联 合 国 A/HRC/44/57

大 会

Distr.: General
18 June 2020
Chinese
Original: English

人权理事会
第四十四届会议
2020 年 6 月 15 日至 7 月 3 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

种族歧视与新兴数字技术：人权分析

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的报告*

概要

在本报告中，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滕达依·阿丘梅分析了在新兴数字技术设计和使用中的种族歧视的不同形式，包括这种歧视的结构和体制层面。她还概述了国家的人权义务和公司打击这种歧视的责任。

* 本报告逾期提交，以纳入最新信息。

GE.20-08060 (C) 170720 200720
2008060

一. 导言

1. 新兴数字技术已从根本上改变了我们的生活方式，因此，它们对人权的影响一直是人权理事会特别程序重要分析的主题。[1] 现有的各种报告阐述了这些技术如何影响广泛的人权，包括意见和表达自由权、和平集会和结社自由权以及赤贫者的人权。联合国人权事务高级专员对新兴数字技术和隐私权进行了分析。[2] 在本报告中，特别报告员的目标是就新兴数字技术与国际人权法下的种族平等和不歧视原则的交叉领域提出类似的有力分析。

2. 根据《消除一切形式种族歧视国际公约》，本报告涵盖基于种族、肤色、世系或民族或族裔出身的种族主义、不容忍、歧视和其他形式的有害排斥和区分对待。包括对土著人民的歧视。在本报告中，特别报告员敦促采取基于平等的办法对新兴数字技术进行人权治理。这需要超越“不分肤色”或“种族中性”的策略。[3] 对法律、社会、经济和政治条件进行不分肤色的分析，致力于公平处事，即使某些个人和群体因蓄意歧视的历史原因而处于不同的地位，也要避免明显的种族或族裔分析，而要对所有个人和群体一视同仁。在新兴数字技术方面，则需要政府官员、联合国和其他多边组织以及私人部门认真关注这些技术的种族化和族裔影响。在本报告中，特别报告员着重探讨了交叉形式的歧视，包括基于性别和宗教的歧视，并提请注意各国和其他利益攸关方一直未能跟踪和解决种族、族裔、性别、残疾状况、性取向和相关理由交叉领域的复合形式歧视。

3. 特别报告员仅简要论述了新兴数字技术对移民、难民和其他非公民的种族歧视影响，因为这些群体将是特别报告员提交大会的另一份报告的重点。

4. 本报告的一项重要结论是，新兴数字技术加剧了现有的不平等并使其更加复杂，其中许多不平等是基于种族、族裔和民族出身等原因。报告中着重指出的例子对新兴数字技术设计和使用中不同形式的种族歧视提出了关切。有时，这种歧视直截了当，明显源自不容忍或偏见。有时，即便没有明确的歧视意图，但由于对不同种族、族裔或民族出身的群体会造成差别影响，也会产生歧视。还有时，直接和间接形式的歧视相结合，可能造成重大的整体或系统性影响，使特定群体受制于种族歧视性结构，影响他们在生活的所有领域获得和享有人权。

5. 例如，在冠状病毒病(COVID-19)大流行的背景下，早期报告显示，大流行对边缘化种族和族裔群体存在差异影响，原因包括这些群体被排除在新兴数字技术的好处之外，或者新兴数字技术的部署方式使这些群体遭受人权侵犯的风险更大。尽管人们普遍认为新兴数字技术的运作是中立和客观的，但在这些技术已经渗透的所有领域中，种族和族裔都影响着对人权的获得和享有。国家有义务防止、打击这种种族歧视并提供补救，公司等私人行为体也有相关责任这样做。

6. 在各种新兴数字技术之中，特别报告员在本报告中侧重于联网预测技术，其中许多涉及大数据和人工智能，并在一定程度上强调基于算法(和算法辅助)的决策。对种族歧视和新兴数字技术的许多现有人权分析揭示了一系列具体问题：网

[1] 见 www.ohchr.org/Documents/HRBodies/SP/List_SP_Reports_NewTech.pdf。

[2] 见 A/HRC/39/29。

[3] 见 https://qz.com/1585645/color-blindness-is-a-bad-approach-to-solving-bias-in-algorithms。

上仇恨事件和利用数字平台协调、资助和建立对种族主义者群体及其活动的支持。在本报告中，特别报告员更进一步，用种族平等和不歧视原则来检验新兴数字技术的结构和体制影响，研究人员、倡导者和其他人认为这种影响令人震惊。令人关切的问题包括，新兴数字技术被普遍用于决定就业、教育、医疗保健和刑事司法的日常结果，这导致系统性歧视风险达到了前所未有的高度。欧洲联盟基本权利署最近的一份报告便着重指出了欧洲联盟这些关切的例子，并就所需的应对措施提供了宝贵的建议。[4]

7. 作为“区分、排序和分类的归类技术”，人工智能系统就其核心而言是“区别对待的系统”。[5] 即使没有会形成模式化看法的显性算法规则，机器学习算法也会复制大规模数据集中内嵌的偏见，能够模仿和复制人类的隐性偏见。[6] 数据集作为人类设计的产物，可能会因“偏差、差距和错误的假设”而产生偏向性。[7] 它们还可能出现“信号问题”，即由于创建或收集数据的方式不平等而导致缺乏人口代表性或代表性不足。[8] 除了不准确、缺失和代表性不佳的数据之外，“脏数据”还包括被有意操纵或被偏见扭曲的数据。[9] 这种数据集可能导致对某些人口的歧视或排斥，特别是基于种族、族裔、宗教和性别身份认同的少数群体。

8. 即使无意造成歧视，使用无害和真正相关的标准也可能造成间接歧视的，因为这些标准也可以成为种族和族裔的代替值。其他关切包括使用和依赖包含历史数据的预测模型——这些数据通常反映了歧视性偏见和不准确的定性——包括在执法、国家安全和移民等背景下。在更基本的层面上，新兴数字技术的设计要求开发者选择如何最好地实现他们的目标，这些选择会导致不同的分布结果。[10] 特别报告员在报告中的一个核心关切便是这种基于种族、族裔和相关理由对个人和群体的人权会产生不同影响的选择。

9. 具体就阶层而言，研究表明，即使决策者、公务员和科学家采用自动化决策意在实现更高效和更公平的决策，但事实显示，他们用来实现这些目标的系统加剧了不平等，并对贫困者造成难以承受的后果。[11] 鉴于种族和族裔上的边缘化社区往往不成比例地生活在贫困条件下，平等和不歧视原则应当成为新兴数字技术促进社会福利和其他社会经济系统的人权分析的核心。极端贫困与人权问题特别报告员最近的一份重要报告描述了数字福利国家在一些国家的兴起，这些国家的社会保障和援助系统由新兴数字技术驱动，这对社会福利产生了严重的负面人权影响。[12] 如稍后的章节所述，特别报告员的评估是，对于目前存在的数字福

[4] 见 https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making.

[5] Sarah Myers West, Meredith Whittaker and Kate Crawford, “Discriminating systems: gender, race and power in AI” (New York, AI Now Institute, 2019), p. 6.

[6] 见 https://philmachinelearning.files.wordpress.com/2018/02/gabbriellejohnson_algorithmic-bias.pdf。

[7] 见 https://foreignpolicy.com/2013/05/10/think-again-big-data。

[8] 同上。

[9] 见 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3403010。

[10] 见 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899。

[11] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York, Picador, 2018).

