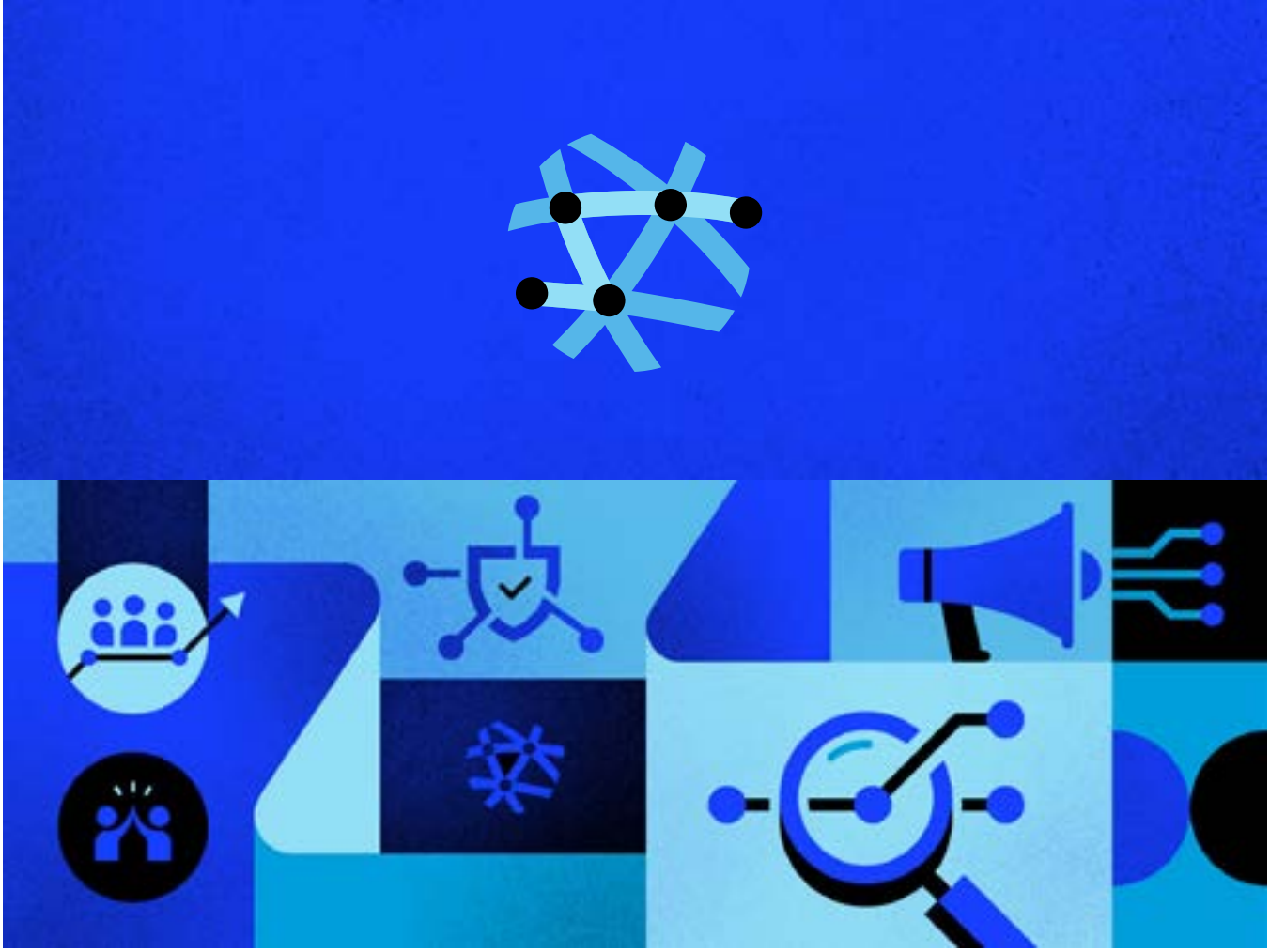


联合国信息完整性全球原则

多方利益攸关方行动建议

联合国

目录

数字时代的信息生态系统

技术发展在短短几十年间彻底改变了传播方式，以此前无法想象的规模将个人和社区联系在一起，为传播知识、丰富文化和可持续发展提供了无可比拟的机遇。技术进步在许多方面提升了人们对信息生态系统完整性的追求——在这一系统中，人们充分享有表达自由，在开放、包容、安全和可靠的信息环境中，人人都能获得准确、可靠、没有歧视和仇恨的信息。

这些技术发展让信息得以大规模传播的同时，也助长了各种行为者以历史上前所未有的数量、速度和病毒式传播方式散布错误信息、虚假信息和仇恨言论，从而危及信息生态系统的完整性。在人工智能技术快速突破的背景下，这些风险包括一系列当前的、新出现的和未来的威胁。

这种对信息空间完整性的侵蚀会削弱人们行使人权的能力，也会阻碍实现和平、繁荣和地球的宜居未来的努力。因此，加强信息完整性是我们时代最紧迫的挑战之一。

信息的完整性意味着一个倡导人权、和平社会和可持续未来的多元化信息空间。它承载着数字时代的希望，希望能够为所有人增进信任、增加知识和个人选择。

促进信息的完整性涉及赋予人们权力，使其能够行使寻求、接收和传递各种信息和思想的权利，以及不受干扰地持有观点的权利。在日益复杂的数字信息环境中，这意味着要使个人能够安全地浏览信息空间，并享有隐私和自由。

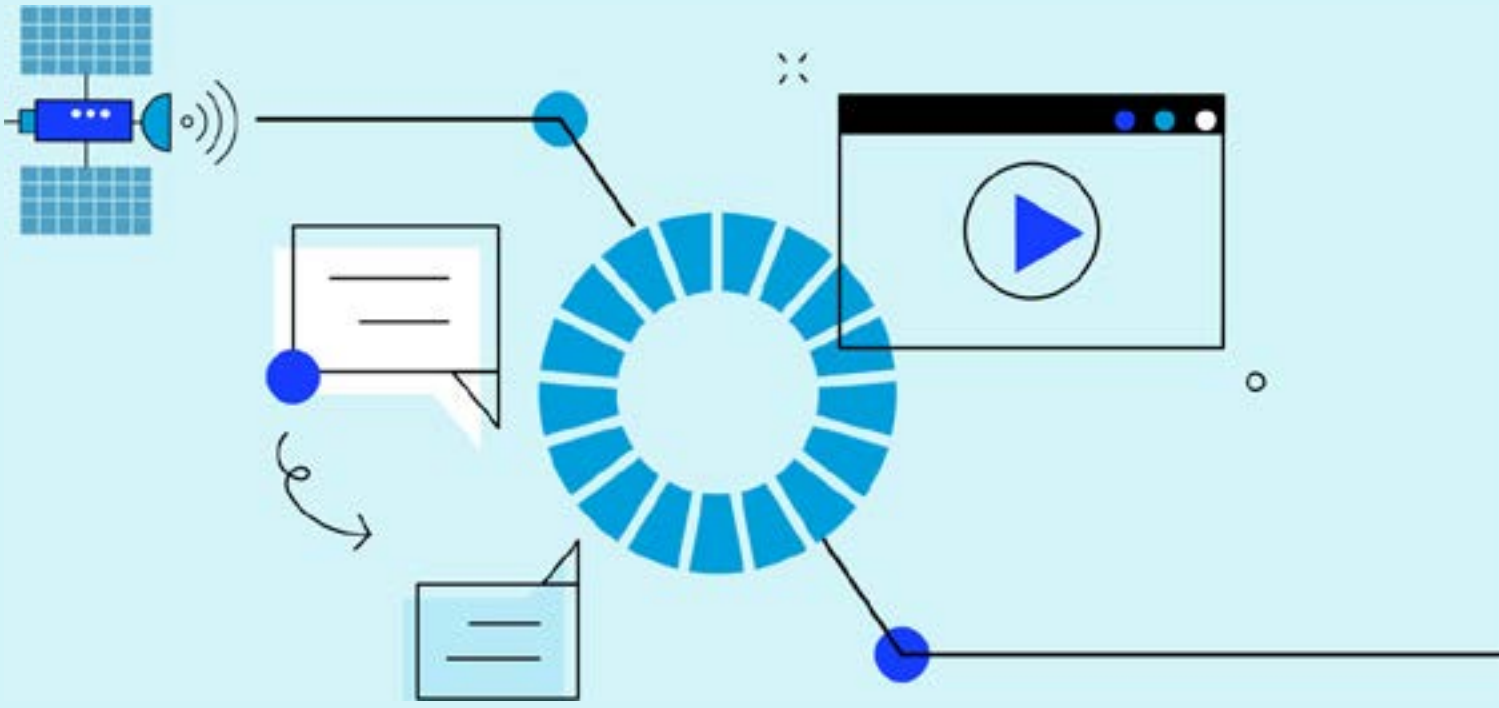
信息完整性与可持续发展目标

努力加强信息完整性对于维护和进一步推进可持续发展目标至关重要。侵蚀信息生态系统的完整性会加剧在实现可持续发展目标方面现有的脆弱性，特别是对全球南方国家而言。

弱势和边缘化群体受到的影响尤为严重。例如，让更多妇女加入全球劳动力大军对于实现可持续发展目标至关重要。然而，除了许多国家存在的歧视性法律和政策外，基于性别的仇恨言论、虚假信息和暴力也被用来系统地压迫妇女，使她们噤声，并将她们排挤出公共领域。这可能对妇女的参与产生破坏性的长期后果，压制妇女的声音，助长自我审查，导致职业和声誉受损，危及在性别平等方面来之不易的进展。

利用信息空间破坏气候行动进一步凸显了挑战的紧迫性。通常由商业利益驱使的协调一致的虚假宣传运动试图否认或怀疑人类引起的气候变化、其原因或影响的科学商定依据，以拖延或破坏实现气候目标的行动。公众人物——活动家、科学家和广播员——因努力提供有关气候危机的信息和应对气候危机而成为仇恨言论、威胁和骚扰的目标。

在可持续发展目标的各个领域，从健康和零饥饿到和平、正义、教育和减少不平等，加强信息完整性的措施将推动实现可持续未来的努力，不让任何人掉队。



实现联合国信息完整性全球原则

联合国与会员国、民间社会（包括青年领导的组织）、媒体、学术界和私营部门的代表就信息完整性问题在所有区域进行了广泛和多样的磋商。利益攸关方通过国家层面的讨论、虚拟会议、双边会议以及全球传播的公开在线表格进行了发言。

