联 合 国 A/HRC/56/68

大 会

Distr.: General
3 June 2024
Chinese
Original: English

人权理事会
第五十六届会议
2024 年 6 月 18 日至 7 月 14 日
议程项目 9
种族主义、种族歧视、仇外心理和相关不容忍行为：
《德班宣言和行动纲领》的后续行动和执行情况

当代形式种族主义、种族歧视、仇外心理和相关不容忍行为

当代形式种族主义、种族歧视、仇外心理和相关不容忍行为特别报告员阿什维尼的报告*

概要

当代形式种族主义、种族歧视、仇外心理和相关不容忍行为特别报告员阿什维尼在本报告中概述过去一年开展的活动，探讨了技术客观而又中立的这种主流假设如何令人工智能得以延续种族歧视。她阐述了人工智能可能助长种族歧视表现的四种交叉方式：数据问题、算法设计问题、故意以歧视方式使用人工智能和问责问题，然后举例说明人工智能在社会各个领域的应用及其在种族歧视方面的影响。她分析了为管理和规范人工智能所做的新的努力，然后概述了相关国际人权法标准。最后，特别报告员就各国应如何管理和监管人工智能技术以防止和应对种族歧视提出了建议。

* 因提交方无法控制的情况，经协议，本报告迟于标准发布日期发布。

一. 导言

1. 本报告根据人权理事会第 52/36 号决议提交，理事会在该决议中请当代形式种族主义、种族歧视、仇外心理和相关不容忍行为特别报告员向其提交年度报告。特别报告员在这份报告中介绍根据任务授权开展的活动，并探讨人工智能与种族歧视这一议题。

2. 为编写本报告，特别报告员向联合国会员国和其他利益攸关方，包括民间社会组织、国际组织和国家人权机构征集资料。特别报告员向所有提交资料的会员国和其他利益攸关方表示衷心感谢。她在编写报告时参考了收到的资料，并愿意就这一重要议题与所有相关利益攸关方保持对话。[1]

二. 活动概述

3. 2023 年 10 月，特别报告员向大会第七十八届会议提交了关于打击美化纳粹主义、新纳粹主义和其他助长当代形式种族主义、种族歧视、仇外和相关不容忍言行的做法的报告以及关于网上种族主义仇恨言论的报告。[2] 2023 年 10 月 31 日至 11 月 14 日，特别报告员对美利坚合众国进行了国别访问。[3]

4. 特别报告员于 2023 年 8 月参加了执行德班宣言和行动纲领问题独立知名专家组第九届会议。2024 年 1 月，特别报告员出席了非洲人后裔国际十年亚太区域会议。2024 年 2 月，她出席了在卡塔尔举行的关于从人权角度实现粮食正义的国际会议，主题为“现实挑战和未来的利害关系”。2024 年 4 月，她出席了非洲人后裔问题常设论坛第三届会议，并在会上作了关于克服教育中的系统性种族主义和历史伤害的发言。

三. 人工智能与种族歧视

5. 特别报告员在本报告中专门论述人工智能与种族歧视的问题。她在提交人权理事会第五十三届会议的报告中阐述了战略愿景和初步优先事项[4]，这一主题符合其概述的数字技术与种族歧视之间的关系这项战略重点。她借鉴了前任任务负责人在新兴数字技术和种族歧视方面的工作[5]，并回应了人权理事会和更广泛的联合国系统对于人工智能治理的关注。[6]

---

[1] 特别报告员的研究和分析工作得到哈佛大学法学院国际人权诊所以及斯坦福大学法学院国际人权和冲突解决诊所和斯坦福种族正义中心的支持。她衷心感谢所有相关人员在本报告编写过程中给予的宝贵支持。

[2] A/78/302 和 A/78/538。

[3] 见 A/HRC/56/68/Add.1。

[4] A/HRC/53/60, 第 50 至 53 段。

[5] 见 A/75/590、A/HRC/44/57 和 A/HRC/48/76。

[6] 例如，见人工智能高级别咨询机构，“中期报告：为人类管理人工智能”(2023 年 12 月)；人权理事会第 53/29 号决议；大会第 78/213 号和第 78/265 号决议。

6. 生成性人工智能的最新发展和人工智能的蓬勃应用继续引发严重的人权问题，包括对种族歧视的关切。生成式人工智能正在改变世界，今后还有可能推动日益剧烈的社会变革。人工智能应用在各个领域迅速普及，特别报告员对此深感关切。这并不是因为人工智能不具备潜在的益处。事实上，人工智能为创新和包容提供了潜在机会。然而，这些技术的发展和演变势头迅猛，基本上未受到制约。特别报告员感到关切的是，旨在管理和规范人工智能的政策和法律措施跟不上技术的发展速度。人工智能的能力十分强大，目前和未来都极有可能延续和加深系统性的种族歧视，扩大区域、国家和社群内部和之间的不平等，而目前为管理和规范人工智能做出的新的努力并未对这种可能性给予充分关注。

7. 如前任任务负责人所述，长期存在一种认为技术中立而又客观的有害观念：

> 公众对技术的惯常看法是，技术在本质上是中立和客观的，有些人指出，这种关于技术客观性和中立性的假设甚至在技术生产者中也仍然很突出。但技术从来都不是中立的——它反映了那些影响其设计和使用的人的价值观和利益，并且在根本上同样受到社会中不平等结构的影响。[7]

8. 特别报告员在本报告中探讨了认为技术客观而又中立的这种主流假设如何令人工智能得以延续种族歧视。

A. 人工智能助长种族歧视表现的交叉方式

9. 人工智能不是单一的模式。事实上，人工智能有多种类型。预测型人工智能被认为是这种技术的“传统”形式，其模型利用历史数据、模式和趋势对未来事件或结果进行有根据的预测。

10. 用于识别印刷字符、人脸、物体和其他信息的人工智能是另一种形式的“传统”人工智能，包含多种识别和区分所获数据中的物体、个人和模式的技术。

11. 生成式人工智能系统是更新形式的人工智能。生成式人工智能用途广泛，可用于多种目的。它包含一整类人工智能系统，其设计目的是在海量训练数据集、神经网络、深度学习架构和用户提示的基础上产生各种输出。生成式人工智能模型能够产生多种类型的输出，包括图像、文本、音频、视频和合成数据。与专门识别现有数据模式的人工智能模型不同，生成式人工智能经过训练，可以模仿机器学习模型训练数据中的模式和特征，创建新的数据点。生成式人工智能的出现将带来许多新的应用，也会造成许多新的人权问题。[8]

12. 上述各类人工智能的应用非常广泛。特别报告员将在下文阐述人工智能应用的更具体实例，以及在种族歧视方面的相关影响。然而，她谨此强调，研究人工智能可能延续种族歧视的各种方式之间存在的共性，十分重要；在关于管理和规范人工智能的法律和政策辩论中尤为需要探讨这一问题。在这种辩论中，必须从系统性种族主义的角度来看待人工智能的影响，系统性种族主义的定义是“国家机构、私营部门和社会结构中一个复杂、相互关联的法律、政策、做法和态度体系的运作，这几个因素结合在一起，可产生基于种族、肤色、血统或民族或族裔

