《我们的共同议程》
政策简报 8

数字平台上的信息完整性

2023 年 6 月

联合国

导言

引文

只有通过更强有力的国际合作，才能应对我们正面临的挑战。将于 2024 年举行的未来峰会是一个契机，可供商定多边解决办法，实现更美好的明天，为今世后代加强全球治理工作 (大会第 76/307 号决议)。作为秘书长，我应邀为峰会的筹备工作献计献策，其形式是在本人题为《我们的共同议程》的报告 (A/75/982) 所载提议的基础上，提出着重于行动的建议，而该报告本身则是对《纪念联合国成立七十五周年宣言》(大会第 75/1 号决议) 作出的回应。本政策简报就是这样一项投入。其中详细阐述《我们的共同议程》中首次提出的想法，同时考虑到会员国随后提出的指导意见以及一年多来政府间和多利益攸关方的协商情况，并以《联合国宪章》、《世界人权宣言》和其他国际文书的宗旨和原则作为根基。

本政策简报的目的

本政策简报重点阐述信息完整性面临的威胁如何影响到在全球、国家和地方问题上取得的进展。在《我们的共同议程》中，我呼吁围绕事实、科学和知识达成有实证依据的共识。为此，本简报概述了行为守则的潜在原则，这一守则将有助于指导会员国、数字平台和其他利益攸关方努力使数字空间对所有人更加具有包容性、更加安全，同时大力捍卫意见和表达自由权以及获取信息的权利。在筹备未来峰会的背景下，正在制定数字平台上的信息完整性行为守则。我希望它将为指导加强信息完整性的行动提供一个黄金标准。

数字平台是改变了世界各地社会、文化和政治互动的重要工具。在世界各地，它们在重要的问题上将有关全球公民联系起来。在我们力争在健康的地球上实现和平、尊严和平等之际，它们帮助联合国直接向人们提供信息并让人们参与进来。它们在危机和斗争时刻给人们带来了希望，放大了以前听不到的声音，并为全球运动注入了活力。

然而，这些平台也暴露了数字生态系统的阴暗面。它们使谎言和仇恨得以迅速传播，在全球范围内造成真正的危害。随着错误信息、虚假信息和仇恨言论从数字空间的边缘激增为主流，对社交媒体让人们联系和接触的潜力所持有的乐观情绪受挫。这一危险怎么强调都不为过。社交媒体推动的仇恨言论和虚假信息可能引发暴力和死亡事件。[1] 传播大规模虚假信息、以破坏通过科学手段确立的事实的能力对人类生存构成了风险 (A/75/982，第 26 段)，并危及民主制度和基本人权。由于生成式人工智能等技术的快速发展，这些风险进一步加剧。在世界各地，联合国正在监测错误信息、虚假信息和仇恨言论如何能威胁在实现可持续发展目标方面取得的进展。显然，一切照旧不是一种选择。

在实践中，
错误信息和虚假信息
之间的区别
可能很微妙，
很难确定

什么是信息完整性?

信息完整性是指信息的准确性、一致性和可靠性。它面临虚假信息、错误信息和仇恨言论的威胁。虽然这些用语没有普遍接受的定义，但联合国实体已制定了工作定义。

促进和保护意见和表达自由权特别报告员认为虚假信息是指“有意传播的造成严重社会危害的不实信息”。[2] 联合国教育、科学及文化组织 (教科文组织) 将虚假信息描述为可能造成具体危害的虚假或误导性内容，不论其动机、意识或行为如何。[3]

就本政策简报的目的而言，错误信息和虚假信息之间的区别在于意图。虚假信息是不仅不准确、而且还旨在欺骗和为造成危害目的而传播的信息。国家或非国家行为体可能在多种情况下传播虚假信息，包括在武装冲突期间，虚假信息可能影响到所有发展领域，从和平与安全到人权、公共卫生、人道主义援助和气候行动。

错误信息是指那些不知道自己在传递谎言的人善意分享的不准确信息的非故意传播。错误信息可能植根于虚假信息，因为蓄意的谎言和误导性的叙述随着时间的推移被武器化、被灌输到公共讨论中并在不知不觉中传递下去。[4] 在实践中，错误信息和虚假信息之间的区别可能很难确定。

根据《联合国关于消除仇恨言论战略和行动计划》中的工作定义，仇恨言论是指“因为个人或群体的身份 (即他们的宗教、族裔、国籍、种族、肤色、血统、性别或其他身份因素) 而攻击他们或对他们使用贬损或歧视性语言的任何言论、文字或行为交流。”[5]

错误信息、虚假信息和仇恨言论是相关、但又截然不同的现象，在如何识别、缓解和解决这些现象方面存在某些重叠和差别领域。这三者都污染了信息生态系统，威胁着人类的进步。[6]

对信息完整性的威胁并不是什么新鲜事。长期以来，一直为了政治或经济利益而传播谎言和仇恨。然而，在数字时代，这些操作可以在以前无法想象的规模上进行。强大的通信工具如今可在全球范围内即时传播内容，造成了一个如此普遍的问题，以至于网上平台本身有时也无法掌握其全部范围。政府没有商定这些术语的定义，这不应造成惰性。我们必须尽一切努力遏制它们造成的危害。

数字平台上的信息完整性

在维护信息完整性的运动中，数字平台应成为不可或缺的参与者。虽然某些传统媒体也可能是错误信息和虚假信息的来源，但此种信息通过数字渠道传播的速度、数量和病毒式扩散使得有必要采取紧急、有针对性的对策。就本简报的目的而言，“数字平台”这一用语是指促进两个或更多用户之间互动的数字服务，涵盖从社交媒体和搜索引擎到消息应用程序在内的各种各样活动。通常情况下，它们收集有关其用户及用户互动的数据。[7]

错误信息和虚假信息是由各种各样的行为者制造的，他们有着各种各样的动机，大体上能够保持匿名。国家和非国家行为体协调一致的虚假宣传运动利用了有缺陷的数字系统宣扬有害的言论，造成严重影响。

许多国家启动了监管数字平台的举措，在过去四年中至少通过或审议了 70 项此类法律。[8] 立法办法的核心通常是采取范围狭窄的补救措施，界定和删除有害内容。一些国家把重点放在删除有害内容上，实行了有缺陷和过于宽泛的立法，实际上压制了国际法所允许的“受保护言论”。其他对策，如全面关闭互联网和取缔平台，可能缺乏法律依据并侵犯人权。

许多国家和政治人物以所谓的对信息完整性的关切为借口，限制信息的获取，抹黑和限制报道，并攻击记者和反对派。[9] 国家行为体还以处理错误信息和虚假信息为幌子，向平台施压，要求它们按自己的命令行事。[10] 表达自由专家强调指出， 国家行为体在这方面负有特殊责任，“不应制造、支持、鼓励或进一步传播虚假信息”(A/77/287，第 45 段）。

即使存在合法的公共利益目的，对表达的监管所蕴藏的内在风险也要求采取精心设计的办法，符合人权法规定的合法性、必要性和相称性要求（同上，第 42 段）。

虚假信息也是一个大事。与国家、政治人物和私营部门签约的“黑公关”和主流公关公司是虚假和误导性内容的主要来源。[11] 除其他外，一个策略是将内容发布到假的克隆版新闻网站上，使文章看起来像是来自合法来源。[12] 这种隐蔽的做法极难跟踪和研究，因此问题的真正规模并不清楚。个人也散布虚假主张，兜售产品或服务以牟利，往往在危机或不安全时期以弱势群体为目标。