这些磋商凸显了对统一建议的需求，这些建议应适用于所有地域和背景，并能满足所有人的要求，特别是关注处于弱势和边缘化境况群体的需求。

为此，《联合国信息完整性全球原则》提供了一个整体框架，指导多方利益攸关方采取行动，建立一个更健康的信息生态系统。该框架由加强信息完整性的五项原则组成，每项原则都包括针对主要利益攸关方群体的建议。

这些原则是：社会信任和韧性；独立、自由和多元化的媒体；透明度与研究；公众赋权；以及健康的激励机制。这些原则的核心都是对人权的坚定承诺。

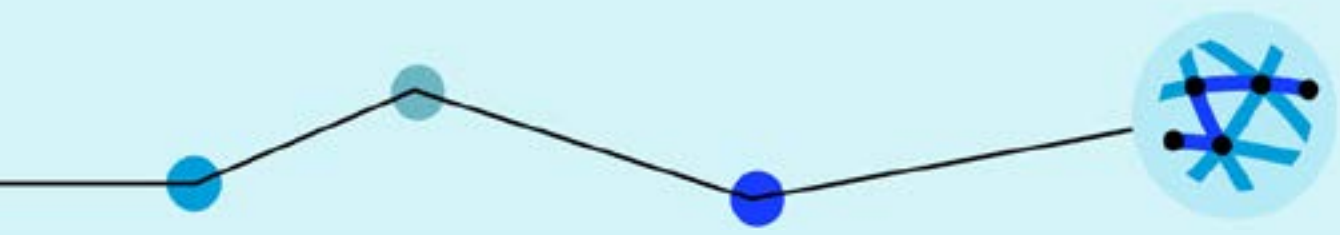
《全球原则》承认并借鉴了各国、民间社会、私营部门和其他利益攸关方已经做出的广　泛努力和取得的进展。它们为在各行各业、各种语言和背景下保护和促进信息的完整性提供了一个统一的出发点，承认全球团结一致，以前所未有的规模、速度和强度做出广泛的响应。

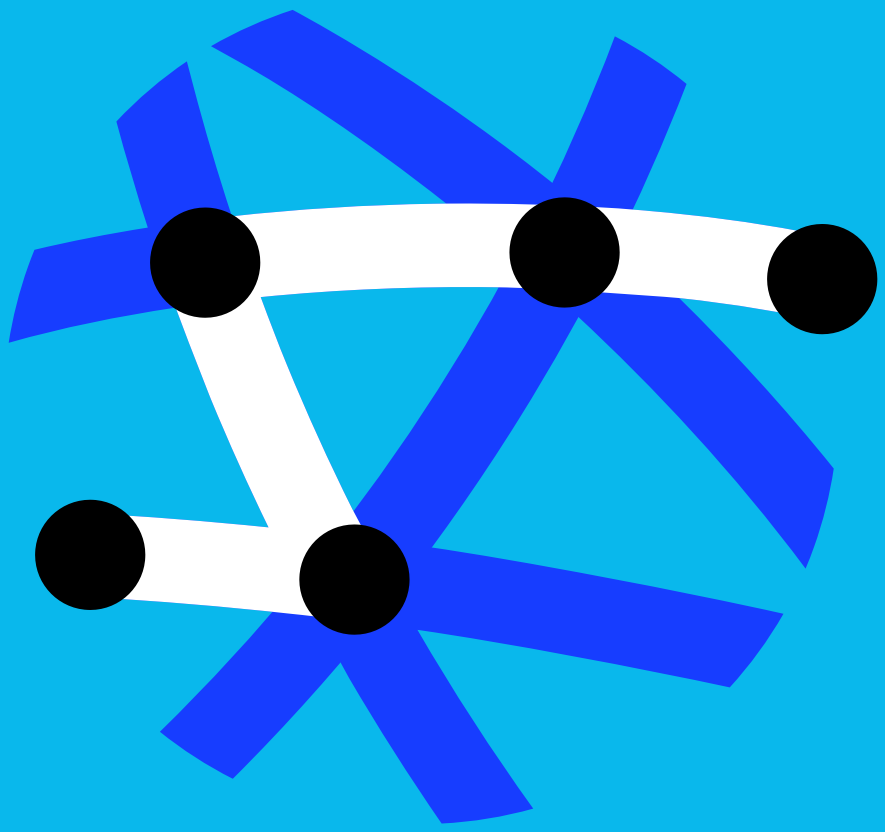
《全球原则》为个人、公共和私营实体，包括联合国系统、各国政府、媒体、民间社会组织以及技术、广告和公共关系部门的营利性公司提供了一个机会，使他们能够与国际法规定的权利和自由保持一致，并为信息的完整性结成广泛的联盟。

《全球原则》以《我们的共同议程》和联合国秘书长的政策简报8《数字平台上的信息完整性》中提出的观点为基础。

除了以包括国际人权法在内的国际法为基础之外，《全球原则》还补充了相关的《联合国工商企业与人权指导原则》、联合国教科文组织《数字平台治理准则》、《联合国关于记者安全和有罪不罚问题的行动计划》、联合国教科文组织《人工智能伦理问题建议书》以及《联合国关于仇恨言论的战略和行动计划》。《全球原则》为联合国会员国考虑《未来契约》和《全球数字契约》提供了资源。

因此，《全球原则》进一步反映了联合国对加强信息完整性的坚定承诺，并旨在指导本组织未来的工作。





5 信息完整性全球原则

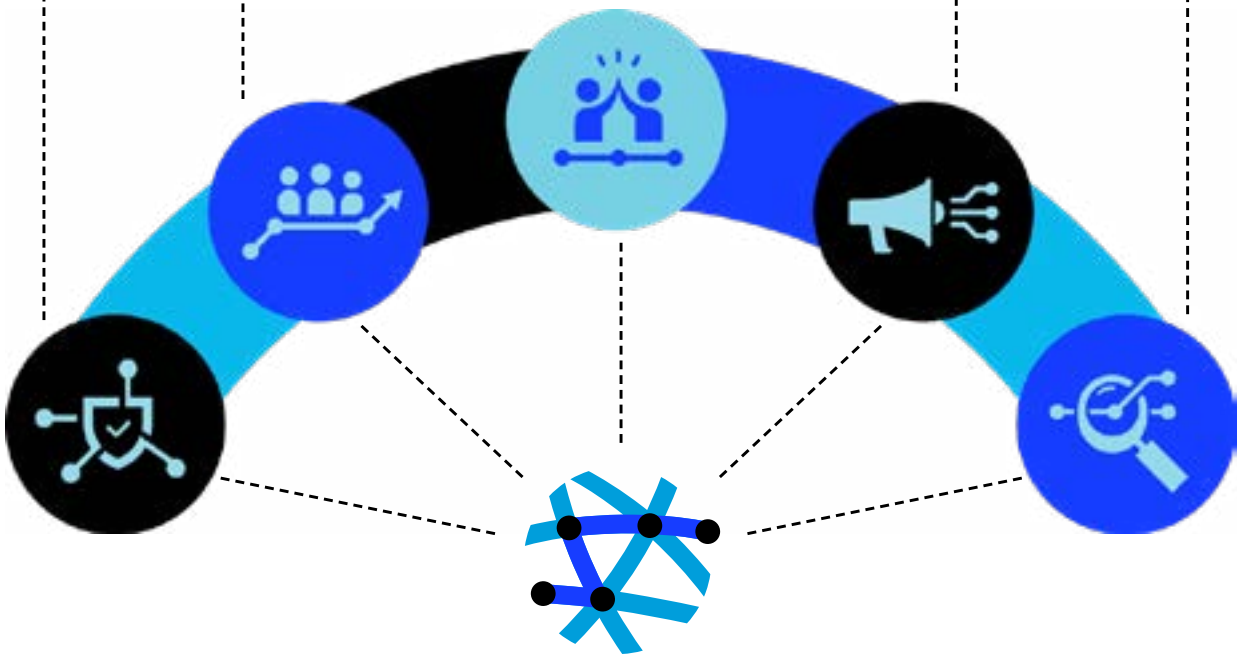
健康的激励机制

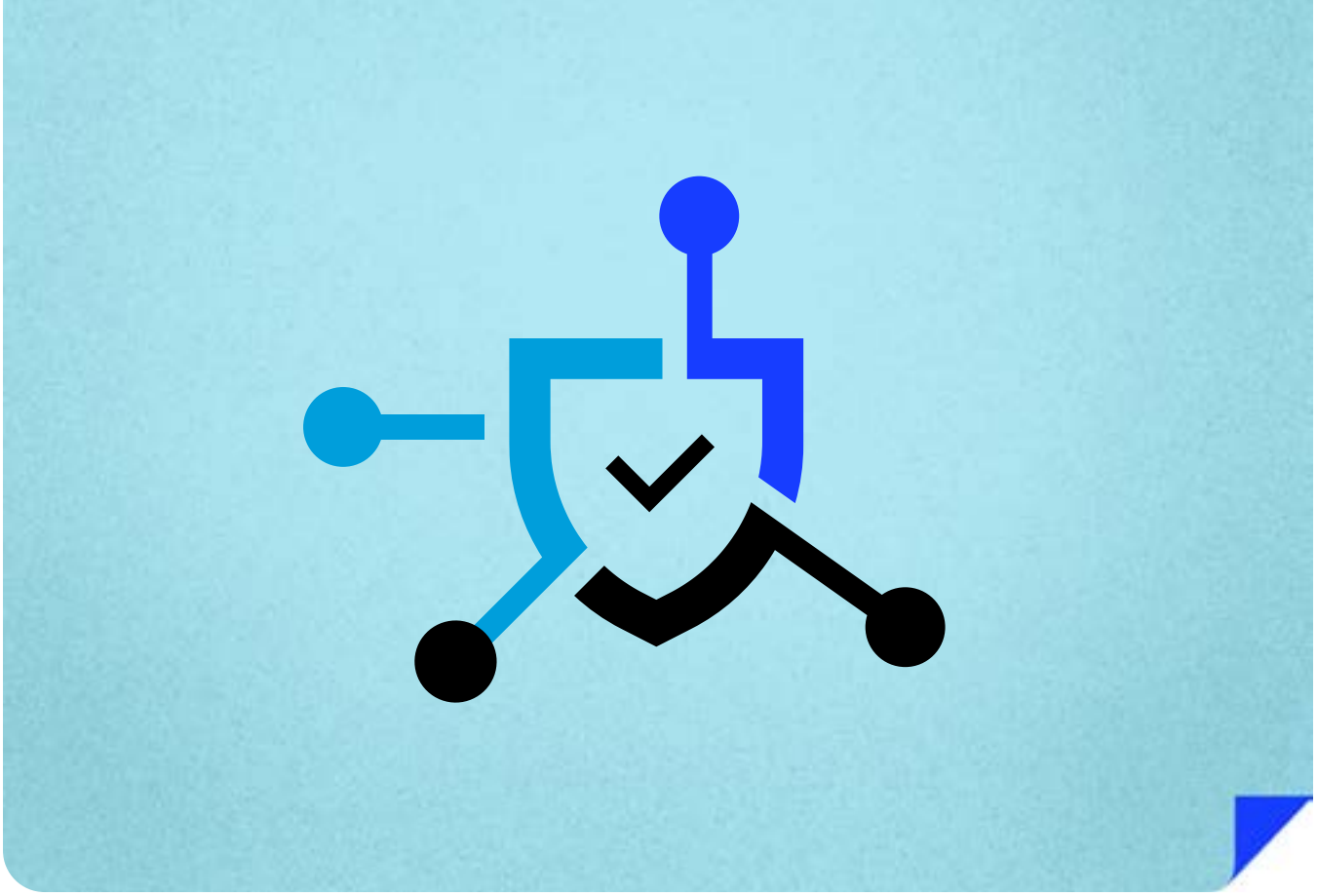
公众赋权

独立、自由和多元化的媒体

社会信任和韧性

透明度与研究





社会信任和韧性

整个社会的信任和韧性是信息完整性的关键组成部分。在此背景下，信任是指人们对所获取信息（包括官方来源和信息）的来源和可靠性的信心，以及对允许信息在整个生态系统中流动的机制的信心。韧性是指社会应对信息生态系统中的干扰或操纵行为的能力。