[7] A/HRC/44/57, 第 12 段。

[8] 澳大利亚人权委员会提交的材料。所有提交材料均将在联合国人权事务高级专员办事处(人权高专办)网站上发布。

的直接或间接、有意或无意、法律上或事实上的歧视、区别、排斥、限制或偏好。”[9] 正如该定义所反映的，系统性种族主义是一种复杂而又往往十分隐蔽地存在于整个社会的现象。系统性种族主义在某一领域的表现形式与其他领域的表现形式相互关联、相互依存、相互强化。研究人工智能助长种族歧视的交叉方式，有助于确定人工智能以何种方式与系统性种族主义的各种表现形式之间相互作用，强化这些表现，并全面加深社会中的系统性种族和族裔压迫。[10]

1. 数据问题

13. 人工智能系统和机器学习算法的兴起导致了数据的大规模数字化。算法使用这些数据做出决策，并参与多个部门的行动。然而，训练算法所使用的数据集通常不完整，或者不能充分代表特定人群。如果特定人群在训练集中的代表性，包括种族和族裔方面的代表性，过高或过低，则可能导致算法偏差。同样，如果训练集包含已经存有偏差的数据，可能产生有偏差的结果。

14. 如果训练数据不足，算法做出的预测也许会对数据中缺乏代表性或代表性不足的群体带有系统性歧视。不仅数据太少会造成算法偏差，算法所依据的数据不具备代表性，也会产成偏差结果。例如，关于美国执法图像数据库的专项研究表明，执法人员使用的人脸识别网络中，非洲人后裔更容易被错误识别。这是由于对这一群体的人脸识别存在误差，而且非洲人后裔在警方照片数据库中的比例过高，这体现了系统性种族主义的历史模式。[11]

15. 历史偏见会影响数据本身。机器学习的一个核心要素是根据过去的数据预测未来。然而，如果过去的数据对某些群体带有偏见，包括种族和民族方面的偏见，计算机模型可能复制和放大这些偏见。使用存有偏见或缺陷的数据为现实生活中的决策提供信息，会进一步针对和伤害边缘化种族和族裔群体，因为在人工智能环境中使用这些数据，会产生更多数据，成为未来决策所参考的信息。这种自我强化的系统会复制和加深现有差距。

16. 数据的最后一个问题是隐私。人工智能系统使用的数据通常包括数据所有者的个人信息。未经同意收集和处理数据侵犯隐私权。还有些情况是未征得同意，便把在一种场合收集的数据，如在卫生保健领域通过保健应用程序等方式收集的数据，共享给其他场合使用，如用于执法。数据泄露和采取黑客手段未经授权获取个人信息的做法带来更多的隐私问题。对于种族边缘化群体成员而言，与隐私权有关的人权关切可能更为严重。这些群体可能会因隐私受到侵犯而受到排斥和歧视，或者人身安全受到威胁。[12]

[9] A/HRC/47/53, 第 9 段。

[10] A/HRC/44/57, 第 43 段。

[11] Nicol Turner Lee, Paul Resnick and Genie Barton, “Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms”, Brookings Institution, 22 May 2019.

[12] Samantha Lai and Brooke Tanner, “Examining the intersection of data privacy and civil rights”, Brookings Institution, 18 July 2022. 另见隐私国际提交的材料。

2. 算法设计问题

17. 人工智能工具的另一种常见偏差源自算法的设计方式。如果设计选择中存在偏差，那么即使算法中输入的数据具有无懈可击的代表性，算法也可能导致偏差结果。关于算法参数和运行方式的决策会带来偏差。算法设计者决定算法使用哪些变量，如何定义分类信息的类别或阈值，以及使用哪些数据来构建算法。设计者所做的选择包括如何衡量特定特征和如何定义算法的成功。有时，算法设计者的背景或视角可能会导致他们在算法设计中植入无意识的偏见，包括种族偏见。[13] 据报告，由于人工智能系统的开发过程中缺乏包容性的协商程序，因此数字技术部门多样性不足的问题更加严重，这也是造成算法设计问题的一个原因。[14]

18. 算法设计中的选择可能对现实生活产生重大的歧视性影响。例如，在构建贷款风险评估算法时，定义和衡量“风险”的方式可能导致歧视性结果。如果算法设计者决定使用信用评分作为评估风险的替代指标，可能对容易获得较低信用评分的人群产生歧视性结果。研究表明，信用评分、种族和其他人口统计指标之间可能存在很强的相关性，使用信用评分对某些群体不利。[15] 许多情况下，这种关联可以被视为现有的系统性种族主义和排斥的产物。算法设计者选择使用信用评分来评估贷款风险，虽然表面上这不是一个歧视性标准，但可能会令某些个人处于不利地位。

3. 用于歧视目的

19. 某些情况下，人工智能可能用于明显的种族主义目的，选择性地对目标群体使用这种技术，导致歧视性结果。例如，有报告称，执法机构从种族歧视的角度蓄意针对特定社区使用人工智能开展监控和过度执法。[16] 此外，当政府和其他人利用这一技术的能力，基于特定群体或个人的种族或族裔身份对其进行监控、定性并将其作为针对的目标时，就可能产生蓄意歧视。[17]

20. 传播虚假信息是出于明显的种族主义目的使用人工智能的另一种方式。政治行为者可以利用人工智能生成文本、图像和视频，包括从种族方面入手，操纵公众舆论和政治进程，使其对自己有利并破坏公众对机构的信任。还有报告称，政府利用人工智能制造不和，为网络审查提供便利。[18]

4. 问责问题

21. 一些人工智能工具无需人类便可独立做出决策，这意味着决策过程是隐蔽的，似乎发生在不透明的“黑箱”之中。此外，算法也许会独立做出决策，因为

[13] Ninareh Mehrabi and others, “A survey on bias and fairness in machine learning”, *ACM Computing Surveys*, vol. 54, No. 6 (2022); 伦敦故事基金会提交的材料；A/HRC/44/57, 第 17 段。

[14] 亚洲网域使命组织提交的材料。

[15] A.R. Lange and Natasha Duarte, “Understanding bias in algorithmic design”, Medium, 6 September 2017.

[16] See Amnesty International, *Decode Surveillance NYC: Methodology* (London, 2022); 亚洲网域使命组织提交的材料。

[17] 亚洲网域使命组织提交的材料。

[18] Tate Ryan-Mosley, “How generative AI is boosting the spread of disinformation and propaganda“, *MIT Technology Review*, 4 October 2023.

人工智能算法一旦获得数据就会不断自我更新。随着时间的推移，人工智能工具在决策中依靠的因素也许不是在编程时设计出来的，而来自于它在数据中自行识别的模式。随着算法将这些新模式纳入其代码和决策过程，依赖该算法的个人也许不再能够“开箱检查”，也无法确定算法究竟使用了哪些标准从而产生了特定的结果。因此，“黑箱”问题使人工智能的推理过程隐蔽而又缺乏透明。[19] 此外，企业实体开发的许多算法由于合同法和知识产权法的原因而无法得到审查，使得问责问题更为严重。[20]

22. “黑箱”问题在系统性种族主义方面的影响令人尤为关切。如上所述，系统性种族主义隐蔽却又极具破坏性，危害整个社会。人们并非总能认识到系统性种族主义背后的驱动力，而在收集种族和族裔分列数据方面持续存在的差距，又使这一现象更为严重。[21] 如果没有有效的问责机制，人工智能很可能成为另一项驱动因素，加剧系统性种族主义这一隐蔽而又具有破坏性的现象。

23. 人工智能的问责问题对遭受种族歧视者寻求有效补救的能力具有重大影响。目前，如果边缘化种族和族裔群体成员因为人的决策而经历不同的结果，法院和其他问责机制能够审查这些行为是否是故意的，是否具有正当理由。[22] 决策者是人的时候，这种评估通常有证据可循。许多情况下，自主决策过程不会像人类决策者那样产生证据线索。[23] “黑箱”问题将加剧遭受种族歧视者在寻求正义方面业已存在的重大障碍。

