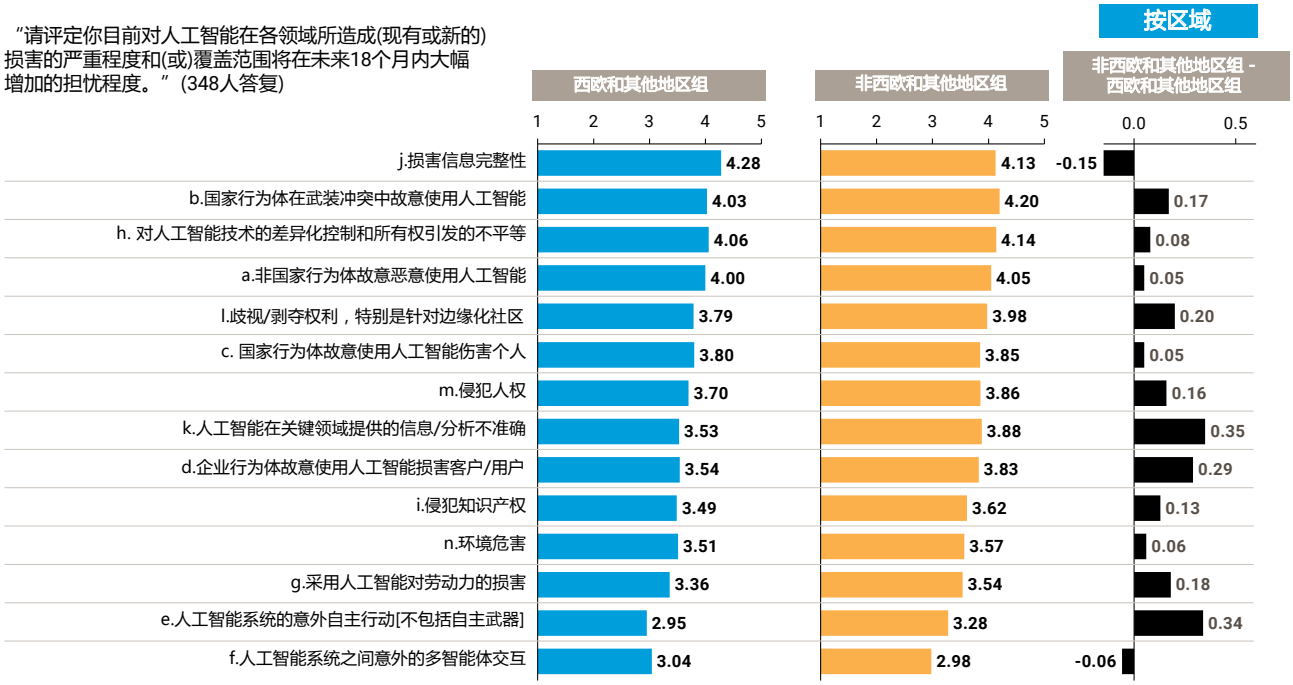
按区域



不包括“不知道”/“没有意见”和空白答复
秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日

在大多数示例领域，非西欧和其他地区组比西欧和其他地区组更加担忧

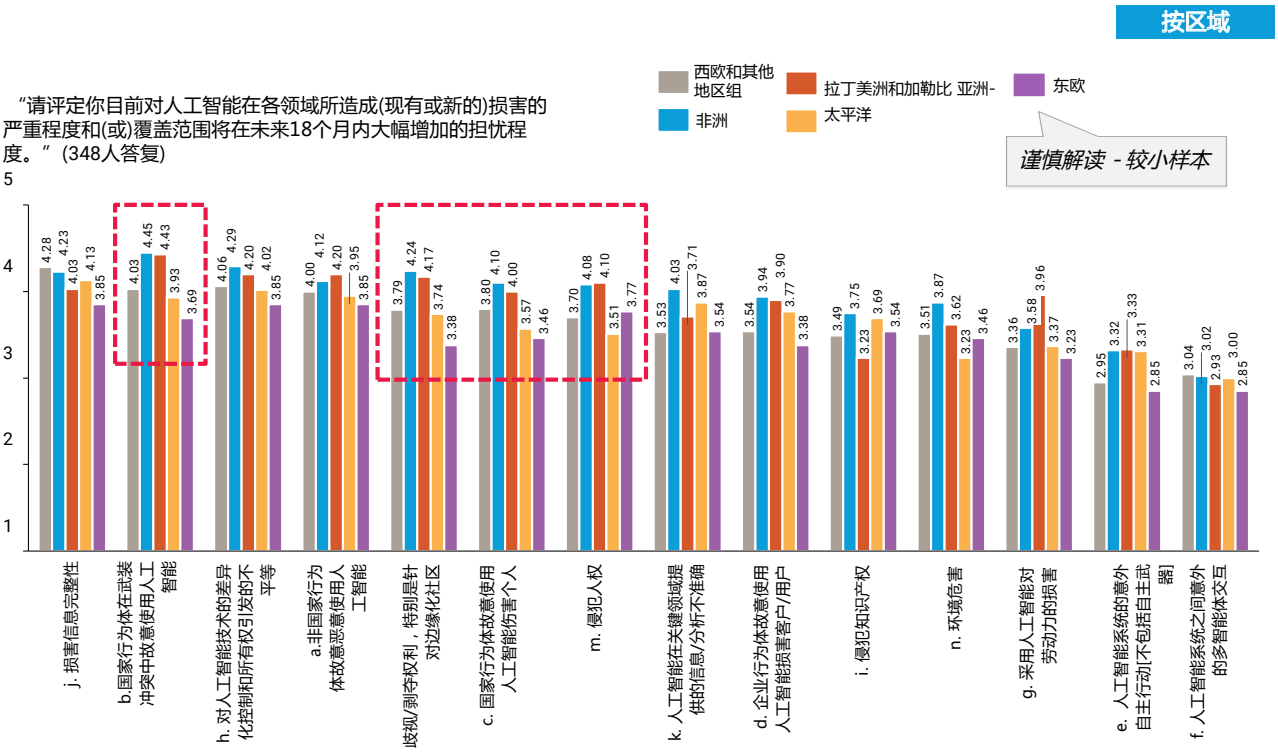
在信息不准确、意外自主行动、企业故意使用方面尤其存在较大差距



各项平均值：1 = 不担忧，2 = 略有担忧，3 = 有些担忧，4 = 担忧，5 = 非常担忧
不包括“不知道”/“没有意见”和空白答复. 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日.