信任和韧性很容易受到国家和非国家行为者行动的影响，这些行为者试图利用信息生态系统来获取战略、政治或经济利益。这些行动有时是经过广泛协调的，会造成一系列伤害，并危及人们批判性地评估科学和事实的能力。

大型科技公司在信息生态系统中拥有巨大的权力，对包括其他企业、广告商、新闻媒体和个人用户在内的利益攸关方在信息互动和获取信息的方式上施加了过大的影响。

人工智能（AI）技术的进步，如生成式人工智能，带来了以最低成本大规模制造信息空间风险的手段。人工智能生成或中介的内容自称是真实或原创的，可信度高，能引起情感共鸣，难以察觉，并能在算法驱动的平台和媒体上迅速传播。这有可能以指数形式造成、加速和加深信任缺失。

应对信息完整性的风险需要强有力的、前瞻性的和创新的数字信任和安全做法，并在不同语言和背景下一致执行。这些做法应反映处于弱势和边缘化境况中的群体的见解，他们受到的潜在伤害尤为严重。
妇女、老年人、儿童、青年、残疾人、原住民族、难民和无国籍人士、LGBTIQ+群体、以及在族裔或宗教上属于少数群体的人尤其需要考虑在内。

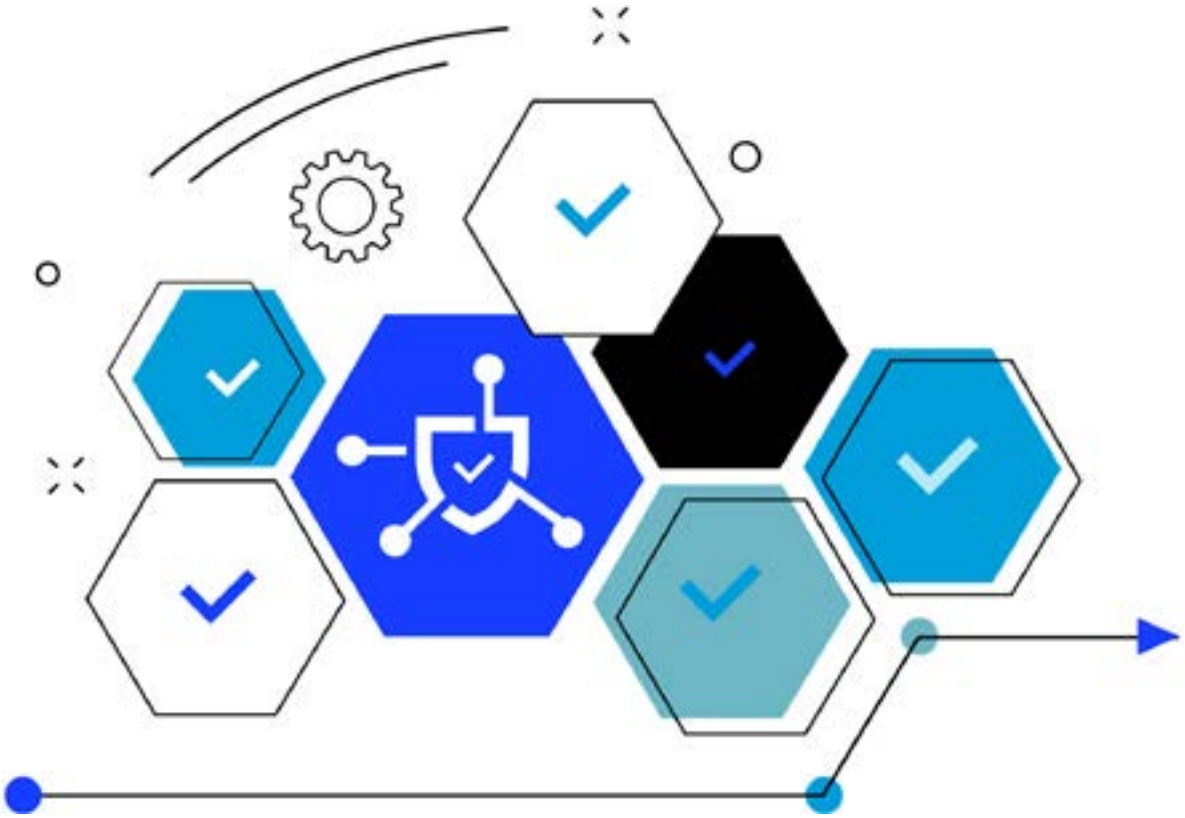
许多青年和儿童的大部分生活时间都是在网上度过的，他们从数字渠道获取大量信息。他们往往已经首当其冲

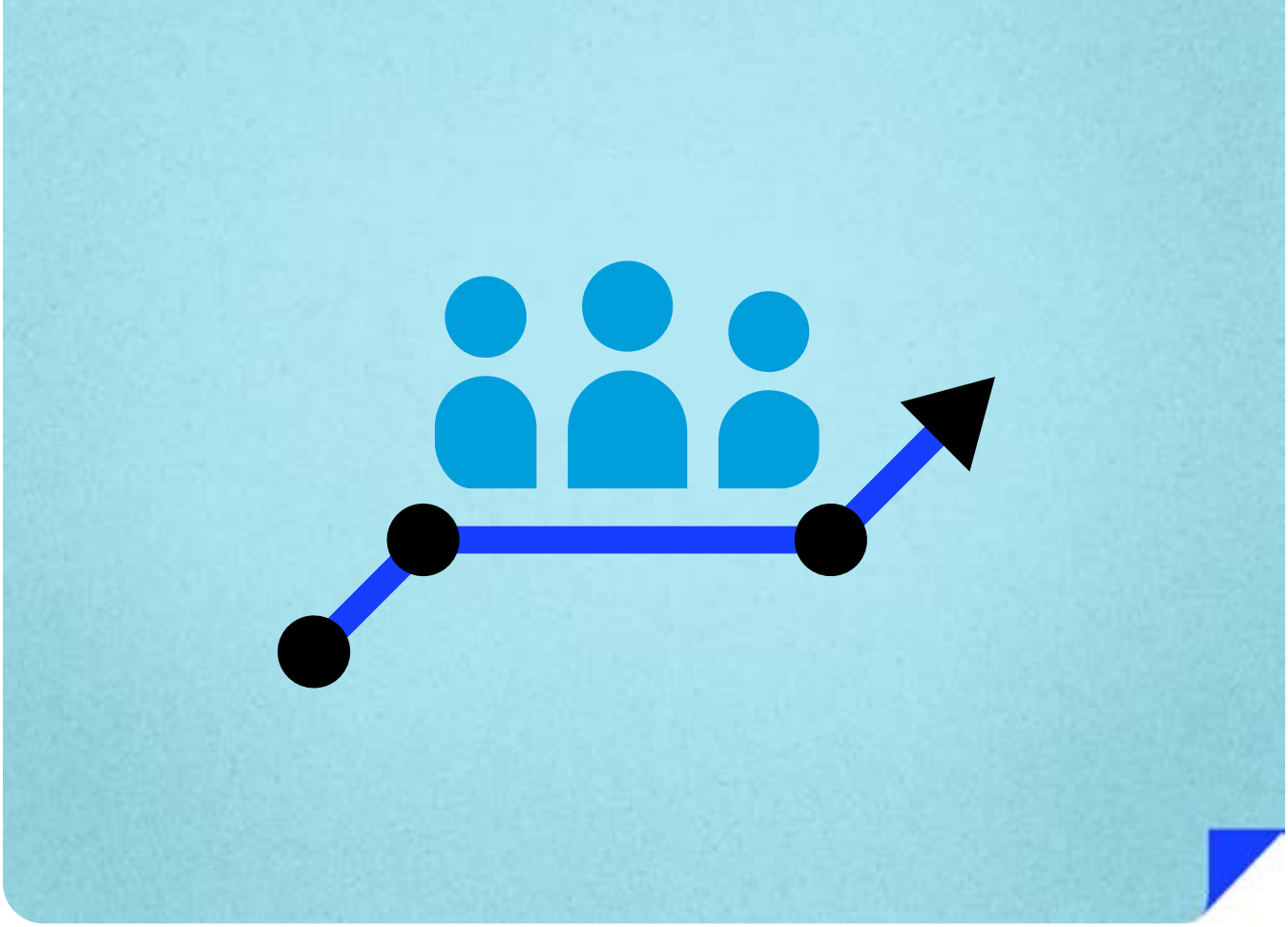
地受到信息空间风险的影响，并将最直接地受到新兴技术和媒体趋势的影响。

当人们能够获得各种信息来源，并感到被包容、平等、有社会经济保障和政治权力时，他们通常更有韧性，更有能力预先防范和驾驭这些风险。如果情况并非如此，这些风险往往会找到更肥沃的土壤来扩散。因此，应对措施应承认潜在的社会需求，以增强长期的韧性。

所有致力于公共利益的利益攸关方都可以利用信息空间为共同利益服务，努力适应传播环境不断变化的现实。在选举、自然灾害和人为危机等关键社会时刻，这一点尤为重要，因为在这些时刻，信息空间面临的风险显而易见，可能加深社会两极分化，削弱人们参与公共生活的能力，在极端情况下，还可能被用来煽动暴力。

