



大会

Distr.: General
13 May 2026
Chinese
Original: English

人权理事会

第六十二届会议

2026 年 6 月 15 日至 7 月 10 日

议程项目 3 和 5

促进和保护所有人权——公民权利、政治权利、
经济、社会及文化权利，包括发展权

人权机构和机制

人工智能系统对善治的影响

人权理事会咨询委员会的研究报告*

概要

在根据人权理事会第 57/5 号决议提交的本报告中，人权理事会咨询委员会探讨了人工智能系统对善治的影响。

咨询委员会指出了人工智能可如何改善公共服务的提供，加强透明度和反腐败工作，并促进包容性参与。委员会审视了在公共治理中部署人工智能所引发的人权风险，特别是大规模监控、算法歧视、自动化决策的不透明、对信息环境的操控、劳动力替代以及环境危害。委员会还探讨了人工智能开发集中在少数行为体手中可能如何固化数字依赖模式，并强调了所有国家在制定适用规范和标准的过程中实现包容与平等参与的紧迫性。

咨询委员会提出了一系列建议，以确保公共治理中的人工智能部署始终符合国际人权义务及善治原则。

* 因提交方无法控制的情况，经协议，本文件迟于标准发布日期发布。



一. 导言

A. 任务

1. 人权理事会在第 57/5 号决议中，请人权理事会咨询委员会编写一份关于人工智能系统对善治的影响的研究报告，特别说明人工智能系统可在哪些领域通过善治促进和保护人权，以及此类系统在哪些领域对善治和人权构成挑战，重点介绍世界各地在开发、部署、使用和管理人工智能系统方面的良好做法，通过善治促进和保护人权，以及确定必要的保障措施。

2. 咨询委员会在第三十二届会议上成立了一个起草小组，该小组由以下成员组成：米兹纳赫·奥迈尔·阿勒奥迈尔、妮科尔·阿姆利纳、约瑟夫·热拉尔·安戈、爱德华·阿莫阿科·阿桑特、白凡锡(报告员)、拉巴赫·布达什、伊格纳西奥·德卡萨、里瓦·甘古利·达斯、亚历山德拉·德伍尔斯基、维罗尼卡·菲克法克、朱厄尔·梅杰、阿芙萨奈·纳迪普尔、任义生、帕特雷茨娅·萨斯纳尔、瓦西利斯·泽维列克斯(主席)、弗兰斯·维尔容。

B. 研究范围

3. 本研究审视了人工智能融入公共治理的过程及其对人权的影响。因此，研究重点并非人工智能治理本身，而是人工智能系统被公共当局部署时，可能如何促进或阻碍人权的实现。研究探讨了人工智能要想支持治理工作兑现提高效率和透明度的承诺，必须具备哪些条件。研究认为，人工智能的部署如果不以人权义务为基础，则可能为系统性的侵犯人权行为创造新途径，而这些侵权行为超越了传统救济机制的范围。

4. 分析回顾了人权理事会自第 7/11 号决议(2008 年)至第 57/5 号决议(2024 年)将善治与人权联系起来的框架的演变过程，表明人们日益认识到二者不仅是互补关系，更是相互构成的关系。分析探讨了人工智能如何对这一既有框架的适应能力提出挑战。尽管人工智能和算法系统能够增强国家履行国际人权法义务的能力，尤其是在改善医疗服务提供、扩大教育机会和加强社会保障机制方面，但此类系统同时也带来了前所未有的监控能力，并有可能固化歧视性模式。此外，透明的行政程序被不透明的计算过程所取代，可能妨碍有效的行政和司法审查，并削弱获得有效救济的权利。

5. 本研究纳入了联合国的讨论及利益攸关方的贡献，包括国家、区域组织、国家人权机构、民间社会和学术机构对咨询委员会调查问卷的 32 份答复。¹ 分析基于有效救济和问责原则，构建了一个全面的、以权利为基础的框架，始终强调必须将人权尽责纳入公共行政中人工智能的采购、设计和部署，确保算法透明度，使受影响的个人能够切实有机会理解和质疑自动化决策，并维持能够发现和应对歧视性偏见、隐私侵犯及程序违规行为的强大制度保障。本研究探讨了人工智能对弱势群体的不成比例的影响，并分析了如果不采取主动措施以确保包容性的设计和有意义的参与，将如何加剧现有的不平等。总体而言，本研究思考了如

¹ 可查阅 <https://www.ohchr.org/en/hr-bodies/hrc/advisory-committee/impact-artificial-intelligence-systems-good-governance>。

何实现人工智能在善治方面的变革性潜力——这种治理以人权、人的尊严和问责制为核心。

C. 善治与人权

6. 人权高专办将善治定义为公共机构以基本上没有权力滥用和腐败的方式处理公共事务、管理公共资源和保障实现人权的过程。² 这一理解源自人权委员会第 2000/64 号决议，委员会在该决议中正式承认了治理、发展与人的尊严之间的关联，确认透明、可问责且回应人民需求的治理对于促进人权至关重要。相应地，正如人权理事会所确认的那样，善治被理解为具有参与性、可问责性、透明性、回应性、有效性，并以法治和人权原则为指导。³

7. 自 2008 年以来，人权理事会逐步完善了将善治与人权相关联的规范框架，确认透明、可问责和参与式的政府是促进人权的基础，强调打击腐败，并呼吁建立一个恪守最高廉洁标准的专业、任人唯贤的公共服务体系。⁴ 最近，理事会还探讨了数字时代的挑战，确认互联网接入和数字技术对于获取信息、公共服务、教育和司法至关重要，同时欢迎在具备适当数据保护和隐私保障的前提下，利用技术创新来预防、发现和打击腐败。⁵

8. 人权理事会第 57/5 号决议标志着将人工智能正式列入善治与人权议程的里程碑。该决议承认人工智能既带来机遇也带来风险，一方面在打击腐败、改善公共服务的提供等方面带来益处，另一方面也带来了包括虚假信息在内的可能损害民主进程的挑战。理事会敦促各国、民间社会、媒体及其他利益攸关方制定战略，在促进人们充分享有人权和基本自由的同时减轻风险。尤为重要的是，理事会呼吁各国以包容和公平的方式，为全人类的利益促进安全、可靠、可信赖的人工智能系统，这体现了这样一项原则，即人工智能的发展应以善治和人权原则(包括透明、公平和问责)为指导。与此同时，秘书长于 2023 年组建了人工智能高级别咨询机构。⁶《全球数字契约》反映了确保人工智能加强而非削弱人权、平等与法治的紧迫性。

二. 人工智能支持的善治：对人权的机遇与风险

A. 人工智能对人权和善治的积极贡献

1. 改善公共服务的提供，实现经济、社会和文化权利

9. 人工智能在治理领域一个被广泛认可的潜力在于其能够改变公共服务的提供方式，从而推动包括经济、社会和文化权利在内的所有权利的实现。人工智能系统可以处理大量数据并自动执行复杂任务，这可能使行政服务更加高效、便捷，