在大多数数字平台目前的商业模式中，一个主导做法取决于“注意力经济”。算法旨在

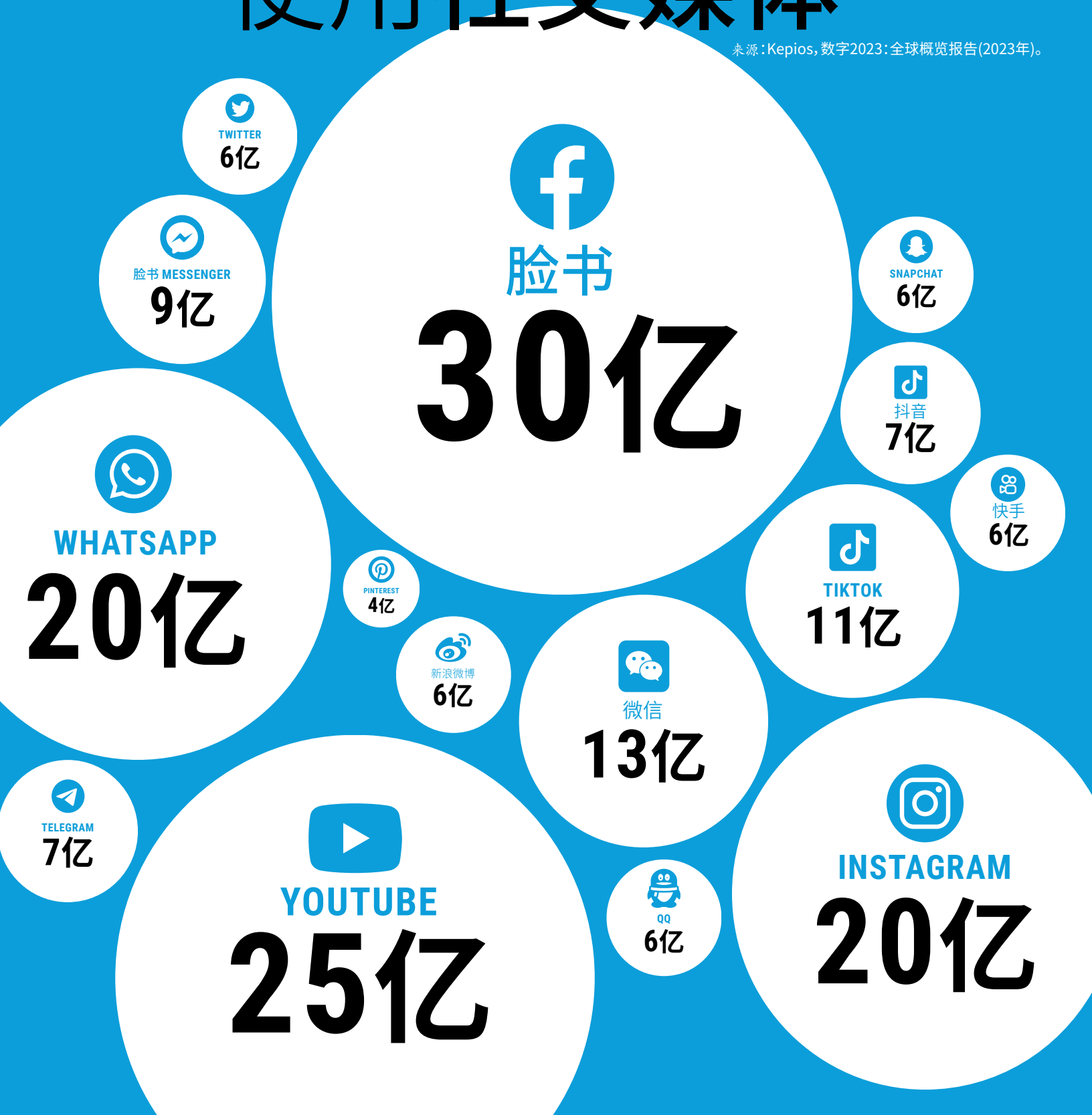
优先考虑能吸引用户注意力的内容，从而使参与度和广告收入最大化。旨在使用户两极分化并产生强烈情绪的不准确和仇恨内容通常是产生最大参与度的内容，其结果是已知有些算法会奖励和放大错误信息、虚假信息和仇恨言论。[13]

面对广告收入下降的情况，数字平台正在寻求注意力经济以外的其他盈利途径。例如，付费验证计划，即账户可以购买以前用于表示真实性的批准印章，引起了对信息完整性的严重关切，因为有可能被虚假信息行为者滥用。[14]

数十亿人

使用社交媒体

来源：Kepios，数字2023：全球概览报告(2023年)。



有什么样的相关国际法律框架?

促进信息完整性必须充分立足于相关的国际规范和标准，包括人权法以及主权和不干涉内政原则。2022 年 8 月，我向大会递交了一份题为“打击虚假信息，促进和保护人权与基本自由”的报告。[15] 在报告中，我列出了适用于虚假信息的国际人权法，包括《世界人权宣言》和《公民及政治权利国际公约》。根据这些国际法律文书，人人享有表达自由权。[16]

《世界人权宣言》第十九条和《公约》第十九条第二项保护表达自由权，包括通过任何媒体，不分国界，寻求、接受和传播各种信息和思想的自由。表达自由这一人权并不局限于受欢迎的信息（A/77/287，第 13 段）。信息自由与表达自由相联系，本身就是一项权利。大会曾指出：“信息自由原为基本人权之一，且属联合国所致力维护之一切自由之关键”（同上，第 14 段）。表达自由和获取信息的自由可能受到某些限制，这些限制符合《公约》第十九条第三项所列特定标准。[17] 各国不能在国际法允许的范围之外增加额外的理由或限制表达。

2012 年通过的《关于禁止构成煽动敌意、歧视或暴力的鼓吹民族、种族或宗教仇恨的拉巴特行动计划》为各国如何最好地执行《公约》第二十条第二项和《消除一切形式种族歧视国际公约》第四条提供了务实的法律和政策指导，其中禁止某些形式的仇恨言论。《拉巴特行动计划》已在不同情况下被会员国采用。[18]

仇恨言论一直是包括灭绝种族罪在内的暴行罪的前兆。1948 年《防止及惩治灭绝种族罪公约》禁止“直接公然煽动灭绝种族”。

大会在 2021 年通过的第 76/227 号决议中强调，一切形式的虚假信息都会对人权和基本自由的享受和实现可持续发展目标产生负面影响。同样，人权理事会在 2022 年通过的第 49/21 号决议中申明，虚假信息会对享有和实现所有人权产生负面影响。

仇恨言论
一直是包括
灭绝种族罪
在内的
暴行罪
的前兆


75%
的联合国维和人员称
错误信息和
虚假信息
影响了其
安全
保障
联合国2022年内部调查

网上错误信息、虚假信息和仇恨言论造成什么样的危害?