活动家、记者、人道主义者、包括维和人员在内的联合国人员、选举工作人员、科学家、医疗专业人员等等都可能成为攻击目标，并可能带来可怕的后果。网络骚扰和其他阴险手段可能导致声音被压制，公民空间被缩小。保护这些人的一致努力至关重要。





健康的激励机制

建立健康的激励机制需要解决当前商业模式对信息生态系统完整性造成的严重影响，因为当前的商业模式依赖于有针对性的广告和其他形式的内容货币化作为主要的创收手段。

这些模式为各种规模的企业，首先是拥有和运营数字平台的技术公司，提供了前所未有的发展机遇，并催生了由无数人推动和造福的创造者经济。这些模式也为利用注意力经济进行虚假信息和仇恨传播的传播者提供了经济激励和机会，技术公司跟踪用户行为以收集数据，并将数据提供给算法，这些算法会优先考虑参与度，以最大限度地提高广告商和创作者的潜在收入。旨在分化和制造强烈情绪的信息往往能产生最多的参与度，结果就是导致算法奖励和放大有害内容。

利用这些商业模式的行为者包括信息操纵者和主流公共关系公司，他们与国家、政治人物和私营部门实体签订合同，提供精心策划的操纵活动，有时是跨国操纵活动。

技术部门将数字广告流程设计得既复杂又不透明，尽量减少人工监督。这对广告技术供应链中的许多参与者都有利，其中大型技术公司获利最大。

这种不透明的设计可能会导致广告预算无意中资助广告商可能无意支持的个人、实体或理念，从而对品牌构成重大风险。这些广告投放还会对广告活动的效果和品牌安全产生负面影响。

在广告技术领域占据主导地位的少数几家公司同时也负责在其拥有的平台上执行广告标准，而这些标准的执行可能是零散且不一致的。

这种对信息生态系统完整性的侵蚀凸显了从根本上转变激励结构的必要性。这可以通过以人权为指导的商业模式来实现，而不是依赖于基于行为跟踪和个人数据的算法驱动的定向程序化广告。

广告商可以通过一种既能加强信息完整性又具有良好商业意义的方式为信息生态系统造福。虽然技术公司不可能轻易放弃现有的商业模式，但通过提高广告商在广告流程中的透明度，以及广告提供商遵守对人权负责的广告政策，可以实现更健康的激励机制。通过加强对透明供应链的控制，广告商也能看到更好的投资回报。

公众赋权

个人在信息生态系统中的赋权要求人们能够控制自己的上网体验，能够对自己选择消费的媒体做出明智的决定，能够自由地表达自己的观点。公众赋权要求人们能够持续地获取各种可靠的信息来源。

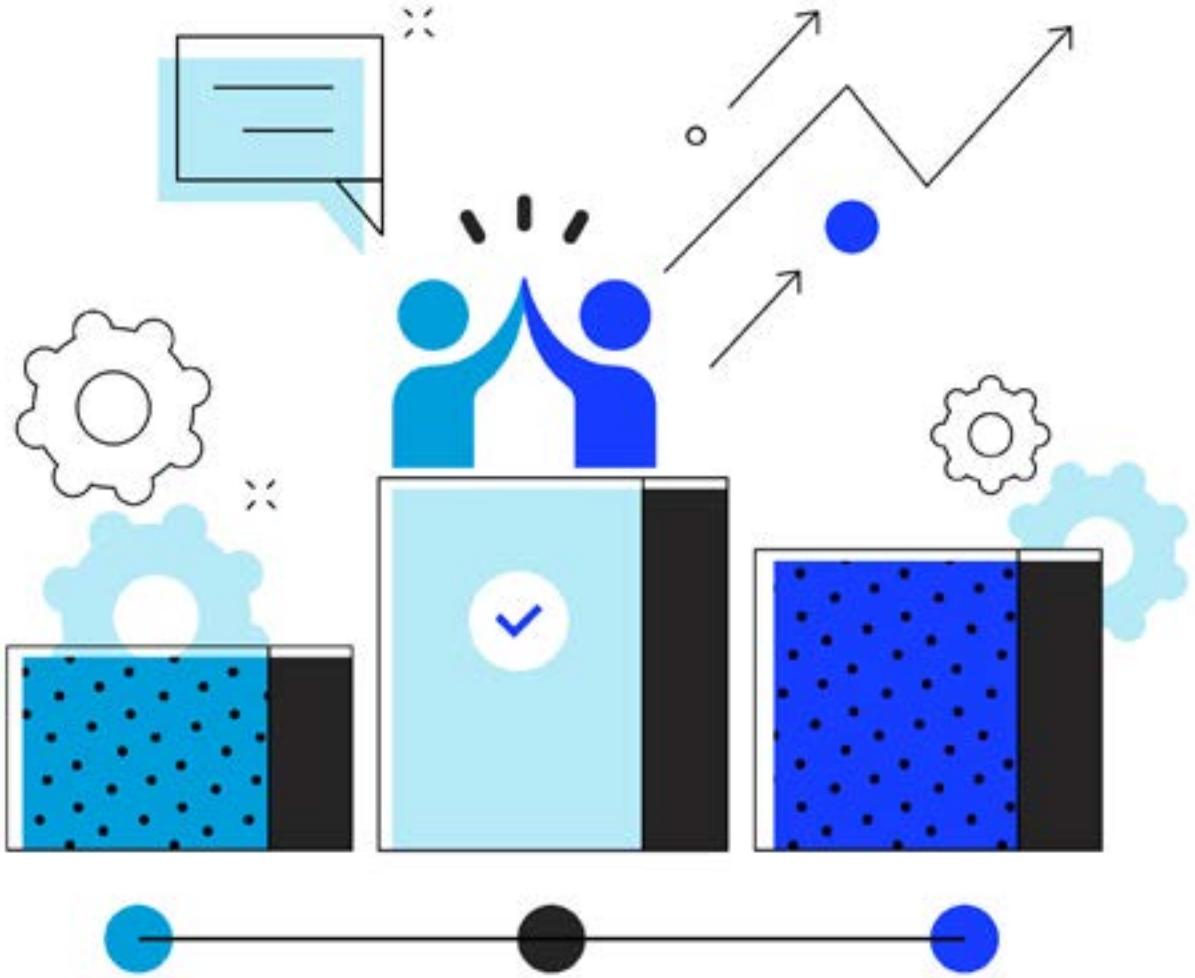
数字空间在许多方面都是包容性参与公共生活的催化剂，将人们跨越地域界限联系在一起，实现共同的进步愿望。如果善加利用，这些空间可以帮助增强个人能力，并赋予那些经常被排斥和边缘化的人以能动性。

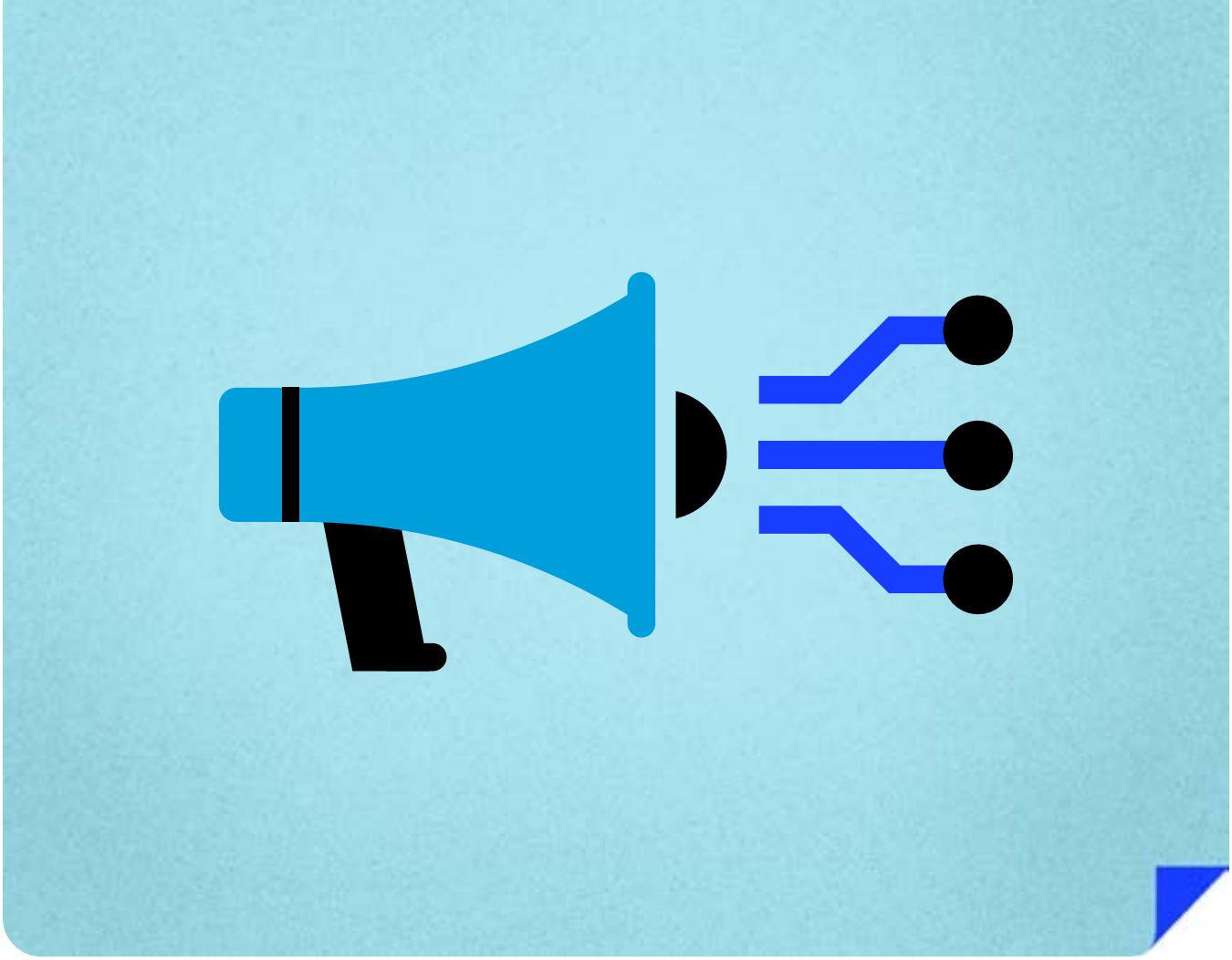
与此同时，数字技术也会阻碍真正的赋权。个人往往无法控制其个人数据的使用方式，也无法控制大型技术公司个性化的算法内容，并且在了解和获取信息提供者在优先考虑和推广特定类型内容时所使用的标准和机制方面面临障碍。