² 见 <https://www.ohchr.org/en/good-governance/about-good-governance>。

³ 例如，见人权理事会第 7/11、19/20、25/8、31/14、37/6、45/9、51/5 和 57/5 号决议。

⁴ 见人权理事会第 7/11、第 19/20 和第 25/8 号决议。

⁵ 见人权理事会第 45/9 和第 51/5 号决议。

⁶ 《为人类管理人工智能：最后报告》(纽约，联合国，2024 年)。

更能回应个人需求。一些国家和国际观察员表示谨慎乐观，认为人工智能在设计、部署和监管得当的情况下，能够在医疗、教育和社会保障等领域大幅提升政府能力。⁷

10. 人工智能驱动的医疗诊断和公共卫生管理创新可以促进健康权的实现，⁸ 这表明人工智能有能力提高卫生服务的可得性和质量——这是实现可达到的最高标准健康权的一项核心义务。⁹ 人工智能可以提高诊断准确度，优化医院运营，简化患者分诊流程，从而使医疗保健系统更加有效和公平。¹⁰

11. 在教育领域，人工智能可以弥合教育差距，加强机构管理。¹¹ 人工智能赋能的教育平台能够根据个人的学习风格和需求实时调整教学，从而更好地包容那些在标准化系统中学习困难的人群——无论其年龄、语言或残疾状况如何，以此解决不平等问题。¹² 将教材转换为替代格式的应用程序可作为残疾人的辅助工具。人工智能还可以通过识别有辍学风险的学生并实施有针对性的干预，为教育管理提供支持，同时让教师能够更多地关注每个学生的个体需求。在以人为本、优先考虑公平获取的前提下，这类工具有望缩小教育成果的差距，并推进善治所倡导的包容、以人为本的服务提供原则。然而，盲目采用人工智能可能会加剧数字鸿沟，削弱人的能动性，最终损害享有包容和公平教育的权利。¹³

12. 人工智能可以改善社会福利项目的管理，从而促进社会保障权和适当生活水准权。¹⁴ 机器学习算法能够快速分析家庭数据，以识别最需要帮助的人群，或发现福利发放中的错误。口译和笔译的自动化可以帮助个人获得司法服务和公共服务。这种自动化如果是透明的，有助于确保有限的资源迅速流向适当的对象，为实现社会和经济权利创造更有效的路径。

13. 申请处理的自动化可以提高机构效率和响应能力，这两点对于善治至关重要。将人工智能整合到报税、福利计算等行政流程中，可以提升效率，使工作人员能够专注于更复杂和更需要判断的工作。¹⁵

14. 在个人服务之外，人工智能的预测分析能力可以推动从被动治理向主动治理的转型。例如，将人工智能纳入灾害管理，有助于在洪水或疾病暴发等危机发生之前进行预测，从而使主管部门能够采取预防措施。¹⁶

⁷ 巴西、新加坡、乌兹别克斯坦和促进参与性民主制度人民团结组织提交的材料。

⁸ 智利、古巴和厄瓜多尔提交的材料。

⁹ 墨西哥城人权委员会提交的材料。

¹⁰ 智利和厄瓜多尔提交的材料。

¹¹ 见 [A/79/520](#)。

¹² 委内瑞拉玻利瓦尔共和国和缅甸国家人权委员会提交的材料。

¹³ 联合国教育、科学及文化组织(教科文组织)，《2023 年全球教育监测报告：技术运用于教育——谁来做主？》(巴黎，2023 年)。

¹⁴ 西班牙提交的材料。

¹⁵ 世界银行，*Artificial Intelligence in the Public Sector: Maximizing Opportunities, Managing Risks* (Washington, D.C., 2020)。

¹⁶ 古巴和日内瓦人权平台提交的材料。

15. 人工智能技术可应用于支持执法、国家安全和边境管控的某些职能。人工智能赋能系统可以协助刑事调查(例如通过识别车牌),并有助于发现非法活动。然而,必须强调的是,在这些领域部署人工智能会引发一些最大的人权关切。这类技术需要健全的法律框架,以确保透明、问责以及对国际人权标准的充分遵守。¹⁷

16. 在实践中,人工智能的这些应用通过加强循证政策制定,为善治做出贡献。人工智能技术增强了公共当局快速做出知情决策并根据预期需求分配资源的能力,从而体现了回应性原则——这是以人为本的治理的核心。尽管预测模型的设计、部署和治理必须适当顾及隐私权、不歧视及其他人权考量,但这些模型在推动公共行政转型以促进经济、社会和文化权利的实现方面,具有巨大潜力。

2. 促进透明、问责和反腐败

17. 人工智能可以作为加强公共行政透明度的重要工具,从而增进公众信任并强化政府机构的问责制。通过缩小日常决策中的个人自由裁量空间,并生成可审计的数字记录,设计良好的人工智能系统能够减少腐败机会,并使不当行为更容易被发现。¹⁸ 在个人数据和隐私得到充分尊重的前提下,数字政府创新有助于预防腐败和促进问责。¹⁹ 更广泛而言,包括人工智能在内的数字技术和新兴技术,在其开发和运用符合相关国际人权法的前提下,可以促进国际合作,以预防和打击腐败。²⁰

18. 人工智能还可以通过加强记录保存和数据分析来加强问责。公共部门的人工智能系统通常会记录每项操作以及决策所依据的数据,从而生成全面的审计追踪。这可以确保决策的可追溯性,遇到申诉或投诉时,为审查提供依据,从而起到遏制不当行为的作用。目前,一些司法管辖区正在推动设计能够解释算法推理的透明人工智能系统,以确保算法治理对法治起到补充而不是破坏的作用。²¹

19. 人工智能可以进一步扩大公民参与和独立监督,从而加强问责。开放数据政策与人工智能驱动的分析相结合,使民间社会、媒体及其他利益攸关方能够对大量公共信息进行审查,以发现不当行为或侵犯权利的迹象。联合国开发计划署强调了以这种方式使用人工智能的价值,包括利用其监测采购活动和非法资金流动,以支持更广泛的治理工作。²²

20. 人工智能并非应对治理挑战的全面方案,但它为推进透明和问责原则提供了新的途径。通过这些应用,公共当局可以采取重要措施,以加强机构廉洁、打击腐败和维护法治。

¹⁷ 见 [A/HRC/60/45](#)。

¹⁸ 经合组织工商业咨询委员会,“Harnessing AI for integrity: opportunities, challenges, and the business case against corruption”(Organisation for Economic Co-operation and Development (OECD), Paris, 2024)。

¹⁹ 见人权理事会第 51/5 号决议。

²⁰ 联合国反腐败公约缔约国会议,第 11/1 号决议。

²¹ 见 <https://e-estonia.com/estonia-and-automated-decision-making-challenges-for-public-administration>。