[12] 见 A/74/493。

利政府，更确切的描述应该是“区别对待的数字福利政府”，因为它们往往任由种族和族裔(等因素)对人权的享有造成歧视性影响。有必要采取紧急干预措施来遏制这些歧视性模式。

10. 在编写报告的过程中，特别报告员受益于以下宝贵贡献：日内瓦大学全球问题研究院、加州大学洛杉矶分校承诺人权研究所、加州大学洛杉矶分校法学院和加州大学洛杉矶分校关键互联网调查中心主办的专家组会议；哈佛法学院网络法律诊所在伯克曼·克莱因互联网和社会研究中心以及纽约大学法学院种族、不平等和法律中心所开展的研究；对研究人员的采访；以及一系列利益攸关方响应征集意见的公开呼吁提供的材料。非机密材料将在本特别任务网页上公布。[13]

二. 新兴数字技术中歧视与不平等的推动因素

11. 任何对新兴数字技术的人权分析都必须首先应对影响其设计和使用的社会、经济和政治力量，以及导致这些技术设计和使用中产生种族歧视的个人和集体人类利益和优先事项。

12. 公众对技术的惯常看法是，它本质上是中立和客观的，一些人指出，这种关于技术客观性和中立性的假设即使在技术生产者中也仍然很突出。但技术从来都不是中性的——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。[14] 例如，2019 年对来自全球 99 家开发商的 189 种面部识别算法的分析发现，“与白种人的面部相比，其中许多算法在识别黑色人种或东亚人种面部照片时的不准确可能性要高出 10 至 100 倍。在搜索数据库以找到一张特定的面孔时，大多数算法在黑人女性中选中不正确照片的比率明显高于在其他人口中的比率。[15] 新兴数字技术复制、强化甚至加剧社会内部和社会之间种族不平等的能力惊人，关于这一点已经不再有任何疑问。许多重要的学术研究已经以具体的方式表明，技术的设计和使用正在各种不同领域产生这种效果。[16] 需要投入更多研究和资金来充分揭示一些人工智能技术(如机器学习)核心的归纳过程是如何削弱平等和非歧视等价值观的。[17]

13. 在生产新兴数字技术的领域和行业中，事实已经证明，对数字中立性或客观性及其克服种族主义能力的错误信念会导致歧视性结果。[18] 即使是为促进新兴数字技术设计和使用中的公平、问责和透明而有所发展的领域，也需要更加关注

[13] 见 www.ohchr.org/EN/Issues/Racism/SRRacism/Pages/Callinformationtechnologies.aspx。

[14] Langdon Winner, *The Whale and The Reactor: A Search for Limits in an Age of High Technology* (Chicago, University of Chicago Press, 1986), p. 29.

[15] 见 www.scientificamerican.com/article/how-nist-tested-facial-recognition-algorithms-for-racial-bias。

[16] 例如，见 Cathy O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York, Penguin, 2016); Ruha Benjamin, *Race After Technology* (Cambridge, United Kingdom, Polity Press, 2019); 以及 Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York, New York University Press, 2018).

[17] 例如，见 Gabrielle M. Johnson, “Are algorithms value-free? Feminist theoretical virtues in machine learning”, *Journal of Moral Philosophy* (forthcoming).

[18] 例如，见 West, Whittaker and Crawford, “Discriminating systems”.

歧视和不公正的更广泛的社会结构。[19] 事实上，解决新兴数字技术使用和设计中的种族歧视的最大挑战之一，是将这一问题单纯或主要视为技术问题，应由计算机科学家和其他行业专业人士通过设计无偏见的数据和算法来解决。技术是社会及其价值观、优先事项甚至不平等的产物，包括与种族主义和不容忍相关的不平等。技术决定论——认为技术能影响社会，但其本身在很大程度上是中立的，不受社会、政治和经济力量的影响——只会保护塑造新兴数字技术的力量及其影响不被发现，不受改革影响。“技术沙文主义”——过度依赖技术可以解决社会问题的信念[20] ——也有类似的效果，对塑造技术和技术成果的价值观和利益提出质疑并予以改变会因此变得复杂。

14. 尽管仍然非常需要对确保平等和不歧视的设计质量进行监督和问责，但要确保这些原则和其他人权原则，首先必须承认问题的核心是政治、社会和经济问题，而不仅仅是技术或数学问题。不平等和歧视，即使是新兴数字技术设计和使用的产物，也不会因为关于平等和不歧视的建模变得更加完美而得到“治愈”。具体而言，这意味着在私人和公共部门，旨在打击新兴数字技术设计和使用中的种族歧视的思想和行动不应成为技术专家的专属领域或近乎专属的领域。相反，这种思考和行动必须更加全面，研究人员和其他在新兴数字技术领域拥有专长的人正是这样主张的。[21] 各国政府和私人部门必须致力于在研究、辩论和决策的所有阶段让种族歧视的政治、经济和社会层面的专家参与其中，以减少新兴数字技术设计和使用中的种族歧视。受影响的种族和族裔少数群体必须在相关进程中发挥决策作用。

15. 私营公司在新兴数字技术的设计和使用方面具有巨大的影响力。在数字平台中，七大“超级平台”——微软、苹果、亚马逊、谷歌、脸书、腾讯和阿里巴巴——占据了全球 70 大平台总市值的三分之二。[22] 尽管这些平台的新兴数字技术影响遍及全球，但对影响力最大的公司主要集中在美利坚合众国的硅谷，而欧洲的份额为 3.6%，非洲为 1.3%，拉丁美洲为 0.2%。[23] 例如，谷歌占据了全球互联网搜索市场 90%的份额。[24] 脸书占据了全球社交媒体市场的三分之二，是全球 90%以上经济体的首要社交媒体平台。亚马逊在全球在线零售活动中占据近 40%的份额。因此，硅谷特定的文化、经济和政治价值观从根本上决定了有多少新兴数字技术在全球运作，包括在远离北美这一小块地区的环境中。

[19] 例如，见 www.tandfonline.com/doi/full/10.1080/1369118X.2019.1593484 和 http://sorelle.friedler.net/papers/sts_fat2019.pdf.

[20] 例如，见 Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World* (Cambridge, Massachusetts, Massachusetts Institute of Technology, 2018).

[21] 例如，见 www.odbproject.org/2019/07/15/critiquing-and-rethinking-fairness-accountability-and-transparency。

[22] 《2019 年数字经济报告：价值创造和捕获——对发展中国家的影响》(联合国出版物，出售品编号 E.19.II.D.17)，第 xvi 页。

[23] 同上，第 2 页。

[24] 同上，第 xvii 页。

16. 除了市场支配地位之外，公司还在政府与其国民之间扮演着重要的中介角色，有能力显著改变人权状况。全球北方强大的全球性公司生产的技术是在非常特定的政治、经济、社会和治理背景下创造的。在其他背景下，例如在全球南方，它可能会产生恶劣影响。脸书在缅甸扮演的角色就是一个例子。[25] 人们还对全球北方追求利润的公司行为体以不受监管、在某些情况下甚至是剥削性的条件从全球南方的个人和国家获取数据表示关切，这些公司行为体无法对此负责。[26]

17. 新兴数字技术部门，如硅谷的数字技术部门，其特点是性别和种族方面的“多样性危机”，[27] 尤其是在最高决策层。该领域的一项重要研究指出，“目前，大规模人工智能系统几乎完全是在少数几个科技公司和一小撮精英大学实验室中开发的，在西方，这些领域往往具有白色、富裕、技术导向和男性的极端特征。这些领域也有歧视、排斥和性骚扰问题的历史。”[28] 这项研究进一步发现，“这绝不仅是一两个恶劣行为体的问题：它指出了人工智能领域及驱动其生产的行业内的排斥模式和在人工智能技术的逻辑和应用中表现出的偏见之间的系统关系。”[29] 这类领域生产的技术过分排斥妇女、种族、族裔和其他少数群体，在使用时很可能会重现这些不平等。在复杂的社会现实和现有系统中生产有效的技术需要理解社会、法律和道德背景，这只能通过纳入多样和有代表性的视角以及学科专业知识来实现。[30]

18. 市场和经济力量对新兴数字技术的设计和使用施加了重大影响，这些技术反过来又对市场，甚至对资本主义本身产生了变革性的影响。[31] 一方面，一些经济影响力有意寻求促进歧视和不容忍。这方面的例子包括富人为鼓吹优越主义意识形态的在线平台提供资金。[32] 另一方面，最强大的市场力量可能主要是想从新兴数字技术中寻求有利可图的结果，而没有明显的种族主义或不宽容的意图。但是证据表明，有利可图的产品会产生种族歧视。如果种族和族裔不平等会影响经济结构——世界各地都是如此——利润最大化通常会与种族和民族不平等并存，并在许多情况下强化或加剧这种不平等。

19. 在很大程度上，在获取和享受新兴数字技术的好处方面的不平等(a) 在国际层面遵循地缘政治不平等的模式，(b) 在个体国家内部遵循种族、族裔和性别不平等的模式。[33]

---

[25] 见 A/HRC/39/64。

[26] 例如，见 https://privacyinternational.org/long-read/3390/2020-crucial-year-fight-data-protection-africa.