非洲及拉丁美洲和加勒比对许多领域的担忧程度最高

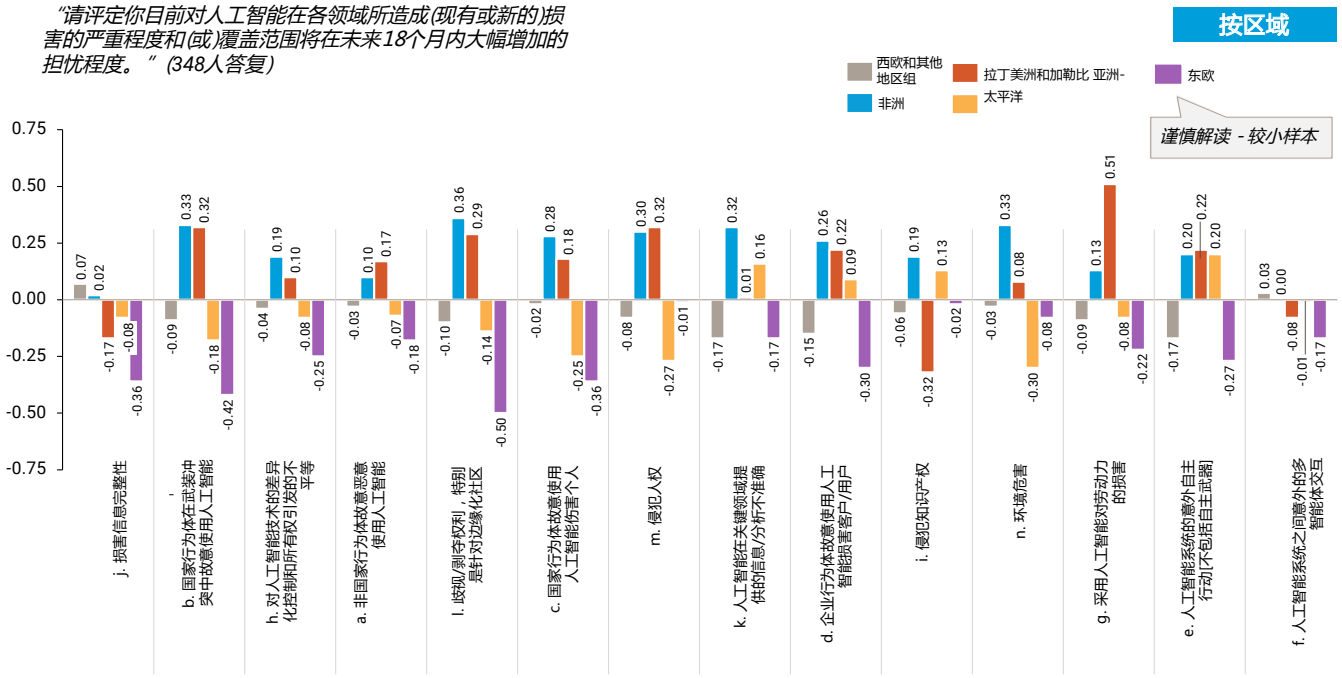
特别是在国家行为体在武装冲突中使用、造成歧视或侵犯人权领域



非洲及拉丁美洲和加勒比对许多领域的担忧程度最高

特别是在国家行为体在武装冲突中使用、造成歧视或侵犯人权领域

"请评定你目前对人工智能在各领域所造成(现有或新的)损害的严重程度和(或)覆盖范围将在未来18个月内大幅增加的担忧程度。"(348人答复)



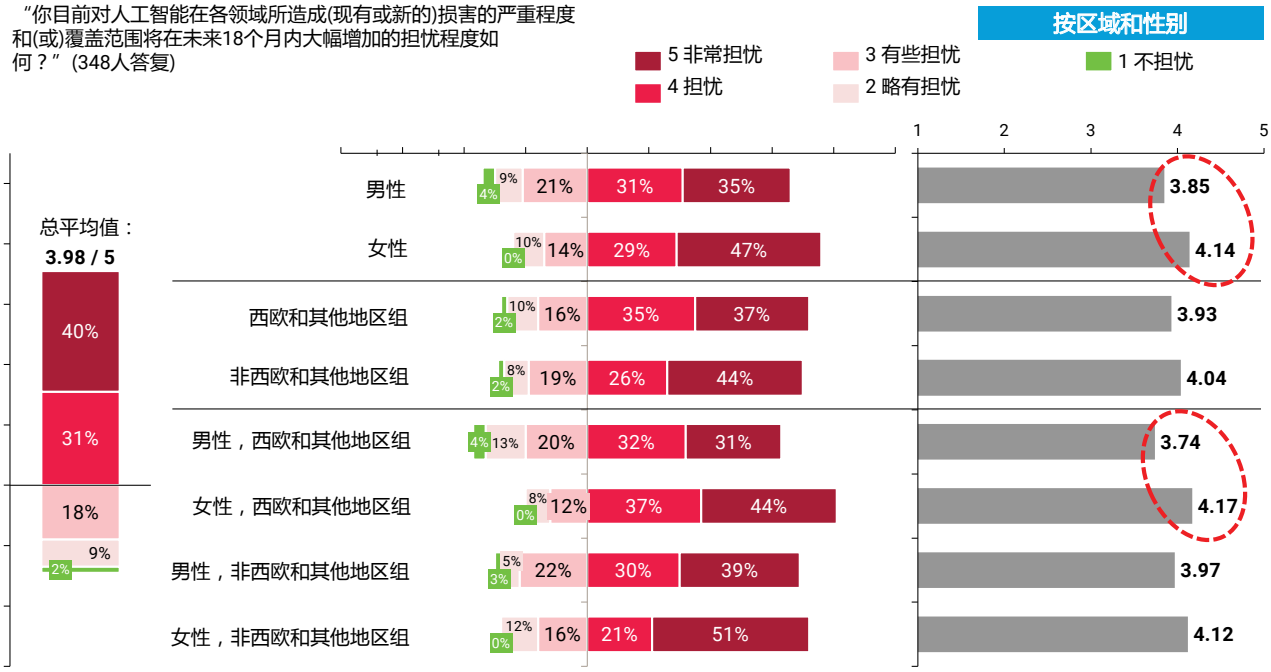
各项平均值：1 = 不担忧， 2 = 略有担忧， 3 = 有些担忧， 4 = 担忧， 5 = 非常担忧

不包括"不知道"/"没有意见"和空白答复. 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日.

71%的答复者对未来18个月内人工智能的危害感到担忧/非常担忧

女性比男性更加担忧，尤其是西欧和其他地区组的女性

"你目前对人工智能在各领域所造成(现有或新的)损害的严重程度和(或)覆盖范围将在未来18个月内大幅增加的担忧程度如何？"(348人答复)