技术公司应授权用户就信任和安全、隐私政策和数据使用的各个方面提供意见和反馈，承认用户的隐私权。他们应加强用户控制和选择，包括与不同供应商提供的一系列服务的互操作性。

s媒体、信息和数字扫盲培训倡议应重点关注增强所有人的能力，特别是关注妇女、老年人、儿童、青年、残疾人以及处于弱势和边缘化境况群体所面临的具体挑战。

虽然互联网连接在不断增长，但世界上仍有三分之一的人处于离线状态。即使对那些上网的人来说，接入不足也会阻碍他们充分利用互联网资源的能力，使他们容易遭受信息空间的风险。随着连通性障碍的迅速减少，需要采取一些措施来增强新互联网用户的权能，并让那些缺乏上网机会的人掌握必要的数字扫盲技能，以获得安全和富有成效的上网体验。





独立、自由和多元化的媒体

只有独立、自由和多元化的媒体才能实现信息的完整性。

新闻自由是法治的基础，是民主社会的基石，使公民能够在知情的情况下进行讨论，追究权力责任，保护人权。只要记者和媒体工作者——包括妇女和那些处于弱势和边缘化境况的人——能够始终自由地进行报道并安全、公开地开展工作，而且所有人都能始终获得多元、可靠的新闻来源，那么新闻就可以被视为是自由的。

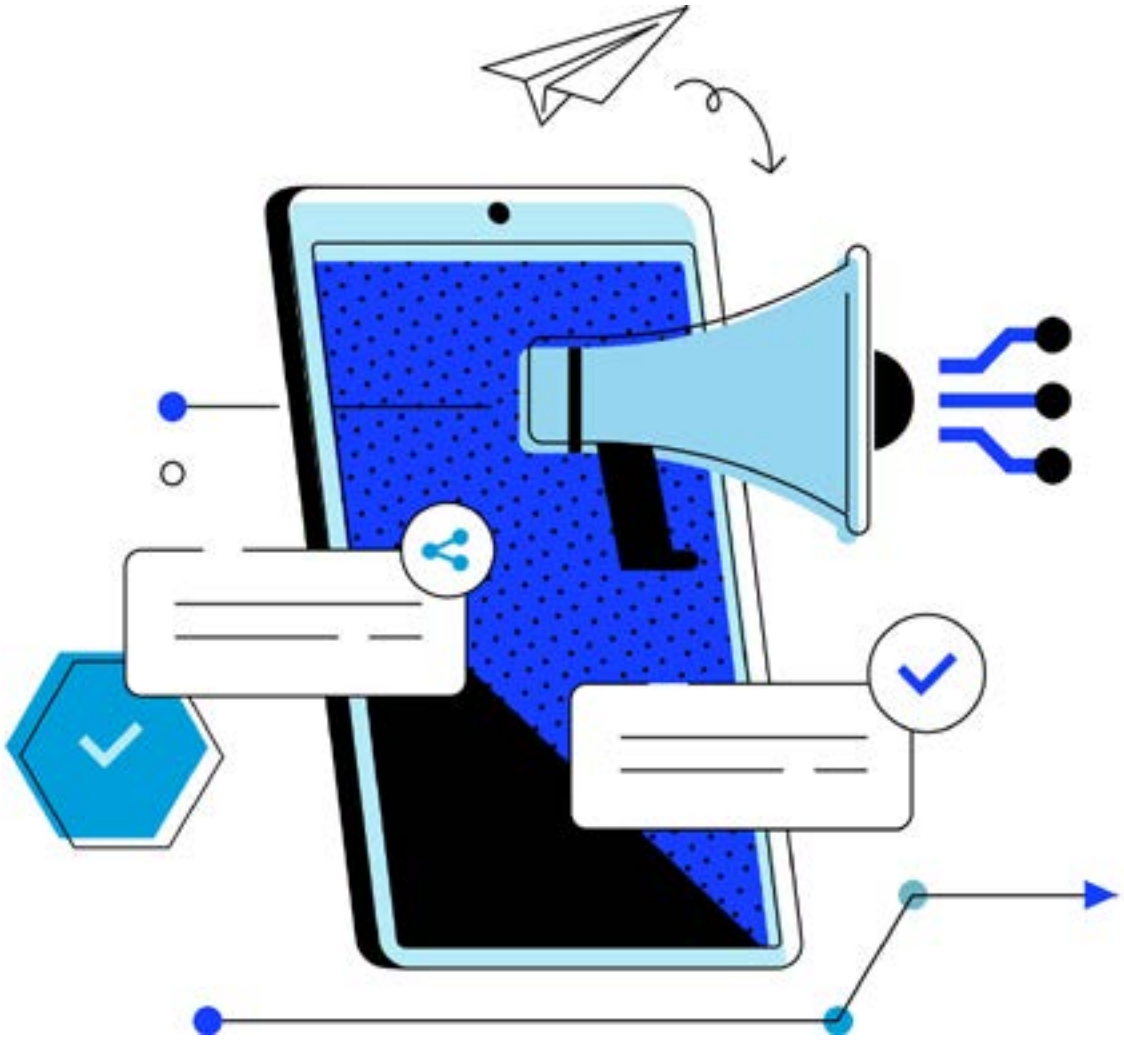
媒体在提供可靠和准确的信息以及降低信息空间的风险方面具有特殊的作用和责任。然而，尽管有表达自由的权利，包括自由、不受审查和不受阻碍的新闻或其他媒体的权利，新闻自由在世界各地面临着重大和持续的威胁。媒体工作者面临在线和离线骚扰、威胁和暴力，有时导致自我审查，增加了职业风险。

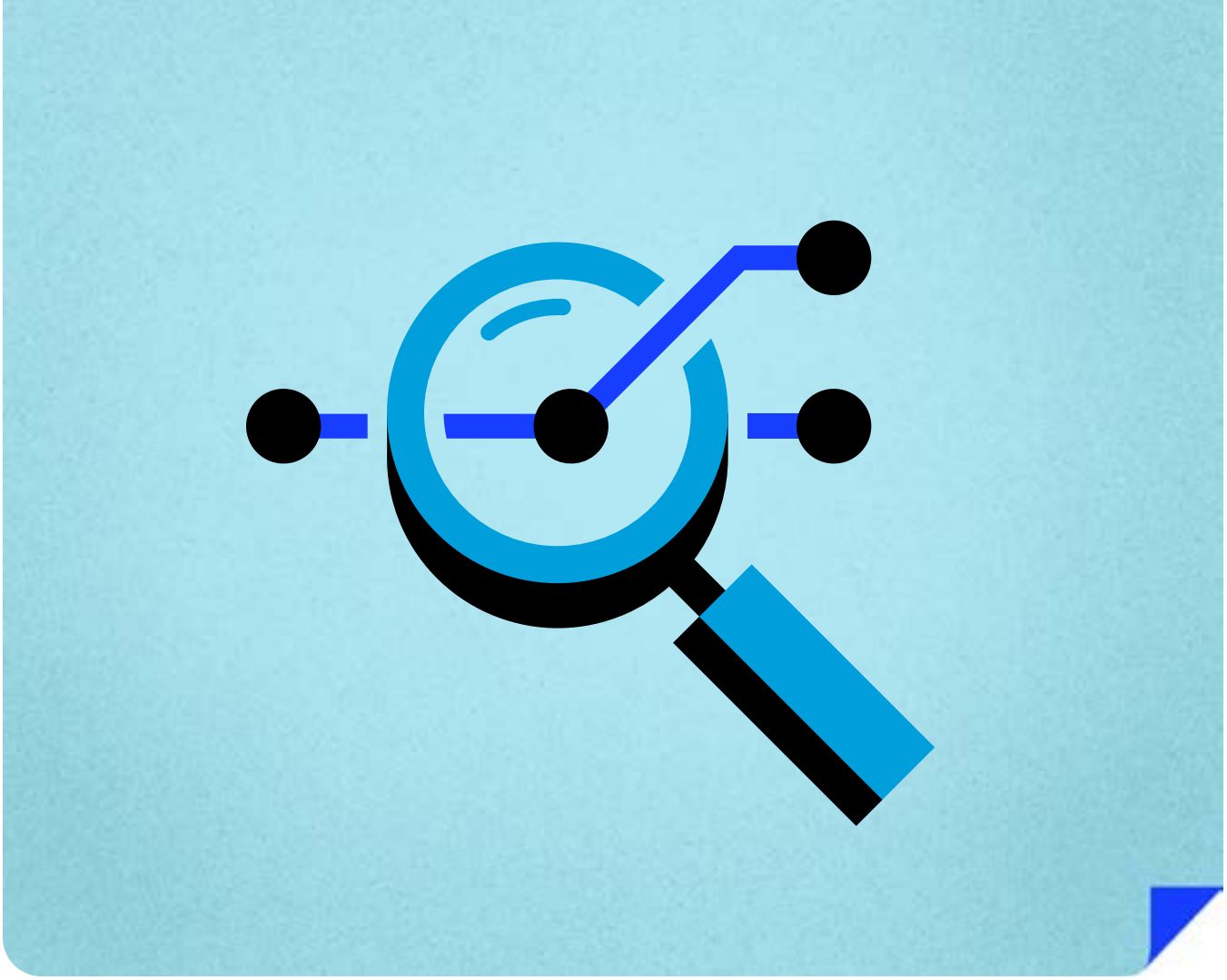
与此同时，由于广告收入向由大型技术公司主导的数字领域转移，新闻业也受到了影响。这些因素使得企业利益进一步加强了对媒体机构的控制，威胁到媒体的

多样性，破坏了地方和公益新闻事业。如果编辑标准得不到强有力的维护，媒体机构就会推动和扩大信息完整性的风险，这些风险可能在在线和离线空间之间交叉传播。

需要采取强有力的紧急应对措施，支持公益新闻组织、记者和媒体工作者，同时承认在媒体基础设施有限的情况下，公民记者为当地选民提供了重要服务。这种应对措施可包括利用当地执行人员提供强有力的、持续的媒体发展援助。

