联合国 A/RES/78/265

大 会

Distr.: General
1 April 2024

第七十八届会议

议程项目 13

联合国经济、社会及有关领域主要大型会议和首脑会议成果的统筹协调执行及后续行动

2024 年 3 月 21 日大会决议

[未经发交主要委员会而通过(A/78/L.49)]

78/265. 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

大会，

重申国际法、特别是《联合国宪章》，并回顾《世界人权宣言》，[1]

又重申其 2015 年 9 月 25 日题为“变革我们的世界：2030 年可持续发展议程”的第 70/1 号决议、2015 年 7 月 27 日题为“第三次发展筹资问题国际会议亚的斯亚贝巴行动议程”的第 69/313 号决议和 2023 年 9 月 29 日大会第 78/1 号决议附件所载的、大会主持召开的可持续发展问题高级别政治论坛通过的政治宣言，

回顾其 2023 年 7 月 25 日题为“快速技术变革对实现可持续发展目标和具体目标的影响”的第 77/320 号决议、2023 年 12 月 19 日题为“信息和通信技术促进可持续发展”的第 78/132 号决议、2023 年 12 月 19 日题为“科学、技术和创新促进可持续发展”的第 78/160 号决议，2023 年 12 月 19 日题为“在数字技术背景下促进和保护人权”的第 78/213 号决议、2022 年 12 月 15 日题为“数字时代隐私权”的第 77/211 号决议、2015 年 12 月 16 日题为“关于信息社会世界首脑会议成果文件执行情况全面审查的大会高级别会议成果文件”的第 70/125 号

[1] 第 217 A (III)号决议。

决议，包括《日内瓦原则宣言》、[2]《日内瓦行动计划》、[3]《突尼斯承诺》[4]和《信息社会突尼斯议程》，[5]以及其 2020 年 9 月 21 日第 75/1 号决议所载的《纪念联合国成立七十五周年宣言》，

表示注意到，国际电信联盟与 40 个联合国机构合作，努力建立“人工智能造福人类”平台，其中包括举办年度峰会和启动国际电信联盟人工智能资料库，以确定负责任和切合实际的人工智能应用方式，推动实现可持续发展目标；联合国教育、科学及文化组织大会于 2021 年 11 月 23 日通过了其《人工智能伦理问题建议书》[6]及其实施计划，包括“准备情况评估方法和伦理影响评估”以及人工智能伦理问题全球论坛；人权理事会 2011 年 6 月 16 日第 17/4 号决议[7]核可了《工商企业与人权指导原则：实施联合国“保护、尊重和补救”框架》；[8]联合国人权事务高级专员办事处在人工智能方面开展工作，

又表示注意到联合国秘书长题为“数字合作路线图”的报告，[9]表示注意到为协调路线图执行工作设立秘书长技术问题特使办公室、秘书长设立一个由多利益攸关方组成的人工智能问题高级别咨询机构以及该机构于 2023 年 12 月 21 日发表中期报告并期待收到其最后报告，

认识到，安全、可靠和值得信赖的人工智能系统——就本决议而言，是指非军事领域的人工智能系统，其生命周期包括前期设计、设计、开发、评价、测试、部署、使用、销售、采购、运行和淘汰等阶段——必须以人为本、可靠、可解释、符合道德、具有包容性，充分尊重、促进和保护人权和国际法，保护隐私、面向可持续发展和负责任，这些系统可以加速和推动在实现所有 17 项可持续发展目标方面取得进展以及通过平衡和统筹兼顾的方式从经济、社会和环境这三个方面实现可持续发展；推动数字化转型；促进和平；克服国家之间和国家内部的数字鸿沟；促进和保护人人享有人权和基本自由，同时坚持以人为本，

又认识到，不当或恶意地设计、开发、部署和使用人工智能系统——例如在没有适当保障措施或不符合国际法的情况下——构成风险，而这些风险可能阻碍在实现《2030 年可持续发展议程》及其可持续发展目标方面取得进展，不利于从经济、社会和环境这三个方面实现可持续发展；扩大国家之间和国家内部的数字鸿沟；强化结构性不平等和偏见；导致歧视；有损于信息的完整性和