[27] 见 www.technologyreview.com/2018/02/14/145462/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our; Noble, *Algorithms of Oppression*; and West, Whittaker and Crawford, “Discriminating systems”.

[28] West, Whittaker and Crawford, “Discriminating systems”, p. 6 (footnote omitted).

[29] 同上，第 7 页。

[30] 同上。

[31] 例如，见 Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York, Public Affairs, 2019).

[32] 例如，见 https://journals.sagepub.com/doi/full/10.1177/1536504218766547。

[33] 例如，见 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3403010。

20. 在国际层面，全球南方国家缺乏全球北方的数字基础设施：全球南方的活跃宽带用户不到全球北方的一半。[34] 在非洲，22%的个人使用互联网，而欧洲为80%。[35] 在所谓的最不发达国家，只有五分之一的人上网，而在所谓的发达国家，有五分之四的人上网。[36] 尽管技术有利于国家应对 COVID-19 大流行，这些好处也没有得到平均分配。最不发达国家不仅最容易遭受 COVID-19 对人类和经济产生的影响，而且在线获取公共卫生信息和利用数字教学、工作和购物平台的数字能力也最不足。[37]

21. 在各个国家中，美国和中国主导着全球数字经济。这两个国家占全球 70 个最大数字平台市值的 90%，其中包括社交媒体和内容平台、电子商务平台、互联网搜索服务、移动生态系统和工业云平台。[38] 目前的预测表明，新兴数字技术将进一步扩大有能力利用这些技术和没能力利用这些技术的国家之间的数字鸿沟。

22. 国家内部也存在数字鸿沟。例如，美国虽然在全球数字经济中占据主导地位，但该国的种族和族裔少数群体从新兴数字技术中获益的机会截然不同。如下文第三章所述，在许多情况下，他们遭受着与新兴数字技术相关的最严重的侵犯人权行为。皮尤研究中心 2019 年的一项调查显示，在美国，黑人和西班牙裔成年人拥有电脑或拥有高速互联网的可能性仍然较低。[39] 有 82%的白人表示拥有台式电脑或笔记本电脑，但拥有电脑的黑人和西班牙裔人分别只有 58%和57%。[40] 在宽带使用方面也存在巨大的种族和族裔差异，白人家里有宽带连接的可能性比黑人或西班牙裔人高出 13%到 18%。[41] 这种沿着种族和族裔坐标的数字鸿沟是显著的。然而，正如研究人员所言，促进对种族和族裔少数群体的数字包容的干预措施绝不能让他们遭受进一步的权利侵犯，包括因隐私和监控问题而遭受的侵犯。[42] 就中国而言，下文第三章举例说明了该国在设计和使用新兴数字技术方面造成的严重人权后果。中国新兴数字技术在全球南方的影响力日益增强，进一步加剧了这些关切。

23. 土著人民也遭到歧视，无法获得新兴数字技术的好处。[43] 加拿大的估计表明，以土著为主的北方人口中有大约一半缺乏南方人口所拥有的高速连接。[44] 澳大利亚的土著数字包容性也很低，尤其是在城市以外地区，2011 年，一些偏远土著社区只有 6%的居民拥有电脑。[45] 到 2015 年，土著人民拥有互联网连接的可能性仍比非土著人低 69%。[46]

---

[34] 《2018 年贸易和发展报告：权力、平台与自由贸易之谬》，(联合国出版物，出售品编号 E.18.II.D.7)，第 viii 页。

[35] 同上。

[36] 见《2019 年数字经济报告》，第 13 页。

[37] 见 https://unctad.org/en/PublicationsLibrary/dtlinf2020d1_en.pdf。

[38] 见《2019 年数字经济报告》，第 xvi 页。

[39] 见 www.pewresearch.org/fact-tank/2019/08/20/smartphones-help-blacks-hispanics-bridge-some-but-not-all-digital-gaps-with-whites。

[40] 同上。

[41] 同上。

[42] 见 https://journals.uic.edu/ojs/index.php/fm/article/view/3821。

[43] 见 https://unesdoc.unesco.org/ark:/48223/pf0000260781。

[44] 见 https://rightscon2018.sched.com/event/EHqs/addressing-the-digital-divide-in-indigenous-communities-in-north-america。

[45] 见 www.creativespirits.info/aboriginalculture/economy/internet-access-in-aboriginal-communities。

[46] 同上。

三. 新兴数字技术设计和使用中的种族歧视举例

A. 明确的不容忍和出于偏见的行为

24. 寻求传播种族主义言论和煽动歧视和暴力的行为体依赖新兴数字技术，而社交媒体平台在其中发挥着关键作用。特别报告员在以往关于新纳粹和其他白人至上主义团体的报告中着重指出过这些趋势，这些团体依靠社交媒体平台招募人员、筹集资金和进行协调。[47] 另一个明显出于偏见使用新兴数字技术的突出例子是，缅甸激进的民族主义佛教团体和军事行动体利用脸书加剧对穆斯林，特别是罗辛亚少数民族的歧视和暴力。[48] 2018 年，脸书首席执行官马克·扎克伯格在美国参议院作证说，在这种情况下，脸书的人工智能系统无法检测到仇恨言论。[49] 这并不是唯一的例子：有一份提交材料也着重指出利用脸书来放大歧视性和不容忍内容，包括煽动针对印度宗教和语言少数群体的暴力的内容。[50]

25. 社交媒体机器人程序——自动账户——被用来转移政治话语和歪曲公众意见。在 70 个国家的样本中，2019 年有 50 个国家的社交媒体利用机器人程序开展操纵活动。[51] 对于那些以新兴数字技术为策略来加剧种族、民族和宗教不和谐和不容忍的群体而言，机器人程序是他们在网上传播种族主义言论或虚假信息的核心所在。有例子表明，在选举之前，系统使用机器人程序特别普遍。例如，在 2018 年瑞典大选之前，研究人员发现，讨论国家政治的推特账户中有 6%是机器人程序，它们发布的与移民和伊斯兰教相关的话题多于真实账户。[52] 同样，2018 年美国大选之前，发布反犹太主义推文的推特账户有 28%是机器人程序，所发推文占所有反犹太主义推文的 43%。[53] 在俄罗斯联邦，通过推特、脸书和其他社交媒体网站上的数百个伪造的在线角色和页面，新兴数字技术被用来在社交媒体上推进族裔和种族分裂。[54] 虽然有些帖子面向族裔少数群体，呼吁种族平等，但许多帖子谴责这类群体，试图加剧种族紧张关系。一些虚拟角色支持白人民族主义团体，煽动对少数种族的歧视和暴力。[55]

---

47 见 A/73/312 和 A/HRC/41/55。

48 A/HRC/42/50，第 71-75 段。

49 见 www.commerce.senate.gov/2018/4/facebook-social-media-privacy-and-the-use-and-abuse-of-data.

