联合国 A/73/348*

大 会

Distr.: General
29 August 2018
Chinese
Original: English

第七十三届会议

临时议程** 项目 74(b)

促进和保护人权：人权问题，包括
增进人权和基本自由切实享受的各种途径

# 促进和保护意见和表达自由权***

## 秘书长的说明

秘书长谨向大会转递促进和保护意见和表达自由权问题特别报告员大卫·凯伊根据人权理事会第 34/18 号决议提交的报告。在本报告中，特别报告员探讨了人工智能技术在信息环境下对人权的影响，特别侧重于意见和表达自由权、隐私权和不受歧视权。

* 由于技术原因于 2018 年 10 月 30 日重发。

** A/73/150。

*** 本文件逾期提交，以反映最新动态。

18-14238 * X (C) 301018 301018

# 促进和保护意见和表达自由权问题特别报告员的报告

## 目录

# 一. 导言

1. 人工智能对全球信息环境的影响越来越深。公司可通过人工智能对搜索结果、新闻推送和广告投放进行“策展”，从而安排用户能够看到什么以及何时看到。社交媒体公司可利用人工智能技术协助对其平台内容进行节制，因此人工智能通常是防止出现违反自身规则内容的第一道防线。人工智能可为人们推荐朋友或“关注”、新闻、可以去的地方、餐厅、商店、旅馆。人工智能以其速度、效率和规模优势，可帮助信通技术领域中的最大公司管理每日上传至其平台的大量内容。人工智能技术可提升全球信息和观点分享的范围和速度，从而带来巨大机会，推动表达自由和信息获取。同时，人工智能的不透明性也有可能干扰个人的自我决断，即本报告所称的“个人自主和能动性”。[1] 所有人权和法治工作者都面临一个巨大的全球性挑战：国家、公司和民间社会如何确保人工智能技术加强和尊重而不是破坏和危害人权？

2. 本报告并非妄自就人工智能与人权问题作出最后结论。相反，本报告试图达到三个目的：定义人工智能背景下对讨论人权问题必不可少的关键术语；确认与人工智能相关的人权法律框架；提出初步建议，确保在人工智能技术发展的过程中考虑人权因素。本报告可作为本人近日提交人权理事会报告(A/HRC/38/35)的参考，其中介绍了基于人权的在线内容节制方法。[2]

# 二. 认识人工智能

## A. 什么是人工智能？

3. 人工智能经常用来指与自动化计算机决策相关的独立性增强、效率提高和规模扩大。人工智能并非仅由单一元素组成，它是由一系列程序和技术构成的，旨在让计算机辅助或替代人工完成具体任务(如决策和解决问题)。[3] 人工智能一词可能词不达意，因为从字面上可以理解为机器能够运用人类智力形成的概念和规则来运转。但其实并非如此。一般来说，人工智能通过迭代重复和尝试，来优化由人类安排的运算任务。但是，该词现已成为文化中的一部分，被公司和政府广泛使用，所以本文也采纳了这个说法。

---

[1] 见 Mariarosaria Taddeo 和 Luciano Floridi, “How AI can be a force for good”, *Science*, vol. 361, No. 6404 (2018 年 8 月 24 日)，可访问 http://science.sciencemag.org/content/361/6404/751.full。

[2] 本报告得益于 2018 年 6 月欧洲联盟资助下在日内瓦举行的一次专家咨询，以及 2017 年和 2018 年撰写 A/HRC/35/38 号文件时专家提出的观点。特别报告员特别鸣谢卡利·奈斯特和阿莫斯·托阿为本报告的基础研究和编写作出的贡献。

[3] 见 AI Now, “The AI now report: the social and economic implications of artificial intelligence technologies in the near term”, 2016. 可访问 https://ainowinstitute.org/AI_Now_2016_Report.pdf；United Kingdom of Great Britain and Northern Ireland House of Lords Select Committee on Artificial Intelligence, “AI in the United Kingdom: ready, willing and able?”，2018 年，第 13 页。

4. 大众文化经常认为，社会正在向广义人工智能方向发展，即计算机将在多领域接近或超过人类智力(“科技奇点”)，但这种能力仍然很遥远。[4] 在可预见的将来，狭义人工智能还将继续扩展。狭义人工智能是指计算机系统根据人类开发的算法执行具体领域中的程序。狭义人工智能包括手机中的语音功能、客服中的聊天机器人、在线翻译工具和自动驾驶汽车、搜索引擎结果和地图服务。机器学习是一种狭义人工智能技术，用于训练算法，使其能够利用数据集来识别和帮助解决问题。例如，基于人工智能的智能家居设备通过不断“学习”收集到的日常语言和语音模式数据，来更准确地处理和回应其用户提出的问题。任何时候，在人工智能系统的设计和传播、人工智能应用程序目标的定义，以及根据应用程序类型选择标记数据集、进行产出分类方面，人类都发挥关键的作用。人类总是决定着人工智能的用途，包括人工智能辅助或替代人类决策的尺度。


5. 人工智能的基础是算法，即人类设计和编写的计算机代码，其中包含的指令可将数据转换成结论、信息或其他产出。长期以来，算法对日常通信和基础设施系统的运行至关重要。现代生活中的大量数据和数据分析能力成就了人工智能。这当然也是私营部门对数据的看法。即可用于算法的数据越多，数据质量越好，算法就越强大和精确。算法系统可以快速分析大量数据，使人工智能程序能够执行以前没有计算工具时人类专属的决策功能。

[4] Article 19 and Privacy International, “Privacy and freedom of expression in an age of artificial intelligence”, London，2018 年，第 8 页。


人的能动性是人工智能的必要组成部分，但鉴于人工智能的鲜明特点，至少需要以人权为视角对三个方面进行审查：自动化、数据分析和适应性。[5]

6. **自动化**。自动化后，可以通过计算工具完成特定任务，消除决策过程中部分流程的人为干预。如果设计时就考虑到避免人为偏见，从人权角度来看，自动化有正面意义。例如，自动入境系统可根据犯罪记录或签证状态等客观特征对入境者进行标记以待审查，从而降低依赖相貌、种族、年龄或宗教方面的主观(且易产生偏见的)评估。自动化还可以人类无法达到的速度和规模处理大量数据，从而服务于公共安全、健康和国家安全。但是，自动化系统可能依赖在设计或执行阶段就存在偏见的数据集，进而会导致歧视性后果。比如，上面提到的犯罪历史或签证数据本身可能包含偏见。过度依赖自动化决策，加上对自动化决策的信心，以及没有认识到存在这一根本问题，可能反过来会影响对人工智能成果的审查，妨碍当事人就人工智能作出的不利决定获得补救。自动化可能会影响过程的透明度和可审查性，妨碍甚至是善意的权威对人工智能决策结果作出解释。[6]

7. **数据分析**。大多数人工智能应用程序需要数量庞大的数据集支持。支持人工智能系统的数据集有很多种，包括从网络浏览习惯到高速公路交通流量等各种数

[5] Council of Europe, Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe study, No. DGI (2017) 12, 2018 年。可访问 https://www.coe.int/en/web/freedom-expression/-/algorithms-and-human-rights-a-new-study-has-been-published，第 5 页.