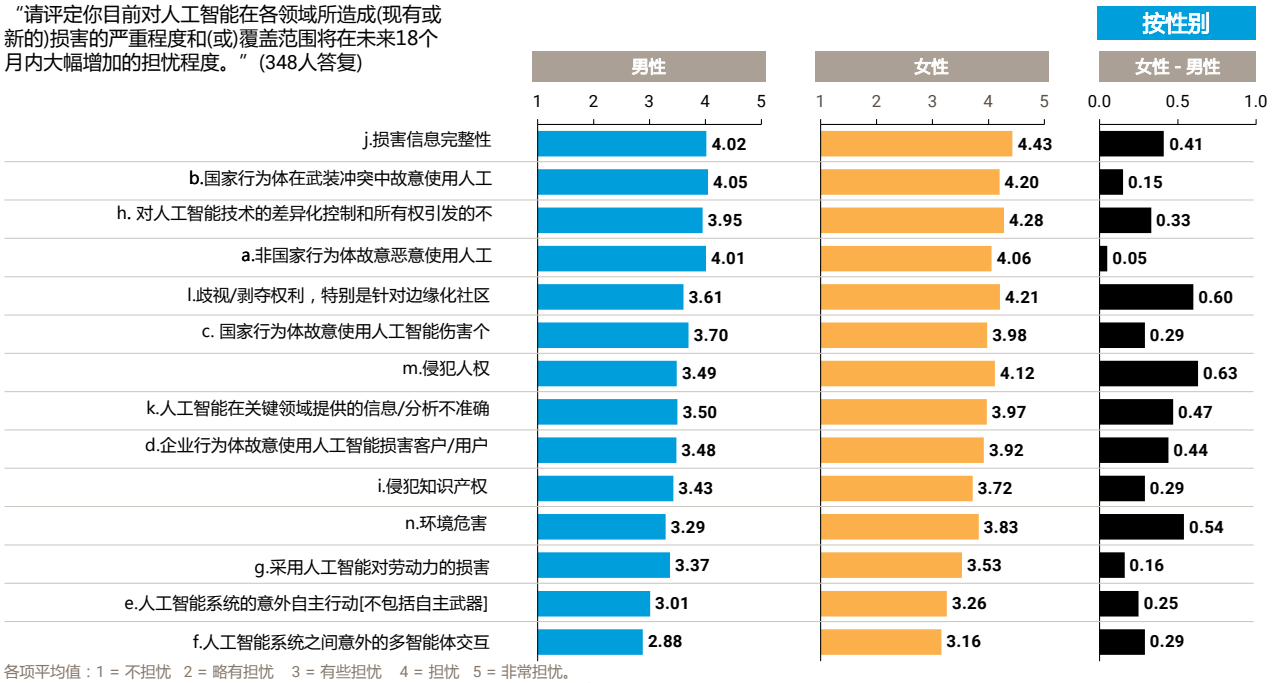


注：不包括"不知道"/"没有意见"和空白答复。

秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日。

在所有示例领域，女性比男性更加担忧

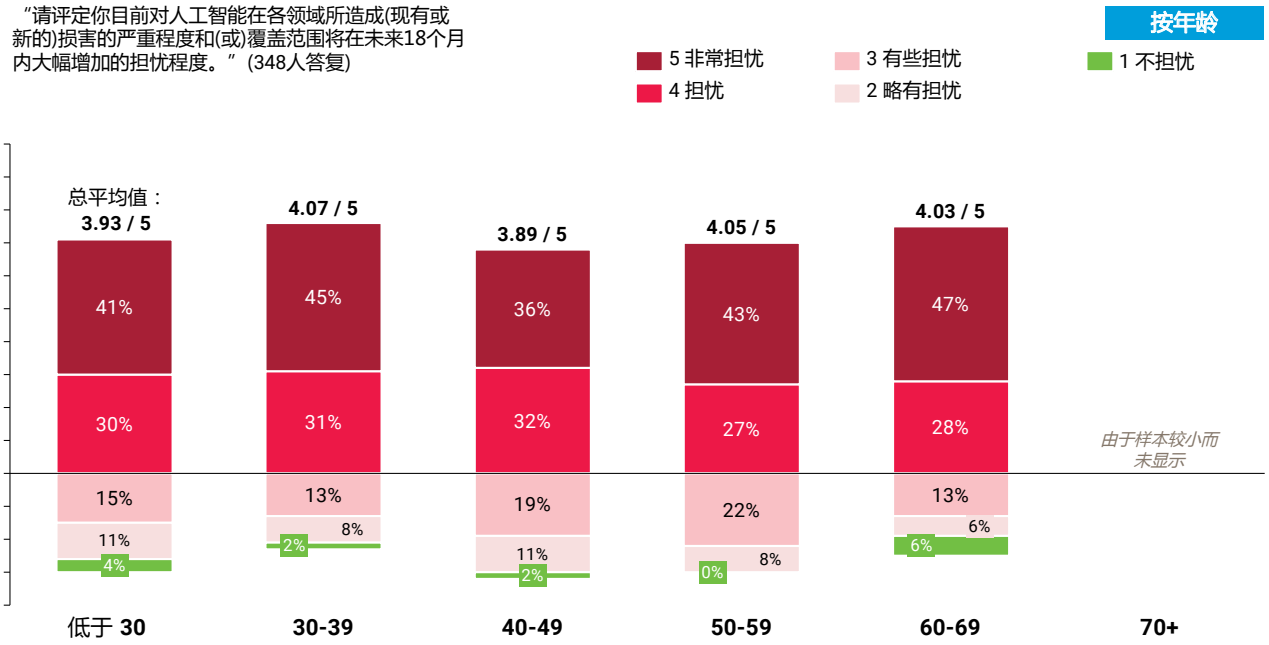
在侵犯人权、歧视和环境领域尤其存在较大差距



各项平均值：1 = 不担忧　2 = 略有担忧　3 = 有些担忧　4 = 担忧　5 = 非常担忧。
不包括"不知道"/"没有意见"和空白答复. 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日。

71%的答复者对未来18个月内人工智能的危害感到担忧/非常担忧

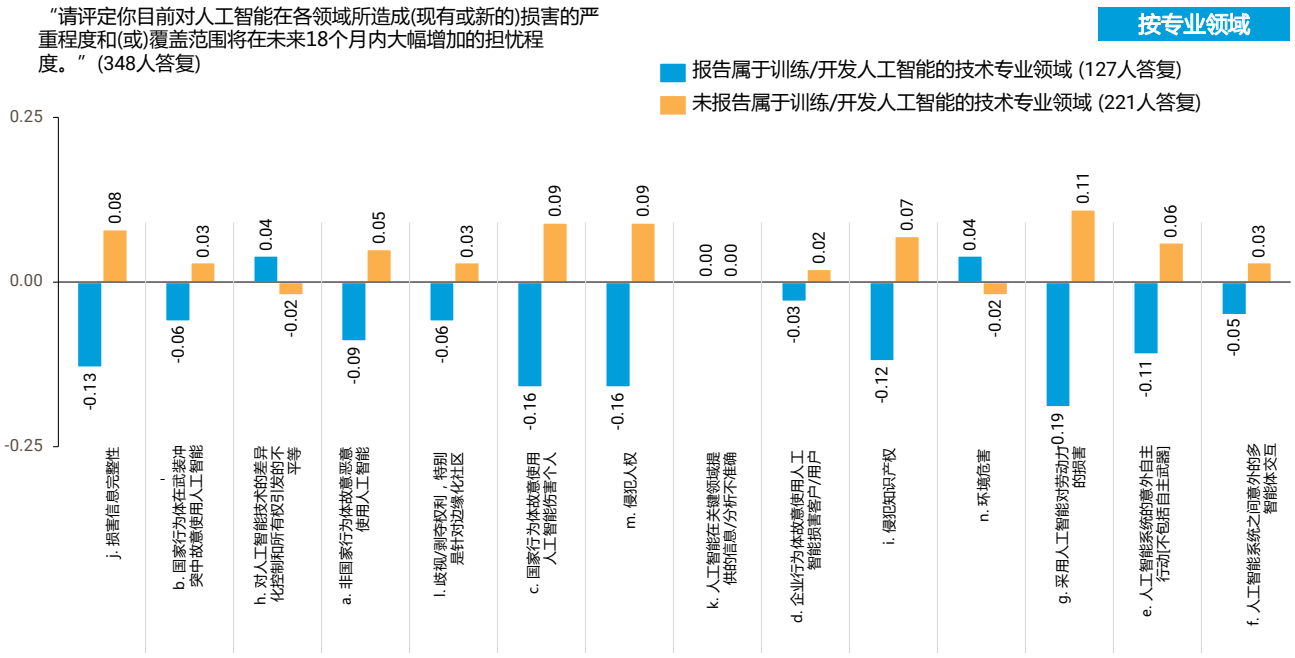
不同年龄答复者的担忧程度差异相对较小



不包括"不知道"/"没有意见"和空白答复。
秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日。

报告属于技术专业领域(训练/开发人工智能)的答复者对大多数示例领域的担忧程度较低

"请评定你目前对人工智能在各领域所造成(现有或新的)损害的严重程度和(或)覆盖范围将在未来18个月内大幅增加的担忧程度。"(348人答复)