国家和技术公司在影响信息流动和政策方面具有相当大的影响力，应加强努力，确保新闻自由和记者的永久安全。





透明度与研究

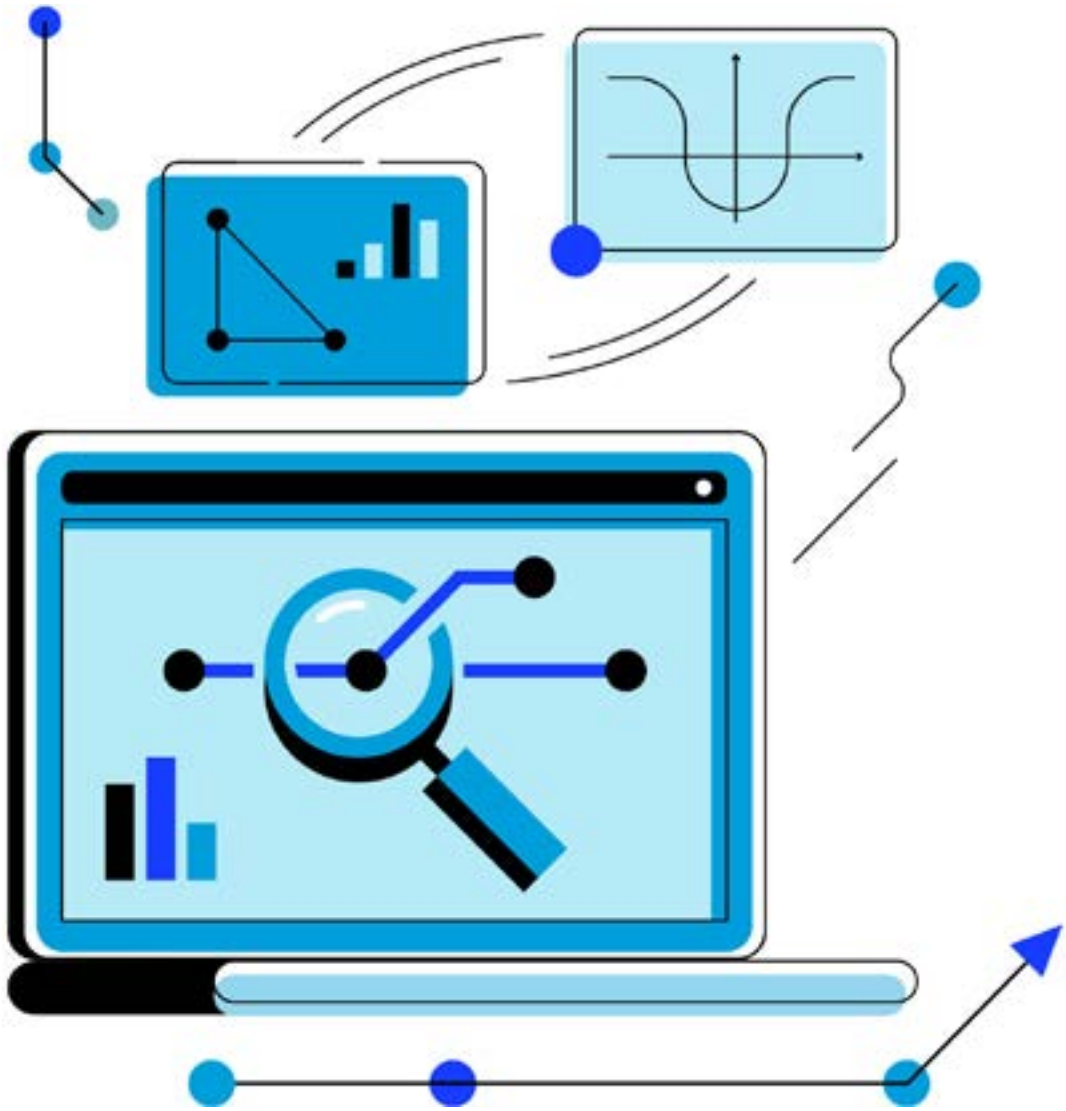
提高技术公司和其他信息提供者的透明度，可以让人们更好地了解信息是如何传播的，个人数据是如何使用的，以及如何应对信息完整性的风险。

然而，权力的不平衡为透明度设置了障碍。少数技术公司可以获取前所未有的大量数据，并与一些媒体所有者一道，在信息生态系统中拥有重要的控制权，有时还与国家、政治和经济行为体关系密切。

此外，大多数技术公司总部所在的少数国家在透明度方面做出的监管选择深深影响着世界其他国家。这些不平衡往往会限制公益研究，阻碍确保公平和满足服务不足、研究不足的环境和社区需求的努力。人工智能技术的部署（其全部影响仍不得而知）为研究和了解信息生态系统增添了更多挑战。

要在全球范围内更细致地了解信息环境，加强有针对性的循证行动，促进信息的完整性，就必须扩大数据和见解的可获得性、质量和可用性。

确保为各类研究人员提供保护隐私的数据访问，将加强填补研究空白和不平等的集体努力。必须保护和支持学者、记者和民间社会在没有恐惧或骚扰的情况下开展重要工作。



行动呼吁

以下建议旨在将五项原则转化为整个信息生态系统利益攸关方的可操作步骤。作为一个整体蓝图，这些建议从国家的法律义务到技术部门的责任，再到媒体和公民社会的最佳做法。

对利益攸关方的建议

- 科技公司
- 人工智能（AI）行为者
- 广告商和其他私营部门行为者
- 新闻媒体
- 研究人员和民间社会组织
- 国家和政治行为者

科技公司

大型科技公司，许多总部设在技术监管有限的地方，拥有巨大的权力。它们从大量收集到的用户行为数据中获利，从而能够塑造跨国信息流，控制全球范围内的数字体验。

为了纠正这种权力失衡，需要建立一个既重视透明度又重视独立监督的框架。用户有权控制自己的数据和在线体验，并有明确的投诉和补救渠道。需要建立问责机制，使技术公司对其产品和服务的设计和使用对人权和社会凝聚力（包括在危机和冲突局势中）造成的后果负责。这就需要对平台架构进行批判性的、透明的评估，以确定侵蚀信息完整性和损害人权的特征。应在保障表达自由和信息获取权的同时，实施防止和减轻此类侵蚀的战略。

虚假信息和仇恨不应产生最大限度的曝光和巨额利润。商业上可行的新商业模式，不依赖于有针对性的程序化广告，可以促进创新、增强用户能力并服务于公众利益。这种多层面的方法可以创造一个更加平衡的信息生态系统，尊重用户权利，营造一个值得信赖的网络环境。

建议

a. 从设计到交付，整合安全与隐私。将健全的安全和隐私政策纳入所有产品和服务的整个生命周期，包括设计、开发、交付和退役的每个阶段，对人类和人工智能生成的媒体一视同仁地适用政策。与独立的第三方组织合作，对所有产品和服务进行持续的人权风险评估，并将评估结果公之于众，以积极主动地将社会风险降到最低，减轻潜在危害，包括在关键社会时刻前后。采取措施，保护弱势和边缘化群体、民间社会成员和其他经常在网上成为攻击目标的人，并赋予他们权力；解决通过使用技术而发生或因使用技术而扩大的基于性别的暴力和其他形式的暴力。创新应对新出现的挑战，包括人工智能技术对信息生态系统完整性造成的潜在风险。确保产品开发各个阶段的人员配置以及信任和安全团队的多样性和包容性。建立内部信息共享程序，确保风险和政策评估在公司各级和各职能部门（包括领导层）得到共享和集体理解。确保始终如一地执行所有信任与安全政策。

b. 重新评估商业模式。评估平台架构是否以及如何助长对信息生态系统完整性的侵蚀和对人权的损害，并在尊重表达自由的前提下采取相应的缓解和补救措施。推广创新的、商业上可行的商业模式，不依赖于有针对性的程序化广告，并服务于公众利益。

c. **保护儿童。制定并执行保护和维护儿童权利的措施，如年龄验证和家长控制。** 实施相关政策和做法，预防和打击通过技术手段或使用技术手段对儿童进行性剥削和性虐待的行为。建立并宣传针对儿童的特别报告和投诉机制。

d. **分配资源。根据风险程度分配足够的、持续的、专门的内部信任与安全资源和专业知识。** 划拨充足的资源，以应对社会文化语言环境和行动语言，以及处于弱势和边缘化境况的群体的不同需求，尤其是在经历冲突或面临不稳定局势的情况下。

e. **确保内容审核的一致性。与独立的第三方组织合作，制定符合国际人权标准的内容审核程序，并确保在各业务领域一致、非任意地执行这一政策。** 为人工和自动内容审核和整理分配足够的资源，并在所有语言和运营环境中一致应用。采取措施处理违反平台社区标准和损害人权的内容，如限制算法放大、贴标签和去中心化。公开提供分类数据，说明内容审核政策的执行情况，以及为不同语言和运营环境的内容审核分配的资源。

f. **坚持劳工标准。** 提供符合国际劳工法和人权法的工作条件，并优先采取各种措施，确保参与信任与安全工作的所有工人（包括内容审核员）的福利、安全和优质培训。

g. 建立独立监督。 定期进行外部人权独立审计，内容包括服务条款和社区标准；信任和安全以及广告政策；风险管理；广告和推荐系统对不同语言和业务环境的影响；内容节制；投诉和申诉程序；透明机制；以及研究人员的数据访问。评估产品和服务对弱势和边缘化群体、性别平等和儿童权利的影响。将这些审计结果公之于众，让所有用户都能获取和理解。

h. 制定行业标准。 与民间社会和其他利益攸关方合作，共同制定行业问责框架，明确界定角色和责任，承诺进行经审计的公开报告和独立监督，并遵守有关隐私、透明度、风险管理、信任和安全的严格标准。为处于弱势、边缘化和脆弱境况中的群体的需求做出具体规定，建立衡量和应对人权风险的有效方法。确保平台和服务之间的合作，认识到风险可能在各种信息空间蔓延，而每个空间都有可能被利用的独特设计缺陷和政策漏洞。

i. 提升危机应对能力。 与在高风险地区开展工作的利益攸关方合作，建立预警和升级程序，在危机和冲突情况下加快及时响应的速度。建立机制，使人们能够及时获取符合公众利益的可靠、准确的信息。

j. 支持政治进程。 在选举和其他政治进程前和整个进程中，对所有产品和服务进行人权风险评估，并公布于众。执行所有相关政策，维护信息的完整性，采取措施解决虚假信息、骚扰和暴力侵害妇女及其他公共生活中的常见目标群体（包括政治候选人）的问题。