[6] Council of Europe, Algorithms and Human Rights，第 8 页。

据。有些数据集包含个人数据，也有许多属于匿名数据。人工智能对此类数据集的使用可能引发严重关切，涉及数据集的来源、准确性和个人对数据集的权利问题；人工智能系统对匿名数据进行去匿名化处理的能力问题；以及训练数据的人工标记可能给数据集植入或灌输的偏见问题。人工智能的数据评估可能判断出相关性，但不一定能判断出因果关系，因此可能导致难以审查的带有偏见和瑕疵的结果。

8. **适应性**。基于机器学习的人工智能系统具有适应性，因为背后的算法能够逐渐发现新问题，找到新的解决办法。根据监督程度的不同，系统可能发现和归纳编写程序或发出指令的人没有预见的规律。这种不可预测性真正有望使人工智能成为一种革命性技术，但它也揭示了人工智能的风险：由于人类逐渐被排除在人工智能系统目标和产出的定义过程，确保透明度和问责以及有效补救变得愈加困难，预见和减轻人权影响也同样变得愈加困难。

## B. 人工智能与信息环境

9. 人工智能对信息环境可产生特别重要的影响，有时候是负面影响。信息环境即由技术、平台和公私部门组成的复杂生态系统，目的是促进信息的数字化获取和传播。无论是在互联网的某个角落，还是在数字设备和技术系统、搜索引擎、社交媒体平台、短信应用程序和公共信息平台，各种算法和人工智能应用程序随处可见。为了与-职责重点保持一致，特别报告员现提出以下三种信息环境背景下引起关切的人工智能应用：

10. **内容显示和个性化**。社交媒体平台和搜索平台日益主导人们访问和分享信息、想法的方式以及新闻传播的方式。算法和人工智能应用程序决定着内容分享的广度、时间和受众。庞大的数据集整合了浏览历史、用户的人口统计信息、语义和情感分析以及许多其他因素，用于构建日益个性化的算法模型，以便对信息进行排序和策展，即确定该信息是展示给用户还是默默排除。付费内容、赞助内容或有主题标签的内容会予以优先显示，其他内容则被排除或排在后面。社交媒体的新闻推送根据特定用户兴趣的主观评估展示内容。所以，用户可能看不到发布到所属平台的某些批评性社会或政治报道和内容。[7] 人工智能正在以一种不透明的方式左右着信息的世界。对用户不透明，甚至经常对策展信息的平台也不透明。

11. 在线搜索是人工智能内容显示和个性化最普遍的形式之一。搜索引擎借助拥有大量个人用户和整体用户数据的人工智能系统，为搜索请求提供结果(并对请求予以补全和预测)。由于排名靠后的内容或被完全排除在搜索结果范围以外的内容不可能会被看到，基于人工智能的搜索应用程序在很大程度上影响着知识的

[7] World Wide Web Foundation, “The invisible curation of content: Facebook’s News Feed and our information diets”, 2018 年 4 月 22 日. 可访问 https://webfoundation.org/research/the-invisible-curation-of-content-facebooks-news-feed-and-our-information-diets/。

传播。[8] 同样，内容集成商和新闻网站[9] 并非根据事件发生远近或重要性向用户显示内容，而是根据人工智能应用程序基于大量数据集对用户兴趣和阅读规律的预测来显示内容。因此，人工智能在很大程度上影响着用户消费什么信息，甚至是消费前对信息的知情权。

12. 内容显示领域的人工智能正在朝着更加个性化的浏览体验方向发展。在信息供应充足的时代，[10] 个性化旨在建立一个有序的互联网，帮助用户找到所需要的信息。个性化的优点可能包括能够以更多语言获取信息和服务，[11] 或者可以及时地获取与个人经历或偏好更匹配的信息。人工智能背景下的个性化还可能导致难以接触不同观点，从而影响人们寻求和分享来自不同意识形态、政治或社会阶层观点和意见的能动性。这样的个性化可能会强化偏见，激励推广和推荐煽动性内容或虚假信息的行为，以维持用户的在线热度。[12] 诚然，各种社会和文化环境都可能会限制个人接触信息的广度。但是，人工智能协助下的个性化通过优化内容扩大热度和传播规模，可能会妨碍个人查找某些类型的内容。这个问题尤其表现在热度低的内容通常被各种算法降低优先级、进而导致用户生成的独立内容被打入冷宫。[13] 精明的行为主体可以利用基于规则的、为吸引参与而优化的人工智能系统，达到提高曝光率的目的；他们可以通过盗用人气标签或使用机器人，来大规模提高访问量，但这种做法损害了信息多样性。

13. **内容监控和删除**。人工智能可帮助社交媒体公司根据平台标准和规则监控内容，包括垃圾邮件检测、哈希值匹配(比如，使用数字指纹识别恐怖主义或儿童剥削的内容)、关键字过滤、自然语言处理(监控内容是否包含违禁关键字或图像)和其他监测。人工智能可用于对违反服务条款的用户帐户进行警告、暂停或停用，也可用于屏蔽或过滤含有违禁数据或内容的网站。社交媒体公司通过使用人工智

[8] Council of Europe, Algorithms and Human Rights，第 17 页。

[9] 比如，见“How Reuters’s revolutionary AI system gathers global news,” MIT Technology Review, 2017 年 11 月 27 日。可访问 https://www.technologyreview.com/s/609558/how-reuterss-revolutionary-ai-system-gathers-global-news/; Paul Armstrong and Yue Wang, “China’s $11 billion news aggregator Jinri Toutiao is no fake,” *Forbes*, 2017 年 5 月 26 日。可访问 https://www.forbes.com/sites/ywang/2017/05/26/jinri-toutiao-how-chinas-11-billion-news-aggregator-is-no-fake/#1d8b97804d8a .

[10] Carly Nyst and Nick Monaco, State-Sponsored Trolling: How Governments are Deploying Disinformation as Part of Broader Digital Harassment Campaigns (Palo Alto, Institute for the Future, 2018)，第 8 页。

[11] World Wide Web Foundation, “Artificial intelligence: the road ahead in low- and middle-income countries”, Washington, D.C., 2017.

[12] Zeynep Tufekci, “YouTube, the great radicaliser”, *New York Times*, 2018 年 3 月 10 日。可访问 https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html; James Williams, *Stand Out of our Light: Freedom and Resistance in the Attention Economy* (Cambridge, Cambridge University Press, 2018).