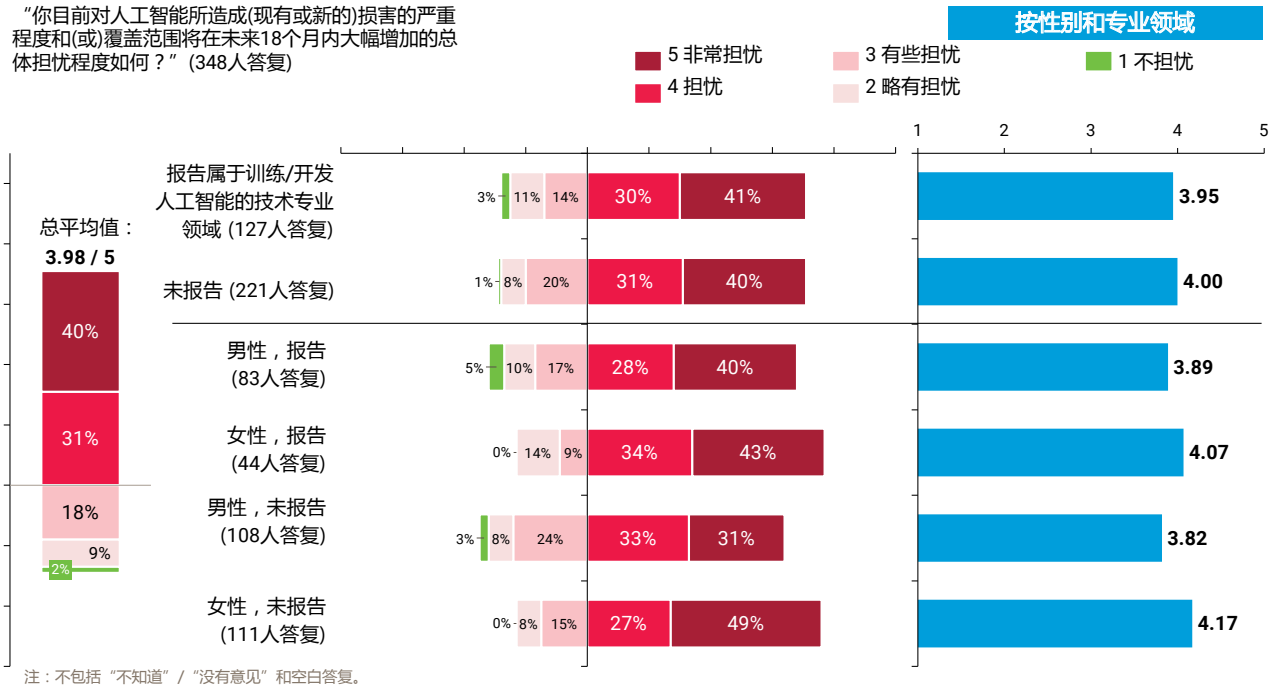


显示为汇总(所有地区)评分与所示区域评分之间的差异，其中：1 = 不担忧， 2 = 略有担忧， 3 = 有些担忧， 4 = 担忧， 5 = 非常担忧。
不包括"不知道"/"没有意见"和空白答复. 秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日。

来自技术专业领域(训练/开发人工智能)的受访者认为影响有限

无论报告的情况如何，男性担忧的程度低于女性

"你目前对人工智能所造成(现有或新的)损害的严重程度和(或)覆盖范围将在未来18个月内大幅增加的总体担忧程度如何？"(348人答复)



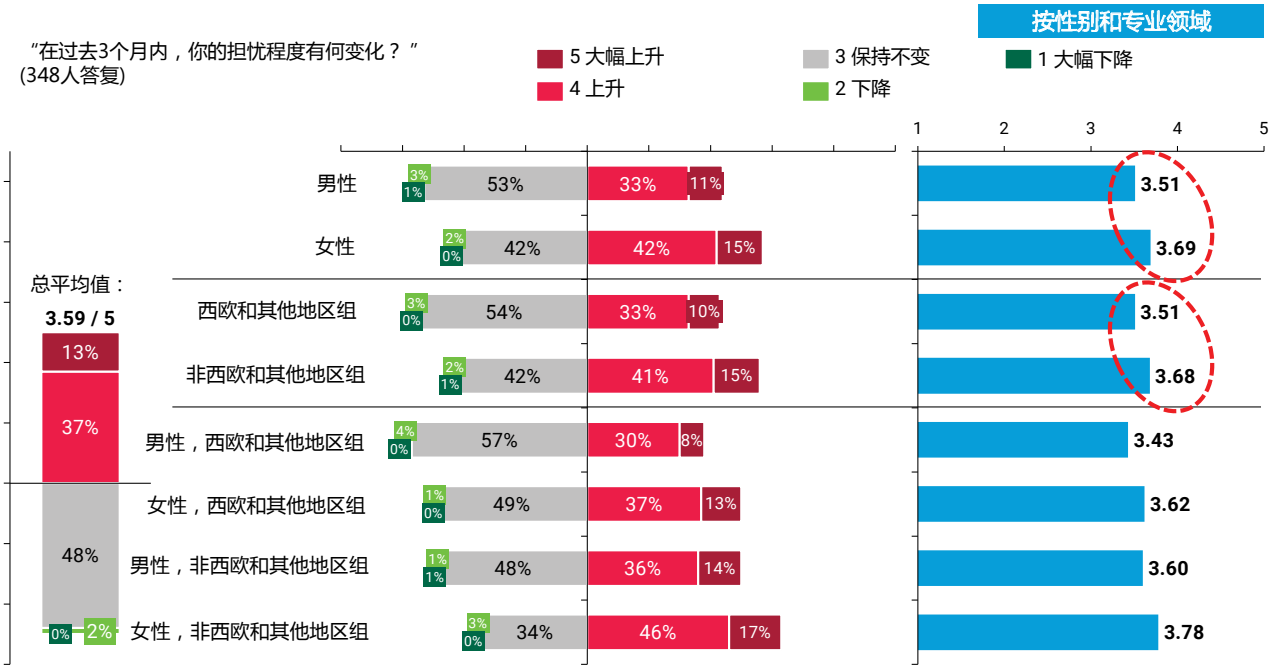
注：不包括"不知道"/"没有意见"和空白答复。

秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日。

过去 3 个月对人工智能危害风险担忧程度看法的变化

50%的答复者在过去3个月内的担忧程度上;
48%保持不变

几乎没有答复者担忧程度下降；更多女性、非西欧和其他地区组的答复者担忧程度上升



注：不包括"不知道"/"没有意见"和空白答复。

秘书长技术问题特使办公室开展的人工智能风险脉动调查, 2024年5月13日至25日。

附件 F：机会扫描答复

应人工智能高级别咨询机构的请求，秘书长技术问题特使办公室进行了全球人工智能机遇扫描调查。[1]在调查中，请各位专家以个人身份（而非代表其机构或雇主）回复其观点。调查分为几个部分，涵盖高收入/中高收入国家和中低收入/低收入国家的机遇，只有报告对中低收入/低收入国家情况具备专门知识的受访者可回答关于这些国家的问题。调查仅询问了人工智能可能产生的积极影响。