50 Avaaz 提交的材料。

51 见 https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf。

52 见 www.semanticscholar.org/paper/Political-Bots-and-the-Swedish-General-Election-Fernquist-Kaati/2af3d1e16d5553dc489d8b44321ea543d571a4a9。

53 见 www.adl.org/resources/reports/computational-propaganda-jewish-americans-and-the-2018-midterms-the-amplification。

54 见 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3304223。

55 同上，第 180 页。

B. 新兴数字技术设计/使用中的直接或间接歧视

26. 新兴数字技术的设计和使用可能基于种族或族裔对于一系列人权的享有造成直接和间接的歧视。

27. 关于工作权，一份提交材料报告称，巴拉圭实施了一个数字就业系统，允许雇主按不同类别对应聘者进行分类筛选，其中一些类别可作为种族的代替值。[56] 此外，该系统只有西班牙语版本，而巴拉圭只有不到一半的农村土著人口讲西班牙语。这种有限的语言无障碍性实际上在族裔方面限制了该系统对求职者的可用性，即使政策制定者本无此意图。

28. 北美和欧洲一些用于甄选成功求职者的算法也因提出歧视性建议而受到批评。这些系统经过培训，可以根据现有“成功”员工的数据集来确定候选人，这些数据集包括一些关于受保护特征的信息，如性别、族裔或宗教。因此，相应的算法系统做出的决策反映出就业中现有的不平等，重复并强化了现有的种族、族裔、性别或其他偏见。这些系统实际造成了直接和间接的种族歧视。[57] 另一方面，如果这些系统禁止考虑受保护的身份因素，例如种族和族裔，它们会削弱国家为促进平等就业机会而可能采取的特别措施或平权行动。

29. 在另一些情况中，引入不直接依赖歧视性输入信息或流程的自动化系统，仍然会减少或淘汰现有职位，在边缘化群体获得工作的机会方面造成间接歧视。有一份提交材料提供了印度一个基于人工智能的卫生管理新项目的例子，该项目将淘汰许多通常由最低种姓或达利特从事的工作。[58] 达利特，尤其是妇女，往往只能在卫生部门找到工作，印度的一些州会优先考虑达利特从事卫生工作。智能卫生系统的实施可能会给达利特的工作和生计造成尤为严重的影响，特别是达利特妇女。鉴于在印度，达利特在社会经济和政治方面遭到更广泛的边缘化，卫生部门的自动化可能从根本上削减那些依赖卫生部门就业机会的人找到工作的机会。

30. 新兴数字技术也对健康权产生了歧视性影响。美国市场上排名前十的医保算法用患者过去的医疗费用来预测未来的费用，这些预测费用被用作医疗保健需求的代替值。[59] 一家领先的医疗服务公司最近对这样一种算法进行了研究，发现它对美国的黑人患者实施了无意但系统的歧视。[60] 该算法原本旨在帮助高风险患者加入保健管理计划，但研究发现，通过使用患者的医疗费用作为其健康需求的代替值，以预测其风险水平，这种算法将种族偏见写入了代码。[61] 由于输入数据中没有种族项，开发者认为该算法是“不分种族”的[62]，对于与白人患病程

[56] 平等权利信托组织提交的材料。

[57] 见 https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making。

[58] 进步通信协会提交的材料。另见 www.apc.org/sites/default/files/gisw2019_artificial_intelligence.pdf。

[59] 见 www.sciencenews.org/article/bias-common-health-care-algorithm-hurts-black-patients。

[60] 见 https://science.sciencemag.org/content/366/6464/447。

[61] 同上。

[62] 见 www.thelancet.com/journals/landig/article/PIIS2589-7500(19)30201-8/fulltext。

度相当的黑人患者，该算法始终给予其较低的风险分数。[63] 该算法未能识别出将近一半与白人患者同样有可能产生复杂医疗需求的黑人患者。其结果是，黑人患者不太可能被转介参与改善健康的干预项目。每年，医院、保险公司和政府机构使用这种算法和类似算法来帮助管理全国超过 2 亿人的医疗服务。[64]

31. 还是美国的一个例子，最近的一项案例研究调查了由全球领先电子健康记录开发商 Epic Systems 公司开发的预测模型。[65] Epic 将人工智能工具直接集成入现有的电子健康记录，通过使用患者的个人信息，包括族裔、阶层、宗教信仰和身体质量指数，以及患者之前的爽约记录，来估计患者爽约的可能性。研究人员指出该工具明显具有歧视弱势患者群体的可能，并表示，“从模型中去除敏感的个人特征并不是彻底消除偏见的方法。”[66] 例如，之前的爽约记录很可能与社会经济地位有关，可能的理由包括患者无力支付交通费或请人照看孩子的费用，或没法为预约请假。由于社会经济地位与种族和族裔之间的相关性，这种记录也可能与种族和族裔相关。[67] 最近的另一项研究还发现，黑人患者更有可能被安排在超额预定的预约时段，因此当他们确实赴约时不得不等待更长时间。[68]

32. 在住房领域，美国的研究显示，脸书的定向广告存在族裔歧视。脸书过去允许广告商通过在其广告定位工具的“人口统计”类别下排除具有某些“族裔亲缘关系”的用户来“缩小受众范围”。[69] 这种定向广告可以用来阻止黑人观看特定的住房广告，这是美国反歧视法所禁止的。据估计，脸书控制着美国 22%的数字广告市场份额[70]，其定向广告是该公司商业模式的核心所在，[71] 已被证明具有种族排斥性。[72] 对这些做法的最佳理解是一种数字标记形式，其定义是“创建并维持一些技术做法，进一步强化对已经边缘化的群体的歧视性做法。”[73] 脸书也在求职领域使用定向广告，引起了类似的关切。

33. 在另一些情况下，无法使用技术——或无法通过技术获取信息——会对不同人群产生差别影响，或者针对特定的种族、族裔或宗教群体，有时是出于歧视。2019 年，包括孟加拉国、刚果民主共和国、埃及、印度、印度尼西亚、伊朗伊斯兰共和国、缅甸、苏丹和津巴布韦在内的多个国家完全关闭了特定地区的互

[63] 见 https://science.sciencemag.org/content/366/6464/447。

[64] 同上。

[65] 见 www.healthaffairs.org/do/10.1377/hblog20200128.626576/full。

[66] 同上。

[67] 见 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3467047。

[68] 同上。

[69] 见 www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race。

[70] 见 www.emarketer.com/content/us-digital-ad-spending-will-surpass-traditional-in-2019。

[71] 见 www.motherjones.com/politics/2019/12/facebook-agreed-not-to-let-its-ads-discriminate-but-they-still-can。

[72] 见 www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin。

[73] 见 www.congress.gov/116/meeting/house/110251/witnesses/HHRG-116-BA00-Wstate-GillardC-20191121.pdf。

联网接入，结果是几乎所有进出这些地区的通信都遭到禁止。[74] 研究人员发现，还有更多有针对性关闭互联网行为涉及少数群体密度较高的地区。[75]

34. 关于获得公平审判的权利，拉丁美洲有多个法院已经开始使用 Prometea 软件，这是一种利用语音识别和机器学习预测来简化司法程序的软件。布宜诺斯艾利斯的地区检察官办公室和法院使用这种人工智能系统实现简单案件中司法决策的自动化，例如关于出租车牌照的纠纷，以及教师关于没有获得学习用品补贴的投诉。[76] 在这种情况下，Prometea 会解读所获得的事实，并根据类似案件的先前判例就法律结果提出建议。决定必须获得法官批准才能正式生效，在 96%的情况中，法官都批准了决定。[77] 切实存在的关切在于，如此高的批准率很可能源自对于技术客观性和中立性的前述假定。哥伦比亚宪法法院使用 Prometea 来过滤监护权申诉，或有关个人宪法权利的申诉，然后决定审理哪些案件。[78] Prometea 和许多其他此类人工智能系统令人关切的问题在于“黑箱”效应——其决策基础不透明，法官、其他法院官员和诉讼当事人(甚至委托开发这些系统的公共机构)很难或不可能确定设计、输入或输出中是否存在偏见。虽然不可能知道 Prometea 对种族和族裔少数群体产生或可能产生的影响，但这类系统很有可能会加强或加剧其所在司法体系中现有的种族和族裔不平等。