[13] 最近，一些技术平台有意从“热度”驱动的个性化转向重视用户在线体验的个性化；见 Julia Carrie Wong, “Facebook overhauls News Feed in favour of ‘meaningful social interactions’”, *The Guardian*, 2018 年 1 月 11 日, available at 可访问 https://www.theguardian.com/technology/2018/jan/11/facebook-news-feed-algorithm-overhaul-mark-zuckerberg. 但是，在人工智能系统如何作出和实施这些评估方面，如果缺乏充分透明度、报告和测量方法，很难评价这种转变是否会对互联网用户体验产生可看得见的效果。

能，过滤违反本公司制订的各种规则的内容(如裸体、骚扰、仇恨言论等)，但这些公司在具体案例中不进行人工干预而完全依赖自动化的程度尚不得而知。[14]

14. 加强人工智能作用的呼声和压力既来自私营，也来自公共部门。公司声称，网上非法、不恰当以及有害内容的数量远远超过了人类能够监控的能力，并认为人工智能是协助人类更好应对这一挑战的工具。根据一些平台的说法，人工智能不仅能够(根据其规则)更有效地识别不恰当和非法内容以供删除(通常由人类管理员进行)，而且比人类决策的准确性更高。与此同时，各国正大力推动使用高效、快速的自动化监控手段，应对各类挑战，包括监控儿童性虐待和恐怖主义内容(这些领域人工智能已经广泛应用)、版权侵权以及删除“极端主义”和仇恨内容。[15] 欧盟委员会 2018 年 3 月“关于进一步改善打击在线非法内容效果的措施的建议”要求互联网平台使用自动过滤器，检测和删除恐怖主义内容，并在某些情况下进行人工审查，以应对自动系统难免造成的错误。[16]

15. 实现内容监控自动化可能会以人权为代价（见 A/HRC/38/35，第 56 段）。人工智能驱动的内容监控有几个局限，包括难以通过语境评估，来思考语言的暗示信息、意义的广泛差异以及语言文化的特殊性。由于人工智能应用程序通常以包含歧视性假设的数据集为基础，[17] 而且当过度监控的成本较低时，这类系统很可能默认删除或暂停没有问题的在线内容或账户，而且可能根据偏见或歧视性的概念而删除内容。因此，弱势群体最有可能因为人工智能内容监控系统而处于不利地位。例如，Instagram 的 DeepText 将“墨西哥人”标识为一种蔑称，因为它

[14] Instagram 的一个工具 Deep Text，可用来判断内容的“毒性”，并允许用户对自己的用词和表情过滤器进行自定义，还可以通过评估用户之间的关系，以进一步确定情景(如了解一条评论是否只是朋友间的玩笑). Andrew Hutchison, “Instagram’s rolling out new tools to remove ‘toxic comments’", Social Media Today, 2017 年 6 月 30 日。可访问 https://www.socialmediatoday.com/social-networks/instagrams-rolling-out-new-tools-remove-toxic-comments。

[15] 据报告，英国开发了一种工具，可在用户上传时自动侦测和删除恐怖主义内容。内政部，“新技术助力打击网上恐怖主义内容”，2018 年 2 月 13 日。见欧盟委员会，《建议欧洲议会和理事会通过数字单一市场版权指令的提案》，COM(2016)593 final，第 13 条；特别报告员给欧盟委员会主席的信，OL OTH 41/2018(2018 年 6 月 13 日)可访问 https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-OTH-41-2018.pdf。

[16] “2018 年 3 月 1 日委员会关于有效打击网上非法内容的措施的建议”(C(2018) 1177 final)，可访问 https://ec.europa.eu/digital-single-market/en/news/commission-recommendation-measures-effectively-tackle-illegal-content-online.另见 Daphne Keller，“对‘2018 年 3 月欧盟委员会关于进一步提高打击网上非法内容有效性的措施的建议’的回应”，斯坦福大学互联网与社会中心，2018 年 3 月 28 日，可访问 http://cyberlaw.stanford.edu/publications/comment-response-european-commissions-march-2018-recommendation-measures-further。

[17] See Aylin Caliskan, Joanna Bryson and Arvind Narayanan, “Semantics derived automatically from language corporate contain human-like biases,” 356 Science 6334, 14 April 2017; Solon Barocas, Andrew D. Selbst, Big Data’s Disparate Impact, 104 Cal. L. Rev. 671 (2016).

的数据集中收录的数据将“墨西哥人”与“非法”相关联，后者是一个包含在算法中的负面编码术语。[18]

16. 人工智能使人很难仔细审查内容操作背后的逻辑。即使算法的内容监控有人工审核补充(大型社交媒体平台认为这种安排因运营规模太大越来越不可行)，人们也倾向于听从机器做出的决定(基于上述客观性假设)，从而妨碍对内容监控结果的质疑，特别是当系统的技术设计没有做到算法透明时。

17. **用户画像、广告和目标定位**。人工智能的进步既受益于互联网数据驱动的商业模式，也反过来促进了这种模式，即个人通过付出个人数据来免费获取内容和服务。公司凭借多年在线监测和“用户画像”积累了大量数据资源，能够为人工智能系统提供丰富的数据集，以开发更精确的预测和目标定位模型。今天，私营和公共部门的广告投放可以精确到个人层面，消费者和选民成为“微观定位”目标，以此回应和利用个人特征。

18. 人工智能驱动的目标定位刺激了对个人数据的广泛收集和利用，增加了通过传播虚假信息操控个人用户的风险。目标群体定位的功能可能会固化歧视，阻止用户得到信息或机会，例如，允许招聘和住房广告屏蔽老年工人、妇女或少数族裔[19]。再如，社交媒体平台通过微观定位，使个人看不到政治信息方面的平等性和多样性，从而创造出一种被策展出来的敌视多元政治话语的世界观。

# 三. 为人工智能制定人权法律框架

## A. 人工智能背景下的人权义务范围

19. 与所有技术一样，人工智能的设计、开发和部署必须符合国际人权法规定的缔约国的义务和私人行为者的责任。人权法规定各国既承担不执行干涉意见和表达自由措施的消极义务，也承担促进发表意见和表达自由并保护行使权利的积极义务。

20. 关于私营部门，各国有义务保证其尊重个人权利，[20] 特别是发表意见和表达自由权，包括保护个人免遭私人行为主体的侵权行为(《公民权利和政治权利国际公约》第二条第一款)。国家履行这一义务可以采取多种方法，包括采取法律手

[18] Nicholas Thompson, “Instagram’s Kevin Systrom Wants to Clean Up The &#%@!Internet,” Wired, 14 August 2017, 可访问 https://www.wired.com/2017/08/instagram-kevin-systrom-wants-to-clean-up-the-internet/。

[19] Julia Angwin, Noam Scheiber, Ariana Tobin, “Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads,” ProPublica and the New York Times, 20 December 2017, 可访问 https://www.propublica.org/article/facebook-ads-age-discrimination-targeting; Julia Angwin, Ariana Tobin, and Madeleine Varner, “Facebook (Still) Letting Housing Advertisers Exclude Users by Race,” ProPublica, 21 November 2017, https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