B. 人工智能的使用及其歧视性影响

24. 特别报告员在本节举例说明人工智能在不同社会领域的使用情况以及在种族歧视方面产生的影响。这些实例是说明性的，并非详尽无遗，举例是为了提供明确证据，表明人工智能已经在助长种族歧视。特别报告员认为，实例中的各类种族歧视表现相互关联并且相互强化，全面加剧了整个社会中系统的种族和族裔压迫。

25. 特别报告员选择以三个领域为例，说明人工智能的歧视性影响：执法、安全和刑事司法系统、教育、医疗保健。关于人工智能在其他领域的使用情况，特别报告员建议参考前任任务负责人关于数字边境的兴起和数字边境和移民执法中的种族歧视和仇外歧视情况描述的报告，以及关于在边境和移民执法中使用数字技术的报告。[24] 特别报告员还请读者参考她提交大会第七十八届会议的关于网上种族主义仇恨言论的报告，其中述及人工智能在社交媒体内容审核中的使用，[25]

[19] Yavar Bathaee, “The artificial intelligence black box and the failure of intent and causation”, *Harvard Journal of Law and Technology*, vol. 31, No. 2 (2018)； A/HRC/44/57, 第 34 段；and Renata M. O’Donnell, “Challenging racist predictive policing algorithms under the Equal Protection Clause”, *New York University Law Review*, vol. 94, No. 3 (June 2019).

[20] A/HRC/44/57, 第 44 段。

[21] A/HRC/47/53, 第 16 段。

[22] Bathaee, “The artificial intelligence black box”.

[23] 同上。

[24] A/75/590 和 A/HRC/48/76。

[25] A/78/538.

以及极端贫困与人权问题特别报告员提交大会第七十四届会议的报告，其中分析了人工智能在社会保障制度中的使用。[26]

1. 执法、安全和刑事司法系统

(a) 自动识别

26. 执法机构使用自动识别工具，把在特定环境中观察到的情况与数据库中潜在的“匹配项”联系起来。最常见的自动识别工具之一是人脸识别技术。人脸识别工具获取某人的视频或照片，将它们输入算法。然后，算法将这些图像与警方存有照片、驾照照片或其他图像的数据库进行比较，以识别此人的身份。[27] 此类工具的设计者运用机器学习的过程，通过展示人脸图像来训练工具的基础模型。目的是训练模型，使之能够分辨人脸的突出特征。[28] 然而，训练模型所使用的图像数据集并非总是具有人口统计学代表性。[29] 在一项关于常用图像数据库的研究中，研究人员发现 18 至 40 岁的男性比例过高，而深肤色人种比例过低。[30] 关于已发布商用版本的人脸识别系统的另一项研究显示，性别分类算法的训练采用了以白人男子图像为主的数据集。[31] 人工智能工具训练集缺乏种族、性别和文化多样性，导致上文所述的一种典型的数据问题。在训练数据中代表性不足的群体，包括遭受交叉形式歧视的群体，更容易被算法错误匹配。

27. 据报告，这些技术对人脸的错误识别导致越来越多的非洲人后裔被逮捕。[32] 促进和保护意见和表达自由权特别报告员和联合国人权事务高级专员指出，人脸识别工具往往会助长非法歧视和种族定性。[33] 尽管存在这样的人权关切，一些国家的执法机构仍然采用了人脸识别系统。例如，据报告，印度政府对这种系统进行了大量投资。德里警方使用的人脸识别系统准确率仅有 2%，而且少数群体被误认和误捕的风险尤为严重。[34] 据报告，巴西执法人员使用有缺陷的人脸识别工具对一些人员实施了错误的指控和逮捕。2019 年的一项研究显示，巴西各城市根据人脸识别技术逮捕的人员 90%是非洲人后裔。[35]

---

[26] A/74/493.

[27] Marissa Gerchick and Matt Cagle, “When it comes to facial recognition, there is no such thing as a magic number”, American Civil Liberties Union, 7 February 2024.

[28] Julia Dressel and Andrew Warren, “Breaking down data analytics and AI in criminal justice”, Recidiviz, 8 March 2022.

[29] 人工智能民享组织提交的材料。

[30] Khari Johnson, “ImageNet creators find blurring faces for privacy has a ‘minimal impact on accuracy’”, VentureBeat, 16 March 2021.

[31] Joy Buolamwini and Timnit Gebru, “Gender shades: intersectional accuracy disparities in commercial gender classification”, *Proceedings of Machine Learning Research*, vol. 81 (2018). See also Gerchick and Cagle, “When it comes to facial recognition, there is no such thing as a magic number”; 人工智能民享组织提交的材料；互联网实验室提交的材料。

[32] Gerchick and Cagle, “When it comes to facial recognition, there is no such thing as a magic number”.

[33] 见 A/HRC/41/35 和 A/HRC/48/31。

[34] Amnesty International, “Ban the scan: Hyderabad”, available at https://banthescan.amnesty.org/hyderabad/.

[35] 巴西专家小组提交的材料。

28. 枪声探测系统是多国执法官员使用的另一种常见的自动识别工具。一种名为 ShotSpotter 的系统在邻近地区放置包含麦克风、GPS、存储器和处理器以及蜂窝功能的传感器。[36] 当传感器探测到可能是枪声的噪音时，算法会进行三角测量，确定噪音源的位置。算法会过滤掉其他可能的声音，然后将音频发给人来审查。[37] 现有资料表明，枪声探测系统在边缘化种族群体居住的社区中部署得尤为密集，[38] 而且其误差率可能非常高。在边缘化种族和族裔群体居住的社区放置枪声探测系统，以及枪声探测系统的不准确性加剧了执法中的系统性偏见。

29. 自动识别技术的使用已经造成改变人生的后果，这样的例子不胜枚举。据报告，2019 年，美国新泽西州一名黑人男子因人脸识别出错而被误捕并入狱 10 天。尽管存在无罪证据，但主管机关在近一年的时间里都没有撤案，他因受到指控而面临长达 25 年的监禁。这一事件对该男子的人生造成重大影响。[39] 据报告，2024 年 2 月，芝加哥的执法人员在接到 ShotSpotter 发出的错误警报后，向一名正在燃放烟花的儿童开火。[40] 使用此类人工智能技术的另一个例子是，据报告以色列国防军采用了“狼群”技术，这是一个巨大的数据库，包含约旦河西岸巴勒斯坦人的图像和所有可用信息，进一步整合了“蓝狼”和“红狼”等各种监视方案。[41] 据报告，以色列国防军在希伯伦老城全面安装了能够识别人脸的人工智能摄像头，接入“蓝狼”程序，这是一个移动应用程序，以色列士兵可以利用巨大的生物识别数据库对约旦河西岸各地的巴勒斯坦人(大多数人未经同意便被录入数据库)进行检测和分类，从而对巴勒斯坦人进行持续监视。以色列国防军严格执行“狼群”制度，加剧了对巴勒斯坦人的长期种族隔离。[42] 这些例子表明，在高风险环境中使用人工智能系统作出重大决策会对人权造成严重影响。

(b) 预测性警务算法

30. 执法部门常用的另一种人工智能技术是预测性警务。预测性警务工具根据位置和个人数据来评估谁将会犯罪，以及犯罪行为将在何处发生。

[36] Alisha Ebrahimji, “Critics of ShotSpotter gunfire detection system say it’s ineffective, biased and costly”, CNN, 24 February 2024.