²² 《新技术促进可持续发展:关于诚信、信任与反腐败的视角》(纽约,2021 年)。

3. 促进包容、参与和平等的工具

21. 另一个积极方面是人工智能在促进治理中的包容与平等方面的潜力。善治要求机构具有包容性，为社会各阶层——包括边缘化群体——提供发声和参与的渠道。人工智能可以帮助克服这些群体面临的障碍。例如，人工智能驱动的辅助技术可以使残疾人能够与他人平等地获取公共服务和信息，从而履行《残疾人权利公约》关于确保信息和服务可及性的义务。这类工具还可以增强那些因语言或识字障碍而难以与政府打交道的社群的信息获取权。

22. 人工智能可以增强人们对民主进程的参与。这类应用可以拓宽参与式民主的范围，为个人——包括那些在政策制定中历来代表性不足的群体——创造更多参与治理的渠道。人权理事会鼓励此类创新，指出数字技术可以促进对公共生活的包容性参与，并呼吁各国弥合数字鸿沟，使边缘化人群能够与他人平等参与。²³ 此外，人工智能赋能的翻译和语言处理工具可以促进政治参与，包括在多语言国家为属于语言少数群体的人参与选举进程提供便利。然而，通过这类工具实现有意义的参与，取决于人们能否公平地获取这些工具。

23. 人工智能还可以帮助识别和解决服务提供中的不平等问题。通过分析行政数据及其他相关数据集，公共当局可以发现服务在哪些方面系统性地忽视了特定群体或地区，例如某些社群长期难以获得医疗服务或应急响应时间较长，从而推动平等和不歧视原则的落实。²⁴ 在以透明方式实施并吸纳社群意见时，这类应用即符合发展权以及“不让任何人掉队”的共同承诺。²⁵ 受影响社群应积极参与人工智能的创建和治理，而不是仅作为政策的被动接受者。然而，这些益处绝非自动实现，而是需要人工智能系统本身在设计上和训练过程中能够防范现有偏见的再现。

B. 人工智能对人权和善治的负面影响

24. 尽管人工智能融入公共治理具有潜力，但在保护人权方面也带来了显著挑战。如果没有健全的法律框架、人权保障机制和人工监督，人工智能系统可能会助长大规模监控、固化系统性偏见、削弱正当程序并扭曲公共话语。若对这些风险置之不理，可能侵蚀善治与法治的根基。将人工智能视为一种中立技术是一种过于简化的观点，因为其设计与应用可能反映并放大现有的权力失衡。

1. 通过大规模监控和数据滥用侵蚀隐私

25. 人工智能在治理领域引发的最直接的人权关切，是其对隐私的影响。²⁶ 人工智能系统通常依赖于收集和处理大量个人数据，这些数据涵盖面部图像、位置轨迹、社交媒体内容及生物特征标识等。当公共当局部署人工智能系统时，会显著扩张监控能力。机器学习技术尤其擅长从看似无关的海量数据集中识别模式，形成所谓的“数据马赛克效应”——通过聚合不同来源的信息，可以构建出个体

²³ 人权理事会第 47/23 号决议。

²⁴ 见 <https://stories.undp.org/bending-the-ai-arc-towards-equity>。

²⁵ 人工智能高级别咨询机构，2023 年，《为人类管理人工智能》。

²⁶ 见 A/HRC/60/45。

的全面画像。²⁷ 这类系统的一个决定性特征在于，它们能够推断出个人未曾明确披露的敏感信息。多个司法管辖区的司法机构已认定，某些大规模画像系统因在缺乏充分透明度或正当理由的情况下，对弱势社群实施隐蔽且不成比例的监控，从而侵犯了基本权利。这凸显出必须建立严格的保障措施，以确保自动化画像不会侵犯隐私或延续歧视。

26. 这些发展引发了对隐私权以及隐私作为一项入门权利(即能够保障表达、结社以及宗教或信仰自由得以行使的先决权利)的功能的严重关切。人工智能增强型监控有可能侵蚀不应受任意侵扰的私生活领域，因为模式识别和大规模数据挖掘可能造成无处不在的监控，超出了传统的必要性和相称性标准所允许的范围。²⁸ 例如，联网视频监控与人脸识别技术相结合，能够在无需持续人工干预的情况下，实时追踪公共空间中的个体，这种监控方式尤其具有侵扰性，因此基于隐私和数据保护理由，必须满足最高程度的正当性要求。

27. 人工智能带来的隐私风险并不局限于摄像头。当局越来越多地运用机器学习来分析税务、社会保障以及公开的线上信息等组合数据集，以识别被视为具有风险的个体。行政调查已得出结论：大规模、自动化的个人数据匹配可能产生带有重大影响的错误决定，这凸显了自动化可能如何侵犯隐私和正当程序。²⁹

28. 更广泛地说，人工智能赋能的监控代表了从基于个人嫌疑的定向拦截向大规模算法筛选的质的转变，有可能使全体人口持续处于监视之下。人权高专办已发出警告，无差别的监控会压制表达和结社自由的行使。³⁰ 无处不在的监控所带来的寒蝉效应——即个人因担心可能受到处罚或面临不利后果而不敢行使合法权利——有据可查，尤其是在表达自由方面。³¹ 当个人知道自己可能受到监控时，他们就不太愿意自由表达或举行集会。预测性分析可能会加剧这些担忧，因为它会依据与种族或社会经济状况相关的替代指标，将监控集中在特定街区或社群，从而在侵蚀隐私的同时强化歧视性做法。

29. 边缘化和弱势群体面临尤为突出的风险。难民和移民在边境越来越多地面临生物特征数据收集和算法风险评估。³² 这类技术可能导致歧视，并损害隐私和正当程序。当监控工具被广泛部署时，人权维护者、记者和政治活动人士同样受到影响。在这种环境下，隐私成为一种集体关切。当监控无处不在时，所有人提出异议的民主空间都会收窄。人权理事会呼吁对任何数字监控措施进行严格的合法性、必要性和相称性检验，³³ 然而许多人工智能赋能的做法仍在过时的法律框架下运行，这些框架并未预见到当前的技术能力。

²⁷ 见 [A/HRC/48/31](#)；以及 David E. Pozen, “The mosaic theory, national security, and the Freedom of Information Act”, *Yale Law Journal*, vol. 115, No. 3 (December 2005)。

²⁸ 见 [A/HRC/41/35](#)。

²⁹ 数字权利研究所提交的材料。

³⁰ 见 [A/HRC/27/37](#)。

³¹ Anita R. Gohdes, *Repression in the Digital Age: Surveillance, Censorship, and the Dynamics of State Violence* (Oxford, Oxford University Press, 2024)。

³² 见 [A/75/590](#)。

³³ 人权理事会第 48/4 号决议。

30. 总体而言，人工智能加剧了隐私保护的紧迫性。³⁴ 正是那些使人工智能系统对治理具有价值的特性，也推动着对个人数据进行愈发广泛的收集和处理，这构成了对隐私权的直接侵犯，并抑制了其他基本权利的行使。在缺乏适当保障措施的情况下，监控的自动化有可能演变为审查和社会的控制自动化，这从根本上与具有民主特征的善治不相容。