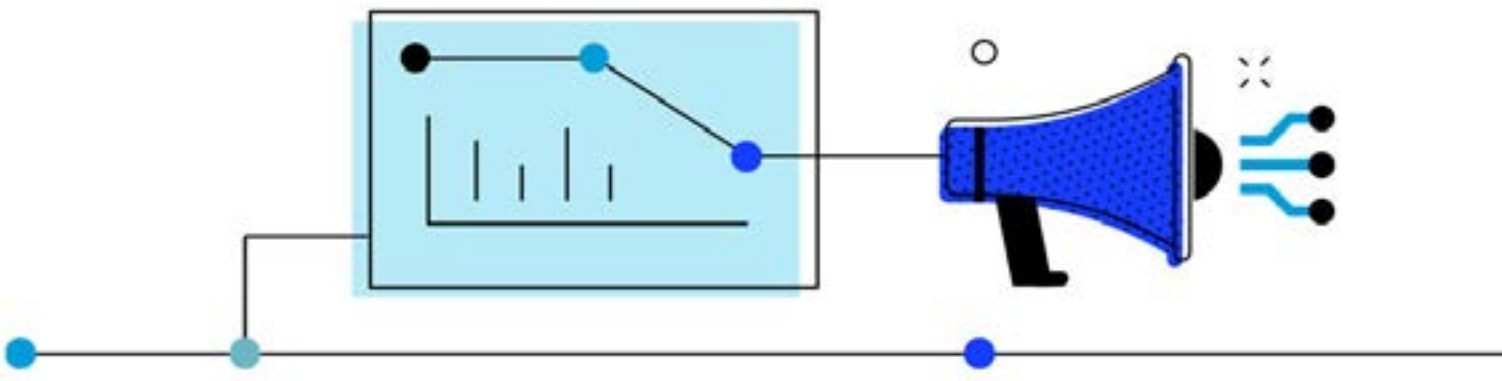
k. 与利益攸关方合作。 积极主动地与各种利益攸关方接触，包括国家、学术界、民间社会、儿童、青年组织和技术界，以便更深入地了解信息生态系统完整性面临的风险，并相应地增强和调整信任与安全机制。

l. 建立健全的投诉机制。 确保用户和非用户的投诉、报告、申诉和补救机制透明、安全、可靠和可获得性，包括为那些处于弱势和边缘化境况的人制定特殊程序。建立并执行防止滥用报告和投诉机制的程序，例如通过协调的不真实行为。

m. 传达明确的政策。 使条款和条件、政策、社区标准和执行程序易于获取、保持一致以及可以理解（包括对儿童而言）。明确有关新闻和政治内容的所有政策、指导方针和规则。

n. 执行广告政策。 制定、宣传并执行明确而有力的广告和内容货币化政策。持续审查现有的出版商和广告技术合作伙伴关系，以评估广告技术供应链中的合作伙伴是否坚持这些政策。每年公开报告政策执行的效果和采取的行动。

o. 展示广告透明度。 明确标注所有广告，使广告商信息、目标定位参数以及任何使用人工智能生成或中介的内容对用户透明。维护完整、可访问、最新且可搜索的广告库，其中包含有关来源或购买者、所花费用以及目标受众的信息。向广



告商和研究人员提供详细数据，说明广告在任何特定时间范围内出现的确切位置，以及围绕广告投放和品牌邻近性的控制和服务的准确性和有效性。就收入来源以及与广告商和内容创作者的分成安排进行透明的报告。对所有政治广告进行明确标注，包括标明由人工智能生成或中介的内容，并提供易于获取的信息，说明广告为何针对目标受众、谁为广告付费以及付费多少。

p. 支持媒体安全和多样性。 为多元化新闻内容的传播创造有利环境，使消费者能够接触到各种媒体来源。支持独立、自由和多元化的媒体，特别是以不同语言和背景开展的地方和公民新闻报道，同时尊重编辑独立性。采取一切措施维护记者和媒体工作者的在线权利。制定明确、透明的规定，帮助保障记者和媒体工作者免受骚扰、虐待和暴力威胁，反映记者面临的风险，尤其是在选举、自然灾害和人为危机等关键社会时刻。更新信任与安全政策和做法，以减少和解决针对女性记者的问题。

q. 提供数据访问。 为研究人员，包括各学科的学者、记者、民间社会和国际组织提供所需的数据，以便更好地了解信息的完整性，为政策和最佳做法提供信息，并加强问责制，同时尊重用户隐私和知识产权。这些数据应进行分类，以便有效研究信息生态系统的完整性，包括社会风险、对不同社区和人群的影响、使用人工智能技术的影响、对实现可持续发展目标的潜在影响以及风险缓解措施的有效性。它应包括以下方面的信息：算法驱动的推荐系统，包括解释如何训练算法对内容进行排序、推荐、分发和标记；删除、禁用或降级账户；以及跨语言和跨背景的信任和安全资源分配。促进以最低成本为研究人员提供可访问、机器可读格式的数据。

r. 确保披露。公开国家对内容删除或放置的要求。披露与事实核查组织的所有合作，包括提供的资金或其他支持；以及向政治机构和候选人提供的资金。

s. 提供控制和选择。提供用户友好型工具、功能和特性，确保知情同意，使人们能够轻松控制自己的在线体验，包括通过与其他服务的互操作性，允许人们有更多选择，并对他们看到的内容以及如何和在哪里使用他们的数据提供知情同意。

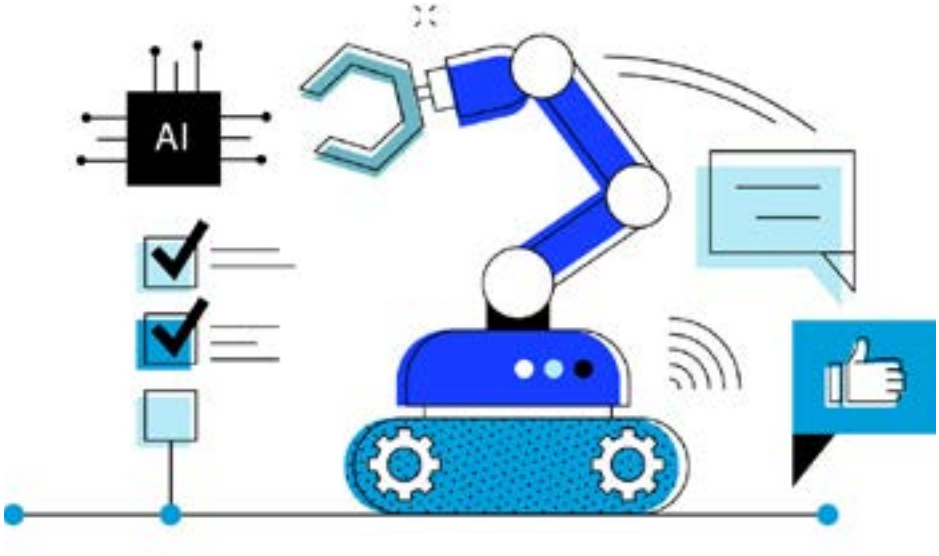
t. 给人工智能内容贴标签。明确标注人工智能生成或中介的内容，在组织层面投资并开发解决方案，以确保用户可以轻松识别此类内容，更广泛地加强而不是削弱用户对信息生态系统完整性的信任。这包括在元数据中标明此类内容为人工智能生成或中介的信息。

u. 确保隐私。确保数据的收集、使用、共享、销售和存储尊重用户隐私，确保用户可以轻松获取有关其个人数据如何被利用（包括用于算法决策）以及其个人数据如何与其他实体共享和从其他实体获取的信息。

v. 促进数字扫盲。支持媒体和信息扫盲活动，提高数字技能，包括提高公众对算法的功能、效果和影响的认识。为所有语言和行动区，特别是脆弱地区，提供扫盲和能力建设资源。为儿童和青年提供与安全相关的培训材料。对扫盲行动的效果进行独立的外部评估，并公布评估结果。

人工智能(AI)行为者

参与人工智能系统生命周期至少一个阶段的政府、私营和公共部门行为者



随着人工智能(AI)技术的快速发展,这些能力将重塑我们的世界。从日常任务自动化到帮助科学发现,潜在的好处是巨大的。然而,在取得进步的同时,我们也亟需确保人工智能的设计、开发、部署和退役安全可靠。

训练数据中的偏见和多样性缺乏会导致人工智能系统生成误导性信息,并使不公平现象长期存在。这种生成真实内容的能力可能会被大规模滥用,给信息生态系统的完整性带来风险。

可以通过优先考虑人工智能技术生命周期的透明度和公平性来降低新出现的风险。政府、技术公司和学术研究机构需要通力合作,确保人工智能在其整个生命周期内安全、负责任地设计、开发、部署和退役。通过共同努力,这些利益攸关方可以确保人工智能技术造福于社会和人类福祉。

建议

a. **确保安全、可靠和可信的人工智能。** 采取措施确保人工智能技术的设计、开发、部署、使用和退役安全、可靠和可信。处理并公开宣传该领域任何可能对信息生态系统的完整性构成风险的创新或进步所带来的影响，包括恶意使用人工智能技术、在没有人为监督的情况下过度依赖人工智能技术，以及在不同地域和社会背景下进一步削弱信任的任何相关可能性。在对公众福祉至关重要的问题上，对人工智能进行可靠、包容的信息来源培训，并采取措施减少培训数据产生的偏见，包括性别和种族偏见。与不同的利益攸关方合作开展人权风险评估，积极主动地将社会风险降到最低，减轻潜在的伤害，包括对妇女、儿童、青年和其他处于弱势和边缘化境况的群体的伤害。

b. **委托独立审计。** 承诺向机构和个人研究人员提供对人工智能模型进行独立审计的途径以及法律和技术安全港，并采取适当的保障措施，如遵守公司漏洞披露政策。确保公众可以获取独立审计的结果、与人工智能系统相关的风险数据——例如可能出现的有害歧视和“幻觉”，即看似真实但完全是编造的内容——以及为预防、减轻和解决潜在危害而采取的措施。

c. **R尊重知识产权。** 尊重知识产权，确保对训练人工智能工具时使用的知识产权（包括原创新闻）进行公平补偿。

d. **显示数据出处。** 通过可见和不可见的形式，如真实性认证、水印和标签，制定和实施有关出处的解决方案和政策。与多方利益攸关方共同努力，实现用户友好型标签的标准化。

e. **支持扫盲。** 在组织层面投资于制定和部署扫盲计划，以提高公众对人工智能模式如何运作以及对全球信息消费者的影响的认识，重点关注信息完整性的风险。

f. **实现用户反馈。** 为用户提供提醒或报告不准确或误导性出处信息的功能，同时保护用户隐私。

广告商

广告商可以对信息生态系统的完整性施加单一的影响，帮助切断那些试图从虚假信息和仇恨中获利的人的经济动机。这样，广告商就能更好地保护自己的品牌，应对重大风险，在根据企业价值观开展业务的同时提高自己的底线。

建议

a. 建立对人权负责的广告。制定保障措施，确保广告不会给信息空间带来风险，并维护人权，包括儿童的权利。避免基于敏感数据和用户特征的歧视性定位做法。通过纳入和排除名单、广告验证工具和人工审核等方法，与支持信息完整性（包括公益新闻）的媒体机构和平台合作发布广告。要求广告技术公司公布网站或频道在盈利前必须遵守的标准。