[37] Jay Stanley, “Four problems with the ShotSpotter gunshot detection system”, American Civil Liberties Union, 24 August 2021.

[38] Ibid.; and MacArthur Justice Center, “ShotSpotter is deployed overwhelmingly in Black and Latinx neighborhoods in Chicago”, available at https://endpolicesurveillance.com/burden-on-communities-of-color/.

[39] Gerchick and Cagle, “When it comes to facial recognition, there is no such thing as a magic number”; and Khari Johnson, “How wrongful arrests based on AI derailed 3 men’s lives”, *Wired*, 7 March 2022.

[40] Adam Schwartz, “Responding to ShotSpotter, police shoot at child lighting fireworks”, Electronic Frontier Foundation, 22 March 2024.

[41] Amnesty International, *Automated Apartheid: How Facial Recognition Fragments, Segregates and Controls Palestinians in the OPT* (London, 2023), pp. 41–45.

[42] Sophia Goodfriend, “Algorithmic State violence: automated surveillance and Palestinian dispossession in Hebron’s Old City”, *International Journal of Middle East Studies*, vol. 55, No. 3 (2023).

31. 预测性警务会加剧长久以来对种族和族裔社区过度执法的现象。[43] 由于执法官员历来关注这些街区，因此，附近的社区成员在警方记录中所占比例过高。这进而又会影响到算法所预测的未来犯罪地点，从而导致相关地区警力增加。[44] 预测性警务也能反映出“黑箱”问题的各个方面，因为算法缺乏透明度，包括所选择的分析数据和预测方法都不够透明。[45]

32. 基于地点的预测性警务算法利用地点、事件和历史犯罪数据之间的联系来预测可能发生犯罪行为的时间和地点。[46] 警方据此制定巡逻计划。当警方在被过度执法的街区记录到新的犯罪行为时，就会产生一个反馈回路，算法所生成的预测对这些街区的偏见越来越严重。简言之，过去的偏见会导致未来的偏见。在大不列颠及北爱尔兰联合王国，政府委托进行的一项关于警务工作中算法偏差的研究表明，将某些地点确定为犯罪“热点”，可能会使警察产生预期，认为这些地区会发生更多的犯罪行为。因此，警察更容易基于偏见而不是真正的公共安全需要，在“热点”地区实施拦截或逮捕。[47] 乌拉圭的研究人员发现，基于地点的预测性警务算法使用的数据可能存在偏差。位置变量可以成为社会经济或种族背景的替代变量，从而引发歧视。[48]

33. 基于人员的预测性警务工具提供了根据个人背景数据预测谁可能将会犯罪的方法。背景数据可以包括人的年龄、性别、婚姻状况、药物滥用史和犯罪记录。与基于地点的工具一样，由于已有的逮捕数据通常受到刑事司法系统中系统性种族主义的影响，这些算法对未来的预测可能出现偏差。社会经济背景、教育水平和地理位置等变量可以成为种族的替代变量，并固化历史偏见。[49] 澳大利亚新南威尔士州警方利用基于算法的“嫌疑人定向管理计划”识别有可能实施刑事犯罪的个人。据报告，该计划被叫停之前，曾导致警方对土著居民和托雷斯海峡岛民实施了过多干预。[50]

[43] Tim Lau, “Predictive policing explained”, Brennan Center for Justice, 1 April 2020; and Jon Fasman, “The black box of justice: how secret algorithms have changed policing”, *Fast Company*, 9 February 2021.

[44] Kristian Lum and William Isaac, “To predict and serve?”, *Significance*, vol. 13, No. 5 (2016); and Australian Human Rights Commission submission.

[45] Lau, “Predictive policing explained”.

[46] Will Douglas Heaven, “Predictive policing algorithms are racist. They need to be dismantled”, *MIT Technology Review*, 17 July 2020.

[47] Ibid. See also Government of the United Kingdom of Great Britain and Northern Ireland, “Report commissioned by CDEI calls for measures to address bias in police use of data analytics”, 16 September 2019.

[48] Juan Ortiz Freuler and Carlos Iglesias, “Algorithms and artificial intelligence in Latin America: a study of implementation by governments in Argentina and Uruguay”, World Wide Web Foundation, September 2018; and Eticas Foundation, “Uruguay’s Ministry of the Interior invests in predictive policing”, 13 September 2021.

[49] Heaven, “Predictive policing algorithms are racist”.

[50] Australian Human Rights Commission submission.

(c) 累犯评估算法

34. 累犯评估工具被用于在刑事司法系统的不同阶段为决策提供信息，包括关于保释、保证金、量刑和假释的决定。[51] 累犯评估工具使用历史数据评估被告以某些方式行事的可能性，特别是是否有可能在未来犯下新的罪行。这些工具利用犯罪记录和被告调查等信息，生成风险评分。[52]

35. 累犯预测工具体现了人工智能带来的助长种族歧视的多种挑战。首先，这些工具面临数据方面的问题。算法训练使用的刑事司法数据反映出长期的种族主义警务行为造成的系统性不平等。[53] 此外，设计时的选择，例如变量的测量或评估方式，也可能导致算法歧视。[54] 此外，算法设计者定义“成功”的方式，会对算法力求突出的内容产生影响。如果一种算法的设计旨在尽可能减少新发犯罪，可能会把刑期较长与重新犯罪率较低关联起来，因为人们在监禁期间无法重新犯罪。然后，算法可能会根据这些模式建议判处较长的刑期。

36. 研究人员曾表示，累犯预测工具不够准确，其误差对种族边缘化群体造成尤为严重的影响。例如，美国的一项研究发现，风险评分在预测暴力犯罪方面极不可靠。据报告，非洲人后裔被误认为可能犯罪的比率几乎是白人的一倍。

(d) 自主武器系统

37. 自主武器系统是指在关键功能上具有自主性的任何武器系统，包括致命性自主武器和非致命性武器。它们在执法和军事领域都得到应用，但在很大程度上仍未受到监管。这些系统可以在没有人类干预的情况下选择、探测、识别和攻击目标。传感器和软件找到与系统算法确定的“目标描述”相匹配的人，便会触发自主武器。自主武器系统对人权具有非常严重的影响，涉及生命权、禁止酷刑和其他虐待以及人身安全权等各项权利。[55]

38. 大会第一委员会听闻，随着世界准备迎接“技术突破”，制定防范措施、应对自主武器和人工智能军事应用所致危险的机会之窗正在迅速关闭。[56] 法外处决、即决处决或任意处决问题特别报告员曾建议人权理事会呼吁所有国家宣布并实施国家禁令，至少暂停试验、生产、组装、转让、获取、部署和使用致命的自主机器人。[57]

---

[51] Julia Angwin and others, “Machine bias”, ProPublica, 23 May 2016.

[52] 同上。

[53] See Heaven, “Predictive policing algorithms are racist”; and Michael Mayowa Farayola and others, “Fairness of AI in predicting the risk of recidivism: review and phase mapping of AI fairness techniques”, in *Proceedings of the 18th International Conference on Availability, Reliability and Security* (Association for Computing Machinery, 2023).

[54] Mehrabi and others, “A survey on bias and fairness in machine learning”.

[55] Amnesty International, “Autonomous weapons systems: five key human rights issues for consideration” (April 2015), p. 5.