2. 算法偏见与歧视：固化不平等

31. 下一个主要关切是算法偏见，即人工智能系统继承或放大了社会偏见及其对平等权 and 不受歧视权的影响。善治要求公共服务和决策必须公正地提供，无论种族、性别、宗教、残疾或其他被禁止的歧视理由，公平对待所有人。然而，人工智能系统并非天生中立，往往反映出潜藏于训练数据或设计中的歧视性模式。因此，在治理中部署此类系统可能产生系统性歧视，而这种歧视相比传统人为决策所产生的歧视，更难被发现，也更难通过现有救济途径提出异议。³⁵

32. 已在多个领域发现公共部门人工智能存在偏见的证据。在刑事司法领域，用于辅助审前释放、量刑和假释决策的风险评估工具引发了关于差异性结果的担忧。在就业相关的筛选系统也发现了类似问题，这些系统可能再现基于性别或残疾的偏见；在金融或行政算法中，代理指标可能无意中对就业空窗期进行惩罚，或将特定类别的纳税人作为目标。在治理背景下，基于反映历史上社会经济差距的数据所训练出的模型，可能会不成比例地将少数族裔拥有的企业或低收入社群标记为审查对象。³⁶ 在为本次研究提交的材料中，利益攸关方一致指出了人工智能加剧歧视和可及性差距的风险，众多案例表明，先前存在的各种不平等——无论是与种族、阶级、性别、残疾还是移民身份相关——均被放大而非缩小。这违背了国家防止歧视性做法和追求实质性平等的义务。³⁷

33. 一个核心难点在于不透明。许多人工智能系统无法以易于理解的方式呈现其推理过程，而且“意图”这一概念对于机器而言并不适用。因此，受影响者甚至可能很难意识到偏见已经发生，更不用说追究责任或寻求救济了。传统的反歧视框架往往以证明存在故意歧视为导向，不适合处理复杂的、数据驱动的伤害。因此，许多法律和人权分析都主张采用基于效果的方法，重点关注差异性影响。³⁸ 当某一受保护群体显示出统计上显著的不利结果时，应触发审查以及修改或撤回相关系统的义务，无论其意图如何。这一方针符合对间接歧视的既定理解。总体而言，算法偏见可能对不歧视原则构成严重威胁，并有可能在技术客观性的表象下强化结构性不平等。

3. 黑箱问题：侵蚀透明度、正当程序和获得救济的权利

34. 人工智能在治理中带来的另一个同样严峻的挑战是算法决策的不透明，通常被称为“黑箱问题”。许多机器学习和神经网络系统无法提供受影响者可理解的

³⁴ 见 [A/HRC/48/31](#)。另见 [A/HRC/61/68](#)，第 25 和第 26 段。

³⁵ 见 [A/HRC/48/31](#)；[A/HRC/54/70](#)，第 24 段。

³⁶ 日内瓦人权平台提交的材料。

³⁷ 爱尔兰、墨西哥、新加坡、西班牙、捍卫伊朗阿塞拜疆人民人权和民主/文化主张协会、韩国进步网 - Jinbonet 和透明国际提交的材料。

³⁸ 见 [A/HRC/56/68](#)。

推理过程。在治理情境下，这种不透明可能与透明、问责和法治原则相冲突，并危及程序性保障(包括理解并对影响权利与义务的决定提出异议的权利、获得公正审讯的权利以及获得有效救济的权利)。数字时代权利的行使要求自动化系统的应用具有透明度和可问责性，因为缺乏可解释性会阻碍诉诸司法。³⁹ 透明对于问责制至关重要，可解释性对于正当程序必不可少，而明确的责任归属是获得有效救济的前提。⁴⁰

35. 国际及国内法院和法庭正在努力应对这些问题。区域性数据保护框架承认个人有权不受完全由自动化系统做出的对其有重大影响的决策约束，并且在此类决策发生时，有权获得关于所涉逻辑的有意义信息。⁴¹ 然而，对于复杂模型而言，什么是有意义的解释仍存在争议。一些人主张提供简化的理由或采取透明度措施，另一些人则提醒，对于深度学习系统而言，事后说明可能过于技术化或仅仅是推断性的。⁴² 尽管存在这些争论，但发展轨迹正指向将可解释性作为一项基线保障。

36. 不透明还造成了问责空白。当人工智能系统造成伤害时，责任可能分散于开发者、供应商以及采购和部署此类系统的公共当局之间。这种不确定性损害了《公民及政治权利国际公约》第二条第三款所规定的获得有效救济的权利，因为受影响的个人难以确定应被追究责任的责任方。这引发了问责问题，包括违反人权义务的责任归属，以及在相关行为体(如部署人工智能系统的国家当局以及设计并提供这些系统的企业)之间的责任分配问题。当公共职能实际上被外包给受商业秘密或知识产权主张保护的专有系统时，问题会更加严重，因为无论是公职人员还是公众，都无法充分了解决策的依据。⁴³ 在公共部门背景下，这种不透明可能会削弱既定的行政问责原则，以及公共当局就影响个人权利的决定说明理由的义务。联合国人权事务高级专员强调，企业需要就其向政府提供的监控及其他高风险技术的设计和运作保持透明。⁴⁴ 人权与跨国公司和其他工商企业问题工作组以及人权高专办强调，需要可解释的人工智能，算法影响评估，并将人权尽责嵌入整个人工智能生命周期。⁴⁵

37. 鉴于这些利害关系，除了事后救济之外，还需要采取系统性的事前预防措施。必要内容包括：强制性的人权影响评估、定期审计、自动化决策系统公共登记册以及独立的监督机构。这种预防性方针反映了一个共识，即在高速运行、数据驱动的环境中，传统的救济机制可能为时已晚。⁴⁶ 与此同时，经合组织的人

³⁹ 见 A/HRC/48/31。

⁴⁰ 人权理事会第 54/18 号决议；以及促进参与性民主制度人民团结组织提交的材料。

⁴¹ 关于通用数据保护的(EU) 2016/679 号条例，第 22 条。

⁴² 经合组织，人工智能原则。可查阅 <https://www.oecd.org/en/topics/sub-issues/ai-principles.html>。

⁴³ 韩国进步网 - Jinbonet 提交的材料。

⁴⁴ 见 A/HRC/48/31。

⁴⁵ 见 A/HRC/59/32 和 A/HRC/59/53。

⁴⁶ 教科文组织，《人工智能伦理问题建议书》(巴黎，2022 年)；Matthew U. Scherer, “Regulating Artificial intelligence systems: risks, challenges, competencies, and strategies”, *Harvard Journal of Law & Technology*, vol. 29, No. 2 (spring 2016)。