在调查中，询问了受访者在多大程度上了解人工智能迄今在增加经济活动、加速科学发现和推动各项可持续发展目标取得进展方面的具体例子。 请专家提供详细信息，包括案例研究、组织名称、数据以及相关文章/出版物/论文的链接。然后，询问了受访者预计未来三年将在这些方面取得多大进展。

为提供另一个视角，询问了受访者预计人工智能何时会在这些方面产生重大影响(有50%的把握/可能性)。其他问题包括：把握某些机会的过程涉及哪些行为体、哪些障碍加深了各国间的人工智能鸿沟、特定群体在利用人工智能机会方面是否面临额外限制以及如何消除这些限制。

调查于2024年8月9日至21日进行，在秘书长技术问题特使办公室和咨询机构网络的基础上拟订了受邀者名单，包括咨询机构深入研究的参与者。此外，国际电信联盟的人工智能造福人类会议以及联合国贸易和发展会议的网络都被广泛用于开展这项调查。总共邀请了1,000多人。120多名受访者对调查作出答复，为了解人工智能带来的机遇提供了丰富而多样化的视角（包括不同地区和性别）。

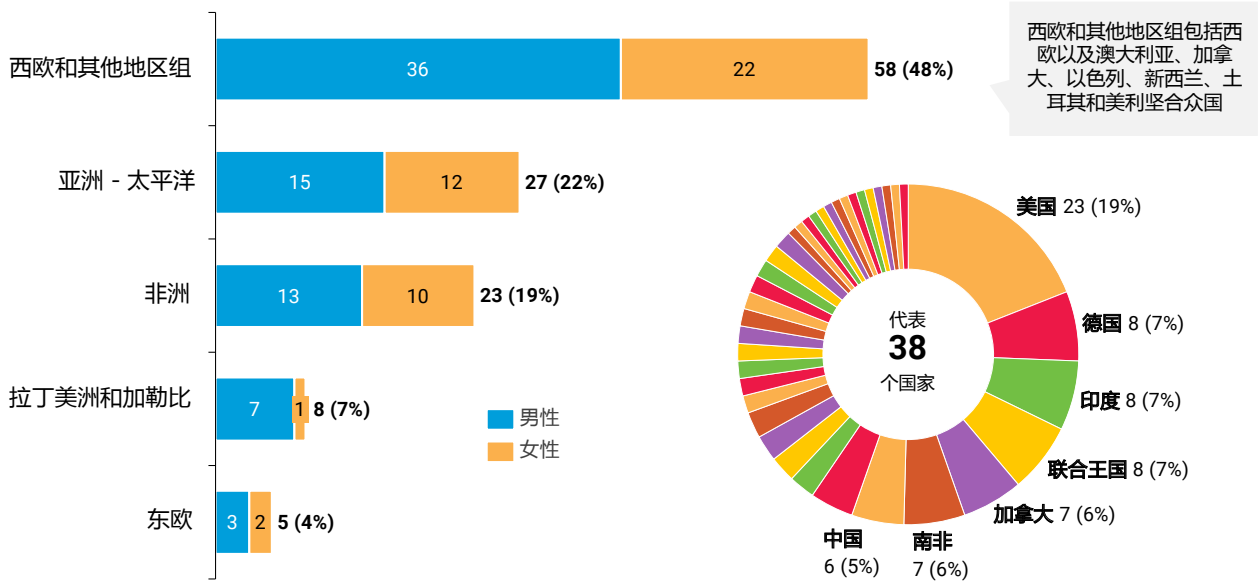
1 没有单独问及可持续发展目标8 (体面工作和经济增长)和可持续发展目标9 (创新、产业和基础设施)，因其与增加经济活动关系密切。也没有单独问及可持续发展目标17(促进目标实现的伙伴关系)。

样本概述

区域代表性：全球参与度高

可对西欧和其他地区组与其他区域的答复进行比较

按国籍所在区域分列的答复者*（121人答复）



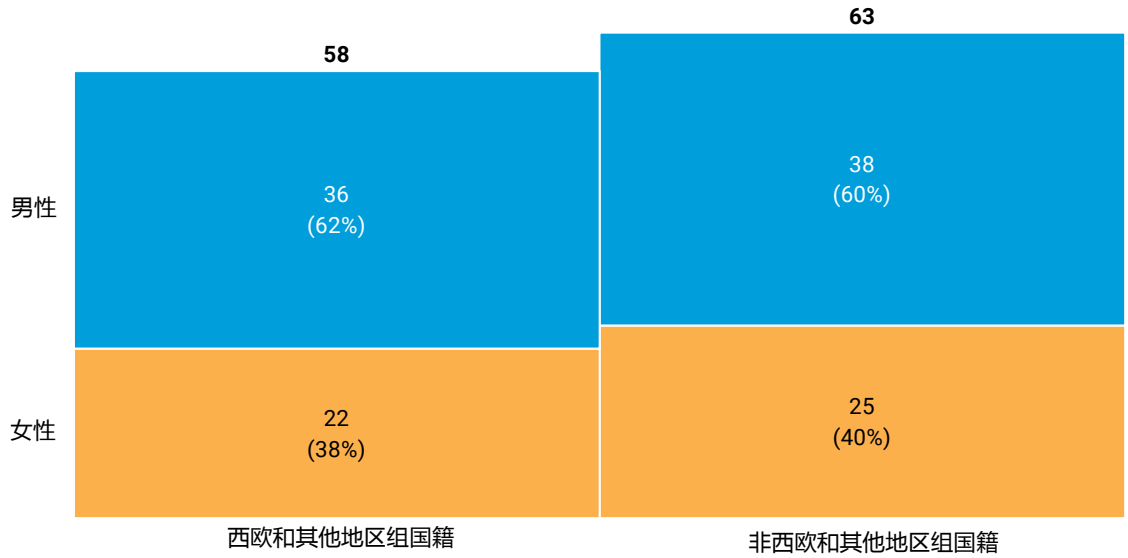
* 9名答复者(7%)表示拥有多个国籍。 如果答复者居住在其中一个国籍国，则使用该国籍进行分析(9人中有8人)，否则使用代表性最低的国籍(9人中有1人)。*信息来源*：联合国秘书长技术事务特使办公室人工智能机遇扫描调查，2024年8月9-21日