[56] 联合国，“发言人告诉第一委员会，如果不采取足够的防范措施，人工智能将从算法演变为军备，威胁全球安全”，2023 年 10 月 24 日。

[57] A/HRC/23/47, 第 113 段。

39. 自主武器系统的使用极有可能导致严重的种族歧视，某些情况下甚至会致命。选择目标的标准可能包括性别、年龄和种族。[58] 目标描述还包括看似中性的标准，如体重或热量特征，但是机器往往反映了编程者和社会的偏见。机器编程也可能带有蓄意的歧视性目标描述。[59] 例如，据报以色列正在使用致命的自主和半自主武器系统。据报告，其中包括对巴勒斯坦人使用遥控四轴飞行器，此外还使用速度和功率超强的自动目标生成系统，以产生“杀人名单”。[60] 据报告，以色列国防军使用的两种人工智能技术系统——“Gospel”和“Lavender”使加沙遭到更严重的破坏并造成重大伤亡，尤其对巴勒斯坦妇女和儿童造成重大伤亡。[61]

2. 卫生保健

(a) 健康风险评分

40. 可以利用人工智能设计健康风险评分，用于各种卫生保健目的，包括医疗诊断和护理规划。在使用这种算法分配医疗资源时，算法设计和人工智能系统的训练数据可能会导致种族歧视效应。有时这种算法被用来识别谁应该有资格获得额外的医疗服务，使用以前的医疗费用作为医疗需求的替代指标。此类决策所依据的数据可能受以下因素的影响：边缘化种族和族裔群体在系统性种族主义背景下曾经缺乏获得适当医疗保健的机会，各群体之间由于在健康的社会经济决定因素方面存在差异而导致疾病模式不同。

41. 美国开发了一种计算器，帮助保健提供者评估剖腹产后阴道分娩成功的可能性。2019 年的一项研究显示，该计算器的基础算法存在偏差。计算器使用两个基于种族的校正因子，导致非洲裔和西班牙裔女性阴道分娩成功率低于具有类似特征的白人女性。由于这些校正因子，该计算器可能会加剧孕产妇健康结果的种族差异，因为临床医生会不愿意为非洲裔和西班牙裔妇女提供阴道分娩机会，导致她们的剖腹产率更高。[62]

(b) 疾病检测

42. 人工智能技术的另一种应用是疾病检测，包括癌症检测。[63] 人工智能系统接受大容量数据集的训练，其中包含数以千计或数以百万计的图像，包括放射扫

[58] Ray Acheson, “Gender and bias”, available at https://www.stopkillerrobots.org/wp-content/uploads/2021/09/Gender-and-Bias.pdf.

[59] Bonnie Docherty, “Expert Panel on the Social and Humanitarian Impact of Autonomous Weapons at the Latin American and Caribbean Conference on Autonomous Weapons”, Human Rights Watch, 8 March 2023.

[60] Marwa Fatafta and Daniel Leufer, “Artificial genocidal intelligence: how Israel is automating human rights abuses and war crimes”, Access Now, 9 May 2024.

[61] Yuval Abraham, “‘Lavender’: the AI machine directing Israel’s bombing spree in Gaza”, *+972 Magazine,* 3 April 2024.

[62] Darshali A. Vyas and others, “Challenging the use of race in the Vaginal Birth after Cesarean Section Calculator”, *Women’s Health Issues,* vol. 29, No. 3 (2019).

[63] 隐私国际提交的材料。

描、病理图像和照片，能够学会区分正常病变和癌症病变。[64] 人工智能的这种应用大大有助于早期癌症检测，在提高医疗系统效率的同时还能够挽救生命。然而，边缘化种族和族裔群体成员也许无法从这些进步中平等受益，因为这些算法不能同样适用于训练数据没有充分代表的患者群体。研究人员提出，用人工智能算法进行皮肤癌检测，深肤色人群的效果较差，因为算法训练使用的公开图像数据集带有偏差，肤色和族裔背景缺乏多样性。[65] 例如，对包含逾 10 万张图像的 21 个开放使用的皮肤病变数据集的调查显示，深肤色的图像明显不足。[66]

(c) 人工智能医疗设备

43. 伴随着人工智能的开发和利用，医疗保健技术，包括医疗保健设备，也出现了其他发展。许多设备受到人工智能的支持，其运行模式本身可能带有种族偏见。例如，英国的一份报告显示，在医疗设备开发的所有阶段，包括涉及算法开发和机器学习的阶段，设备的运行模式中都会植入偏见。在冠状病毒病大流行(COVID-19 疫情)期间，使用脉搏血氧仪测量血液中的低氧水平，导致深肤色人群的血液含氧量被高估。[67]

3. 教育

(a) 学业和职业成功算法

44. 芬兰和美国等国家在教育领域使用预测分析工具，根据数据、统计算法和机器学习来确定学生未来取得成功的可能性。[68] 这些算法使用的数据包括出勤率、成绩、表现和在线活动。这些算法旨在帮助教育工作者指导学生决定他们的教育和职业走向。尽管预测分析工具旨在帮助教育工作者改善学生的成绩，可是由于算法设计和数据选择的原因，结果通常认为少数族裔在学业和职业生涯中取得成功的可能性较小。根据这些评分，教育工作者可能会引导边缘化种族和族裔的学生放弃能够最大限度地发挥其潜力并提供最佳机会打破排斥循环的一些教育和职业选择，或者在这些学生身上投入较少资源。

(b) 评分算法

45. 评分算法通常使用历史评分数据来评估学生的表现。这些数据可能会因教育机构中系统性种族主义的历史模式而出现偏差。为学生预测评分的算法会复制数

[64] Likhitha Kolla and Ravi B. Parikh, “Uses and limitations of artificial intelligence for oncology”, *Cancer*, 30 March 2024.

[65] David Wen and others, “Characteristics of publicly available skin cancer image datasets: a systematic review”, *The Lancet Digital Health*, vol. 4, No. 1 (2022).

[66] 同上。另见隐私国际提交的材料。

[67] 隐私国际提交的材料。

[68] Stina Westman and others, “Artificial intelligence for career guidance – current requirements and prospects for the future”, *International Academic Forum Journal of Education*, vol. 9, No. 4 (2021); and Kelli A. Bird, Benjamin L. Castleman and Yifeng Song, “Are algorithms biased in education? Exploring racial bias in predicting community college student success”, *Journal of Policy Analysis and Management*, 31 January 2024.

据中的偏差，在排除教师意见的时候尤为如此。[69] 评分算法可能严重影响到学生的机会，包括进入大学或毕业后的就业机会。因此，带有种族偏见的自动决策可能会限制边缘化种族和族裔学生的机会，削弱教育作为瓦解系统性种族主义的一项工具的潜力。

46. 英国在使用评分算法方面提供了一个具有警示意义的例子。2020 年，高级水平(A-level)考试因 COVID-19 疫情而被取消。教师被要求预测学生的成绩，以替代考试成绩。国家评分监管机构随后采用了一种算法，根据每所学校的历史分数数据对预测分数进行标准化处理。40%的学生分数因此被下调，他们大多在低收入地区的学校就读。与之相反，该算法为自费的独立学校的大量学生调高了分数。政府面对争议，取消了算法所作的标准化处理。然而，这一事件对大学招生过程造成严重干扰。[70]