人工智能原则将透明、可解释和人工监督作为基本要求，并呼吁建立能够让受到不利影响的人员对结果提出异议的机制(原则 1.3)。

4. 对表达自由、信息获取和民主参与的威胁

38. 人工智能对表达自由和公共领域诚信的影响亟需关注，因为善治的前提是公众知情以及思想交流和公平透明政治进程所需的多元化环境。一个显著问题是人工智能生成的虚假信息大量扩散，这些以极低成本、大规模生产的虚假或误导性文本、图像和视频足以以假乱真，威胁到人们对可靠信息的获取。⁴⁷ 例如，在多个国家的选举前夕，观察员记录到旨在抹黑候选人或抑制选民投票率的协同网络和深度伪造视频。⁴⁸ 当虚假信息通过加密通讯软件传播时，情况变得更加复杂，因为难以提供可靠的替代信息。⁴⁹ 对这些技术的恶意应用导致人们对整个信息环境的信任度下降。一旦人们意识到虚假内容可以被伪造得真假难辨，即使是真实的事实也可能被当作潜在的伪造品，从而影响循证决策和公众对政府信息的信任。此外，政治行为体和其他相关方可以通过分析个人数据，提供定制化的说服力内容，以影响投票行为或压制参与。此类做法，特别是基于侵入性画像的做法，可能侵犯隐私权，并损害意见的自由形成，而意见自由是思想与表达自由的核心要素，也是有意义参与公共事务的前提。⁵⁰ 此外，人工智能驱动的平台利用推荐算法，通过持续放大带有偏见的叙事，最大化用户参与度，久而久之，这会在用户毫无察觉的情况下塑造其观念，引发对精神自主权的担忧。这些动态通过瓦解共同的事实基础、加剧两极分化并阻碍理性的政策辩论，从而对治理造成侵蚀。

39. 有些危害不仅涉及表达自由，还关乎人身平等与安全。人权理事会已表示关切，担心信息操纵(包括通过数字技术进行的信息操纵)可能助长暴力和冲突，并呼吁各国鼓励私人行为体在维护人权的同时应对虚假信息。⁵¹ 人工智能还可以通过生成大量绕过基本过滤器的变体，放大仇恨或歧视性内容，从而加剧了目标群体在尊严、安全和平等地位方面的风险。

40. 另一个维度涉及审查与内容审核。国家可能部署人工智能以大规模监控和限制线上内容。移除非法律材料可能是合法的，但自动过滤器往往存在过度删除的倾向，且通常无法考虑语境、讽刺或出于公共利益发表的言论等因素，从而有可能对合法表达造成不合理的干预，并产生寒蝉效应。对表达自由的任何限制，包括为打击虚假信息而采取的措施，必须严格遵守《公民及政治权利国际公约》第十九条第三款的要求，满足合法性、必要性和相称性原则。应对虚假信息的措施不得成为不当限制合法表达的借口。⁵²

⁴⁷ 美利坚大学提交的材料。

⁴⁸ 自由之家，“Freedom on the Net 2019: the crisis of social media” (Washington, D.C., 2019)。

⁴⁹ A/HRC/61/68，第 74 段。

⁵⁰ 见 A/73/348 和 A/HRC/47/25。另见人权事务委员会，第 34 号一般性意见(2011 年)，第 9 和第 10 段。

⁵¹ 人权理事会第 55/10 号决议。另见 A/HRC/61/68，第 70 段。

⁵² A/HRC/61/68，第 21 段。

41. 这些现象累积起来，会侵蚀人们对信息环境的信任。当个人不再能相信他们所读或所见内容的真实性，也不再能确信线上互动的真实性时，社会凝聚力就会下降，从而削弱善治的一个重要基础。此外，这种不信任往往会扩展到所有形式的信息，给依赖数字通信的新闻媒体、研究机构和政府机构带来重大挑战。应对这些挑战需要采取多利益攸关方合作的方法。⁵³ 国家、私人平台和民间社会必须共同努力，制定适当的保障措施。这些措施包括：确保政治广告的透明度、对人工智能生成内容进行标注、纳入适当的人类身份验证、投资于媒体和数字素养教育以增强社会韧性，⁵⁴ 以及采取措施以促进推荐系统的公平性和多元化。为追求这些目标而对表达施加的任何限制，都必须符合关于合法性、必要性和相称性的国际标准。

42. 人工智能对表达自由和民主参与可能产生的负面影响是广泛且系统性的。它们冲击了公众进行知情讨论和追究当局责任的能力，从而削弱了民主治理的架构。然而，这些结果并非不可避免，通过持续的多利益攸关方努力，在遵守人权和法治原则的前提下，维护人的能动性、多元主义和信任，便可以避免这些结果。在诸多手段中，纳入适当的“人类身份验证”保障机制对于维护对数字公共空间的信任至关重要，这能确保民主进程中的参与反映的是真正的人类意见。⁵⁵ 此类措施有助于保护公民参与的真实性，同时减轻与虚假信息、协同造假行为以及对决策过程的不当影响有关的风险。

5. 自动化与工作权：岗位流失与不平等加剧

43. 人工智能驱动的自动化可能对《世界人权宣言》第二十三条和《经济社会文化权利国际公约》第六条所载的工作权构成重大挑战。虽然从历史上来看，技术进步在淘汰某些工作类型的同时也会创造新的工作类型，但当代人工智能部署的速度和范围是前所未有的，其影响已从常规的手工劳动延伸到文职、创意和决策类型的工作。人工智能驱动的岗位流失所带来的影响分布不均。面临高自动化风险的行业往往雇用了不成比例的边缘群体成员。女性在文书和行政支持岗位上的比例通常过高，因此面临因自动化而失业的风险要大得多。⁵⁶ 此外，这种后果不仅局限于发达经济体。传统上依赖劳动密集型产业作为经济发展路径的发展中国家面临特殊风险，因为自动化可能会削弱此类发展战略所依赖的劳动力成本优势。联合国的分析警告称，发展中国家有相当大比例的就业可能易受自动化影响，从而引发对贫困、经济稳定和全球不平等扩大的担忧。⁵⁷

44. 除经济后果外，人工智能导致的大规模岗位流失还牵涉到人的尊严和社会凝聚力。对大多数人而言，工作不仅是收入来源，也是身份认同、人生意义和社会参与的基础。让有工作意愿的劳动者突然变得冗余，可能产生深远的心理和社会代价，包括疏离感和自我价值感的丧失。这些动态要求采取前瞻性的治理应对措施

⁵³ 见 A/HRC/47/52。

⁵⁴ 人权理事会第 57/5 号决议。

⁵⁵ Bryan Ford, “Identity and personhood in digital democracy: evaluating inclusion, equality, security, and privacy in pseudonym parties and other proofs of personhood” (2020).