网上错误信息、虚假信息和仇恨言论引起了全球公众的严重关切。142 个国家答卷者所提供调查数据的研究结果显示，全世界 58.5% 的互联网和社交媒体常规用户担心在网上遇到错误信息，年轻人和低收入阶层的人的脆弱感比其他群体大得多。[19] 如今的青年是数字土著，他们比其他群体更有可能建立网上连接，这使他们成为有史以来数字连接最多的一代。[20] 在世界各地，每半秒就有一名儿童第一次上网，这使他们有可能接触网上仇恨言论和受到伤害，在有些情况下影响他们的精神健康。[21]

网上错误信息、虚假信息和仇恨言论的影响在世界各地都可以看到，包括在卫生、气候行动、民主和选举、性别平等、安全和人道主义应急等领域。在 2021 年的一项问卷调查中，75% 的联合国开发计划署国家办事处认定信息污染是一个重大问题。教科文组织最近委托对 800 多份学术、民间社会、新闻和企业文件进行审查后发现，信息污染对信任、安全、民主和可持续发展产生了严重影响。[22]

142个国家 答卷者所提供调查数据显示，**全世界**

58.5%

的互联网和社交媒体常规用户

担心在网上遇到

错误信息

错误信息和虚假信息可能是危险的，甚至可能是致命的，特别是在危机、紧急情况或冲突时期。在冠状病毒病 (COVID-19) 大流行期间，有关病毒、公共卫生措施和疫苗的大量错误信息和虚假信息开始在网上流通。[23] 某些行为者利用这种混乱来达到自己的目的，反疫苗活动家将用户带到兜售假治疗方法或预防措施的网站。[24] 许多 COVID-19 的受害者在接触到网上错误信息和虚假信息后拒绝接种疫苗或采取基本的健康预防措施。[25]

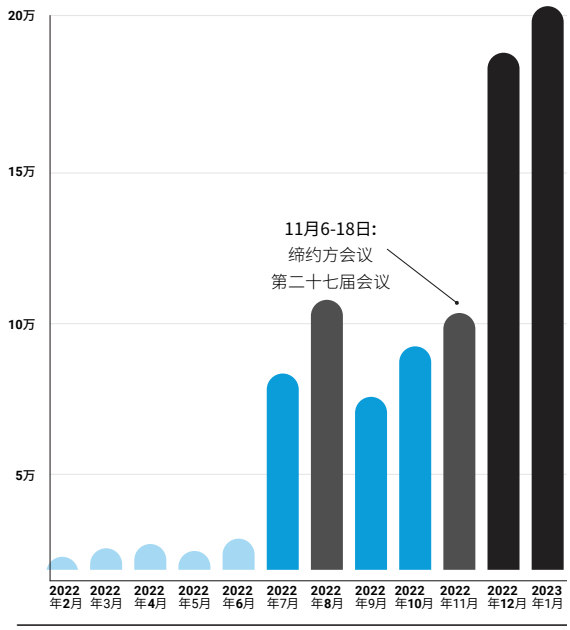
在本已动荡的社会和政治环境中，虚假信息可能同样证明是致命的。在 2022 年的一份报告中，促进和保护意见和表达自由权特别报告员审查了武器化信息在制造混乱、助长仇恨、煽动暴力和延长冲突方面的影响。[26] 在 2022 年发布的另一份报告中，特别报告员认为，虚假信息可能 “涉及针对少数群体、妇女和任何所谓的‘其他人’的偏见和仇恨言论，不仅对直接目标构成威胁，而且危及

包容性和社会凝聚力。在突发情况、危机、关键政治时刻或武装冲突期间，虚假信息会加剧紧张局势，扩大分歧”。[27]

网上危害的一些最严重影响是在被数字平台忽视的情况下产生的，甚至是在平台拥有很高渗透率的情况下产生的。处于冲突之中的国家，或处于其他动荡环境中的国家，往往是利润较低的市场，没有为其分配足够的资源用于审核内容或协助用户。虽然传统媒体仍然是冲突地区大多数人的重要新闻来源，但在数字平台上传播的仇恨也引发和助长了暴力。[28] 一些数字平台因其在冲突（包括目前在乌克兰发生的战争）中的作用而受到批评。[29]

图一

推特上#climatescam的每月使用次数



来源：全球传播部，采用Talkwalker提供的数据。

同样，有关气候紧急情况的错误信息和虚假信息正导致迟迟无法为确保地球的宜居未来而采取紧迫行动。气候错误信息和虚假信息可被理解为虚假或误导性内容，削弱了人类引起的气候变化的存在、其原因和影响的科学商定依据。协调的运动企图否认和极力贬低政府间气候变化专门委员会的科学共识或分散对它的注意力，并企图阻挠为实现2015 年《巴黎协定》所载目标而采取的紧急行动。气候科学否定主义者[30] 虽然属于少数，但却经常发声，他们继续拒绝共识立场，并在一些数字平台上拥有过大的存在。例如，在 2022 年，民间社会组织的随机模拟显示，脸书的算法正在以牺牲气候科学为代价推荐气候否定主义者的内容。[31] 在推特上，#climatesscam 标签的使用次数从 2022 年上半年的每月不到 2 700 次飙升至 7 月的 80 000 次，到 2023 年 1 月的 199 000 次。这一短语还被该平台列为搜索“气候”的热门结果之一。[32] 2022 年 2 月，政府间气候变化专门委员会首次批评气候虚假信息，称“故意破坏科学”导致“对科学共识的误解、不确定性、风险和紧迫性被忽视并引发异议”。[33]

一些化石燃料公司通常采用“漂绿”策略，误导公众相信某个公司或实体为保护环境做得更多，为破坏环境做得更少。这些公司并非单独行动。广告和公关提供商、广告技术公司、新闻媒体和数字平台都在推动和支持旨在迷惑公众和转移人们对化石燃料业责任的注意力的各种努力。[34] 创造漂绿内容的广告和公关公司以及分发这种内容的第三方从这些保护化石燃料业免受审查和免于被追究责任的努力中集体赚取了数十亿美元。公关

70%
的联合国维和人员称
错误信息和虚假信息
对其工作产生了
严重、关键
或中度影响

公司为煤炭、石油和天然气公司开展了数百项宣传活动。[35]

错误信息和虚假信息正在对民主产生深远影响，削弱对民主制度和独立媒体的信任，阻碍对政治和公共事务的参与。在整个选举周期中，接触到虚假和误导性信息会导致选民丧失作出知情选择的机会。错误信息和虚假信息的传播会破坏公众对选举机构和选举进程本身（比如选民登记、投票和结果）的信任，并可能导致选民漠不关心或拒不接受可信的选举结果。事实证明，国家和政治领导人是虚假信息的有效来源，他们故意和有策略地散布谎言，以维持或获得权力，或破坏其他国家的民主进程。[36]

边缘化和弱势群体也经常成为错误信息、虚假信息和仇恨言论的目标，导致他们在社会、经济和政治上进一步受到排斥。女候选人、选民、选举官员、记者和民间社会代表成为网上性别化虚假信息攻击的目标。[37] 这些攻击破坏了政治参与，削弱了民主制度和人权，包括这些群体的表达自由和获取信息的机会。[38] 对于国际社会而言，这必须成为日益紧迫的优先事项，这不仅仅是因为 2024 年世界各地将有 20 多亿选民参加投票。

错误信息和虚假信息也会在平台和传统媒体之间和内部交叉传播，如果未在源头发现，甚至会变得更加复杂，难以跟踪和解决。虚假信息可能是因政治和企业利益而被笼络、且受意识形态影响的媒体机构的一个蓄意策略。[39] 与此同时，数字平台的兴起导致值得信赖的独立媒体急剧衰退。新闻受众和广告收入大量转移到互联网平台——这一趋势因 COVID-19 大流行而加剧。在一些地区或国家，出现了“媒体灭绝”或“新闻沙漠”现象，即社区失去值得信赖的地方新闻来源，[40] 这导致信息生态系统被污染。“新闻清洗”指赞助内容被包装得看起来像报道的新闻故事，在发布到数字平台上时往往没有充分标明，从而使它披上了合法外衣。一旦被其他媒体报道、被政客引用或在不同平台上广泛分享，信息来源就会变得越来越模糊，新闻消费者也无法将其与客观事实区分开来。