在西欧和其他地区组与非西欧和其他地区组的样本中，男性都占到约60%

保持一致性意味着按性别进行单变量分析，区域结果不会立即受到影响

按国籍所在区域分列的答复者*（121人答复）

按性别和专业领域

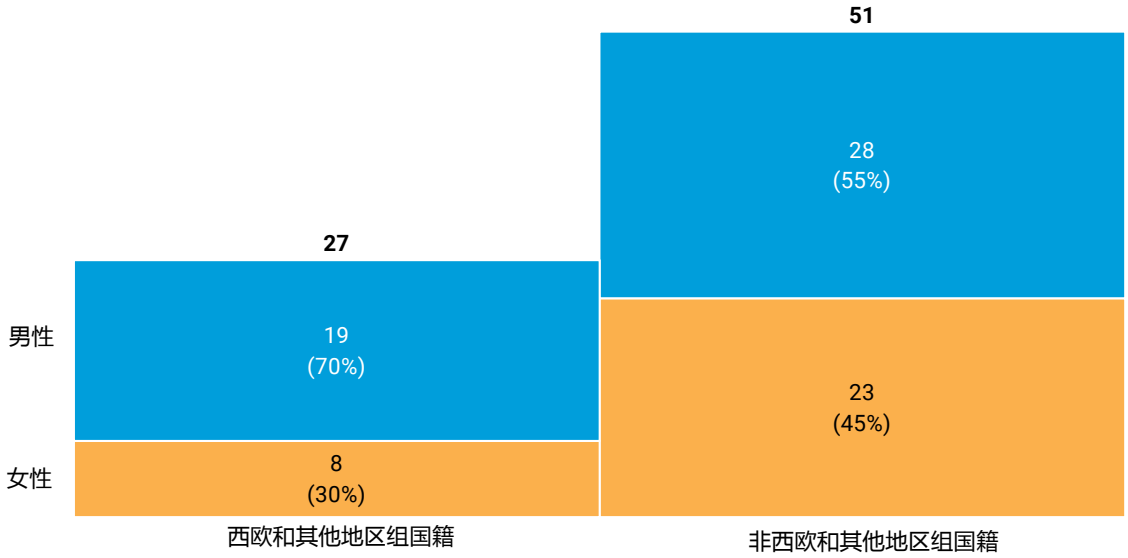


* 9名答复者(7%)表示拥有多个国籍。 如果答复者居住在其中一个国籍国，则使用该国籍进行分析(9人中有8人)，否则使用代表性最低的国籍(9人中有1人)。
信息来源：截至2024年8月21日的人工智能风险脉动调查结果（121人答复）。

熟知发展中国家情况的受访者样本分布不太均衡

报告者对中低收入国家背景的具体知识，按国籍地区分类* (78 人答复)

按性别和专业领域



* 9名答复者(7%)表示拥有多个国籍。 如果答复者居住在其中一个国籍国，则使用该国籍进行分析(9人中有8人)，否则使用代表性最低的国籍(9人中有1人)。信息来源：联合国秘书长技术事务特使办公室人工智能机遇扫描调查，2024年8月9-21日。

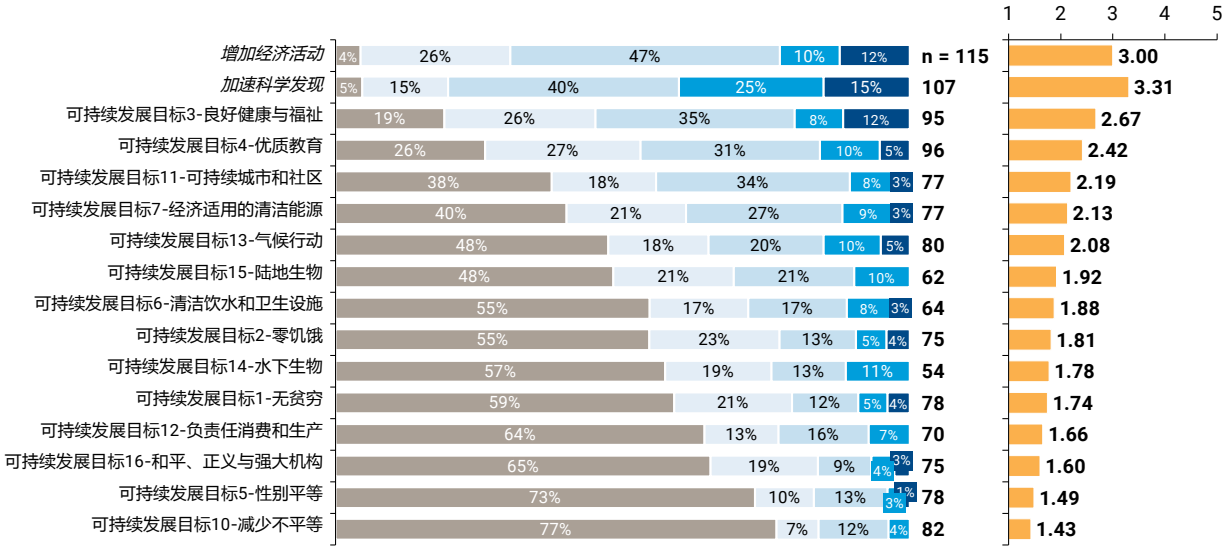
对人工智能迄今产生的积极影响的看法

迄今对增长和科学领域产生了积极影响，但对大多数可持续发展目标的积极影响较少

迄今对高收入/中高收入国家的影响

*"你在多大程度上了解人工智能目前或最近在高收入/中高收入国家为.....**作出直接贡献的具体例子**？"*

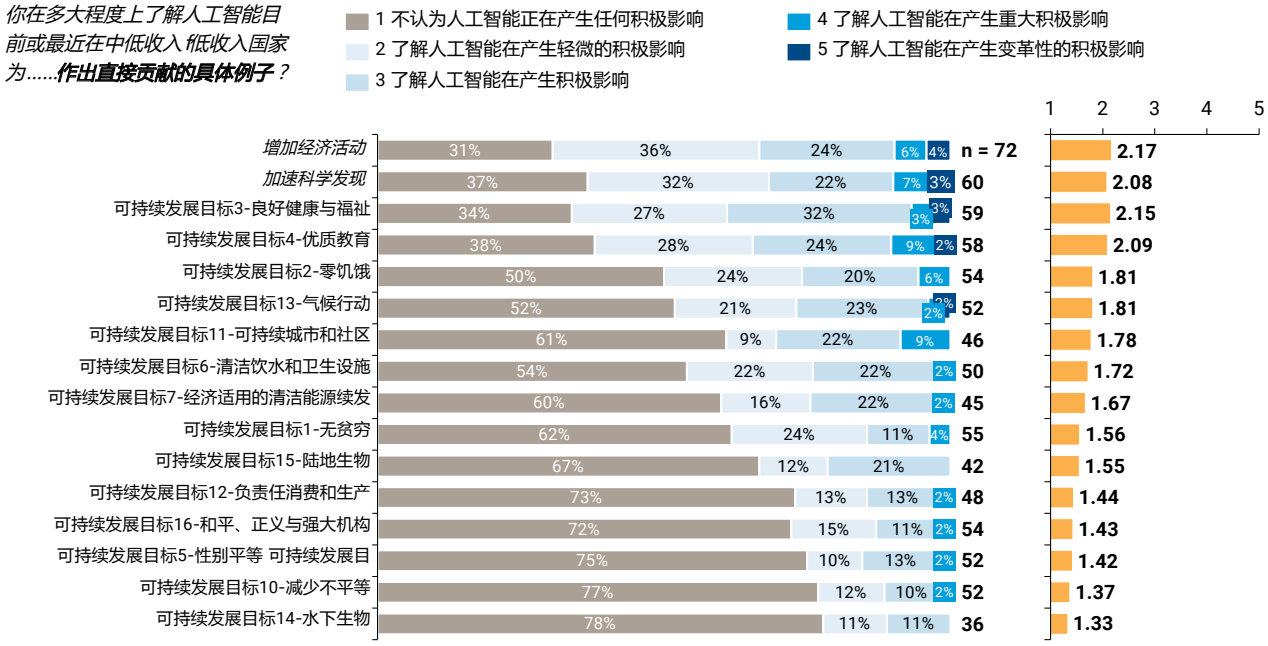
1 不认为人工智能正在产生任何积极影响
2 了解人工智能在产生轻微的积极影响
3 了解人工智能在产生积极影响
4 了解人工智能在产生重大积极影响
5 了解人工智能在产生变革性的积极影响