⁵⁶ Paweł Gmyrek 等人, “Generative AI and jobs: a global analysis of potential effects on job quantity and quality”, 劳工组织第 96 号工作文件(日内瓦, 2023 年)。

⁵⁷ 《2020 年世界社会报告：剧变世界中的不平等》(联合国出版物, 2020 年)。

施。为确保不让任何一个人掉队，必须采取协调一致的行动来维护工作权，包括获得体面工作的机会，以此作为人工智能时代善治的一个基本要素。

6. 能源消耗与环境权利：对可持续性的威胁

45. 公共当局部署人工智能系统会带来显著且常常被忽视的物质足迹，对享有清洁、健康和可持续环境的权利产生影响。⁵⁸ 与大众认知相反，人工智能并非一种完全无形的技术，其基础设施对资源的消耗极大。训练大规模模型和执行实时数据分析所需的计算需求呈指数级增长，这显著增加了能源消耗。预测表明，到2030年，数据中心的能源需求将急剧上升，随之而来的风险是为了满足基础电力负荷而增加对化石燃料的依赖。⁵⁹ 这种趋势有可能破坏全球气候承诺，并加剧局部空气污染，对公众健康和生态系统构成风险。

46. 除碳排放外，人工智能基础设施的热管理还对淡水资源造成了巨大压力。⁶⁰ 数据中心的冷却需要消耗大量水资源，常常与当地社区的用水需求形成竞争，尤其是在水资源紧张的地区。在此背景下，善治面临的一个重大挑战是环境数据的普遍不透明。大型技术提供商往往没有披露关于取水量和耗水量的精确指标，这阻碍了公共当局和民间社会确保问责的能力，并侵犯了公众获取信息的权利，而这些信息对于在资源管理中实施有效的参与式治理至关重要。

47. 此外，人工智能的生命周期固化了环境不公与全球不平等的模式。人工智能运算所需的硬件推动了对关键矿产的开采，这些矿产通常来自发展中国家，开采条件可能违反劳工权利和环境标准。这类硬件的快速淘汰加速了电子垃圾的产生，电子垃圾的负担则不成比例地由发展中国家承担。这种动态反映了一种数字鸿沟，即人工智能带来的经济利益集中在发达国家，而环境外部性和资源退化则被转嫁给发展中国家，从而加深了现有的全球差距。因此，各国若要利用人工智能推进善治，就必须将这些环境和分配后果作为其人权义务中不可分割的一部分。

7. 司法系统和羁押环境中的人工智能：维护公平与人的尊严

48. 在司法工作中，对部署人工智能的担忧尤为突出。人工智能正日益融入司法系统的行政工作中，包括案卷管理、翻译、判决书的自动匿名化处理、电子立案，并用于法律搜索引擎，以及协助法官和律师撰写案件摘要或分析判例。⁶¹ 法院和司法当局也开始探索使用人工智能赋能的风险评估工具，协助就保释、假释和量刑作出决定。⁶² 虽然此类应用在案件积压严重的情况下可能提高效率，但它们提出了一个根本性问题：核心司法职能可以在多大程度上托付给算法系统。当负责筛选训练数据、设计算法或监督法院人工智能系统的不是司法机构本

⁵⁸ 智利、摩洛哥、沙特阿拉伯、捍卫伊朗阿塞拜疆人民人权和民主/文化主张协会、数字权利研究所、人道法高级研究中心和日内瓦人权平台提交的材料。

⁵⁹ 国际能源署，Energy and AI: World Energy Outlook Special Report (2025)。

⁶⁰ Joyeeta Gupta, Hilmer Bosch and Luc van Vliet, “AI’s excessive water consumption threatens to drown out its environmental contributions”, The Conversation, 21 March 2024.

⁶¹ 见 A/79/296。

⁶² 欧洲委员会，European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment (Strasbourg, 2019)。

身，而是行政或立法部门时，尤其是在此类职能被外包给私营公司时，司法独立可能受到损害。⁶³ 公正审判的精髓在于人的审慎判断和可问责性。虽然人工智能可以辅助司法程序，但法官在运用自由裁量权和识别每个案件的独特性方面发挥着不可替代的作用。因此，司法界领导人强调，在司法事务中，人(而非机器)必须始终是最终的决策者。⁶⁴ 在所有涉及人工智能参与的案件中，严格的人为监督、透明度以及当事方能够理解和质疑人工智能发挥的作用，至关重要。当被告人无法对影响其自身的决定提出质疑时，人工智能系统的不透明性可能对《公民及政治权利国际公约》第九条和第十四条所保障的人身自由与安全权以及公正审判权产生不利影响。认识到这些风险后，一些司法管辖区已开始将供司法当局使用的人工智能系统归类为高风险系统，并禁止使用完全依赖画像或人格特质来预测犯罪的人工智能风险评估。⁶⁵ 因此，将人工智能融入司法程序必须采取应有的谨慎态度，以维护司法独立、程序公正和公众对司法工作的信任。

49. 在监狱环境中，引入人工智能驱动的系统引发了额外的人权关切。一些新建的设施(通常被称为“智慧监狱”)采用面部识别、生物特征监测和预测性分析来加强安全性和提高运营效率。此类工具或许有助于减少暴力和违禁品偷运，但它们可能因最大程度地减少人际接触和专业裁量权的行使，而削弱羁押环境中的人性化程度。对被剥夺自由者进行持续的算法监控可能加剧其孤立感，而在羁押管理中过度依赖自动化，可能因去除了关怀和个体评估所不可或缺的人的要素，而削弱改造工作的效果。《联合国囚犯待遇最低限度标准规则》(《纳尔逊·曼德拉规则》)强调有意义的人际接触的重要性，规定没有这种接触的长期单独监禁构成一种残忍、不人道或有辱人格的待遇(规则 43 和 44)。因此，在羁押场所一味地推行自动化，可能与国际人权法规定的以人道和尊重人固有尊严的方式对待所有被剥夺自由者的义务相冲突。

三. 建设尊重权利的人工智能以促进善治

A. 人工智能发展中的数据偏见与全球不平衡：对人权的挑战

50. 一个根本性的挑战在于数据和算法中存在偏见，而人工智能的开发集中于少数参与者和特定环境，进一步加剧了这一问题。人工智能系统往往反映出其设计理念以及训练数据中固有的价值观、假设和局限性。目前，人工智能行业和研究领域仍由少数技术先进的国家和几家实力雄厚的企业主导，导致这些系统常常只体现较为单一的视角。这引发了一个担忧：人工智能技术可能无法契合或充分代表多元的文化、社会和区域背景。

51. 这一现象被描述为一种数字殖民主义——少数几家主要技术供应商及其规范性框架占主导地位，使得历史上的排斥模式长期延续。⁶⁶ 人工智能行业尚处起

⁶³ 见 A/79/296；教科文组织，《司法机构人工智能与法治全球工具包》(巴黎，2024)。

⁶⁴ 教科文组织，《人工智能伦理问题建议书》。另见教科文组织，《法院和法庭使用人工智能系统的指南》(巴黎，2025 年)。

⁶⁵ 欧洲议会和欧盟理事会 2024 年 13 月 6 日关于制定人工智能统一规则的(EU)2024/1689 号条例。

⁶⁶ Jake Okechukwu Effoduh and Miracle Okumu Mudeyi, “TWAILing AI governance: will third world countries be left behind?” TWAIL Review (7 August 2025).