[20] “联合国工商业与人权指导原则”，A/HRC/17/31, 原则 3，见 https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf; A/HRC/38/35, 第 6-8 段。

段限制或影响人工智能应用的开发和实施，采取政策手段管理公共部门从私营公司采购人工智能应用，采用自律和共同监管计划，以及加强私营公司能力建设，使其在公司活动中承认和重视发表意见和表达自由权。

21. 公司也有责任根据人权法指导其人工智能技术的建设、采纳和运用（见A/HRC/38/35,第 10 段）。《联合国工商业与人权指导原则》为包括社交媒体和搜索服务公司在内的所有公司提供了“全球预期行为标准，无论其在何处运营”（原则11）。为了使报告中的结论适用于人工智能领域(同上，第 11 段)，《指导原则》要求各公司至少做出高层政策承诺，在所有人工智能应用程序中尊重用户的人权(原则 16)；避免通过使用人工智能技术造成或助长不利的人权影响，并防止和减轻与其业务相关的任何不利影响(原则 13)；对人工智能系统进行尽职调查，以识别和关注实际和潜在的人权活动(原则 17-19)；开展预防和缓解战略(原则 23)；不断审查与人工智能有关的活动，包括通过让利益攸关方和公众参与审查(原则 20-21)，并提供便于落实的补救措施，补救人工智能系统对人权的不利影响(原则 22、29 和 31)。

## B. 意见自由权

22. 持有意见而不受干涉的自由是一项绝对权利，载于《公约》第十九条第一款和《世界人权宣言》第十九条。这项自由权利“不允许任何例外或限制”，无论是否经由“法律或其他权力”。[21] 特别报告员在 2015 年提交人权理事会的关于数字通信中的加密和匿名的报告中指出(A/HRC/29/32)，在数字时代，信息的存储、传输和保护方式对行使持有意见权产生了独特的影响。搜索查询、浏览活动、电子邮件和文本通信，以及保存在云中的文档和记忆，这些数字活动和记录共同构成了用户观点的基础(同上，第 12 段)。国家和非国家行为体都可能干涉这些形成和持有观点的机制和过程。

23. 持有意见权的一个基本要素是“形成并通过推理展开意见的权利”。[22] 人权事务委员会的结论是，这项权利要求个人在形成信仰、意识形态、反应和立场时不受不当胁迫。[23] 因此，强迫性精神干预、思想灌输活动(如“再教育”集中营)或暴力威胁，这些旨在迫使个人形成特定观点或改变其观点的行为，都违反了第十九条第一款。委员会发现，胁迫性“诱导优惠待遇”可能达到某种说服程度，以至构成对形成和持有意见权的干预(见 CCPR/C/78/D/878/1999)。

24. 技术和内容策展的结合，引发了哪种类型的强制或诱导可以被视为干涉形成意见权的新问题。内容策展一直以来都在影响个人形成意见的能力：例如媒体将

[21] 人权事务委员会，“关于意见和言论自由的第 34 号一般性意见(2011 年)”，[9] 可访问 http://www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf; Manfred Nowak，“联合国公民权利和政治权利公约：公民及政治权利国际公约评论第 441 页 (1993)”。

[22] Nowak，《联合国公民权利和政治权利公约》。

[23] Yong-Joo Kang 诉大韩民国, CCPR/C/78/D/878/1999, 联合国人权事务委员会(HRC), 2003 年 7 月 16 日。

特定故事提升到头版，意图塑造和影响个人对当天重大新闻的认知。商业广告也设法诱导受众对特定产品和服务的好感，培养人们对这些产品和服务的欲望。

25. 人工智能的使用扩展和增强了互联网内容策展的传统，为用户提供了更加复杂和高效的内容个性化及内容策展手段，其规模超出了传统媒体的覆盖范围。由于某些人工智能辅助的内容策展模式已占据主导地位，人们担忧其可能影响个人形成和发展意见的能力。例如，绝大部分网上搜索都使用少数科技公司的引擎。公司对搜索市场的垄断使用户很难选择不使用特定算法对搜索结果的排列和策展，还可能诱使用户相信(正如公司的意图那样)所生成的结果是关于某一特定主题的最相关或最客观的信息。在如何通过人工智能制定和执行搜索标准方面缺乏透明度，也可能强化如下假设，即在特定平台上生成的搜索结果是事实的客观体现。

26. 因此，市场主导地位在人工智能辅助策展领域产生的问题，考验着人们对内容策展如何影响(或不影响)形成意见能力的历史判断。因为都是新问题，再加上普遍缺乏干预意见权的判例，人们对于当代数字环境中人工智能辅助内容策展如何影响人权，存有更多的疑问而不是答案。尽管如此，这些问题应该推动权利研究，探讨人工智能辅助内容策展的社会、经济和政治影响。公司至少应该提供有意义的信息，说明他们如何制定和实施平台内容的策展和个性化标准，包括在相关人工智能系统的设计和开发中发现社会、文化或政治偏见的政策和流程。

## C. 表达自由权

27. 《公约》第十九条第二款保障“寻求、接受和传递各种信息和思想”的广泛权利，无论国界或媒体类型，这种权利必须受到保护和尊重。表达自由权的享有与行使其他权利密切相关，是民主制度有效运作的基础。因此，该权利隐含的义务包括促进媒体多样性和独立性，以及保护获取信息的机会。[24]

28. 与形成和持有意见的权利不同，表达思想和获取信息的权利在有限的情况下可以受到限制(《公约》第十九条第三款)。限制必须符合合法性标准，即限制须由符合清晰和准确标准的法律公开规定，并由独立的司法当局加以解释；必须符合必要性和相称性标准，即限制是实现当前合法利益所必需的最不具侵扰性的措施，并且不危害权利的实质；必须符合正当性标准，即限制必须是为了实现明确列出的合法利益，即保护他人的权利或声誉、国家安全或公共秩序、或公共卫生或道德(见 A/HRC/38/35 第 7 段)。在这一框架内，表达权利也可以根据《公约》第二十条第二款受到限制——该款要求各国禁止“鼓吹民族、种族或宗教仇恨以

[24] 促进和保护意见和表达自由权问题特别报告员、欧洲安全与合作组织媒体自由问题代表、美洲国家组织表达自由问题特别报告员和非洲人权和民族权委员会表达自由和获取信息问题特别报告员，“关于表达自由和‘假新闻’、假信息和宣传的联合宣言”，2017 年 3 月 3 日。可访问 https://www.osce.org/fom/302796。另见人权委员会，关于意见和表达自由的一般评论 34 号(2011)；A/HRC/29/32, 第 61 段 和 A/HRC/32/38, 第 86 段。

致构成煽动歧视、敌视或暴力”——但限制仍然必须满足合法性、必要性和正当性的累积条件[25]。

29. 内容监控中的决策本身就十分复杂，引入自动化流程后可能更加复杂。与人类不同，目前的算法无法评估文化背景、察觉讽刺或进行必要的批判性分析以准确识别“极端主义”内容或仇恨表达，[26] 因此更有可能默认屏蔽和限制内容，这就有可能损害个人用户的表达权以及他们不受限制或审查地获取信息的权利。