(c) 教育中的大型语言模型

47. 生成式人工智能工具依靠大型语言模型生成新颖的内容，包括文本、音乐、图像和视频。大型语言模型被应用于教育环境中，可以帮助所有年龄段的学生提高学业成绩。研究表明，语言模型偏向于英语，英语是互联网上使用最广泛的语言，也是大多数人工智能研究人员和技术人员的工作语言。此外，世界上约 6,000 种语言中，只有少数几种语言拥有可用于训练人工智能模型的高质量数据资源。为了解决这一差距，一些公司已开始开发多语言的语言模型。但是，多语言模型的表现不如英语模型。在教育中使用大型语言模型可能会使语言背景在基础数据资源中不具代表性的学生处于不利地位，可能从种族角度造成不成比例的影响。[71]

48. 关于是否应禁止学生使用基于大型语言模型的人工智能生成工具，而不是将其纳入课程的问题，一直存在争论。有些教育部门也在采取措施，试图限制学生使用基于大型语言模型的人工智能生成工具。一些教育机构使用人工智能工具来检测学生使用人工智能的情况。使用这种可能带有算法偏见的工具来检测作弊情况，有可能带来进一步的偏见，对边缘化种族和族裔群体的学生造成伤害。如果教育机构没有建立公平的申诉程序，这种伤害必然会更为严重。[72]

---

[69] Benjamin Herold, “Why schools need to talk about racial bias in AI-powered technologies”, *Education Week*, 12 April 2022.

[70] Bryan Walsh, “How an AI grading system ignited a national controversy in the U.K.”, Axios, 19 August 2020; and Daan Kolkman, “‘F**k the algorithm’? What the world can learn from the UK’s A-level grading fiasco”, London School of Economics Impact Blog, 26 August 2020.

[71] Felix Richter, “The most spoken languages: on the Internet and in real life”, Statista, 21 February 2024; Emily M. Bender, “The #BenderRule: on naming the languages we study and why it matters”, The Gradient, 14 September 2019; Gabriel Nicholas and Aliya Bhatia, “Lost in translation: large language models in non-English content analysis”, Center for Democracy and Technology, 23 May 2023; A. Bergman and Mona Diab, “Towards responsible natural language annotation for the varieties of Arabic”, in *The 60th Annual Meeting of the Association for Computational Linguistics: Findings of ACL 2022* (Association for Computational Linguistics, 2022); and BigScience Workshop, “A 176B-parameter open-access multilingual language model” (ArXiv, 2022).

[72] See Regina Ta and Darrell M. West, “Should schools ban or integrate generative AI in the classroom?”, Brookings Institution, 7 August 2023; and Robert Topinka, “The software says my student cheated using AI. They say they’re innocent. Who do I believe?”, *The Guardian*, 13 February 2024.

(d) 教育机构中的人脸识别

49. 如上所述，尽管有证据表明人脸识别技术在操作中存在种族偏见，却已被世界各地许多教育机构所采用。人脸识别系统被用于自动化考勤、加强校园安保、监考，甚至用于记录学童的情绪以监控他们的学习状况。这种做法往往没有得到充分的人权尽责管理或监管监督。例如，巴西越来越多的学校正在采用人脸识别工具来简化流程、跟踪出勤情况和加强安保。[73] 然而，据报告，这些项目实施前，市政府和州政府都没有开展人权影响评估研究，也没有分析与人脸识别软件相关的歧视风险。[74]

50. 在教育环境中使用人脸识别软件造成了种族歧视的影响。有些案例，如荷兰王国报告的一个案例显示，非洲裔学生必须用灯光照在脸上，才能被人工智能系统识别，参加重要的考试。这种经历不仅影响学生平等接受教育的权利，而且造成摩擦和排斥，因为边缘化种族和族裔群体的学生会认为该系统不是为他们所设计。在学校记录和监控孩子的情绪对所有学生的隐私都有很大影响，并可能延续种族偏见。人们发现，这些系统对非洲人后裔和白人的面部表情有不同的解释，更多地把非洲人后裔的表情解读为轻蔑和愤怒等负面情绪。[75]

C. 规范和管理人工智能的新举措

51. 各国已经开始采取措施规范和管理人工智能，效果可期。特别报告员在本节重点介绍了其中一些举措。所作的分析并非详尽无疑，主要基于各国和民间社会团体提交的材料，以及她为本报告开展的国别工作和研究。

1. 各国的举措

52. 各国已采取措施，通过具有约束力的法律规定和自愿性政策标准，在许多情况下双管齐下，在国家层面对人工智能进行规范和管理。例如，巴西正在审议关于规范技术空间的法律条款，[76] 包括规范人工智能的法律条款。巴西政府还通过了若干相关政策文件，例如“互联网上的种族主义：制定数字政策的证据”这份文件，据报告，该文件由种族平等部编写，载有应对算法偏见的措施，包括如何应对与种族有关的偏见。[77] 特别报告员欢迎各国努力制定专门的、有约束力的监管规定，并辅之以相关政策标准。然而，她收到令人不安的信息称，在制定关于人工智能监管的法律条款时，没有与非洲人后裔开展有效协商，非洲人后裔的参与不足，各类不同的标准及国家当前的做法之间也缺乏整体一致性和连贯性。[78]

53. 另一个例子是美国。据报告，美国已采取措施制定人工智能的使用标准，既有具有约束力的标准，也有自愿标准。特别报告员近期访问美国后，欣见该国签

[73] 互联网实验室提交的材料。

[74] 同上。

[75] 同上。

[76] 巴西提交的材料。

[77] 巴西专家小组提交的材料。

[78] 同上。

署了《关于安全、可靠和可信地开发和使用人工智能的第 14110 号行政命令》，并欢迎其中提到在人工智能使用方面存在偏见和歧视的风险。特别报告员编写本报告时，收到了关于美国人工智能监管的进一步资料，包括就“人工智能权利法案”这一自愿标准开展的工作，弗吉尼亚、加利福尼亚和纽约等各州为规范人工智能所做的努力，以及促进企业实体自愿承诺开发安全、可靠和透明的人工智能的举措。[79] 特别报告员对上述努力表示欢迎，但感到关切的是，尽管有大量研究涉及美国数字商业产品中深刻的算法种族偏见，第 14110 号行政命令却没有明确提及种族歧视和偏见。[80]

54. 据报告，加拿大也在同时制定具有约束力的标准和自愿标准。《人工智能和数据法》目前还处于草案形式，据报告，其中包含对高风险人工智能系统具有约束力的管控法规。此外，加拿大还制定了自愿性标准，包括《关于负责任地开发和管理先进生成式人工智能系统的自愿行为守则》。加拿大还开发了算法影响评估工具，旨在帮助政府部门和机构评估并减轻人工智能的风险，包括与歧视和偏见有关的风险。[81]

55. 特别报告员收到了关于澳大利亚、中国、印度和日本等其他国家的资料，据称这些国家已采取措施管理并规范人工智能，包括采取政策措施，以及在某些情况下制定具有约束力的立法。[82]

2. 区域举措

56. 关于区域举措，特别报告员欣见欧洲联盟及其成员国提供的关于通过《人工智能法案》的信息。[83] 她承认，该法案是一项具有约束力的监管标准，通过使各国法律标准与其条款保持一致，将在欧盟区域产生重大影响。特别报告员欣见《人工智能法案》纳入了种族问题，对高风险人工智能的使用设立人权保障，禁止人工智能的某些应用，并为受到使用高风险人工智能系统影响的人提供补救机制。据报告，《2020-2025 年欧洲联盟反种族主义行动计划》述及因使用人工智能等新技术而产生的种族歧视问题，并建议在欧盟各类不同标准之间实现一定程度的政策一致性，特别报告员对此也表示欢迎。[84] 然而，特别报告员收到令人深感关切的信息，在移民和边境管理及执法这两个领域，该法案所规定的保护措

[79] 亚洲网域使命组织提交的材料。See also Kay Firth-Butterfield, Karen Silverman and Benjamin Larsen, “Understanding the US ‘AI Bill of Rights’ – and how it can help keep AI Accountable”, World Economic Forum, 14 October 2022; United States, Office of Science and Technology Policy of the White House, “Blueprint for an AI bill of rights: making automated systems work for the American people”, white paper, October 2022; and United States, White House, “Fact sheet: Biden-Harris Administration secures voluntary commitments from eight additional artificial intelligence companies to manage the risks posed by AI”, 12 September 2023.