步阶段的国家往往采用由境外开发的、主要面向技术发达经济体设计的人工智能系统或监管框架。这类解决方案有时在本地政策制定者或受影响社群几乎未参与的情况下便被部署。在实施过程中，影响力的不平衡有可能加剧现有不平等。当技术未能针对特定文化和社会背景进行调整时，“一刀切”的人工智能方法可能会在不经意间凌驾于本地知识体系之上，特别是土著人民的知识体系。在此类情况下，从外部引入的单一化人工智能框架可能会使本地价值观和专业技能被边缘化。⁶⁷ 数据和算法偏见对人权的影响直接且重大。存在偏见或歧视性的算法可能产生不平等、不公正的结果，实质上违反了平等和不歧视原则。如果不加以解决，这类系统可能会固化所谓的“自动化的不公正”，即基于有缺陷或有偏见的数据，不透明的算法决策剥夺了个人获得公平对待、服务或机会的权利。此类伤害很可能对属于边缘化和弱势群体的个人造成不成比例的影响，从而进一步扩大现有差距。

52. 要应对这些挑战，必须将人权原则、公平和多元化融入人工智能开发的各个阶段。⁶⁸ 仅靠技术层面的调整是不够的，在人工智能系统开发团队的人员多元化以及此类系统的核心目标方面，都需要进行更广泛的改革。在人工智能的整个生命周期——从数据收集、算法设计到部署——确保多元参与和包容性设计，对于减少偏见和促进问责至关重要。通过纳入更广泛的视角，人工智能系统以及规范其运行的治理框架才能更好地反映多样化的社会需求与价值观。

53. 基层正在涌现出令人鼓舞的人工智能民主化举措。非洲、亚洲等地区的许多数据科学家正在开展合作，创建适合低资源语言和本地需求的数据集和工具。这些努力旨在让人工智能更具包容性和文化适当性，以应对那些不考虑多元背景的单一化方法。国际上的标准制定工作也开始反映这些关切。⁶⁹ 2021 年，教科文组织呼吁各国确保人工智能的开发尊重人权、体现多元化并维护公平。⁷⁰ 落实这些原则需要质疑默认的设计假设，并让本地专家参与开发过程。人工智能系统应与多元的法律和文化价值观相一致，而不是不经意地强加外部规范。此外，国际人权框架本身也必须以具有文化敏感性的方式适用于人工智能。一些学者提醒说，不顾及地方背景而仅仅依赖国际人权法，可能体现出一种隐含的个人主义取向，这种取向无法充分反映某些文化背景中盛行的以集体或社群为基础的权利观念。因此，在解释和实施关于数据保护或算法透明度的全球规范时，应当既要维护基本人权，又要允许地方层面的问责机制存在合理的文化和法律差异。

B. 确保有意义的参与和包容

54. 在利用人工智能促进善治方面，一个关键挑战是在关键决策过程中缺乏多元利益攸关方的切实参与。人工智能的研发集中在少数几个国家，高层次的人工智能战略和伦理准则通常由这几个国家的专家起草，其他地区的意见投入不足。受人工智能驱动的公共服务或监控系统影响最大的群体，包括普通民众、边缘化社群成员，以及小国或技术较落后的国家，常常被排除在政策制定过程之外。这种

⁶⁷ 促进参与性民主制度人民团结组织提交的材料。

⁶⁸ 见 A/HRC/59/32。

⁶⁹ 见 A/HRC/53/42。

⁷⁰ 《人工智能伦理问题建议书》。

排斥反映了信息技术治理中更广泛的民主赤字。如果全球认可的人工智能原则未能纳入世界大多数人口的优先关切，这些原则将很难付诸实践。

55. 世界各国的人工智能就绪度和治理能力存在显著差异。一些国家政府，尤其是发达国家，已经建立了国家及区域层面的人工智能框架，并积极参与国际论坛；而许多其他国家则刚刚开始制定人工智能政策或指南。放大代表性不足群体的声音，区域合作是一个途径。在某些地区，国家已开始通过区域组织开展合作，就人工智能伦理及其在治理中的应用形成共同立场。这些立场既体现了共同的本地价值观，也融入了普遍的人权原则。这表明，发展中国家可以通过集体努力，影响正在形成的有关人工智能的全球规范。然而，此类倡议需要得到更广泛的国际社会的支持。

56. 在有意义的参与在国内层面同样至关重要。当政府将人工智能应用于公共行政——无论是自动化福利资格认定、实施智慧城市监控，还是使用人工智能辅助的司法工具——受影响的个人和社群，除了享有知情权和透明度之外，还应当对这类系统的设计和使用拥有有意义的发言权。个人应当能够参与影响其自身权利的决策，这是一项普遍的法律原则。因此，这一原则同样适用于在治理中采用人工智能的情况。然而，在许多国家，公民对技术政策的参与仍然有限或流于表面。

57. 为填补这一治理缺口，各国每当计划采用具有重大社会影响的人工智能系统，就应建立正式的磋商和监督机制。在实施任何可能影响人权或公共利益的人工智能驱动系统之前，应建立公众磋商程序。应确保社群和民间社会在监督机构中的代表性，并为政府发起的人工智能项目设立伦理审查委员会。通过吸纳受影响社群的代表、人权组织和独立专家，这类机构和委员会能够更有效地评估部署人工智能所引发的社会影响，并就必要的保障措施提出建议。这些群体能够提供关于人权影响和文化背景的宝贵知识，而这恰恰是技术官僚式路径常常忽略的。若要人工智能服务于善治的目标，就必须在人工智能拟服务对象的切实参与下对其进行治理，从而增强公众信任，确保人工智能的部署符合公共利益。

C. 界定各利益攸关方在基于权利的人工智能实施中的责任

58. 要构建一个维护人权、防止侵害的人工智能生态系统，需要所有利益攸关方在公平、主权⁷¹ 和社会正义原则的指导下采取协调一致的行动。国家在确保人工智能的开发和部署符合国际人权规范方面承担首要责任。这一责任首先体现为建立健全的法律和监管框架，包括对公共部门的算法决策提出透明度、问责和监督要求。⁷² 政府还应投资于公共教育和数字素养，使个人了解人工智能如何运用于治理中，并知晓自身在此背景下享有的权利。发展中国家需要足够的政策空间和支持，以便根据本国具体的社会优先事项来治理人工智能。鉴于许多发展中国家缺乏独立为算法系统建立有效治理框架的资源，国际合作应侧重于通过技术转让、监管培训及南南知识交流来建设实质性能力，从而使所有国家都能利用人

⁷¹ 在此语境下，主权是指国家(特别是发展中国家)不受不当的外部依赖，对人工智能相关政策决定、数据治理和数字基础设施进行切实、自主控制的能力。关于人工智能开发的集中化所带来的挑战以及数字依赖的风险，参见上文第 50 至 54 段。

⁷² 见 A/HRC/48/31。

工智能推动善治。因此，全球标准应保持基于原则并具有灵活性，允许各国根据自身的制度和文化背景加以实施。⁷³ 各国政府必须将人工智能治理牢固地建立在普世人权价值观的基础之上，同时积极引导技术发展服务于本国和本地的需求。