虚假信息也对联合国的工作产生直接影响。驻地协调员、特使、调解人和维和人员对虚假信息对本组织业务安全、效力和交付能力的影响表示关切。在 2022 年的一项问卷调查中，70% 的联合国维和人员表示，错误信息和虚假信息对其工作产生了严重、关键或中度影响，75% 的人表示，这些信息对其安全保障产生了影响。错误信息和虚假信息也可能被用来攻击人道主义人员和阻碍冲突地区的救生行动。

图二

信息完整性和可持续发展目标

如下所示，信息完整性所面临的威胁可能对实现可持续发展目标产生负面影响。

错误信息和虚假信息继续对消除贫穷的努力和全球经济产生影响。经济困难也会助长两极分化和仇恨谎言的传播，包括关于边缘化群体的谎言的传播。例如，生活成本危机一直尤其是虚假信息传播的沃土，虚假信息错误地指责转向可再生能源导致能源成本飙升或失业。

信息完整性所面临的威胁会加剧全球饥饿，包括加剧冲突、气候变化、灾害、贫困和不平等。虚假信息会转移并干扰人们对冲突给全球粮食安全带来的挑战的关注。

在冠状病毒病(COVID-19)大流行期间，有关错误信息和虚假信息造成的信息疫情削弱了公共卫生措施及疫苗接种工作。接触有害内容对儿童健康和福祉造成的威胁持续存在。

错误信息、虚假信息和仇恨言论可能对获得优质教育的机会产生不利影响，对包括年轻妇女和女童在内的边缘化群体而言尤为如此。获取信息和旨在增强韧性的数字媒体素养运动将在限制网上危害的社会影响方面发挥关键作用。

基于性别的仇恨言论和虚假信息企图通过使妇女噤声并将她们赶出公共领域来系统地征服她们。仇恨言论和虚假信息可能造成毁灭性的后果，从压制妇女的发言权和助长自我审查，到损害职业和声誉，以及煽动人身暴力。

20亿人生活在没有安全管理的饮用水服务的环境中。关于饮用水和环境卫生安全的错误信息和虚假信息可能对健康造成危险的后果。[a]

气候错误信息和虚假信息大部分是化石燃料业播种的，正在破坏向更清洁能源生产形式的紧急过渡，导致为所有人提供可持续未来的正在关闭的窗口缩小。

研究表明，错误信息、虚假信息和仇恨言论对经济造成了有害影响。[b]

错误信息、虚假信息和仇恨言论以及对这些现象采取的过于宽泛的应对措施，可能对创新产生有害影响，包括限制边缘化群体的潜力，使数字空间不那么平等和包容。

网上传播的错误信息、虚假信息和仇恨言论正在使社会两极分化，并以本已边缘化和脆弱的社区为攻击目标，可能导致他们在社会、经济和政治上进一步受到排斥。

使城市和社区更具可持续性的努力可能遭到虚假信息破坏，虚假信息否认人类活动对环境的影响，或转移人们对此种影响的关注。

倡导循环经济和促进零浪费做法的倡议背后的活动家已经成为网上仇恨言论和虚假信息攻击的目标。

气候虚假信息及其助长的惰性正在破坏采取紧急行动应对气候危机的努力，包括阻碍从污染性化石燃料向可再生能源的关键转变以及对气候复原力的紧急投资。

错误信息和虚假信息可能对养护和可持续利用海洋和海洋资源的努力产生负面影响。

致力于保护陆地生物的环保活动家已成为网上仇恨和虚假信息宣传的目标，并带来了现实生活中的后果。气候错误信息和虚假信息正在破坏采取气候行动的努力。[c]

虚假信息和仇恨言论被用来影响选举和公共言论，并被用来制造混乱。它们被用来削弱对手，阻挠建立和平的努力，煽动暴力，延长冲突，破坏对法治的信任。促进和平和包容性社会的努力以及联合国在支持和平与安全方面的作用因此受到严重影响。[d]

错误信息、虚假信息和仇恨言论可能有碍于为实现可持续发展目标建立有意义的伙伴关系，而转用于解决这一问题的资源可能削弱不让任何人掉队的努力。

a 世界卫生组织和联合国儿童基金会 (儿基会)，《2000-2020 年家庭饮用水、环境卫生和个人卫生方面进展状况：可持续发展目标实施五年》(2021 年，日内瓦)。

b 见 Roberto Cavazos 和 CHEQ 公司，“互联网上不良行为者的经济成本：假新闻，2019 年”；及伦敦经济咨询公司，“谎言的代价：评估与 COVID-19 有关的网上错误信息在人和财政方面对联合王国造成的影响”，2020 年 12 月。

c 全球见证组织，最后的防线：引发气候危机以及袭击土地和环境维权者行为的产业 (2021 年)。

d A/77/288。

全球增加

1亿

月活跃用户
所用月数

来源：Similar Web，采用Sensor Tower提供的数据

INSTAGRAM

30个月

TIKTOK

9个月

CHATGPT

2 个月

我们如何能加强信息完整性?

错误信息、虚假信息和仇恨言论不是存在于真空中。当人们感到被排斥和被忽视、面临经济差距的影响以及感到政治幻想破灭时，它们就会传播。各种对策应解决这些现实世界的挑战。努力实现可持续发展目标对于建设一个可以恢复信任的世界至关重要。

在制定对策时，重要的是不要忽视数字平台给世界带来的巨大价值。平台彻底改变了实时大众通信，在自然灾害和大流行疫情期间推动了救生信息的传播。它们有助于动员各方支持联合国为之奋斗的目标，往往证明是融入和参与公共生活的积极力量。它们将地域不同的人群联系起来，这些人原本被排斥在外，包括那些患有罕见疾病的人，并将致力于使世界变得更美好的众多活动家联系起来。

监管对策

数字平台是否可以并且应该对其托管的内容承担法律责任的问题一直是漫长辩论的主题。在有些情况下，针对诽谤、网络欺凌和骚扰的现行法律已被有效用于抵制对信息完整性构成的威胁，不会对表达自由施加新的限制 (A/77/287，第 44 段)。

此外，近期开展了一些立法工作来应对区域和国家层面的问题。这些工作包括欧洲联盟于 2022 年通过的框架，包括《数字服务法》、关于政治广告透明度和针对性的倡议以及《反虚假信息行为守则》。《数字服务法》为欧洲联盟内部的用户、数字平台和网上运营企业制定了新的规则。这些措施旨在打击非法网上内容、商品和服务，并为用户提供一种机制，既可以标记非法内容，也可以质疑对他们不利的审核决定。这些措施要求数字平台提高透明度，特别是推荐算法的使用和性质方面的透明度，并要求大型平台为研究人员提供数据访问权。

《反虚假信息行为守则》为在欧盟打击网上虚假信息传播而为网上平台和广告部门制定了原则和承诺，其签署方同意予以落实。[41] 其中包括自愿承诺帮助虚假信息非货币化，为此既要防止传播含有虚假信息的广告，又要避免在含有虚假信息的内容旁边放置广告。签署方还同意更清楚地标记政治广告，以及赞助商、广告支出和展示期的详细信息，并创建可搜索的政治广告数据库。此外，他们承诺分享在其平台上发现的关于用于传播虚假信息的恶意操纵行为 (如假账户、机器人驱动的放大、冒充和恶意深度伪造) 的信息，并定期更新和实施对付这些行为的政策。其他承诺的重点是使用户能够识别、了解和