35. 在刑事司法背景下，世界不同地区的警察部门使用新兴的数字技术进行预测性警务，其中人工智能系统从多种数据来源获取信息，如犯罪记录、犯罪统计数据和社区人口统计。[79] 其中许多数据集反映了现有的种族和族裔偏见，因此尽管这些技术被假定具有“客观性”，甚至被认为有可能减轻它们所补充或取代的人类行为体的偏见，但其运作反而加剧了种族歧视。此外，警察部门在以族裔或种族少数群体为主的贫困社区往往会更多使用预测技术。

36. 例如，大不列颠及北爱尔兰联合王国使用一个被称为帮派暴力矩阵的数据库，该数据库已被证明具有歧视性。[80] 警察据称会根据个人的种族、性别、年龄和社会经济地位对其做出假设，这进一步强化了这些成见。[81] 结果是，矩阵中 78%的人是黑人，另有 9%来自其他族裔少数群体，而警方自己的数字显示，严重青年暴力的责任人中只有 27%是黑人。警方还与就业中心、住房协会和教育机构等其他机构共享该矩阵，导致基于个人假定的帮派关系对其实施歧视。根据信息共享方式的性质，这为侵犯隐私权创造了机会，并可能导致住房和就业权利因歧视受到影响。那些名字出现在“矩阵”中的人“多次遭遇表面上缺乏任何法

[74] 见 www.hrw.org/news/2019/12/19/shutting-down-internet-shut-critics。

[75] 见 www.accessnow.org/cms/assets/uploads/2020/02/KeepItOn-2019-report-1.pdf。

[76] 见 www.bloombergquint.com/businessweek/this-ai-startup-generates-legal-papers-without-lawyers-and-suggests-a-ruling。

[77] 见 www.giswatch.org/2019-artificial-intelligence-human-rights-social-justice-and-development。

[78] 见 www.ambitojuridico.com/noticias/informe/constitucional-y-derechos-humanos/prometea-inteligencia-artificial-para-la (西班牙文)。

[79] 进步通信协会提交的材料。

[80] A/HRC/41/54/Add.2，第 40 段。

[81] 见 www.amnesty.org.uk/files/reports/Trapped%20in%20the%20Matrix%20Amnesty%20report.pdf。

律依据的拦截和搜查。”[82] 一些人报告说，警察已经拦截和搜查了他们 200 次，其他人报告称多达 1000 次，有些人称每天都会遭到多次拦截。这影响到个人隐私不受干涉的权利，以及不因种族歧视遭受任意逮捕的权利。

37. 另一个例子是，一份提交材料中强调指出，预测性警务正在成为美国洛杉矶等城市所谓预防犯罪战略中所使用的地方治安方法。[83] 直到最近，洛杉矶警察局一直在使用一种叫做 PredPol 的技术来检查 10 年以来的犯罪数据，包括犯罪的类型、日期、地点和频率，以预测未来 12 小时内犯罪可能发生的时间和地点。这些由警察收集和分类的数据，既是加强对黑人和拉丁社区监控的产物，也是其原因。预测性警务再现并加剧了警察队伍中现有的偏见，同时由于使用了据称为中性的算法决策而披上了客观性的伪装。尽管洛杉矶警察局已暂停使用 PredPol, 但并未否认会使用其他可能引发类似关切的预测性警务产品。

C. 种族歧视性结构

38. 世界各地的例子表明，不同新兴数字技术的设计和使用可能被有意无意地结合在一起，从而产生种族歧视性结构，基于特定群体的种族、族裔或民族血统，结合其他特征，从整体上或系统性地损害这些群体对人权的享有。换言之，不应仅仅认为新兴数字技术能够削弱各项人权的获得和享有，还应认识到，这些技术能够在系统上或结构上制造和维持种族和族裔排斥。在这个小标题下，特别报告员回顾了现有的歧视性结构和可能造成歧视的结构的一些例子，着重指出生物特征数据系统、种族化监视和种族化预测分析在维护这些结构中的普遍作用。

39. 中国使用生物识别和监控技术来追踪和限制少数民族维吾尔族人的行动和活动，侵犯了该群体成员的平等权和不受歧视权等权利。[84] 维吾尔族人经常遭遇警察的无端拦截，在警方检查站手机遭到扫描，这侵犯了他们的隐私权。强制收集维吾尔族人的大量生物特征数据，包括脱氧核糖核酸样本和虹膜扫描。根据可信的报告，国家“利用面部识别技术和全国各地的监控摄像头，根据维吾尔族人的外貌对他们进行专门的监控，记录他们的来往行踪，以便进行搜索和审查”[85]。还有报告指出，在进行这种监视和数据收集活动的同时，还以打击宗教极端主义为借口将大量少数民族隔离关押在政治“再教育营”，没有对被拘留者提出指控或进行审判。[86] 所呈现的画面是一种系统性的族裔歧视，这种歧视得到了众多新兴数字技术的支持，确切地说，是新兴数字技术使之成为可能，侵犯了维吾尔族人的广泛人权。

[82] 见 www.stop-watch.org/uploads/documents/Being_Matrixed.pdf。

[83] 阻止洛城警局监控联盟提交的材料。

[84] 见 CERD/C/CHN/CO/14-17。

[85] 见 www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html；以及 A/HRC/41/35，第 12 段。

[86] 见 CERD/C/CHN/CO/14-17。

40. 肯尼亚和印度在获取公共服务方面实施了生物识别技术，分别名为 Huduma Namba 和 Aadhaar。[87] 方案内容包括收集各种形式的生物特征数据，包括指纹、视网膜和虹膜图案、声纹和其他识别特征。两国的特定种族和族裔少数群体试图通过这些系统获得公共服务时发现自己被排除在外，另一些人则面临后勤障碍和漫长的审查过程，实际上，这可能导致这些人事实上无法获得其有权获得的公共服务。所涉公共服务有印度的养老金和失业福利，以及肯尼亚的所有基本政府服务，包括投票、出生登记和结婚登记、纳税和领取财产契约。印度最高法院判决要求凭借 Aadhaar 号码领取政府福利的法规有效。尽管同一项判决禁止私人实体将 Aadhaar 用于非政府业务，如银行、就业和移动电信等，但在实践中这一要求仍然普遍存在。此外，残疾人——包括族裔和种族少数群体中的残疾人——因不能提供指纹或虹膜扫描而遭受歧视。尽管法律为这类人员规定了特殊机制，但他们仍然面临后勤障碍，因为许多中心没有接受过培训，不知道如何在没有生物特征数据的情况下为他们登记。[88] 如果没有严格的保护，公共服务的数字身份识别系统会不成比例地排斥种族和族裔少数群体，特别是那些公民身份没有保障的人。[89]