59. 当国家将教育、医疗和供水等公共服务的提供外包给私人行为体时，国家的人权义务仍然完全适用。在利用人工智能系统提供公共服务时，应根据既定原则适用人权义务，包括：在适用情况下将相关行为归因于国家，且在任何情况下，国家均有责任针对私人行为体履行应有的审慎义务。国家必须确保此类部署不会损害服务的可得性、可及性、可接受性和质量，也不会引发歧视或其他负面人权影响。⁷⁴ 这属于国家积极人权义务的范畴，包括：建立并执行同样适用于人工智能驱动的私人服务的监管标准；开展持续的监测与评估；确保在发生侵犯人权行为时进行有效调查；以及为权利受到影响的个人提供有效救济。无论服务是通过何种技术手段提供的，国家在这方面的义务都完全适用。⁷⁵

60. 政府采购或部署的人工智能技术有很多是由私营部门开发的。根据《工商企业与人权指导原则》，所有工商企业，包括设计和供应人工智能系统的企业，都负有尊重人权的责任。在实践中，企业应在其人工智能开发全过程中践行人权尽责。⁷⁶ 这包括：与利益攸关方进行磋商，评估算法偏见、歧视及其他形式滥用等潜在风险，并调整设计选择以减少伤害。企业还必须解决其人工智能工具中存在的系统性偏见以及少数地区视角占主导地位的问题。这可以通过推进员工队伍多样化以及丰富模型训练所使用的数据集来实现，以求更好地反映全球多样性。在此基础上，如果能够真正吸纳本地研究人员并对本地背景予以应有的尊重，将有助于确保人工智能产品能够更好地契合更广泛的社会和文化环境。

61. 民间社会——包括非政府组织、学术机构、媒体、社区团体和青年——在确保人工智能在治理中的部署立足于人权方面发挥着不可或缺的作用。这些行为体应被赋予权能，并获得参与决策过程的正式渠道。民间社会和学术界的专业知识有助于制定必要的保障措施，例如要求对公共部门使用的人工智能系统进行独立的算法审计或建立社区反馈机制。民间社会组织在提高公众对人工智能的认识和素养方面也发挥着重要作用。通过深入浅出地阐明人工智能对人权的影响，这些组织让社群能够理解相关影响并要求追究责任。

62. 包括联合国在内的国际组织有责任建立包容性的框架，以指导如何使用人工智能，使其符合人权并服务于善治。全球层面的努力必须明确纳入发展中国家的视角。来自发展中国家和边缘化社群的声音警告称，国际倡议应避免以全球治理为名复制现有的权力等级结构。因此，国际规范的制定，无论是通过不具约束力的准则还是未来可能具有法律约束力的文书，都应通过包容、透明的程序进行，使所有国家拥有平等的发言权。为加强人工智能与人权的一致性，可能需要探索新的全球治理机制。

⁷³ 教科文组织，《人工智能伦理问题建议书》。

⁷⁴ 见 [A/HRC/59/53](#)。

⁷⁵ 经济、社会及文化权利委员会第 24 号一般性意见(2017 年)，第 11 段；非洲人权和民族权委员会，第 7 号一般性意见，第 43 段。

⁷⁶ 见 [A/HRC/59/32](#)。

四. 建议

A. 国家

63. 各国应通过全面的立法和监管框架，对人工智能在公共行政中的使用加以规范。这些框架应规定在部署之前及之后定期进行人权影响评估，要求算法具有透明性和可解释性，建立独立的监督机构以审计人工智能系统是否存在偏见、是否准确以及是否符合国际人权法，并确保在部署人工智能导致侵犯人权时进行有效调查和提供救济。

64. 各国应通过建立透明的磋商程序，积极吸纳民间社会、技术专家和受影响社区参与，以保障公众切实参与有关在公共部门引入人工智能系统的决策，并确保治理技术反映多元的社会需求和价值观。

65. 各国应严格规范人工智能赋能的监控技术的使用，防止对隐私权、表达自由权和结社自由权的侵犯，确保任何部署均符合合法性、必要性和相称性原则，并接受有效、独立的司法监督。

66. 各国应通过提升媒介与数字素养、要求政治广告透明化、要求对人工智能生成内容进行标识、探索或允许使用适当的人类身份验证机制以及支持多元化等方式，维护民主进程和信息生态系统的完整性，同时确保打击虚假信息的措施与表达自由保持适当平衡，不致造成不当的审查。

67. 各国应确保将人工智能融入司法程序和惩戒机构时，保留人的监督、程序公平和固有的人类尊严，禁止完全依赖自动化系统做出影响人身自由的决定，同时在羁押场所保持有意义的人际接触。

68. 各国应积极应对人工智能驱动的自动化带来的社会经济影响，为此应调整教育体系，提供再培训计划，并加强社会保障机制，以维护工作权利并确保公正转型。

69. 各国应评估并减轻人工智能部署对环境造成的影响，包括数据中心大量的能源和水资源消耗以及电子废弃物的产生，确保追求人工智能赋能的治理不会损害享有清洁、健康和可持续环境的权利，也不会将环境负担转嫁给发展中国家。

B. 国际社会

70. 国际社会应确保人工智能的全球治理具有包容性和公平性，为此应保障发展中国家和全球南方的代表能够切实参与国际标准制定进程并拥有平等的发言权，同时防止强加那些不符合当地实际情况或延续数字依赖模式的统一监管框架。

71. 国际社会应提供能力建设、技术转让和资金支持，使所有国家，特别是低收入国家，能够建立并实施有效且符合本国国情的人工智能治理框架，还应将实施有力的人权保障措施作为对数字项目提供国际援助的条件。

72. 国际社会应探讨设立专门的联合国机制，例如关于新的和新兴数字技术及其人权影响问题的特别报告员或工作组，以监测相关发展并就保护措施提出建议。还应考虑在联合国的主持下设立一个专门机制，使各国就人工智能对人权的影响承担具体义务。

C. 工商企业

73. 设计、开发或部署人工智能系统的工商企业应尊重人权，包括在人工智能的整个生命周期内践行人权尽责，评估并减轻算法偏见、歧视和侵犯隐私的风险，与受影响社群进行磋商，并建立或参与有效的申诉机制，使个人能够报告损害并获得补救。

74. 工商企业应促进包容和公平的人工智能，为此应推动员工队伍和训练数据集的多样化，以反映全球多样性，确保人工智能工具能够适应不同的社会、文化和区域背景，并避免采取那些会加深历史偏见或加剧数字不平等的做法。

D. 所有利益攸关方

75. 所有利益攸关方，包括国家、国际组织、工商企业、民间社会、学术界以及受影响社群，应通过透明、包容的多利益攸关方进程开展合作，以制定维护人权的治理措施。此类合作应辅以对人工智能社会影响的独立监测、便捷的救济渠道和通用工具(例如算法影响评估框架和审计标准)，以确保人工智能治理始终以人权为中心，能有效回应所有人的需求并尊重人的尊严。

76. 所有利益攸关方应通过提供资金和能力建设举措，积极支持民间社会组织，特别是发展中国家的民间社会组织，以提升公众认知、监测人工智能带来的影响，并确保那些受新兴技术影响最深的群体能够参与塑造这些技术的部署。