30. 在人工智能管理的系统中，信息和思想的传播由不透明的力量控制，系统的优先考虑可能不是为媒体的多样性和独立声音创造有利环境。与此相关的是，人权事务委员会认为，各国应该“采取适当行动……防止私人控制的媒体集团在垄断情况下过度主导或集中媒体，从而损害信息来源和观点的多样性”。[27]

31. 用户也很难了解人工智能平台和网站使用的游戏规则。用户不清楚在线人工智能和算法应用程序的使用范围和程度，从而无法理解信息何时、以何种标准被传播、限制或锁定为目标。为解决这个问题，一些搜索引擎作出了一点让步，例如把被赞助的搜索结果选择性标记出来，[28] 也有社交媒体平台标出政治人物的付费广告，这种做法可能会稍有助于用户理解信息环境的规则，但是这些方法既无法反映也无法解决算法对信息环境影响的规模。

32 即使告知用户人工智能系统的存在、范围和运行，人工智能系统本身也可能阻碍增强透明度和适用性方面的努力。目前人们尚未能研发出成熟、可扩展的方法，来审查在线自动决策的技术架构并使之透明化。[29] 这意味着，个人表达权利往往会受到负面影响，而且难以查明或理解受影响的原因、方式或依据。

## D. 隐私权

33. 隐私权往往是实现意见和表达自由的必由之路。[30] 《公约》第十七条保护个人的“私生活、家庭、住宅和通信均不得受到任意或非法干涉”，且“其荣誉和名誉也不得受到非法攻击”，还规定“对于此种侵扰或破坏，人人有受法律保护之权利”。人权高专办和人权理事会强调，任何干涉隐私的行为都必须符合合法性、必要性和相称性的标准。(A/HRC/27/37 第 23 段，人权理事会第 34/7 号决议第 2 段)。

34. 人工智能决策系统依赖于数据的收集和利用，包括环境数据、非个人数据、可识别个人信息。输入人工智能系统的绝大部分数据介于两者之间，即从个人数

[25] 人权委员会，关于意见和表达自由的一般评论 34 号(2011)第 50 段。

[26] “欧洲委员会，算法和人权”，第 21 页。

[27] 人权委员会，关于意见和表达自由的一般评论 34 号(2011)第 40 段。

[28] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York, New York University Press, 2018)。

[29] Mike Ananny and Kate Crawford, “Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability”, *New Media and Society*, vol. 20, No. 3 (13 December 2016)。可访问 http://journals.sagepub.com/doi/abs/10.1177/1461444816676645?journalCode=nmsa。

[30] 见 A/HRC/29/32 第 16 段，大会第 68/167 号决议和人权理事会第 20/8 号决议。

据中推测或提取的数据，或者匿名化的个人数据(通常匿名化不够彻底)。公司会使用来自在线用户画像和数字指纹中生成的数据，或从第三方(如数据代理商)购买的数据集，或从大量聚合数据集中生成新数据，来投入人工智能系统。人工智能消费产品和自动化系统经常配有传感器，可以生成和收集附近大量的个人数据，[31] 社交媒体平台可以使用人工智能方法推断和生成用户未提供或确认的敏感信息，如性取向、家庭关系、宗教信仰、健康状况或政治派别。

35. 人工智能挑战了传统的同意、目的、使用限制、透明度和问责等概念，而这些概念恰恰是国际数据保护标准的基石。[32] 由于人工智能系统通过利用现有数据集和创建新数据集来工作，所以在人工智能环境中，个人了解、理解和控制数据使用方式的能力失去了实际意义。一旦人工智能系统将数据用于其他目的，数据就失去了原始背景信息，增加了个人数据不准确或过期的风险，还剥夺了用户纠正或删除数据的能力。现在这类数据正被人工智能系统用于做出重要决策，有些决策会对人们的生活产生深远影响，[33] 然而个人几乎没有渠道控制由个人信息中衍生的数据，即使是匿名化技术也仍然存在不足。

## E. 不歧视的义务

36. 不歧视是人权法的一项固有原则，不仅为缔约国确保公民不受歧视地享受所有其他人权的义务增设了条件，而且正如《公约》第二十六条所载，还是法律面前人人平等和法律平等保护的独立保障。各国有明确义务“禁止任何歧视并保证所有的人得到平等的和有效的保护，以免受基于种族、肤色、性别、语言、宗教、政治或其他意见、国籍或社会出身、财产，出生或其他身分等任何理由的歧视”。因此，第十七条和第十九条包括个人在持有和形成意见、表达和获取观点和信息方面不受歧视的权利，以及行使隐私权和保护个人数据方面不受歧视的权利。

37. 人工智能嵌入和固化的偏见和歧视有可能导致行使意见和表达自由权时受到歧视。内容监控算法可能考虑不到文化、语言或性别背景和敏感性，或者内容涉及的公共利益。[34]由人工智能驱动的新闻推送可能会固化和强化歧视态度，而

[31] Article 19 and Privacy International, “Privacy and freedom of expression”.

[32] 人权事务委员会，第 16 号一般性意见(1988 年)：第十七条(隐私权)，第 10 段。

[33] Article 19 and Privacy International, “Privacy and freedom of expression”，第 19 页。

[34] 例如，这导致了删除具有特定文化意义的历史照片：见 Julia Carrie Wong, “Mark Zuckerberg accused of abusing power after Facebook deletes ‘napalm girl’ post”, *The Guardian*, 9 September 2016。可访问 https://www.theguardian.com/technology/2016/sep/08/facebook-mark-zuckerberg-napalm-girl-photo-vietnam-war; see also A/HRC/38/35，第 29 段。

人工智能的用户画像和广告系统则明显助长了基于种族、宗教和性别的歧视。[35] “自动完成”这一人工智能功能也产生了种族歧视的结果。[36]

38. 有多种因素导致了人工智能系统的内置偏见，增加了歧视的可能性，包括人工智能系统的设计方式、训练人工智能系统的数据来源和范围、开发人员内置在数据集中的社会和文化偏见、人工智能模型本身，以及人工智能模型结果在现实中的实施方式。例如，面部识别应用程序的输入数据主要来自白人男性，而对于女性和肤色较深人群，错误率高达 20%。[37] 例如，当这种系统用于搜索引擎的图像结果分类时，系统潜在的歧视性就会转变为具体的干预，影响到个人行使搜索、接收和传递信息以及自由集会或结社的权利。

## F. 获得有效救济的权利

39. 人权法保障个人有权获得合格的司法、行政或立法当局给予的救济(《公约》第二条第三款)。救济方式必须为权利受到侵犯者所知晓并能方便利用；必须包括对所称侵权行为进行快速、彻底和公正的调查；[38] 并且必须能够制止正在进行的侵权行为 (A/HRC/27/37，第 39-41 段)。