注：不包括"不知道"/"没有意见"和空白答复。在每个问题后分列了具体答复人数。未就可持续发展目标8、9和17提问。
联合国秘书长技术事务特使办公室人工智能机遇扫描调查，2024年8月9-21日。

受访者报告称，人工智能对较低收入国家各方面产生的影响较少

迄今对中低收入/低收入国家的影响

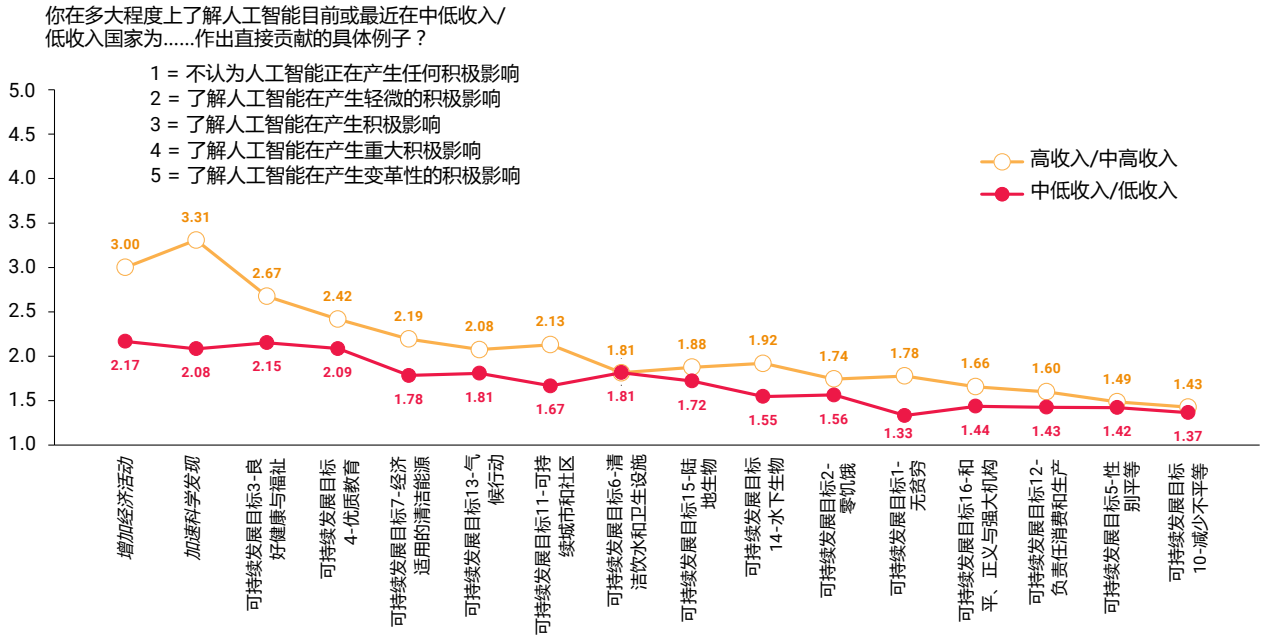


注：不包括“不知道”/“没有意见”和空白答复。在每个问题后分别了具体答复人数。未就可持续发展目标8、9和17提问。

信息来源：联合国秘书长技术事务特使办公室人工智能机遇扫描调查，2024年8月9-21日。

受访者报告称，人工智能对较低收入国家各方面产生的影响较少

经济增长和科学领域的差距最为明显



注：不包括“不知道”/“没有意见”和空白答复。

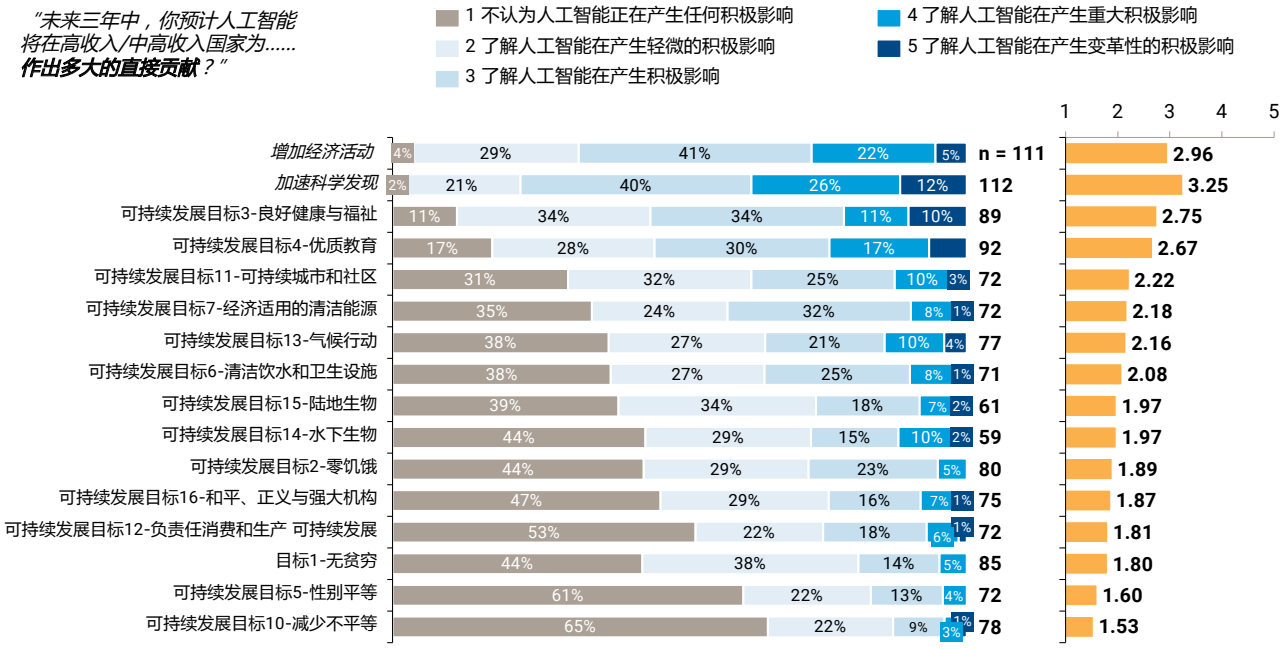
信息来源：联合国秘书长技术事务特使办公室人工智能机遇扫描调查，2024年8月9-21日。