[2] 见 A/C.2/59/3，附件。

[3] 同上。

[4] 见 A/60/687。

[5] 同上。

[6] 联合国教育、科学及文化组织，《大会记录，第四十一届会议，2021 年 11 月 9 日至 24 日，巴黎》，第一卷，《决议》，附件七。

[7] 见《大会正式记录，第六十六届会议，补编第 53 号》(A/66/53)，第三章，A 节。

[8] A/HRC/17/31，附件。

[9] A/74/821。

获取信息的途径；削弱对人权和基本自由的保护、促进和享有，其中包括个人隐私不受非法或任意干涉的权利；增加发生事故的潜在风险和来自恶意行为者的多重威胁，

还认识到，人工智能系统的设计、开发、部署和使用迅速加快以及技术变革日新月异，这对加快实现可持续发展目标具有潜在影响，因此强调指出迫切需要：就安全、可靠和值得信赖的人工智能系统达成全球共识；推动开展包容性国际合作，制定和使用有效、具有国际互操作性的保障措施、做法和标准，以促进创新，防止对安全、可靠和值得信赖的人工智能系统的治理不成体系；认识到目前存在人工智能和其他数字方面的鸿沟，国家之间和国家内部的技术发展水平参差不齐，认识到发展中国家在跟上迅速加快的步伐方面面临独特挑战，这对可持续发展造成障碍，认识到需要消除发达国家与发展中国家之间在条件、可能性和能力方面的现有差距，因此，又强调指出迫切需要加强能力建设以及对发展中国家的技术和财政援助，以消除国家之间和国家内部的数字鸿沟，支持发展中国家有效、公平、切实参与人工智能系统治理方面的国际进程和论坛并拥有代表权，

又认识到，人工智能系统治理是一个不断发展变化的领域，需要随着技术和我们对技术的认识不断提高，继续讨论可能采取的适当治理办法，这些办法必须以国际法为基础，具有互操作性、灵活性、适应性、包容性，能够满足发达国家和发展中国家的不同需求和能力，并使所有人受益，

1. **决心**弥合国家之间和国家内部的人工智能鸿沟和其他数字鸿沟；

2. **决心**推广安全、可靠和值得信赖的人工智能系统，以在全面实现《2030 年可持续发展议程》[10] 方面加快取得进展，进一步弥合国家之间和国家内部的人工智能鸿沟和其他数字鸿沟；强调指出需要为安全、可靠和值得信赖的人工智能系统制定标准，以促进而不是阻碍数字化转型和公平获取这些系统所带来的惠益，从而实现所有 17 项可持续发展目标和从经济、社会和环境这三个方面实现可持续发展，并应对其他共同的全球挑战、特别是发展中国家面临的挑战；

3. **鼓励**会员国并邀请来自所有区域和国家的多利益攸关方，包括私营部门、国际和区域组织、民间社会、媒体、学术界和研究机构、技术界及个人，在各自角色和职责范围内，制定和支持与安全、可靠和值得信赖的人工智能系统有关的监管和治理办法和框架，在各级创造一个有利的生态系统，包括促进创新、创业和按照共同商定条件传播知识与技术，同时认识到各国政府和多利益攸关方之间的有效伙伴关系和合作对于制定此类办法和框架是必要的；

4. **促请**会员国并邀请其他利益攸关方采取行动，与发展中国家合作并向其提供援助，以实现包容和公平地获得数字化转型以及安全、可靠和值得信赖的人工智能系统所带来的惠益，包括为此：

[10] 第 70/1 号决议。

(a) 扩大所有国家、特别是发展中国家对数字化转型的参与，以利用安全、可靠和值得信赖的人工智能系统所带来的惠益并有效参与其开发、部署和使用，包括为此开展与人工智能系统有关的能力建设，同时认识到促进知识共享活动和按照共同商定条件转让技术是能力建设的一个重要方面，并强调指出需要消除人工智能鸿沟和其他数字鸿沟，提高数字素养；

(b) 通过强化伙伴关系，增强数字基础设施的连通性和技术创新的获取，以帮助发展中国家在人工智能系统的整个生命周期内有效参与，加快人工智能系统以包容方式对社会作出积极贡献，包括促进全面实现《2030 年议程》及其可持续发展目标，同时确保世界各地的人工智能系统在其整个生命周期内都是安全、可靠和值得信赖的；

(c) 增强发展中国家、特别是最不发达国家的能力，应对主要的结构性障碍，消除在获得新技术和新兴技术以及人工智能创新所带来惠益方面存在的障碍，以实现所有 17 项可持续发展目标，包括以加强伙伴关系等方式扩大使用科学资源、可负担的技术、研究和开发；

(d) 力求提供更多资金，促进在数字技术和安全、可靠和值得信赖的人工智能系统方面开展与可持续发展目标相关的研究和创新，并在所有区域和国家开展能力建设，以促进这方面研究并从中受益；

(e) 营造基于创新的国际环境，提高发展中国家发展技术专长和实力的能力，利用数据和计算资源，制定国家一级监管和治理方法、框架和建立采购能力，并在各级建立包容有利的环境，以提出基于安全、可靠和值得信赖的人工智能系统的解决方案；

(f) 紧急动用各种执行手段，例如，按照共同商定条件进行技术转让，通过能力建设消除人工智能鸿沟和其他数字鸿沟，以及根据发展中国家的国家需求、政策和优先事项，向发展中国家提供与人工智能系统有关的技术援助和融资；