41. 许多国家正在试验将新兴的数字技术纳入其福利系统[90]，其方式加强了种族歧视性结构。澳大利亚实施了在线遵约干预系统，俗称机器债务。[91] 这种自动决策系统使用机器学习算法来识别政府福利津贴的疑似超额支付事件，并要求那些被标记为领取了超额福利津贴的人提供文件证明。在 2016 年和 2017 年的六个月期间，该系统每周发出约 20,000 封债务信函。一项调查估计，由于系统流程和数据存在缺陷，20%至 40%的债务函为“虚报”。国家将证明他们不欠国家任何债务的举证责任转移到福利领取者身上。由于澳大利亚土著人享受的福利额度比澳大利亚白人更高[92]，他们承担了该系统缺陷的最大代价，同时，鉴于他们面临的障碍，他们也最不具备质疑该系统的条件。最近荷兰对司法程序中的一次人权干预凸显了类似的关切，在该国，利用新兴数字技术提供社会福利导致该国最贫穷和最弱势群体的人权受到侵犯。[93] 在那里，种族和族裔少数群体同样面临着不成比例的社会经济边缘化，引发了关于阶级歧视也是种族歧视的紧迫关切。

[87] 见 A/74/493。

[88] 见 https://timesofindia.indiatimes.com/city/kolkata/court-relief-in-disabled-womans-aadhaar-battle/articleshow/68961357.cms。

[89] 关于获得公民身份方面的种族歧视的人权分析，见 A/HRC/38/52。

[90] 见 A/74/493。

[91] 见 www.unswlawjournal.unsw.edu.au/forum_article/new-digital-future-welfare-debts-without-proofs-authority and www.ombudsman.gov.au/__data/assets/pdf_file/0022/43528/Report-Centrelinks-automated-debt-raising-and-recovery-system-April-2017.pdf。

[92] 见 www.aihw.gov.au/reports/australias-welfare/australias-welfare-2019-data-insights/contents/summary。

[93] 见 www.ohchr.org/Documents/Issues/Poverty/Amicusfinalversionsigned.pdf。

42. 随着各国越来越多地使用新兴数字技术来计算风险和对需求进行分类，例如丹麦、新西兰、联合王国和美国等[94]，各国政府必须作为优先事项对新兴数字技术对种族或族裔少数群体产生有差异影响的潜力进行更严格的审查。由于在进行福利制度数字化的社会，有一些群体因其种族和族裔而遭到边缘化、歧视和排斥，因此除非国家积极采取预防措施，否则这些制度几乎肯定会加剧这些不平等。如果没有紧急干预，数字福利国家就有可能成为歧视性的数字福利国家。

43. 有时，尽管种族歧视性结构仅限于特定部门，例如刑事司法，但它们从整体上削弱了受影响者的人权，并强化了他们在社会中遭受的结构性压迫。美国就是这种情况，新兴数字技术延续和复制了刑事司法中的种族歧视性结构。在那里，新兴数字技术不仅在警务中常见，而且在司法系统中也很常见，造成了对种族和族裔少数群体的歧视性后果。美国有几个州在刑事司法程序的每一步都使用人工智能风险评估工具。开发人员希望这些系统能够提供客观的、基于数据的司法结果，[95] 但是这些算法通常依赖于“在有证据表明存在有缺陷、带有种族偏见、有时甚至非法的做法和政策的时期产生的数据。”[96] 由于这些算法影响判决，它们可能侵犯在法律面前人人平等的权利、获得公平审判的权利以及免受任意逮捕和拘留的权利。这种风险评估权衡的因素通常包括：先前的逮捕和定罪记录、父母的犯罪记录、邮政编码和所谓的“社区混乱”。[97] 正如一项研究的作者指出的那样：“这些因素反映了过度监管、黑人和棕色人种社区的执法行为、种族种姓制度导致的更广泛的社会经济劣势模式，而不是目标人群的行为”。换言之，数据更能预示被告所在社区的种族劣势和警力，而不是个人的行为。[98]

四. 采取结构性和交叉性人权法方针处理新兴数字技术设计和使用中的种族歧视：分析和建议

44. 从人权角度来看，新兴数字技术构成了巨大的监管和治理挑战。在许多情况下，造成歧视性和相关后果的数据、代码和系统是复杂的，并受到保护可不受审查，包括不受合同法和知识产权法的审查。在某些情况下，甚至连计算机程序员自己也无法解释其算法系统的运行方式。这种“黑箱”[99] 效应使得受影响群体难以克服通过法律程序证明存在歧视通常需要的沉重举证负担，这还是在假定首先法院程序可用的情况下。另一方面，要求负责创造和实施新兴数字技术的公司证明其系统符合人权原则，不会产生种族歧视性结果的法律规定很少，甚至没有。

[94] A/74/493，第 27 段。

[95] 见 www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[96] 见 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423.

[97] 纽约大学种族、平等和法律中心提交的材料。

[98] 同上。

[99] Frank Pasquale, *The Black Box Society: the Secret Algorithms that Control Money and Information* (Cambridge, Massachusetts, Harvard University Press, 2015).

45. 正如促进和保护意见和表达自由权特别报告员所强调的那样[100]，国际人权法绝不是解决本报告所指出问题的灵丹妙药，但它将在查明和解决人工智能的社会危害以及确保对这些危害问责方面发挥重要作用[101]。管理新兴数字技术的道德方针必须符合国际人权法，各国必须确保这些道德方针不会取代发展和执行具有法律约束力的现有义务。在这一节中，特别报告员解释了国际人权法下的直接、间接和结构性种族歧视的概念和理论，概述了它们在新兴数字技术方面对各国规定的义务。这些义务也影响到非国家行为体，如技术公司，它们在许多方面对这些技术的控制权要多于国家。本节还包括具体实施所提出规范和义务的非详尽建议清单。

A. 新兴数字技术设计和使用中法律禁止的种族歧视的范畴

46. 特别报告员回顾，国际人权法的前提是，所有人都应享有与生俱来的所有人权，不得以任何理由受到歧视。禁止种族歧视已经获得国际法强制性规范[102] 和普遍义务的地位。[103] 根据国际人权法，各国在若干不同的条约体制中进一步阐述了种族平等和不歧视义务；平等和不歧视原则已写入所有核心人权条约。[104]《公民及政治权利国际公约》第二十六条规定，法律应禁止任何歧视，并保证人人享受平等而有效之保护，以防因种族、肤色、性别、语言、宗教、政见或其他主张、民族本源或社会阶级、财产、出生或其他身份而生之歧视。《经济社会文化权利国际公约》也禁止基于这些理由的歧视。[105]

47. 《消除一切形式种族歧视国际公约》第一条第一款将种族歧视定义为基于种族、肤色、世系或民族或人种的任何区别、排斥、限制或优惠，其目的或效果是取消或损害在政治、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、享有或行使。《公约》的目的远不止是一种形式上的平等观念。国际人权框架中的平等是实质性的，要求各国采取行动打击有针对性的或蓄意的种族歧视，也要打击事实上或无意的种族歧视。[106]

48. **在新兴数字技术背景下，各国必须根据国际人权法对禁止种族歧视进行结构性解读。决定国家如何定义因新兴数字技术的特定使用而产生的种族歧视是一项重要职能，在这方面必须采用人权法的定义，国家应要求私营部门采用的方针参考这些定义。这意味着，它们不仅必须述及新兴数字技术的使用和设计中明显的种族主义和不容忍现象，还必须涵盖这类技术的设计和使用所导致的间接和结构**

[100] A/73/348，第 19-60 段。

[101] 见 https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

[102] 人权事务委员会关于紧急状态期间《公约》条款克减问题的第 29 号一般性意见(2001 年)，第 8 段和第 13(c)段。另见 A/CN.4/727，第 59 段。

[103] 巴塞罗那电车、电灯及电力有限公司案，判决，《1970 年国际法院案例汇编》，第 3 页起，见第 32 页第 33 段。

[104] A/HRC/32/50，第 10-14 段。

[105] 见第二条第二款。

[106] 消除种族歧视委员会，关于《公约》中特别措施的含义和范围的第 32 号一般性建议(2009 年)，第 6-7 段。

性的种族歧视，后者同样重要。打击种族歧视的义务涵盖上文第三部分所述的种族歧视性结构和其他形式的直接和间接歧视。各国在治理和监管新兴数字技术时必须摒弃“无关肤色”的方针，这种方针忽视了种族和族裔少数群体特定的边缘化，将与此类技术有关的问题和解决方案概念化，而没有考虑到它们可能对这些群体产生的影响。