40. 人工智能系统经常干扰救济权。首先，这些系统几乎生来就不具备个人通知功能。在信息环境下几乎所有人工智能技术的应用中，个人甚至都不知道存在影响其意见和表达权的算法决策过程以及决策的范围和程度。第二，也是更具挑战性的是人工智能系统本身的可审查性。算法决策背后的逻辑对于训练有素的系统底层专家也可能并非显而易见。虽然逻辑上可以假设人工智能系统透明度越高，就能进行更严格的审查，但算法透明并不一定等同于决策过程容易解释清楚。算法可能会掩盖已经作出的重要决定，或因其过于复杂和依赖具体情况而难以解释。另外因为在信息环境中，公司经常更新算法，使得情况更加复杂；[39] 同样，机器学习应用程序也可能随着时间的推移改变自己的规则和算法。

41. 使这些问题雪上加霜的是，救济制度本身有自动化的趋势，在这种趋势下，无论是关于内容监控的决定，还是关于人工智能技术对人权的不利影响，个人用

[35] Julia Angwin, Madeleine Varner and Ariana Tobin, “Facebook enabled advertisers to reach ‘Jew haters’”, ProPublica, 14 September 2017。可访问 https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters; Ariana Tobin, “Why we had to buy racist, sexist, xenophobic, ableist and otherwise awful Facebook ads,” ProPublica, 27 November 2017。可访问 https://www.propublica.org/article/why-we-had-to-buy-racist-sexist-xenophobic-ableist-and-otherwise-awful-facebook-ads。

[36] Paris Martineau, “YouTube’s search suggests racist autocompletes”, The Outline, 13 May 2018。可访问 https://theoutline.com/post/4536/youtube-s-search-autofill-suggests-racist-results?zd=1&zi=3ygzt6hw。

[37] Joy Buolamwini, “The dangers of supremely white data and the coded gaze”, presented at Wikimania 2018, Cape Town。可访问 https://www.youtube.com/watch?v=ZSJXKoD6mA8&feature=youtu.be。

[38] 联合国人权事务委员会，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 15 段。

[39] Barry Schwartz, “Google: we make thousands of updates to search algorithms each year”, Search Engine Roundtable, 5 June 2015。可访问 https://www.seroundtable.com/google-updates-thousands-20403.html。

户的申诉都会由人工智能技术考虑和确定。[40] 鉴于自动回复程序缺乏自由裁量、背景分析和程序内置的独立判断，人们担心申诉救济机制能否实现有效救济。[41]

## G. 人工智能的立法、监管和政策应对措施

42. 许多国家正在制定国家人工智能战略，以探索和制定能最大程度提高人工智能对其公民潜在利益的政策和举措。[42]虽然尚未有国家提出全面的人工智能法律或法规，但有理由谨慎对待相关立法，因为法律法规可能不适合此类创新领域的情况，还可能为了弥补细节的缺失，制定得过于严格或过于宽松。行业监管或许更合适；当然现有的法律法规，例如数据保护方面的法律法规，可能已经相当灵活和有效，不再需要进一步立法。

43. 同时，各国应确保人工智能的发展符合人权标准。任何制定人工智能领域国家政策或法规的工作都应确保考虑人权问题。[43] 尤其是意见和表达自由权，往往被排斥在涉及人工智能的公共讨论和政治辩论之外，因为在讨论人工智能中的人权问题时，往往侧重于服务提供中的偏见和歧视。

44. 由于有效人工智能系统的开发取决于大型数据集的获取以及对技术能力的长期投资，私营部门实体可能在开发、生产和能力方面占据主导地位，导致公共部门越来越依赖公司来访问人工智能系统。公司利益和政府利益在实际上和公众眼中都有可能变得日益交织。在信息环境中尤其如此，在这种环境中，政府往往是社交媒体平台、搜索引擎和其他技术的用户，而不是提供商。公共利益和私人利益的结合本身并不造成人权干涉，但会引起对透明度和问责的关注。随着人工智能在私营部门的发展，一个非常现实的风险是，国家将把越来越复杂和繁重的审查和监督任务委托给公司。

45. 任何国家如果希望明确制定人工智能领域的法律或政策，都应该同时面向公共和私营部门的人工智能应用，而不应仅关注公共部门的人工智能监管。正如欧洲委员会得出的结论，“与算法治理和/或监管相关的问题属于公共政策特权，不应只留给私营部门解决。”[44] 国家可以采取措施，增强公司透明度和披露责任，通过数据保护的强力立法，解决人工智能领域的相关问题。

46. 公共和私营部门正在制定日益增多的倡议，探索如何在智能系统的采购、设计、部署和实施中加入伦理规范。本人强烈鼓励在这些工作中加入人权关切。私

[40] Council of Europe, Algorithms and Human Rights，第 24 页。

[41] Pei Zhang, Sophie Stalla-Bourdillon and Lester Gilbert, “A content-linking-context model for ‘notice-and-take-down’ procedures”, *WebSci ‘16*, May 2016。可访问 http://takedownproject.org/wp-content/uploads/2016/04/ContentLinkingModelZhangStallaGilbert.pdf。

[42] Tim Dutton, “An overview of national AI strategies”, Medium, 28 June 2018。可访问 https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd。

[43] 例如，令人关切的是，大不列颠及北爱尔兰联合王国的一个议会委员会发布了一份长达 200 页的报告，其中甚至一次都没有提到人权。见 United Kingdom of Great Britain and Northern Ireland House of Lords Select Committee on Artificial Intelligence, “AI in the United Kingdom”。

[44] Council of Europe, Algorithms and Human Rights，第 44 页。

营部门关注伦理和公共部门推动伦理往往意味着对人权监管的抵触。[45] 虽然伦理为应对人工智能领域的特殊挑战提供了重要框架，但并不能取代人权，因为人权保障是各国的法律义务。即使公司和政府正在制定伦理规范和指导，也应确保在人工智能业务的所有方面都切实纳入人权考虑和责任。[46]

# 四. 人权指导下的人工智能

47. 本任务负责人的近期报告为公司列出了若干法律和实际考虑因素，要求公司将人权原则置于其内容监管政策的核心，并详细说明了实质标准和程序，帮助公司确保在其业务的各个方面遵守《工商企业与人权指导原则》规定的人权责任。这一框架也为此处提出的人工智能技术方法提供了框架。下文提议的实质标准和程序既适用作为人工智能系统设计、部署和实施方的公司，也适用于有义务在采纳和使用人工智能系统时不干涉人权的国家。这些标准和程序旨在确保人权法在人工智能的发展中居于核心地位。所提供的标准和程序贯穿了两项基本原则：保护和尊重个人能动性和自主权——这是行使意见和表达自由权的关键先决条件之一；以及政府和行业进行有意义的披露的重要性，也就是要通过开放的创新举措向公众解释人工智能技术并促进公众审查。