对未来三年内人工智能的预期积极影响的看法

对增长、科学、健康、教育的预期影响——对其他领域的影响较少

未来三年对高收入/中高收入国家的预期影响

*"未来三年中，你预计人工智能将在高收入/中高收入国家为.....**作出多大的直接贡献？**"*

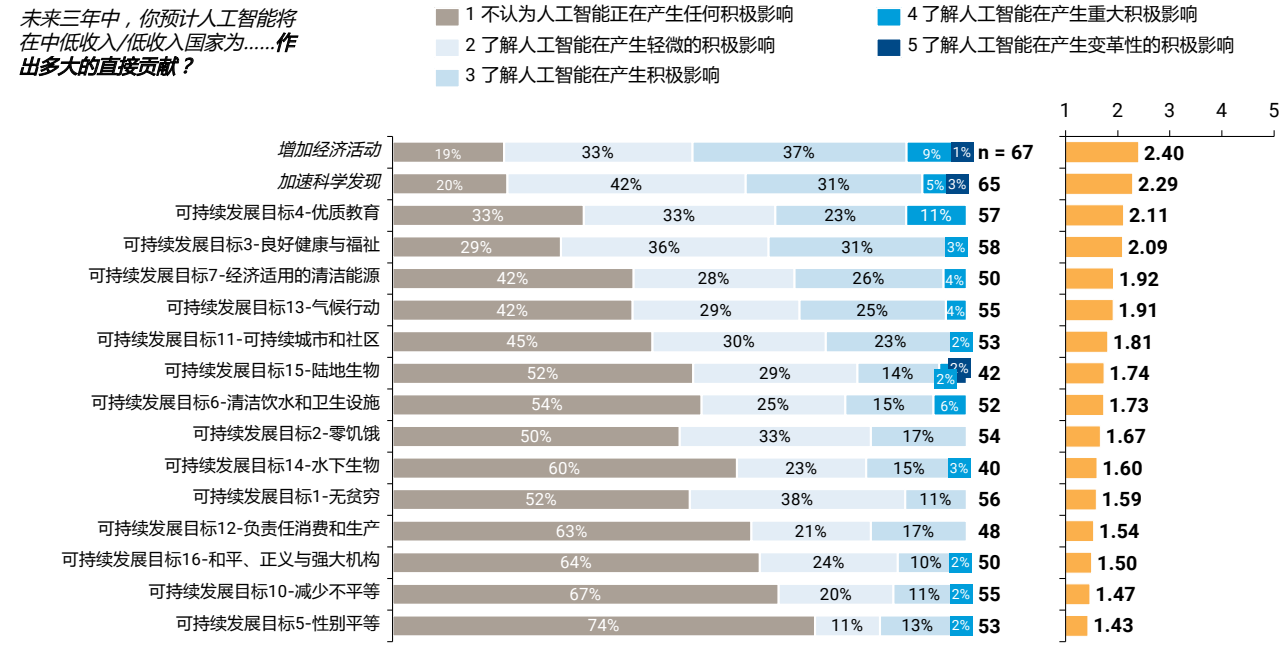


注：不包括"不知道"/"没有意见"和空白答复。在每个问题后分列了具体答复人数。未就可持续发展目标8、9和17提问。
信息来源：截至2024年8月21日的人工智能风险脉动调查结果（121人答复）。联合国秘书长技术事务特使办公室人工智能机遇扫描调查，2024年8月9-21日。

一些受访者预计会对较低收入国家产生影响，但影响更为有限

未来三年对中低收入/低收入国家的预期影响

*未来三年中，你预计人工智能将在中低收入/低收入国家为.....**作出多大的直接贡献？***

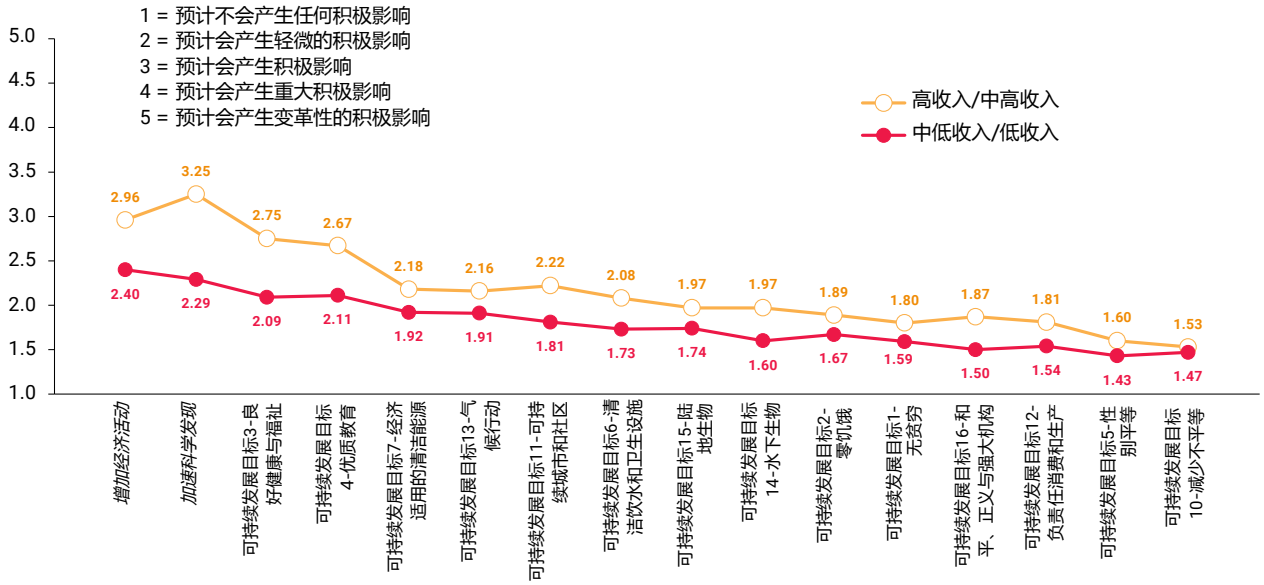


注：不包括"不知道"/"没有意见"和空白答复。在每个问题后分列了具体答复人数。未就可持续发展目标8、9和17提问。
信息来源：截至2024年8月21日的人工智能风险脉动调查结果（121人答复，其中78人回答了关于中低收入/低收入国家的问题）。联合国秘书长技术事务特使办公室人工智能机遇扫描调查，2024年8月9-21日。

预计对较低收入国家各方面的影响较少

经济增长、科学、健康和教育领域的差距最为明显

对“未来三年中，你预计人工智能将为……作出多大的直接贡献？”的平均评分。按国家收入分组，其中：



注：不包括“不知道”/“没有意见”和空白答复。在每个问题后分列了具体答复人数。未就可持续发展目标8、9和17提问。

信息来源：截至2024年8月21日的人工智能风险脉动调查结果（121人答复，其中78人回答了关于中低收入/低收入国家的问题）。联合国秘书长技术事务特使办公室人工智能机遇扫描调查，2024年8月9-21日。

使用think-cell绘制的图表

附件 G：缩略语表

东盟	东南亚国家联盟
粮农组织	联合国粮食及农业组织
G20	二十国集团
G7	七国集团
原子能机构	国际原子能机构
民航组织	国际民用航空组织
劳工组织	国际劳工组织
海事组织	国际海事组织
气专委	政府间气候变化专门委员会
国际电联	国际电信联盟
经合组织	经济合作与发展组织
人权高专办	联合国人权事务高级专员办事处
贸发会议	联合国贸易和发展会议
开发署	联合国开发计划署
教科文组织	联合国教育、科学及文化组织
难民署	联合国难民事务高级专员公署
反恐办	反恐怖主义办公室
世卫组织	世界卫生组织
知识产权组织	世界知识产权组织

捐赠者

咨询机构衷心感谢以下政府和合作伙伴的财政和实物捐助，没有他们机构将无法履行其职责：

捷克共和国政府
欧洲联盟
芬兰政府
德国政府
意大利政府
日本政府
荷兰王国政府
沙特阿拉伯王国政府
新加坡政府
瑞士政府
阿拉伯联合酋长国政府
大不列颠及北爱尔兰联合王国政府
Omidyar Network 基金
法语国家及地区国际组织

United Nations | AI Advisory Body