标记虚假信息，加强与事实核对人员的合作，并为研究人员提供更好的数据访问权。对这些新机制的真正考验将是其执行情况。

《反虚假信息行为守则》的一个主要目标是提高平台的透明度。2023 年 2 月，《行为守则》签署方公布了其首份关于它们如何履行承诺的基线报告。这些报告使人深入了解到在多大程度上防止广告收入流向虚假信息行为者和其他被发现的操纵行为，包括在几个欧洲国家为操纵有关乌克兰战争的舆论而作出的大规模协调努力。[42]

数字平台的对策

数字平台在规模、功能和结构方面非常多样化，并采取了各种各样的对策来应对危害。一些较大的平台已公开承诺维护《工商企业与人权指导原则》，[43] 但在政策、透明度和执行方面仍存在差距。一些平台没有执行自己的标准，在不同程度上允许和放大了谎言和仇恨。[44] 为推动平台盈利模式而创建的算法旨在故意最大限度地提高参与度并垄断注意力，有可能把用户推向两极分化或挑衅性的内容。

大多数数字平台都设有某种自我监管制度、审核或监督机制，但围绕内容删除政策和做法的透明度仍然是一个挑战。[45] 对不同地区和不同语文的这些机制的投资极其分散，主要集中在全球北方，平台执行自身规则的情况也是如此。最近的一项问卷调查发现，各个平台将审核工具和监督机制翻译成当地语文的工作未完成。[46] 与此同时，英语以外其他语文的审核往往被外包，而且资源严重不足。[47] 审核人员的证词引发了令人不安的问题，涉及虐待、劳工标准和二次创伤。[48] 审核人员报告说，他们经常接触到暴力和令人不安的内容，并只有几秒时间确定举报的帖子是否违反公司政策。自动化内容审核系统可以发挥重要作用，但因用于训练它们的数据和结构的原因，可能会受到偏见的影响。英文系统的错误率也很高，其他语文的成功率甚至更低。许多数字平台聘用了信任和安全、人权和信息完整性团队，但这些专家往往不是在产品开发的最早阶段加入的，而且在采取成本节约措施期间往往是首批被裁员者。

数据访问权

研究人员的数据访问权也是全球范围内的一个紧迫优先事项。现有的研究和资源仍然严重偏向美利坚合众国和欧洲。已发布的关于对世界其他地区影响的研究报告有限，有几个值得注意的例外情况，包括关于联合国非洲维持和平和缅甸问题国际独立实况调查团的报告[49] 以及一些调查性报道和记者的报道。[50] 部分原因是研究人员无法访问平台及其数据。对平台提供的有限数据进行有效研究所需的工具也往往是在考虑营销的情况下设计的，而且大多极为昂贵。这些平台如果从“按请求访问”办法转变为“默认披露”，并为隐私提供必要的保障，将使研究人员能够正确评估危害。

增强用户能力

民间社会团体和学术界就如何在保护表达自由的同时最好地处理错误信息、虚假信息和

仇恨言论进行了广泛的研究。一些人强调需要自下而上的解决方案，使互联网用户能够限制网上危害对他们自己社区的影响，并将权力从平台手中下放。

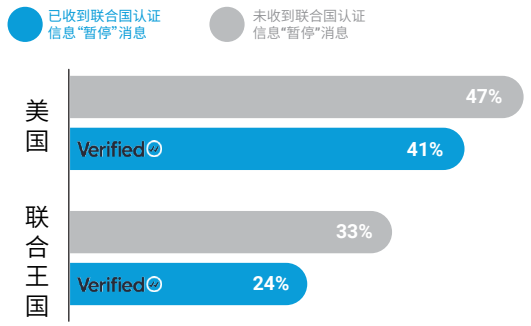
应鼓励平台用户，包括边缘化群体，使其融入和参与政策空间。青年尤其拥有丰富而深刻的专门知识。作为数字土著，年轻人（特别是年轻妇女）和儿童本已经常成为错误信息、虚假信息和仇恨言论的目标，并将直接受到新兴和新的平台的影响。年轻的用户可以根据经验谈论各种提案的不同影响及其潜在缺陷。他们还积极促进网上宣传和事实核对工作。[51]

经改善的批判性思维技能可以使用户增强抵御数字操纵的能力。具体而言，数字素养教会用户更好地评估他们在网上遇到的信息，并以负责任的方式加以传递。各种联合国实体在这一领域拥有宝贵的经验。联合国认证信息倡议[52]成功地部署了一系列策略，包括针对用户发送消息，预先警告用户（即在用户遇到谎言之前警告他们）以及数字素养运动。

图三

联合国的运动有效打击错误信息和虚假信息

分享假新闻的可能性 (2021 年)[a]



a 基于麻省理工学院 2021 年 3 月开展的研究。

抑制办法

大多数数字平台当前的业务模式将参与度置于人权、隐私和安全之上。这包括为谋利而使个人数据货币化，尽管越来越多证据表明这一商业模式造成社会危害。

一些民间社会团体和研究人员探讨了使网上错误信息、虚假信息和仇恨言论去货币化并因此抑制其制造和传播的途径，并指出，虽然表达自由是一项基本人权，但从中获利却并非如此。[53] 所提建议力求解决虚假信息的获利能力，确保内容货币化完全透明和进行独立风险评估，并阻止参与网上广告的人推动虚假信息。

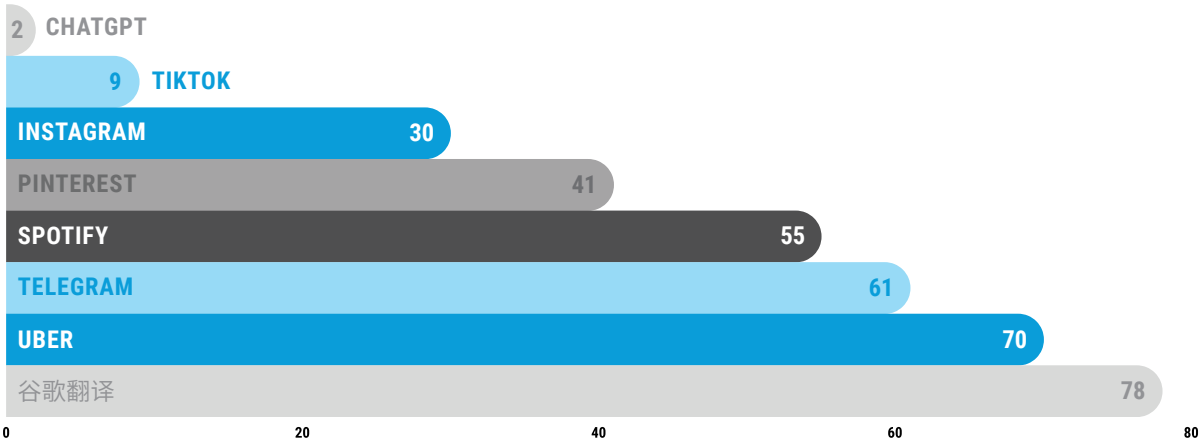
与错误信息、虚假信息和仇恨言论一起宣传的品牌有可能损害其宣传活动的成效并最终损害其声誉。广告商可制定明确政策，以避免不经意间为错误信息、虚假信息和仇恨言论供资并使其合法化，并帮助消除其盈利能力。执行措施可包括管理最新的包括和排除名单以及使用广告验证工具。广告商还可敦促数字平台加快保护信息完整性的行动，不通过助长仇恨和散布虚假信息的媒体机构做广告。[54]