## A. 人工智能系统的实质标准

48. 公司应围绕普遍人权原则确定具体标准、规则和系统的设计方向。(A/HRC/38/35，第 41-43 段)面向公众的条款和规则应由公司内部的政策承诺加以补充——即承诺将人权考虑纳入公司的各项业务，特别是人工智能和算法系统的开发和部署。公司应考虑如何为人工智能工程师制定专业标准，将人权责任转化为技术设计和运营决策的指导。制定伦理守则和相应的制度架构可以补充人权承诺，但不能取而代之。公共和私营部门机构发布的守则和准则应强调人权法是人工智能中个人保护的基本规则，而伦理框架可以在特定情况下帮助进一步发展人权内容和应用。

49. 公司和政府必须明确告知用户，信息环境中有哪些决定是由自动化系统作出的，有哪些决定是经过了人工审核，还要告知用户自动化系统所用逻辑的一般性要素。同时还要告知用户他们向私营部门提供的个人信息(包括明确提供或通过使用服务或网站而提供)何时会作为数据集的一部分被人工智能系统加以利用，以便用户决定是否同意信息收集，以及同意披露哪些类型的数据[47]。就像使用闭路电视摄像机需公开告知一样，使用人工智能系统，应以明确、易懂的方式(通过弹出式窗口等创新方法)向用户主动告知人工智能程序正在收集他们的信息用于做出决定，还要告诉用户该程序的运行机制和对用户产生的后果等有意义的信息。

---

[45] Ben Wagner, “Ethics as an escape from regulation: from ethics-washing to ethics-shopping?”, in *Being Profiling: Cogitas Ergo Sum*, Mireille Hildebrandt, 编辑。(Amsterdam University Press (forthcoming)。

[46] Article 19 and Privacy International, “Privacy and freedom of expression”，第 13 页。

[47] 人权事务委员会，关于隐私权的第 16 号一般性意见(1988)。

50. 透明度不仅限于向用户披露他们使用的平台和在线服务存在人工智能技术。公司和政府需要在人工智能价值链的每个环节都保持透明。透明度无需复杂也能行之有效；即使是简单地解释人工智能系统的目的、政策、输入和输出也有助于公共教育和讨论。[48] 与其纠结如何向普通受众解释复杂的技术流程，公司应该努力通过提供非技术性的指引来实现透明度。为此，重点应该是向用户告知人工智能系统的存在、目的、组成和影响，而不是向他们提供源代码、培训数据以及输入和输出信息。[49]

51. 要想全面实现信息环境下人工智能系统影响的彻底透明，需要披露的信息包括：人工智能系统删除了多少信息，人工智能建议删除的内容有多少被人类管理员批准，内容删除受到质疑的频率以及质疑被接受的频率。应该向用户提供关于内容显示趋势的汇总数据和解释内容优先顺序的案例研究。彻底透明的关键环节是披露政治和商业广告的来源和受益人。使用人工智能系统的公私部门还要披露人工智能系统的各种局限，包括信任措施、已知漏洞和适当的使用限制。[50]

52. 解决人工智能系统中普遍存在的歧视问题是公司和政府面临的一个重大挑战，如不能正视和解决歧视性要素和影响，人工智能技术不仅不能发挥作用，而且还会很危险。在如何解决人工智能系统中的偏见和歧视问题上，公司和政府有足够先进思想和资源可以借鉴；一般来说，需要在投入和产出两个层面隔离歧视行为并行使问责。这至少需要解决采样误差(如果数据集不具有社会代表性)、清理数据集以删除歧视性数据，以及采取措施，补偿“含有历史性和结构性歧视模式”[51] 且人工智能系统可能从中派生出歧视性替代变量的数据。对人工智能系统中的歧视性结果进行主动监控也是避免和降低对用户人权的负面影响的主要方法。

## B. 人工智能系统相关进程

53. **人权影响评估**。实现贯穿人工智能生命周期始终的彻底透明意味着公司和政府要采取措施，让系统从构思到执行阶段接受审查和质疑。通过进行人权影响评估，政府和公司可以证明其致力于解决人工智能系统的人权影响。人权影响评估应在采购、开发或使用人工智能系统之前进行，包括自我评估和外部审查。智库 AI Now 提出了一项公共机构算法影响评估，其中规定政府需对人工智能系统进

[48] Aaron Rieke, Miranda Bogen and David Robinson, “Public scrutiny of automated decisions: early lessons and emerging methods” (Omidyar and Upturn, 2018)，第 5 页。

[49] Rieke, Bogen and Robinson, “Public scrutiny of automated decisions”，第 8 页。

[50] Amnesty International and Access Now, “Toronto declaration: protecting the right to equality and non-discrimination in machine learning systems”, art. 27 (d), 2018. Available at 可访问 https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/。

[51] Iason Gabriel, “The case for fairer algorithms”, Medium, 14 March 2018 2018 年 3 月 14 日。可访问 https://medium.com/@Ethics_Society/the-case-for-fairer-algorithms-c008a12126f8。

行内部审查，同时促进外部审查，以测试和验证各种假设和结论[52]。公司也应按照类似思路进行评估。

54. 公共部门向私人供应商采购人工智能技术时，必须在采购前进行公共协商，就人工智能系统的设计和实施征询公众意见。在开发、采购或使用人工智能系统和技术之前，公司和政府必须同民间社会、人权团体、相关地方性群体、历史上被边缘化或代表不足群体的代表进行有意义和持续协商。

55. **审计**。促进人工智能系统的外部审查是确保透明化过程严谨性和独立性的关键。因此，现有的独立审计应在采购前开展人权影响评估，作为提高人工智能系统透明度和问责的重要机制。鉴于事关专有技术保护需求，私营部门反对在人工智能领域进行审计。虽然这些担忧可能不无道理，但特别报告员同意 AI Now 的观点，特别是当公共部门使用人工智能应用程序时，供应商拒绝公开系统的运行情况会影响到公共机构履行问责义务。

56. 无论如何，目前有很多创新性建议，可确保在尊重专有秘密的前提下审查人工智能技术。比如，可以设想通过算法生成零知识证明，以证明算法符合某些属性，从而避免了对深层算法[53] 的审查；或者，可以向第三方专家披露算法，由其在保密条件下予以代管，这样一来就可以在不向公众公开算法[54] 的前提下进行公共利益审查。可以允许政府的电信或市场竞争主管部门在保密条件下查看人工智能系统，比如澳大利亚和新西兰在对赌博机的监管方面，公司必须提供算法系统以配合审计审查。[55] 一些学术文献也提出关于人工智能审计[56] 的其他创新建议。

57. 上述几种机制可能会在执行时面临若干挑战，特别是在信息环境下，但公司应致力于促成对人工智能系统的审计。政府应致力于通过政策或立法干预，加强审计的有效性，要求公司确保人工智能代码可审计，保证审计留痕，藉此提高针对受影响用户的透明度。

58. **用户自主**。人工智能不得于无形中取代、操纵或干扰个人在信息环境中形成和持有个人意见和获取和表达想法的能力。尊重用户自主权，至少应确保用户拥有知情权、选择权和控制权。广泛应用的人工智能应用程序隐藏在后台，模糊了内容显示、个性化、内容监控、用户画像和目标定位的程序，损害了用户的意见

[52] Dillon Reisman and others, “Algorithmic impact assessments: a practical framework for public agency accountability” (AI Now, 2018).可访问 https://ainowinstitute.org/aiareport2018.pdf。

[53] Council of Europe, Algorithms and Human Rights，第 36 页。

[54] Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (Cambridge, Harvard University Press, 2015).