49. 特别报告员回顾《德班行动纲领》第 92 至 98 段，敦促各国收集、汇编、分析、传播和公布按种族或族裔分列的可靠统计数据，以解决与新兴数字技术的设计和使用有关的个人和群体种族不平等问题。特别报告员敦促各国采取立足人权的方针处理数据，确保在数据收集和存储过程中注意分类、基于自我认同、透明、确保隐私、参与和问责。[107] 确定和解决直接和间接形式的歧视需要(按照人权标准)收集能够揭示新兴数字技术差别影响的数据。然而，许多国家未能收集或要求收集此类数据。事实上，一些欧洲联盟国家禁止收集分类数据，以便识别和纠正基于族裔或种族的歧视。[108] 这些禁令阻碍了这些国家履行其防止和打击种族歧视的义务，它们应该进行改革。在这方面，最近联合王国的种族差异审计是积极进展的一个例子。[109]

50. 按照国际人权法的规定，消除种族歧视需要进行交叉性分析。以下关于交叉性的定义很好地说明它的重要性：

> “交叉性”的概念旨在表现两种或多种形式的歧视或压迫系统之间相互作用的结构性和动态后果。它具体涉及种族主义、父权制、经济不利地位和其他歧视性制度如何造成了不同层面的不平等，进而确定了男女之间、不同种族群体以及其他群体的相对地位。此外，它还涉及具体行为和政策如何沿着这些交叉轴线制造障碍，从而助长了剥夺权能的动态。[110]

51. 消除种族歧视委员会已经澄清，《消除一切形式种族歧视国际公约》适用于多重和交叉形式的歧视。[111] 此外，《公约》禁止种族歧视的规定应结合《消除对妇女一切形式歧视公约》(第一条)、《残疾人权利公约》(第二条)和《联合国土著人民权利宣言》(第 2 条)一并适用，这些文书同样禁止或谴责直接和间接形式的歧视。

52. 各国同时还应努力打击与种族和族裔歧视交叉的其他形式的歧视，应认识到，国家义务要求收集和分析分类数据，以便更好地了解遭受多重和交叉形式歧视的群体和个人的人权状况。在新兴数字技术的背景下，这意味着反种族歧视干预措施必须包括对性别、残疾状况和其他受保护类别的有意义的关注。非洲人后裔问题专家工作组最近的一份报告提供了例子说明在该领域十分重要的交叉分析的基本特征。[112]

---

[107] 见 A/HRC/42/59。

[108] 见 https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making。

[109] A/HRC/41/54/Add.2，第 16-19 段。

[110] 见 www.un.org/womenwatch/daw/csw/genrac/report.htm.

[111] 消除种族歧视委员会，第 32 号一般性建议。

[112] 见 A/HRC/42/59。

B. 防止和打击新兴数字技术设计和使用中的种族歧视的义务

53. 《消除一切形式种族歧视国际公约》阐明了一些一般性国家义务，在新兴数字技术的具体背景下也必须履行这些义务。《公约》确立了所有缔约国的法律承诺，即不对个人、群体或机构采取任何种族歧视行为或做法，确保国家和地方的所有公共主管部门和公共机构均遵守此项义务行事。相反，缔约国必须立即以一切适当方法实行消除一切形式种族歧视的政策。[113]《公约》还要求缔约国采取有效措施，对政府及全国性和地方性的政策进行审查，并对任何法律规章足以造成或持续不论存在于何地的种族歧视者，予以修正、废止或宣告无效。[114] 此外，缔约国应于情况需要时在社会、经济、文化及其他方面，采取特别具体措施确保属于各该国的若干种族团体或个人获得充分发展与保护，以期保证此等团体与个人完全并同等享受人权及基本自由。[115]

54. 根据《公约》第 7 条，各国承诺立即采取有效措施尤其在讲授、教育、文化及新闻方面以打击导致种族歧视之偏见。在最近的其他一些报告中，特别报告员阐述了各国打击种族主义和仇外言论和行为，包括在线言论和行为的人权义务。[116] 这些义务同样适用于本报告分析的问题：在新兴数字技术的背景下，各国必须采取有效措施，发现和打击在此类技术的设计和使用过程中对获取公民、政治、经济、社会和文化权利种族歧视。[117]

55. **各国防止和消除新兴数字技术设计和使用中的种族歧视的义务要求解决上文第二部分讨论的各个部门的“多样性危机”。国家必须与私营公司合作，包括基于具有法律约束力的框架，制定必要的特别措施，确保种族和族裔少数群体在与新兴数字技术设计和使用相关决策的所有方面都得到有意义的代表。必须在新兴数字技术的各个阶段都实现真正的权力转移，而不仅仅是让妇女和少数种族和族裔群体来装点门面。权力转移的核心——甚至在私营部门内部——在于更深入地参与研究和知识生产并为其提供资金，这些研究和知识生产专门旨在从跨学科的角度加深对新兴数字技术设计和使用中的歧视的理解**。种族和数字批判性研究中心的研究人员提供了例子。[118]

56. **各国必须采取迅速有效的行动，防止和减少在新兴数字技术使用和设计中产生种族歧视的风险，包括要求若公共主管机关采用基于此类技术的系统，则必须进行种族平等和不歧视人权影响评估。影响评估中必须提供与种族或族裔边缘化群体的代表共同设计和共同实施的切实机会。纯粹基于自愿或甚至主要基于自愿的平等影响评估方针是不够的；强制性方针必不可少**。例如，欧洲委员会最近在这方面的进展[119] 就值得称赞。既不能忽视种族歧视，也不能将种族和族裔少数

[113] 第二条第一款(子)项。

[114] 第二条第一款(寅)项。

[115] 第二条第二款。

[116] 见 A/73/305，A/73/312 和 A/HRC/38/53。

[117] 《公约》第五条明确规定禁止种族歧视适用于获取和享受公民、政治、社会、经济和文化权利。

[118] 见 https://criticalracedigitalstudies.com.

[119] 部长委员会关于算法系统的人权影响的 CM/REC(2020)1 号建议，2020 年 4 月 8 日。

群体排除在决策之外。有时，为防止公共主管部门在设计和使用新兴数字技术时产生种族歧视性结果和其他侵犯人权行为，可能需要彻底禁止使用这些技术，直到其危害风险得到充分缓解。旧金山市禁止地方政府使用面部识别软件的决定就是这方面的一个好榜样。

57. **为了履行平等和不歧视义务，各国必须确保公共部门在使用新兴数字技术方面的透明度和问责制，并允许开展独立分析和监督，包括仅使用可审计的系统。**加拿大最近开展了改革，对使用新兴数字技术实施公共部门问责制，提供了这方面重要第一步的例子。[120]

58. **各国必须确保以具有法律约束力的国际人权原则(包括禁止种族歧视的原则)为基础，制定道德框架和准则，对新兴数字技术进行灵活、实用和有效的监管和治理。《保护机器学习系统中的平等权和不歧视权的多伦多宣言》体现了具有约束力的国际人权法和人工智能治理道德准则或原则之间应该存在的共生关系。**[121]《多伦多宣言》强调国际人权法规定的平等和不歧视的约束性，并为这些法律的实际执行提供了可操作的指导方针。

私营公司、联合国和其他多边机构

59. 虽然国际人权法仅对国家具有直接的法律约束力，但为了履行这方面的法律义务，各国必须确保包括公司在内的私人行为体实施的种族歧视能够获得有效补救。[122] 根据《消除一切形式种族歧视国际公约》，各国必须颁布特别措施，在公共和私人领域实现和保护种族平等。[123] 这应该包括对涉及新兴数字技术的公司进行严密监管。

60. 《工商企业与人权指导原则》阐明，私营公司有尊重人权的责任，包括通过人权尽责调查。人权尽责调查要求：评估实际和可能的人权影响；综合评估结果并采取行动；跟踪有关反映；并通报如何消除影响。[124] 正如将《指导原则》应用于数字技术的“技术中的工商企业与人权项目”(B-Tech 项目)所强调的那样，尽责调查应适用于新产品的概念化、设计和测试阶段——以及作为新产品支持的基础数据集和算法。[125]《多伦多宣言》为机器学习系统的组织人权尽责调查确定了三个核心要素或步骤：(a) 识别可能出现的歧视后果；(b) 防止和减少歧视并跟踪应对措施；以及(c) 就识别、防止和减少歧视的努力保持透明。正如最近的一份报告所强调的那样，预防性人权尽责调查方法必须建立在“多学科团队中，

[120] 见 www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

[121] 见 www.torontodeclaration.org/declaration-text/english.