(g) 促进安全、可靠和值得信赖的人工智能系统的获取以及设计、开发、部署和使用，以期从经济、社会和环境这三个方面实现可持续发展；

5. **强调**必须在人工智能系统的整个生命周期内尊重、保护和增进人权和基本自由，促请所有会员国并在适用情况下促请其他利益攸关方不要使用或停止使用无法按照国际人权法运作的、或对享有人权构成过度风险的人工智能系统，特别是对于那些处于弱势的人而言，并重申人们在线下享有的权利也必须在线上得到保护，包括在人工智能系统的整个生命周期内；

6. **鼓励**所有会员国酌情根据本国优先事项和具体情况以及在实施各自在国家一级的不同监管和治理办法和框架时，并在适用情况下鼓励其他利益攸关方，以包容和公平的方式促进安全、可靠和值得信赖的人工智能系统，造福所有人，并为这些系统营造有利环境，以应对世界上最大的挑战，包括从经济、

社会和环境这三个方面实现可持续发展，同时具体考虑到发展中国家以及不让任何一个人掉队，为此：

(a) 根据各自的国家一级政策和优先事项、相关的国家以下各级政策和优先事项以及国际法规定的义务，促进制定和实施国内监管和治理办法和框架，以支持利用负责任和包容各方的人工智能创新和投资促进可持续发展，同时推广安全、可靠和值得信赖的人工智能系统；

(b) 鼓励采取有效措施，促进创新，以便在人工智能系统的设计和开发期间以及在部署和使用之前，对脆弱性和风险采取具有国际互操作性的识别、分类、评估、测试、预防和缓解措施；

(c) 鼓励纳入反馈机制，以使最终用户和第三方可在开发、测试和部署人工智能系统后，以循证方式发现和报告技术漏洞，并根据情况发现和报告滥用人工智能系统的行为和人工智能事件，以解决这些问题；

(d) 提高公众对人工智能系统的核心功能、能力、局限性和适当民用领域的认识和了解；

(e) 促进建立、实施和披露风险监测和管控机制、确保数据安全机制(包括个人数据保护和隐私政策)以及在人工智能系统的整个生命周期内酌情开展的影响评估；

(f) 加大在制定和实施有效保障措施方面的投入，其中包括实体安全、人工智能系统安全以及人工智能系统整个生命周期内的风险管控；

(g) 鼓励开发和部署有效、可获取、适应性强、具有国际互操作性的技术工具、标准或做法，包括可靠的内容认证和来源识别机制——例如在技术上可行和适当的情况下使用水印或标识，以使用户能够识别信息操纵行为，区分或确定真实数字内容和人工智能生成内容或经过处理的数字内容——并提高对媒体和信息的认知度；

(h) 促进为人工智能系统的训练和测试制定和实施有效、具有国际互操作性的框架、做法和标准，以加强决策，帮助保护个人免遭一切形式的歧视、偏见、滥用或其他伤害，并在人工智能系统的整个生命周期内避免强化或固化具有歧视或偏见的应用程序和结果，例如，为此分析和减轻数据集中存在的偏见并以其他方式打击算法中的歧视和偏见，同时又不会无意或过度地影响其他用户和受益方的正向发展、获取和使用；

(i) 在适当和相关的情况下，鼓励实施适当的保障措施，以尊重知识产权(包括受版权保护的内容)，同时促进创新；

(j) 在测试和评估系统时维护隐私和保护个人数据，并遵守相关国际、国家和国家以下各级法律框架中规定的透明度和报告要求，包括关于在人工智能系统整个生命周期内使用个人数据方面的要求；

(k) 针对作出或协助作出影响最终用户的决定的人工智能系统，在其整个生命周期内促进透明度、可预测性、可靠性和可理解度，其中包括提供通知和解释以及推动人类监督，例如，通过审查自动化决定和相关流程，或在适当和相关情况下以人类决策替代，或为那些因人工智能系统的自动化决策而受到不利影响的人提供有效补救措施和追责手段；

(l) 在人工智能系统的整个生命周期内，加强对制定和实施有效保障措施、尤其是对风险和影响评估的投资，以保护人权和基本自由的行使，并减轻对充分切实享有人权和基本自由的潜在影响；

(m) 推广可以促进、保护和保全语言和文化多样性的人工智能系统，在这些系统的训练数据中以及在其整个生命周期内考虑到使用多种语言，特别在大语言模型方面；

(n) 在人工智能系统的整个生命周期内发挥作用的各个实体按照共同商定条件加强信息共享，以确定、理解人工智能系统方面基于科学和证据的最佳做法、政策和方法，并将其用于采取行动，以实现效益最大化，同时在人工智能系统(包括先进人工智能系统)的整个生命周期内减轻潜在风险；