[80] A/HRC/56/68/Add.1，第 54 段。

[81] Canada, Innovation, Science and Economic Development Canada, “The Artificial Intelligence and Data Act (AIDA) – companion document”, 13 March 2023；亚洲网域使命组织提交的材料。

[82] 亚洲网域使命组织提交的材料。

[83] 欧洲联盟和西班牙提交的材料。

[84] 欧洲联盟提交的材料。另见 https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html。

施可出现例外。[85] 据报告，尽管这两个领域都存在严重的历史性种族歧视，而且允许平行法律框架同时发展的做法具有内在缺陷，但这种例外仍然存在。[86] 这种做法有可能加深现有的种族等级制度，并延续欧洲联盟各成员国在移民和边境管理及执法中严重侵犯人权的状况。

3. 国际举措

57. 特别报告员了解联合国为促进人工智能管理而采取的措施。她欢迎秘书长设立人工智能高级别咨询机构，也欢迎该机构近期发布了中期报告。然而，她感到遗憾的是，该报告没有具体提及种族偏见和歧视的风险。特别报告员欣见联合国人权事务高级专员办事处(人权高专办)通过“B-Tech 项目”等将人权问题纳入关于人工智能等新兴技术的国际对话。除了联合国的工作之外，特别报告员还了解到促进对话和管理的其他国际举措，包括经济合作与发展组织和七国集团的倡议。[87]

58. 国际组织能够很好地促进国际合作、技术援助和研究，以确保人工智能的管理方式不会加剧国家之间已经存在的严重不平等，这种不平等多数是殖民主义和奴隶制遗留的问题。技术基础设施的显著差异可能会导致各国在使用人工智能工具时面临不同挑战。在人工智能技术及其带来的歧视的问题上，关注的焦点一直是全球北方的国家，这可能导致在理解人工智能将如何影响全球南方的文化、宗教和其他少数群体方面出现空白。[88] 北方最发达的国家可能会影响关于人工智能的辩论和对话，从而延续全球力量失衡的状况，并限制南方国家获取潜在利益的能力。

D. 国际人权法框架

59. 人工智能技术应以国际人权法标准为基础。《消除一切形式种族歧视国际公约》对种族歧视予以最全面的禁止。如《公约》第一条第一款所体现的，各国在起草《公约》时纳入了对种族歧视的广义定义，即基于种族、肤色、世系或民族或人种的任何区别、排斥、限制或偏好，其目的或效果为取消或损害政治、经济、社会、文化或公共生活任何其他方面人权及基本自由在平等地位上的承认、享受或行使。

60. 《消除一切形式种族歧视国际公约》缔约国承诺努力实现没有任何形式种族主义的国内和国际社会。为了促进在实质上实现种族平等，《公约》第二条要求缔约国确保不参与任何种族歧视行为或导致种族不平等其他计划。此外，如果存在种族主义、种族不平等或种族歧视，缔约国有义务立即采取有效行动。采取行动的义务是一项绝对义务。缔约国防止种族不平等和种族歧视的义务既包括预

[85] 隐私国际提交的材料；and Access Now, “The EU AI Act: a failure for human rights, a victory for industry and law enforcement”, 13 March 2024.

[86] 见 A/HRC/48/76。

[87] 亚洲网域使命组织提交的材料。

[88] Danni Yu, Hannah Rosenfeld and Abhishek Gupta, “The ‘AI divide’ between the Global North and the Global South”, World Economic Forum, 16 January 2023.

防，也包括补救。《公约》对种族歧视予以最全面的禁止，其他条约也保护人们免受这种形式的歧视。

61. 实现种族平等和确保不歧视的义务涵盖政府政策和影响的所有领域，包括人工智能技术的设计和应用。考虑到《消除一切形式种族歧视国际公约》和其他人权条约禁止种族歧视的范围，人工智能造成的种族歧视是否属于蓄意，与缔约国的行动义务无关。《公约》缔约国有责任努力实现一个没有任何形式种族歧视的国内和国际社会，也关系到各国采取何种方式，预防和解决国家内部和国家之间在人工智能技术惠益分配方面的不平等问题。

62. 各国还必须确保所有种族和族裔群体充分享有《消除一切形式种族歧视国际公约》第五条载列的所有人权。第五条规定：法律面前人人平等，其中特别包括在法庭上和其他一切司法裁判机关中享有平等待遇的权利；人身安全和受国家保护免遭暴力或身体伤害的权利，不论这种伤害是由政府官员还是由任何个人团体或机构所致；和平集会和结社自由的权利；享有公共卫生、医疗保健、社会保障和社会服务的权利；接受教育和培训的权利。其他人权条约，包括《公民及政治权利国际公约》和《经济、社会、文化权利国际公约》，也规定了上述权利和保障以不歧视方式适用这些权利的条款。

63. 如上所述，国际人权法的其他条款也规定各国有责任解决人工智能技术的歧视性影响。在缺乏人权保障的情况下收集和使用数据，会引起严重的隐私问题，而对于边缘化种族和族裔群体而言，这些问题可能会更加严重。因此，特别报告员希望提醒各国注意，《公民及政治权利国际公约》第十七条规定个人隐私不受任意或非法干涉，并赋予各国确保相关法律保护的义务。该公约其他条款也适用于与人工智能技术有关的种族歧视表现。使用人工智能，包括在执法环境中使用人工智能，可能影响人身自由和安全，还可能对边缘化种族和族裔群体造成致命后果。《公约》第六条概述了人所固有的生命权，并规定各国有义务就此提供法律保护。第七条规定，不得对任何人施以酷刑或残忍、不人道或有辱人格的待遇或处罚。第九条规定，人人有权享有人身自由和安全，不得对任何人实施任意逮捕或拘禁。第十四条明确规定，在法院和法庭面前人人平等。第二十六条规定保护少数群体不受歧视。《公约》第二条第一款规定了确保《公约》所有条款得到不歧视适用的义务。国际人权法框架中也包括关于在移民和边境管制以及社交媒体中使用人工智能的规定。本任务的前几份报告探讨了这些问题。[89]

64. 国际人权法规定，所有可能遭受种族歧视的人都有权获得补救，这适用于人工智能导致歧视的情况。《消除一切形式种族歧视国际公约》第六条确立了通过国内主管法庭和其他国家机关获得有效保护和补救的权利。此外，大会还确认了严重侵犯人权行为受害者获得补救和赔偿权利的五大要素：恢复原状、补偿、康复、满足和保证不再发生。[90]

65. 企业实体在人工智能的设计和应用中发挥着重要作用。这些实体是开发人工智能的主要行为者，而且经常受政府委托在公共部门环境中部署人工智能。《工商企业与人权指导原则》概述了政府的相关义务以及政府和企业的相关人权责