独立媒体

数十个国家的新措施继续破坏新闻自由。教科文组织旗舰产品《世界表达自由和媒体发展趋势》系列中的 2022 年全球报告称，

图四

ChatGPT与其他受欢迎的应用程序相比增加1亿月活跃用户所用的月数



来源：Similarweb，采用 Sensor Tower 提供的数据。

在前一个五年中，世界上 85% 的人口经历了本国新闻自由度下跌的情况。[55] 由于仍有 27 亿人处于离线状态，[56] 进一步的优先事项是加强独立媒体，促进事实核对举措的普及，并为了公众利益支持可靠和准确的报道。真正的公开辩论依赖事实，表述清晰，并以符合道德、独立的方式加以报道。有道德的记者在接受高质量培训并获得适当工作条件后，就拥有了在错误信息和虚假信息面前恢复平衡的技能。他们可以提供重要的服务，关于重要问题的准确、客观和可靠的信息。

具有前瞻性

我们在当前环境下寻求保护信息完整性的解决方案，即使在此之际，我们也必须确保建议具有前瞻性，应对新兴技术和尚未出现的技术。Open AI 的 ChatGPT-3 平台于 2022 年 11 月推出，到 2023 年 1 月获得了 1 亿用户，使其成为历史上增长最快的消费者应用程序，[57] 许多其他公司竞相开发竞争性工具。尽管人工智能在应对全球挑战方面拥有几乎无法想象的潜力，但其最近取得的进步 (包括图像生成器和视频深度伪造) 同样具有强大的潜力，威胁到信息完整性，人们对此感到严重和紧迫的担忧。最近的报告和研究已表明，生成式人工智能工具生成错误信息、虚假信息和仇恨言论，并将其作为令人信服的事实呈现给用户。[58]

我的技术特使正在领导评估生成式人工智能和其他新兴平台的影响的工作。在这样做的时候，我们必须从过去的错误中吸取教训。数字平台在没有充分认识到或评估对社会和个人的潜在损害的情况下就向世界推出。我们现在有机会确保历史不会随着新兴技术的出现而重演。硅谷的“快速行动，打破常规”哲学时代必须结束。用户隐私、安全性、透明度和设计保障安全必须从一开始就纳入所有新的技术和产品。

联合国的对策

联合国和平行动和国家办事处等方面也正在采取步骤，监测、分析和应对错误信息和虚假信息对联合国执行任务构成的威胁。《联合国消除仇恨言论战略和行动计划》为本组织在国家和全球两级应对仇恨言论提供了战略指导。2023 年 2 月，教科文组织主办了“互联网促进信任”会议，以讨论一套监管数字平台的全球准则草案，该草案将于今年晚些时候定稿。[59]

这些倡议和办法合起来有助于为制定联合国行为守则的基本原则指明前进方向。

推特上

#CLIMATESCAM

的**每月**使用次数

10万

2022年11月

缔约方会议第二十七届会议期间的最高点

200

2022年2月

2023年1月

发帖数量达到顶峰

199 300

努力制定联合国行为守则

我将提出联合国数字平台上的信息完整性行为守则，它将建立在以下原则的基础上：

- 对信息完整性的承诺
- 尊重人权
- 支持独立媒体
- 提高透明度
- 增强用户能力
- 加强研究和数据访问权
- 加强对策
- 更有力的抑制办法
- 增强信任和安全

这些原则是从本政策简报中讨论的核心思想中提炼出来的，与我关于全球数字契约的政策简报一致并相互关联。将请会员国在国家一级执行该行为守则。将继续与利益攸关方进行协商，以进一步完善行为守则的内容，并确定具体方法来落实其各项原则。

行为守则可借鉴以下建议：

对信息完整性的承诺

(a) 所有利益攸关方都应避免为任何目的使用、支持或放大虚假信息和仇恨言论，其中包括追求政治、军事或其他战略目标，煽动暴力，破坏民主进程或针对平民、弱势群体、社区或个人；

尊重人权

(b) 会员国应：

㈠ 确保应对错误信息、虚假信息和仇恨言论的措施符合国际法，包括国际人权法，不被滥用于以下目的：阻止观点或意见的任何合法表达，包括为此全面关闭互联网或取缔平台或媒体机构；

㈡ 采取监管措施，保护数字平台用户的基本权利，包括执法机制，对技术公司提出的要求要完全透明；

(c) 所有利益攸关方都应遵守《工商企业与人权指导原则》；

支持独立媒体

(d) 会员国应保障自由、有活力、独立和多元的媒体环境，有力保护记者和独立媒体，并支持以当地语文设立、资助和培训独立的事实核对组织；

(e) 新闻媒体应确保准确和合乎道德的独立报道，并辅之以符合国际劳工和人权规范和标准的高质量培训和工作条件；

提高透明度

(f) 数字平台应：

㈠ 确保算法、数据、内容审核和广告切实具有透明度；

㈡ 公布和宣传关于错误信息、虚假信息和仇恨言论的便于查阅的政策，并报告关于平台服务的协调虚假信息的普遍性以及打击此类行动的政策的效力；

(g) 新闻媒体应确保资金来源和广告政策切实具有透明度，并明确区分编辑内容和付费广告，包括在向数字平台发布时；

增强用户能力

(h) 会员国应确保公众获得准确、透明和来源可靠的政府信息，特别是符合公共利益的信息，包括可持续发展目标的所有方面；

(i) 数字平台应确保以透明的方式增强用户能力和保护用户，让人们对他们看到的内容和如何使用他们的数据有更多的选择。它们应使用户能够证明身份和真实性，而不需要在金钱和隐私之间作出权衡，并建立透明的用户投诉和报告程序，辅之以独立、广为宣传和便于利用的投诉审查机制；

(j) 所有利益相关方都应投资于强有力的数字素养运动，使所有年龄段的用户都能更好地了解数字平台的工作方式，他们的个人数据可能如何得到使用，并识别和应对错误信息、虚假信息和仇恨言论。应特别注意确保年轻人、青少年和儿童充分认识到他们在网上空间的权利；

加强研究和数据访问权

(k) 会员国应投资于并支持对不同国家和不同语文的错误信息、虚假信息和仇恨言论的普遍性和影响进行的独立研究，特别是在服务不足的情况下和英文以外的语文，使民间社会和学术界能够自由和安全地运作；

(l) 数字平台应：

㈠ 允许研究人员和学者访问数据，同时尊重用户隐私。应使研究人员能够收集关于错误信息、虚假信息和仇恨言论所针对的个人和群体的实例和定性数据，以更好地了解危害的范围和性质，同时尊重数据保护和人权；