(o) 鼓励开展研究和国际合作，以了解、平衡和应对与人工智能系统在弥合数字鸿沟和实现所有 17 项可持续发展目标方面所发挥作用有关的潜在惠益和风险，包括扩大开源人工智能系统等数字解决方案所发挥的作用；

(p) 促请会员国采取具体措施，消除性别数字鸿沟，确保特别关注利用、可负担性、数字素养、隐私和在线安全，加强数字技术(包括人工智能系统)的使用，并将残疾、性别和种族平等视角纳入政策决定及其指导框架的主流；

(q) 鼓励开展研究和国际合作，以制定措施确定和评估人工智能系统的部署对劳动力市场的影响；提供支持，减轻对劳动力，特别是对发展中国家、尤其是最不发达国家的劳动力产生的潜在负面影响；促进旨在开办各种方案，以开展数字培训，建设能力，支持创新，让更多人获得人工智能系统带来的惠益；

7. **又认识到**数据是人工智能系统开发和运行的根本；强调公平、包容、负责和有效的数据治理，数据生成、获取和基础设施的改善，以及数字公共产品的使用，对于利用安全、可靠和值得信赖的人工智能系统的潜力促进可持续发展至关重要，并敦促会员国分享数据治理方面的最佳做法，在数据治理方面促进国际合作、协作和援助，以在可行的情况下提高各种方法的一致性和互操作性，从而推动安全、可靠和值得信赖的人工智能系统实现可以信赖的跨境数据流动，使其开发更加包容、公平、有效和惠及所有人；

8. **确认**必须继续讨论人工智能治理领域的发展动态，以使国际做法跟上人工智能系统及其用途的演进步伐；并鼓励国际社会继续努力促进包容性研究、摸底和分析，以了解人工智能系统和快速技术变革在现有技术、新技术和新兴技术的开发方面以及对加速实现所有 17 项可持续发展目标的潜在影响和各种应用，使所有各方受益，并且为人工智能设计方、开发方、评估方、部署方、用

户和其他利益攸关方制定、促进和实施有效、具有国际互操作性的保障措施、做法、标准和工具提供参考借鉴，以使人工智能系统做到安全、可靠和值得信赖；强调指出各国政府、私营部门、民间社会、国际和区域组织、学术界和研究机构及技术界和所有其他利益攸关方需要酌情继续共同努力；确认所有社群、特别是发展中国家的社群需要以更有凝聚力、更有效、更协调和更包容的方式投入和参与安全、可靠和值得信赖的人工智能系统的包容治理；

9. **鼓励**私营部门遵守相关国际和国内法律，并按照《联合国工商企业与人权指导原则：实施联合国“保护、尊重和补救”框架》行事；确认必须以更加包容公平的方式获得安全、可靠和值得信赖的人工智能系统带来的惠益；认识到需要加强合作，包括公共部门、私营部门、民间社会、学术界和研究机构以及技术界之间和内部的合作，以在安全、可靠和值得信赖的人工智能的整个生命周期内提供和促进公平、开放、包容和没有歧视的营商环境、经济和商业活动、竞争性生态系统和市场；鼓励会员国制定政策和法规，以促进在安全、可靠和值得信赖的人工智能系统和相关技术方面开展竞争，包括为此支持和推动为小企业、企业家和技术人才创造新的机会，并通过关键投资、特别是对发展中国家的关键投资，促进在人工智能市场上开展公平竞争；

10. **促请**联合国系统各专门机构、基金、方案、其他实体、机构和办事处以及相关组织在各自任务和资源范围内，通过适当的机构间机制，继续评估和加强对策，以协作、协调和包容的方式利用人工智能系统带来的机遇并应对这方面的挑战，其中包括就潜在影响和应用开展研究、摸底和分析，使所有各方受益，报告在应对各种问题方面取得的进展和遇到的挑战；与发展中国家合作并协助它们进行能力建设，获取和分享安全、可靠和值得信赖的人工智能系统带来的惠益，以实现所有 17 项可持续发展目标和从经济、社会和环境这三个方面实现可持续发展；强调指出需要消除国家之间和国家内部的人工智能鸿沟和其他数字鸿沟；

11. **回顾**其题为“未来峰会的方式和范围”的 2022 年 9 月 8 日第 76/307 号决议和 2023 年 9 月 1 日第 77/568 号决定，并在这方面期待制定一项全球数字契约；

12. **期待**大会在 2025 年全面审查自信息社会世界峰会以来取得的进展；

13. **肯定**联合国系统根据其任务规定，在根据国际法、特别是根据《联合国宪章》、《世界人权宣言》和《2030 年可持续发展议程》就安全、可靠和值得信赖的人工智能系统达成全球共识方面作出独特贡献，其中包括推动包容各方的国际合作，促进发展中国家在审议中的融入、参与和代表性。

2024 年 3 月 21 日
第 63 次全体会议