[122] 例如，见人权事务委员会关于《公约》缔约国的一般法律义务的性质的第 31 号一般性意见(2004 年)，第 8 段。

[123] 消除种族歧视委员会，第 32 号一般性建议，第 23 段。另见经济、社会及文化权利委员会关于经济、社会和文化权利方面不歧视的第 20 号一般性意见(2009 年)，第 11 段；消除种族歧视委员会，关于《公约》第一条第一款(世系)的第 29 号一般性建议(2002 年)，第七条。

[124] A/HRC/17/31，附件，原则 17。

[125] 见 www.ohchr.org/Documents/Issues/Business/B-Tech/B_%20Tech_Project_revised_scoping_final.pdf。

这些团队能够从产品开发开始，在整个生命周期的各个阶段识别人工智能中的盲点，并发现特定环境的系统性偏见”。[126]

61. 各国必须确保涉足新兴数字技术的公司的人权伦理框架与具有约束力的国际人权法义务，包括平等和不歧视义务挂钩，并以此为参照。确实存在这样一种风险，即公司会为了看起来合乎道德的公共关系利益而随意提及人权，甚至有时并没有采取有意义的干预措施来实施人权原则。虽然在公司治理文件中越来越多地提到人权，甚至平等和不歧视，[127] 但仅仅提到这些并不能确保问责制。**同样，实施《工商企业与人权指导原则》框架，包括通过 B-Tech 项目等举措，必须纳入具有法律约束力的义务，禁止种族歧视，并提供有效的补救措施。**

62. 由技术公司颁布基于道德的方针存在一个固有问题，即如果道德承诺没有与工作场所的责任结构直接挂钩，那么它们对软件开发实践几乎没有影响。[128] 从人权角度来看，依靠公司自我监管是一个错误，也是对国家责任的摒弃。推动公司切实保护人权(特别是边缘化群体的人权，他们并不占有商业主导地位)的激励因素可能与利润动机直接对立。当风险很高时，对股东的信托义务往往比考虑那些群体的尊严和人权更重要，因为这些群体并没有让公司负责的办法。此外，即使公司用意良好，也可能在制定和应用道德准则时主要关注技术问题，而不是从更广泛的全社会、基于尊严的人权框架的视角出发。

63. **各国必须借助国际人权法对种族歧视的禁令来确保开展企业人权尽责调查。**如果欧洲委员会提出的公司强制性尽责调查[129] 能够确保切实落实和执行人权，那将是一项有希望的发展实例。

C. 为新兴数字技术设计和使用中的种族歧视提供有效补救的义务

64. 国际人权体系运作的前提是，违反国际人权法的行为导致违反者有义务向违法行为的受害人提供充分和有效救济。[130] 侵犯人权行为的受害者，包括种族歧视性侵权行为的受害人，拥有获得全面补救的相应权利，包括通过司法或政府决定的赔偿。《消除一切形式种族歧视国际公约》第六条在这方面是明确的：缔约国应保证在其管辖范围内，人人均能经由国内主管法庭及其他国家机关对违反本公约侵害其人权及基本自由的任何种族歧视行为，获得有效保护与救济，并有权就因此种歧视而遭受的任何损失向此等法庭请求公允充分的赔偿或补偿。产生这一要求是因为，权利要切实具有意义，就必须规定有效的补救办法来纠正侵权行为。

[126] 见 www.institut-fuer-menschenrechte.de/fileadmin/user_upload/Publikationen/ANALYSE/Analysis_Business_and_Human_Rights_in_the_Data_Economy.pdf。

[127] 见 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482。

[128] 见 https://ainowinstitute.org/AI_Now_2018_Report.pdf。

[129] 见 https://responsiblebusinessconduct.eu/wp/2020/04/30/european-commission-promises-mandatory-due-diligence-legislation-in-2021.

[130] 例如，见《公民及政治权利国际公约》第二条；以及《消除一切形式种族歧视国际公约》，第六条。

65. 在对新兴数字技术设计和使用中的种族歧视进行有效救济方面，各国必须确保有效救济涵盖所有方面，包括诉诸司法、保护免遭侵权行为、保证停止和不再发生侵权行为，同时打击有罪不罚现象。[131]《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》规定了对侵犯人权行为进行救济和赔偿的五大要素：恢复原状、补偿、康复、抵偿和保证不再发生。[132] 这些要素中的每一个都在确保整体和有效的救济方面发挥着不同的作用，这与过渡时期正义的概念密切相关。[133]

66. 恢复原状旨在将受害人恢复到发生严重违反国际人权法行为之前的原有状态。[134] 赔偿包括支付经济上可以估量的损害，包括身心伤害、失却的社会福利、物质损害、精神伤害和产生的费用。[135] 康复包括提供医疗和心理护理以及法律和社会服务。[136] 抵偿是赔偿和救济的一个内容宽泛的要素。在适用的情况下，抵偿可包括终止侵权行为、披露真相、恢复尊严、承担责任、纪念伤害和确保对责任方进行制裁的措施。[137] 最后，保证不再发生是有助于不再发生的赔偿和救济措施。最常见的进行结构改革和加强国家机构，确保文职政府的充分监督和对人权的适当尊重。[138]

67. 各国必须向新兴数字技术设计和使用中的种族歧视的受害人确保恢复原状、赔偿、康复、抵偿和保证不再重犯。各国还应参考寻求真相、正义、赔偿和保证不再发生问题特别报告员关于制定和执行赔偿措施的指导，以及土著人民权利专家机制的指导。[139]

68. 关于救济和赔偿的现有人权框架也是承诺打击新兴数字技术使用和设计中的种族歧视的私营公司的重要资源。在其他情况下，私人行为体在为种族歧视提供赔偿方面发挥了重要作用，包括对其在这种歧视中扮演的角色承担责任。[140] 微软、苹果、亚马逊、谷歌、脸书、腾讯和阿里巴巴等私营公司在向与其技术和产品相关的种族歧视受害人提供恢复原状、赔偿、康复、抵偿和保证不再重犯方面可以发挥重要作用。

---

[131] 消除种族歧视委员会，关于在刑事司法系统的司法和运作中防止种族歧视的第 31 号一般性建议(2005 年)；以及人权事务委员会，第 31 号一般性意见。

[132] 大会第 60/147 号决议，附件，第 18 段。

[133] A/69/518，第 20 段。

[134] 大会第 60/147 号决议，附件，第 19 段。

[135] 同上，第 20 段。

[136] 同上，第 21 段。

[137] 同上，第 22 段。

[138] 同上，第 23 段。

[139] 见 A/HRC/EMRIP/2019/3。

[140] A/74/321，第 62 段。