㈡ 确保民间社会充分参与应对错误信息、虚假信息和仇恨言论的努力；

加强对策

(m) 所有利益相关方都应：

㈠ 分配资源，以处理和报告错误信息、虚假信息和仇恨言论的来源、传播和影响，同时尊重人权规范和标准，并进一步投资于不同国家和不同情况下的事实核对能力；

㈡ 组建关于信息完整性的联盟，汇集不同的专门知识和办法，以帮助弥合地方组织和在全球范围内运作的技术公司之间的差距；

㈢ 促进培训和能力建设，以了解错误信息、虚假信息和仇恨言论的表现形式，并加强预防和缓解战略；

更有力的抑制办法

(n) 数字平台应摆脱将参与度置于人权、隐私和安全之上的商业模式；

(o) 广告商和数字平台应确保广告不被放置在网上错误信息或虚假信息或仇恨言论旁边，并确保不推广包含虚假信息的广告；

(p) 新闻媒体应确保所有付费广告和软文内容都明确标出这一点，并且没有错误信息、虚假信息和仇恨言论；

增强信任和安全

(q) 数字平台应：

㈠ 通过所有产品的设计确保安全和隐私，包括为此为内部信任和安全专门知识提供充足资源，以及在不同国家和不同语文间协调一致地适用政策；

㈡ 投资于在业务所在国使用的所有语文的人类和人工智能内容审核系统，并确保内容报告机制透明，加快响应速度，特别是在冲突环境中；

(r) 所有利益攸关方都应立即采取紧急措施，确保以安全、可靠、负责任、符合道德和人权的方式使用人工智能，并解决该领域最近取得的进步对错误信息、虚假信息和仇恨言论传播的影响。

今后的步骤

- 联合国秘书处将就制定联合国行为守则的事宜、包括后续行动和执行机制与一系列利益攸关方进行广泛协商。这可能包括设立一个由公认专家组成的独立观察站，以评估承诺遵守行为守则的行为者所采取的措施，也包括设立其他报告机制。
- 为支持和充实该守则，联合国秘书处可开展深入研究，以在全球范围内加强对信息完整性的认识，特别是在世界上研究不足的地区。
- 秘书长将在联合国秘书处建立专门能力，以加强对影响联合国执行任务和实质性优先事项的网上错误信息、虚假信息和仇恨言论采取的对策。这种能力将在专家监测和分析的基础上，制定有针对性的传播战略，以预测威胁，并(或)在威胁升级为在线和离线危害之前迅速应对，并支持联合国工作人员和会员国的能力建设。它将支持会员国、数字平台和其他利益攸关方努力遵守和执行最终敲定的守则。

结论

对国际社会而言，加强数字平台上的信息完整性是个当务之急。从卫生和性别平等到和平、正义、教育和气候行动，限制错误信息、虚假信息和仇恨言论影响的措施将促进实现可持续未来和不让任何人掉队的努力。即使在国家一级采取行动，这些问题也只能通过加强全球合作才能充分解决。本政策简报中概述的核心思想表明，加强信息完整性的道路必须以人权为基础，涉及多个利益攸关方和多个层面。这些思想已被提炼为若干原则，在制定联合国数字平台上的信息完整性行为守则时需要加以考虑，该守则将为在大力维护人权的同时加强信息完整性提供蓝图。我期待与会员国和其他利益攸关方协作，将这些原则转化为实际承诺。

附件

与会员国和其他相关利益攸关方的协商

本政策简报中的构想借鉴了题为《我们的共同议程》的报告 (A/75/982) 中概述的建议，该报告得益于与会员国、联合国系统以及世界各地的思想领袖、年轻人和民间社会行为体的广泛协商。会员国和其他利益攸关方在大会 25 次讨论过程中就《我们的共同议程》进行了大量细致的反思，本政策简报尤其对这些反思作出了回应。

在发布本政策简报之前，与会员国进行了协商，包括为此向新闻委员会做了非正式简报，并邀请所有非委员会成员参加。还与民间社会伙伴、学术界、专家和包括技术公司在内的私营部门进行了讨论。

在未来峰会举行前，将在制定行为守则的过程中进行广泛协商。

尾注

1 A/HRC/42/50；A/77/287；A/HRC/51/53；联合国，“防止灭绝种族罪行问题特别顾问艾丽斯·瓦伊里穆·恩德里图发表声明，谴责埃塞俄比亚境内战斗最近升级”，新闻稿，2022 年 10 月 19 日；联合国人权事务高级专员办事处 (人权高专办)，“缅甸：联合国专家称，社交媒体公司必须抵制军政府的线上恐怖活动”，新闻稿，2023 年 3 月 13 日；人权高专办，“联合国专家：言论自由不代表能在社交媒体上自由传播种族仇恨”，声明，2023 年 1 月 6 日；促进和保护意见和表达自由权特别报告员，“#JournalistsToo：女记者发声”，2021 年 11 月 24 日；以及人权高专办，“斯里兰卡：专家对倒退的措施感到失望，呼吁重新进行联合国审查并努力确保追究责任”，新闻稿，2021 年 2 月 5 日。

2 A/HRC/47/25，第 15 段。

3 Kalina Bontcheva 和 Julie Posetti 编，《平衡行为：在尊重表达自由的同时打击数字虚假信息——宽带委员会关于“表达自由和应对互联网上的虚假信息”的研究报告》(日内瓦，国际电信联盟 (国际电联)；巴黎，教科文组织，2020 年)。

4 见联合国，“打击虚假信息”，可查阅 www.un.org/en/countering-disinformation，以及 A/77/287。

5 可查阅：https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf。

6 联合国目前正在进行一项研究，以审查错误信息和虚假信息与仇恨言论之间的相互联系和关系，以及这些相关但却不同的现象在概念和运作层面上的趋同和分歧之处。

7 欧盟委员会在“塑造欧洲的数字未来：网络平台”中提供了网络平台的定义，2022 年 6 月 7 日，可查阅：https://digital-strategy.ec.europa.eu/en/policies/online-platforms。

8 见人权高专办，“审核在线内容：抵制危害还是压制异见？”，2021 年 7 月 23 日。

9 见联合国，“打击虚假信息”，以及 A/77/287。

10 A/HRC/47/25。

11 Stephanie Kirchgaessner 和其他人，“已揭露：黑客和虚假信息团队干涉选举”，《卫报》，2023 年 2 月 14 日。

12 Alexandre Alaphilippe 和其他人，“分身——为俄罗斯宣传服务的媒体克隆”，欧盟虚假信息实验室，2022 年 9 月 27 日。

13 联合国经济学家网络，“可持续发展新经济学：注意力经济”。

14 推特，“关于推特蓝标”；和 Meta，“测试 Meta Verified，以帮助创建者建立存在”，2023 年 3 月 17 日。

15 A/77/287。

16 截至 2023 年 2 月，有 173 个会员国已成为《公民及政治权利国际公约》缔约国。

17 对表达自由的限制必须符合下列既定条件：合法性，即限制必须由法律规定，以充分准确地区分合法和非法言论；必要性和相称性，即限制显然对权利的行使构成的负担最小，而且实际上保护或可能保护有关的合法国家利益；合法性，也就是说，要想合法，限制必须只保护《公民及政治权利国际公约》第十九条第三项所列举的那些权益。

18 这些情况包括科特迪瓦、摩洛哥和突尼斯的视听通信，以及联合国中非共和国多层面综合稳定团对煽动暴力行为的监测。

19 Aleksi Knuutila、Lisa-Maria Neudert 和 Philip N. Howard，“谁害怕假新闻？在 142 个国家建模对错误信息的

风险认知”，哈佛肯尼迪学院，《错误信息评论》，第 3 卷，第 3 号 (2022 年 4 月)。

20 国际电联，《衡量信息社会》(2013 年，日内瓦)。

21 联合国儿童基金会 (儿基会)，“保护上网儿基”，2022 年 6 月 23 日。可查阅 www.unicef.org/protection/violence-against-children-online。

22 教科文组织，关于数字治理以及对信任和安全构成的挑战的工作文件。可查阅：www.unesco.org/en/internet-conference/working-papers。

23 见 Julie Posetti 和 Kalina Bontcheva，“虚假信息疫情：解密 COVID-19 虚假信息”，政策简报 1(巴黎，教科文组织，2020 年) 和“虚假信息疫情：剖析对 COVID-19 虚假信息的反应”，政策简报 2(巴黎，教科文组织，2020 年)。