[55] Council of Europe, Algorithms and Human Rights，第 34 页。

[56] Christian Sandvig and others, “Auditing algorithms: research methods for detecting discrimination on Internet platforms”, paper presented at Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a pre-conference at the 64th annual meeting of the International Communication Association, Seattle, 22 May 2014 在国际通信协会第 64 届年度会议会前会议(题为“数据与歧视：将关键关切转换为高效调查”)上发表的文件，西雅图，2014 年 5 月 22 日。

自由、表达自由和隐私权。公司应警惕人工智能应用程序中重视商业或政治利益，牺牲透明度和个人选择的负面人权影响。

59. **通知和许可**。公司必须确保用户充分了解基于算法的决策程序对其使用平台、网站或服务的影响。这可以通过教育宣传、弹出式窗口、插页和其他提示方法来实现，以告知用户人工智能正在影响其对搜索引擎、新闻网站或社交媒体平台的使用体验。国家制定的披露规定可以起到保护知情权和许可权的作用。用户还有权知道其信息何时会被人工智能应用程序收集，其信息是否会被纳入数据集进而被输入至人工智能程序，以及信息的使用、存储和删除条件。

60. **补救**。人工智能系统对人权的负面影响必须是能够补救的，而且必须由涉事公司进行补救。建立有效补救程序的先决条件是确保用户知道他们受制于特定算法的决定(包括人工智能系统建议但经人工复核的决定)，并且确保他们知道决定形成的原理。此外，公司应确保补救措施的申请会经过人工审查，以便适当制衡人工智能系统和并确保问责。还要披露数据，说明人工智能技术作出的决定触发补救机制的频率。

# 五. 结论和建议

61. 特别报告员在本报告中探讨了人工智能对意见自由和言论自由权的现有和潜在影响，认为人工智能目前是信息环境的关键部分，对个人享受权利既带来利益也构成和风险。特别报告员提出了一个概念框架，用以思考面对不断提升的技术能力，政府和公司维护这些权利的义务和责任，同时在人工智能技术的力量、触角和范围不断扩张的背景下，提出了可供政府和公司执行的若干具体措施，以确保人权得到尊重。

## 对各国的建议

62. 在采购或部署人工智能系统或应用程序时，各国应确保公共部门机构始终遵守人权原则。其中包括在采购或部署人工智能系统前，进行公众意见征询并进行人权影响评估或公共机构算法影响评估。应特别关注人工智能技术对种族和宗教少数群体、政治反对派和活动家的不同影响。政府使用的人工智能系统应定期接受外部独立专家的审计。

63. 各国应确保私营部门在设计、部署和落实人工智能系统时将人权列为首要考虑。这包括更新和将现有法规(特别是数据保护规定)适用于人工智能领域；采取监管或联合监管措施，要求公司必须对人工智能技术进行影响评估和审计，并确保外部问责机制[57] 切实有效。为保护人权，在必要时可以对特定人工智能应用程序适用行业监管。如果这种限制引入或方便了对表达自由的干预，各国应确保限制对实现《公约》第十九条第三款规定的正当目标是必要和相称的。还应通过广

[57] Wagner, “Ethics as an escape from regulation”.

泛征询公众意见，包括通过民间社会、人权团体、被边缘化或代表不足的最终用户代表参与，制定人工智能方面的法规。

64. 各国应创造一个促进多元信息环境的政策和立法环境。其中包括采取措施确保人工智能领域的市场竞争。这些措施包括制定反技术垄断法规，防止人工智能技术和力量集中在少数主导公司手中；制定提高服务和技术互操作性的法规；以及通过促进网络中立性和设备中立性的政策。[58]

### 对公司的建议

65. 旨在制定人工智能技术伦理问题准则或守则的一切努力都应以人权原则为基础。公私部门在开发和部署人工智能技术的同时，应提供机会让民间社会发表意见。在公司政策和技术指引中，公司应该向工程师、开发人员、数据技术人员、数据清理人员、程序员和人工智能生命周期中的其他工作人员重申，应以人权责任指导一切业务运营，伦理原则可起到辅助作用，帮助在人工智能设计、部署和实施的具体场景中适用人权原则。特别是，平台的服务条款应基于普遍人权原则。

66. 公司应该明确说明人工智能技术和自动化技术在其平台、服务和应用程序中的部署位置和使用方式。为了使用户理解并处理人工智能系统对其享受人权的影响，通过创新方式提醒用户至关重要，包括告知用户他们正被置于人工智能决策程序的控制之下，人工智能正在影响内容显示或进行内容监控，以及用户个人信息正在被纳入数据集，进而会被输入至人工智能系统。公司还应公布关于删除内容的数据，包括删除内容被质疑和质疑被支持的频率；以及内容显示趋势的数据、关于用户商业和政治画像的案例研究和教育资料。

67. 公司必须在人工智能系统的输入和输出两个层面上预防歧视并承担责任。具体做法是确保负责设计和部署人工智能系统的团队的多样化和非歧视态度，在选择数据集和设计系统时注重避免偏见和歧视，包括解决采样误差，清理数据集以删除歧视性数据，以及采取措施弥补歧视性数据。还必须积极监控人工智能系统的歧视性后果。

68. 在新的人工智能系统的设计和部署阶段，包括把现有系统部署到新的全球市场时，应进行人权影响评估，征询公众意见。征询公众意见应在产品或服务最终形成或推出之前进行，以确保征询意见的意义。征询对象应包括民间社会、人权捍卫者和被边缘化或代表不足的最终用户代表。人权影响评估和公众意见征询的结果应予以公开。

69. 公司应确保所有人工智能代码完全可审计，并在监管要求之外，通过采取创新方法，促进人工智能系统的外部独立审计。人工智能审计的结果应予以公开。

[58] Autorité de régulation des communications électroniques et des postes, *Devices, the Weak Link in Achieving an Open Internet* (2018). 可访问 https://www.arcep.fr/uploads/tx_gspublication/rapport-terminaux-fev2018-ENG.pdf。

70. 个人用户必须能够就人工智能系统对其人权的不利影响获得补救措施。公司应建立人工审查和补救系统，以及时回应所有用户针对人工智能系统的投诉和申诉。应定期发布数据，说明人工智能系统被投诉和申请补救的频率，以及可提供的补救措施种类和有效性。