[89] A/75/590 和 A/78/538。

[90] 《严重违反国际人权法和严重违反国际人道主义法行为受害人获得补救和赔偿的权利基本原则和导则》，第 15 至 23 段。

任。《指导原则》规定，各国必须提供保护，防止包括工商企业在内的第三方在其领土和(或)管辖范围内侵犯人权。各国应采取行动，通过确保有效的政策、立法、法规和裁决等方式，提供这种保护。《指导原则》规定，企业有责任防止、减轻和补救其可能造成或促成的侵犯人权行为，并对相关商业活动开展人权尽责管理。[91] 此外，《指导原则》还规定了政府义务和企业责任，以确保与企业相关的侵犯人权行为可获得补救，对上述其他标准中规定的补救权形成补充。人权高专办的“B-Tech 项目”为在技术领域实施《指导原则》提供指南和资源，包括围绕人工智能开展了具体工作。[92]

四. 结论和建议

66. 前任任务负责人曾向各国和包括企业实体在内的其他利益攸关方发出明确呼吁，要避免采取“不分肤色”的办法管理和规范人工智能等新兴技术。她敦促各国在规范这些技术时认识到结构性种族主义并以关键的人权标准为依据。然而，人工智能的管理和规范在很大程度上仍然存在不足，对种族偏见的关注不够，也未能体现国际人权法标准。尽管前任任务负责人曾明确而及时地呼吁采取“不分肤色”的办法，而且在随后的四年里，各方对系统性种族主义的认识有所提高，但是这种情况持续至今。认为技术客观而又中立的假设依然普遍，正在驱使各方不顾这种技术在种族歧视方面的影响，争先恐后地将人工智能融入社会，而没有适当考虑是否有必要这样做。虽然人工智能确实具有积极潜力，包括促进平等和包容，但它并不是解决所有社会问题的灵丹妙药，必须对其加以有效管理，以平衡利益和风险。

67. 对人工智能的全面有效监管是实现这种微妙平衡的核心要素。虽然对人工智能的有效监管至关重要，但是各国和其他各方也可以采取更多措施来有效应对这些技术的种族歧视影响。基于人权对公众开展新兴技术的教育和培养人工智能素养也非常重要。当个人和群体了解了人工智能并知悉他们在数字空间中的人权时，就有能力负责任地使用这些知识，并成为有洞察力和责任感的受众，能够改善人工智能系统的问责。

68. 各国应当：

(a) 以更大的紧迫感应对人工智能监管的挑战，铭记这些技术的研发速度，以及这些技术在社会各领域延续种族歧视的多种方式；

(b) 在全面理解系统性种族主义的基础上，以国际人权法包括以禁止种族歧视的规定为依据，制定人工智能监管框架。此类框架不应建立在各自为政的基础上，而应考虑到不同的法律文书，包括专门的人工智能立法、隐私法、信息自由的规定、反歧视立法和部门规章，以实现全面有效的监管，防止和解决人工智能的种族歧视影响；

[91] 另见联合国防止灭绝种族罪行和保护责任办公室及经济和社会研究委员会，人权、大数据和技术项目，University of Essex, “Countering and addressing online hate speech: a guide for policy makers and practitioners”, policy paper, July 2023; A/74/486, 第 44 和 45 段。

[92] 见人权高专办，“B-Tech 项目：人权高专办和工商业与人权”，可查阅 https://www.ohchr.org/en/business-and-human-rights/b-tech-project。

(c) 考虑自愿标准在人工智能监管框架中可以发挥的作用。自愿标准可以为实际执行法律标准提供可操作的准则。然而，人工智能监管不应该完全依赖自愿标准，因为这些技术对人权具有重大影响，包括与种族歧视有关的影响；

(d) 在监管框架内规定一项具有法律约束力的义务，在所有人工智能技术的开发和部署中开展全面的人权尽责评估，包括制定评估种族和族裔偏见的明确标准。人权尽责评估应包括数据测试规程和阈值，以防止算法偏见，包括种族和族裔偏见。这些工作应在新技术部署之前完成，在教育、执法和医疗等公共环境中尤应如此；

(e) 考虑禁止使用已被证明具有不可接受的人权风险的人工智能系统，包括违反禁止种族歧视规定的人工智能系统；

(f) 确保监管框架含有相关规定，在全面实行人权尽责管理的基础上，在人工智能被视为可以使用的情况下，保障自动决策过程充分透明，包括保障信息获取权；

(g) 制定明确和便于利用的上诉程序，负责评估和处理人工智能的种族歧视影响并进行人工审查。应确保能够公平利用此类上诉程序；

(h) 建立机制，在人工智能技术导致侵犯人权包括种族歧视的情况下，使受影响的个人和群体能够获得补救，确保恢复原状、赔偿、康复、满足和保证不再发生；

(i) 避免监管标准中出现任何例外情况，可能导致违反国际人权法禁止种族歧视的规定；

(j) 确保在制定和实施人工智能法规以及开发和使用人工智能技术时，以切实有效的方式征求所有边缘化种族和族裔群体利益攸关方以及相关部门专业人员的意见；

(k) 加强收集所有相关部门的分类数据，以获得必要信息，在了解系统性种族主义的基础上制定人工智能法规，解决人工智能系统中的数据问题，更好地监测和评估人工智能技术对边缘化种族和族裔群体的影响；

(l) 采取基于人权标准的数据办法，在数据收集和存储工作中确保分类、自我认同、透明、隐私、参与和问责；[93]

(m) 建立健全的机制，对人工智能工具开展监督和持续监测，包括定期审计其影响，以确保相关法规得到遵守，解决受影响个人或社区提出的任何关切，应对人工智能模型随时间推移可能产生的偏见；

(n) 开展国际合作、能力建设和研究，确保在各国之间更公平地分配人工智能的惠益，以避免加深殖民主义和奴隶制遗留的不平等现象；

(o) 立足于人权开展公共教育，宣传如何以可接受的、负责任的方式使用人工智能技术，提高人工智能素养，内容包括专门提高对人工智能的种族歧视影响的认识。

[93] 见 A/HRC/42/59 和 A/HRC/44/57。

69. 企业实体应当：

(a) 在人工智能产品设计、开发和部署的各个阶段开展人权尽责评估；

(b) 确保在设计、开发和部署人工智能产品时，与边缘化种族和族裔群体、相关社会领域的专业人员以及对系统性种族主义有专门了解的人员开展切实有效的磋商；

(c) 制定规程，确保具有人权影响的产品的算法决策完全透明，确保共享相关信息；

(d) 确保持续监测人工智能产品是否存在种族偏见；

(e) 为所有参与人工智能设计、开发和部署的人员开展种族歧视方面的培训，包括关于隐性偏见和系统性种族主义的培训。培训材料应借鉴国际人权法标准和关于人工智能技术种族歧视影响的研究；

(f) 协助开展立足于人权的公共教育，提高人工智能素养的普及程度。

70. 联合国及其独立人权机制应当：

(a) 促进各利益攸关方就人工智能技术及其监管问题开展有效的对话和辩论；

(b) 将人工智能技术在种族歧视方面的影响作为重点，明确纳入人工智能高级别咨询机构的工作；

(c) 确保关于人工智能技术的出版物和指南以国际人权法(包括禁止种族歧视的规定)为基础，并明确承认人工智能技术设计和部署中的种族偏见是一个严重的全球性问题；

(d) 发挥作用，监测人工智能技术对人权的影响，包括与种族歧视有关的影响；

(e) 支持开展国际合作、能力建设和研究，努力在各国之间更公平地分配人工智能的惠益。