24 打击数字仇恨中心，疫情暴利商人：反疫苗生意 (2021 年)。

25 Michael A Gisondi 和其他人，“致命的信息疫情：社交媒体和 COVID-19 错误信息的力量”，《医学互联网研究期刊》，第 24 卷，第 2 号 (2022 年 2 月)。

26 A/77/288。

27 A/77/287，第 6 段。

28 2018 年，人权理事会任命的一个独立国际实况调查团宣布脸书是“缅甸仇恨言论的主要平台”(A/HRC/42/50，第 72 段)。

29 见联合国新闻，“仇恨言论：日益增长的国际威胁”，2023 年 1 月 28 日，以及“联合国专家警告说，数字技术、社交媒体前所未有地助长仇恨言论”，2022 年 10 月 20 日。

30 见 John Cook,“解构气候科学否定说”，载于《通报气候变化研究手册》，David C. Holmes 和 Lucy M. Richardson 编 (联合王国，切尔滕纳姆，爱德华 • 埃尔加出版社，2020 年)。Cook 报告称，Abraham 等人 (2014 年) 总结了包含否定主义主张 (比如有关卫星测量显示变冷或低气候敏感度估计的主张) 的论文如何在科学文献中遭到有力驳斥。同样，Benestad 等人 (2016 年) 试图复制唱反调的论文中的结论，并发现了一些缺陷，如不当的统计方法、错误的二分法以及根据错误理解的物理学得出的结论。

31 全球见证组织，“气候鸿沟：脸书的算法如何放大气候虚假信息”，2022 年 3 月 28 日。

32 全球传播部使用 Talkwalker 的数据进行的分析。

33 Jeffrey A. Hicke 和其他人，“北美洲”，载于政府间气候变化专门委员会，《2022 年气候变化：影响、适应和脆弱性》，政府间气候变化专门委员会第六次评估报告第二工作组提供的资料 (联合王国，剑桥，剑桥大学出版社，2022 年)。

34 Mei Li、Gregory Trencher 和 Jusen Asuka，“英国石油公司、雪弗龙、埃克森美孚和壳牌公司的清洁能源声明：讨论、行动和投资不相称”，《公共科学图书馆：综合》，第 17 期，第 2 号 (2022 年 2 月)。

35 Robert J. Brulle 和 Carter Werthman，“公关公司在气候变化政治学中的作用”，《气候变化》，第 169 卷，第 1-2 号 (2021 年 11 月)。根据非营利监督机构“全球虚假信息指数”，技术行业广告商在 2021 年向 98 个载有英语气候虚假信息的网站提供了 3 670 万美元。“打击数字仇恨中心”(一个宣传团体)2022 年 11 月的一份报告显示，石油和天然气公司过去两年仅在谷歌上的搜索广告方面就花费了 2 370 万美元资金，其中近一半资金针对有关环境可持续性的搜索词。InfluenceMap 的研究发现，2020 年，美利坚合众国 25 个石油和天然气行业组织在脸书平台上发布了 25 147 则误导性广告，总支出为 9 597 376 美元。截至目前，应对措施一直与问题的规模不相称。

36 见大会第 76/227 号决议；人权理事会第 49/21 号决议；和欧盟对外行动署，“对付虚假信息、外国信息操纵和干扰”，2021 年 10 月 27 日。

37 Lucina Di Meco,“使厌女症货币化：性别化的虚假信息以及对全球妇女权利和民主的破坏”，#ShePersisted，2023 年 2 月。

38 见 Andrew Puddephatt,“社交媒体和选举”，通信和传播讨论笔记 (Cuadernos de Discusión de Comunicación e Información)，第 14 期 (蒙得维的亚，教科文组织，2019 年)；以及 Julie Posetti 和其他人，《不寒而栗：网上对女记者施暴行为的全球趋势》，研究讨论文件 (教科文组织，2021 年)。

39 欧盟虚假信息实验室，“‘媒体’在制造和传播虚假信息活动中的作用”，2021 年 10 月 13 日。

40 见联合国新闻，“社交媒体对传统、值得信赖的新闻构成‘生存威胁’：教科文组织”，2022 年 3 月 10 日；以及 Anya Schiffrin 和其他人，“为新闻业的蓬勃发展寻找资金：支持媒体生存能力的政策选择”，世界表达自由和媒体发展趋势 (巴黎，教科文组织，2022 年)。

41 欧盟委员会，“塑造欧洲的数字未来：2022 年反虚假信息行为守则”，2022 年 7 月 4 日。

42 见欧洲联盟委员会负责价值观和透明度的副主席 Věra Jourová 在欧洲联盟委员会的发言，“《反虚假信息行为守则》：新的透明度中心首次提供关于网上虚假信息的见解和数据”，每日新闻，2023 年 2 月 9 日。可查阅：https://ec.europa.eu/commission/presscorner/detail/en/mex_23_723。

43 可查阅：https://unglobalcompact.org/library/2。

44 打击数字仇恨中心和人权运动，“数字仇恨：社交媒体在放大有关 LGBTQ+ 人士的危险谎言方面发挥的作用”，2022 年 8 月 10 日。

45 见 Andrew Puddephatt,“让阳光照进来：数字时代的透明度和问责制”，世界表达自由和媒体发展趋势 (巴黎，教科文组织，2021 年)。

46 谁的知识？牛津互联网研究所和互联网与社会中心，互联网语文状况报告 (2022 年)。

47 A/HRC/38/35。

48 Billy Perrigo,“脸书非洲血汗工厂内幕”，《时代周刊》，2022 年 2 月 17 日。

49 A/HRC/42/50。

50 值得注意的例子包括 Maria Ressa，《如何勇敢抵制独裁者》(纽约，哈珀柯林斯出版社，2022 年)；和 Max Fischer，《混乱机器》(纽约，利特尔和布朗公司，2022 年)。

51 见儿基会，“年轻记者对 COVID-19 信息进行事实核对”。

52 见 https://shareverified.com/。

53 全球虚假信息指数是一个非营利团体，负责跟踪与虚假信息放在一起的广告。联合国一直是这种做法的受害者，全球虚假信息指数已发现，儿基会的广告与反疫苗的文章放在一起，联合国难民事务高级专员公署的广告与反难民的内容放在一起。

54 有意识广告网络，宣言。可查阅：www.consciousadnetwork.com/the-manifestos/。

55 教科文组织，新闻是一种公益：世界表达自由和媒体发展趋势——2021/2022 年全球报告 (巴黎，2022 年)。

56 国际电联，“2021 年事实和数字：29 亿人仍处于离线状态”，2021 年 11 月 29 日。会员国将在 2024 年举行的未来峰会上讨论《全球数字契约》，其中将概述为所有人创造一个开放、自由和安全的数字未来的共同原则 (见 www.un.org/techenvoy/global-digital-compact)。

57 Krystal Hu,“ChatGPT 创下用户群增长最快的记录 - 分析师笔记”，路透社，2023 年 2 月 2 日。

58 见打击数字仇恨中心，“谷歌的新人工智能聊天机器人 Bard 上的错误信息”，2023 年 4 月 5 日；以及 Tiffany Hsu 和 Stuart A. Thompson,“虚假信息研究人员针对人工智能聊天机器人发出了警报”，《纽约时报》，2023 年 2 月 13 日。

59 准则草案可查阅：www.unesco.org/en/internet-conference。