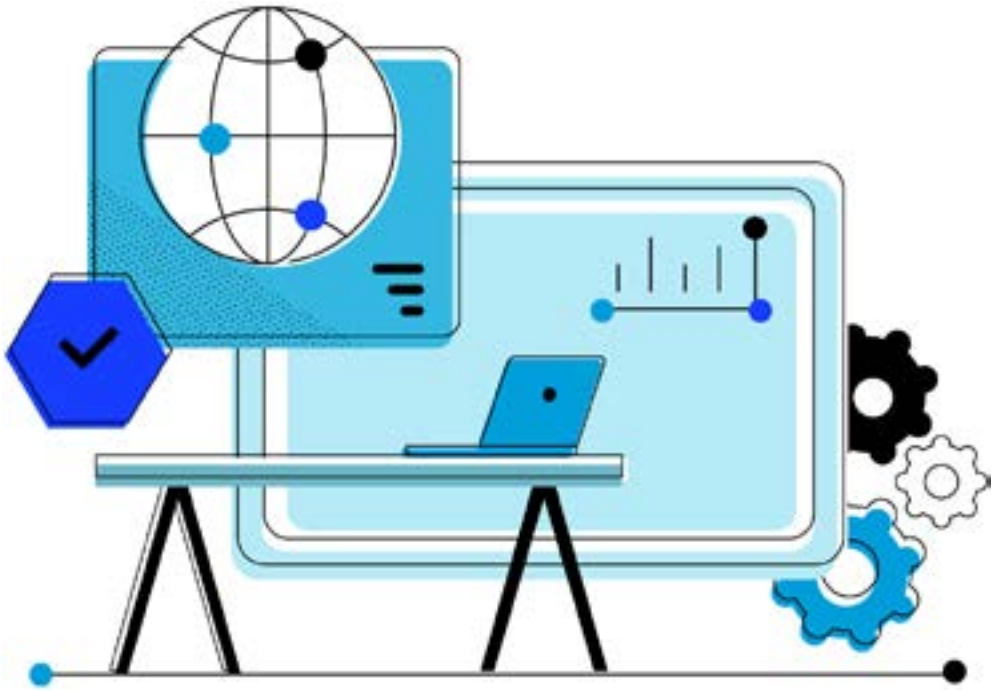
b. 利用行业标准。利用行业标准制定明确的政策，最大限度地降低信息完整性风险，帮助确保品牌安全。

c. 形成联盟。跨行业合作，并与民间社会合作，及时分享有关信息完整性的最佳做法和经验教训，包括评估广告的影响，以及系统性地降低广告和内容货币化带来的风险和潜在危害。

d. 要求提供数据。持续建立全面详细的广告邻近性概览，要求提供显示广告出现位置的细粒度数据，并在广告投放前进行适宜性审查。对广告活动进行全面审核。

e. 强制透明。要求广告技术公司采用透明标准，对广告技术供应链进行端到端验证，并与客户和研究人员共享完整的广告活动数据，包括日志级的投放和屏蔽数据。

f. 进行审核。要求广告技术公司对广告交易供应合作伙伴进行独立的第三方审计和审查 。



其他私营部门行为者

不直接涉及技术部门的更广泛的私营部门实体的行动也会影响信息空间，既削弱信息完整性，也支持信息完整性。企业有责任尊重人权，包括表达自由权和知情权，并可与其他利益攸关方结成合作伙伴关系，帮助实现更健康的信息生态系统。

建议

a. 维护完整性。维护人权，包括言论和意见自由的权利，不为经济或任何其他战略目标而故意传播或支持危及信息生态系统完整性的风险。

b. 投资扫盲。与相关民间社会合作并利用其专业知识，在组织层面对人员的媒体和信息素养进行投资。

新闻媒体

独立、自由和多元化的媒体在向公众提供有关公共利益的信息、促进公民参与和推动当权者问责方面发挥着至关重要的作用。

然而，对媒体独立性、自由和多样性的直接和间接威胁，以及地方和公共利益新闻业的衰落，都会破坏这些重要职能。如果不严格遵守专业标准，新闻媒体就会损害信息的完整性。通过合乎道德的报道和编辑做法以及对透明度的承诺，并辅之以高质量的培训和工作条件，新闻记者提供了不可或缺的服务，并能在信息生态系统的完整性面临风险时帮助恢复平衡。

建议

a. 报道信息的完整性。 投资于数据驱动型和调查型新闻报道的能力建设，主动报道并向公众通报信息生态系统完整性面临的风险。采用强有力的编辑流程和标准，包括在信息来源方面，以帮助维护和确保媒体消费者的信任。建立事实核查机制，供公众参考。

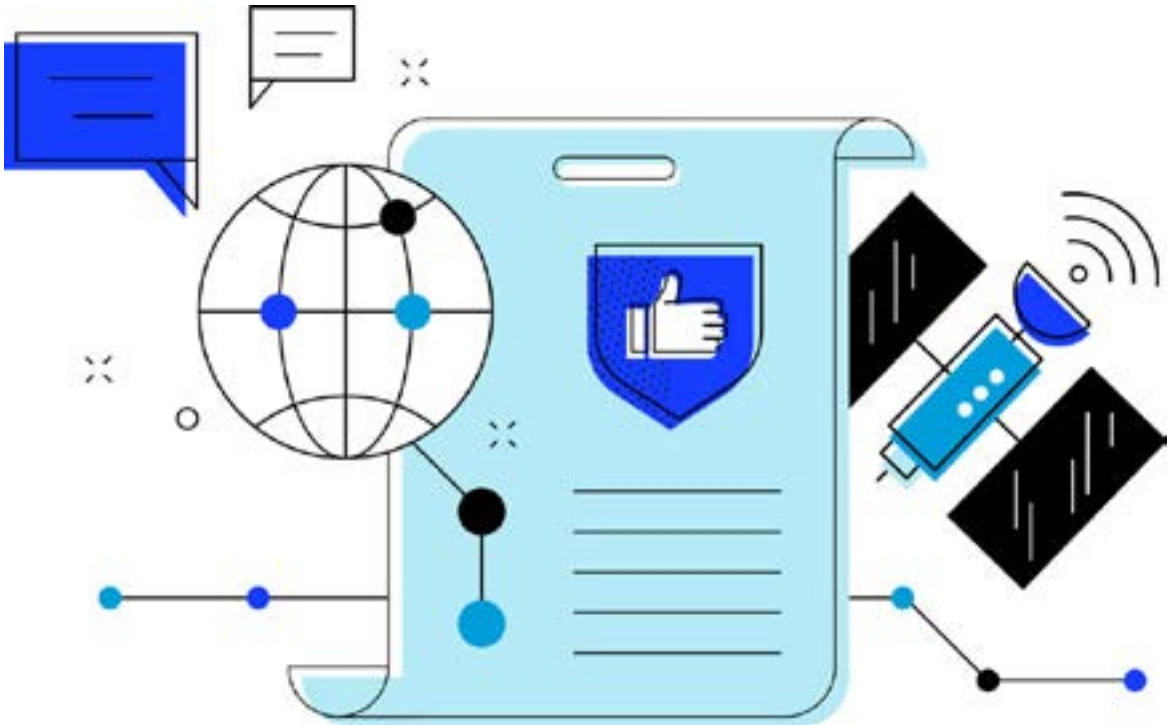
b. 提供危机响应。 在紧急和危机情况下，当信息生态系统的完整性面临更大风险时，承诺向公众提供免费的及时信息。

c. 保持专业和道德标准。 承诺并遵守为公众利益而制作的全球公认的专业和伦理新闻准则和做法，强调公正性和编辑独立性，并积极采用自律问责机制。定期提供高质量的培训，以提高报道的道德性、准确性和公正性，更新技能， 促进创新，适应传播环境的变化，包括采用“解决方案或建设性”的新闻报道方法。公开资金来源、所有权结构和财务激励机制，使个人能够更好地了解他们选择和消费的新闻。

d. 合乎伦理地使用人工智能。 建立健全的人工智能技术伦理使用政策，包括在出版或广播时明确标注人工智能生成或中介的材料。这包括在元数据中标明这些内容是人工智能生成或中介的信息。

e. 建立透明的、对人权负责的广告。 采取措施确保广告不会给信息空间带来风险。明确区分新闻、观点和赞助内容，确保观点文章的资金来源和潜在利益冲突的透明度。明确标注所有付费的、人工智能生成或中介的广告和广告内容。提供有关广告收入来源的透明报告以及清晰、易懂的广告政策和做法。

f. 尊重劳动标准。 建立符合国际劳动法和人权法的工作条件，优先考虑有助于确保记者福利和安全的举措（包括在数字空间），并特别关注对女性记者和媒体工作者的歧视、虐待、骚扰和暴力威胁。



研究人员和民间社会

研究人员和民间社会组织在理解和应对信息生态系统完整性所面临风险的多方面影响上发挥着关键作用。他们的努力有助于揭示信息空间的风险，加强宣传工作的证据基础，促进韧性，特别是对处于弱势和边缘化境况的群体而言。合作伙伴关系和知识交流对于弥合研究见解与有效解决方案之间的差距至关重要。

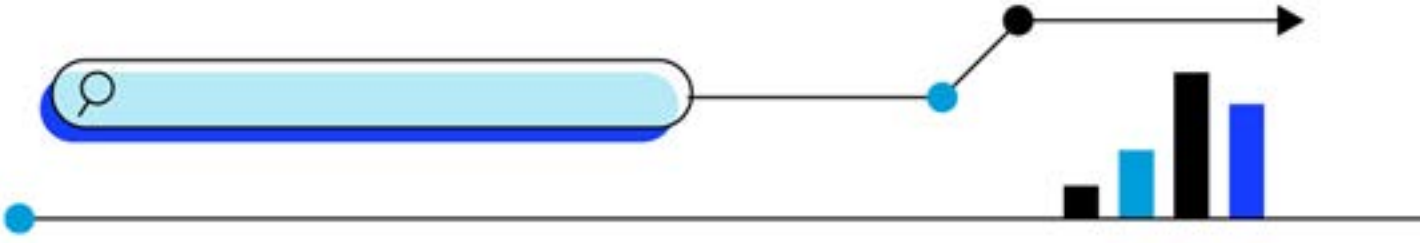
建议

a. 合作。与不同地域和背景的利益攸关方合作，分享有效和符合道德规范的方法，以加强信息生态系统的完整性。

b. 维护完整性和道德标准。维护人权，不故意传播或赞助危害信息完整性的行为。以道德、透明和注重隐私的方式开展所有研究。

c. 促进开放式获取。采取开放存取措施，免费提供研究成果，促进跨学科合作。

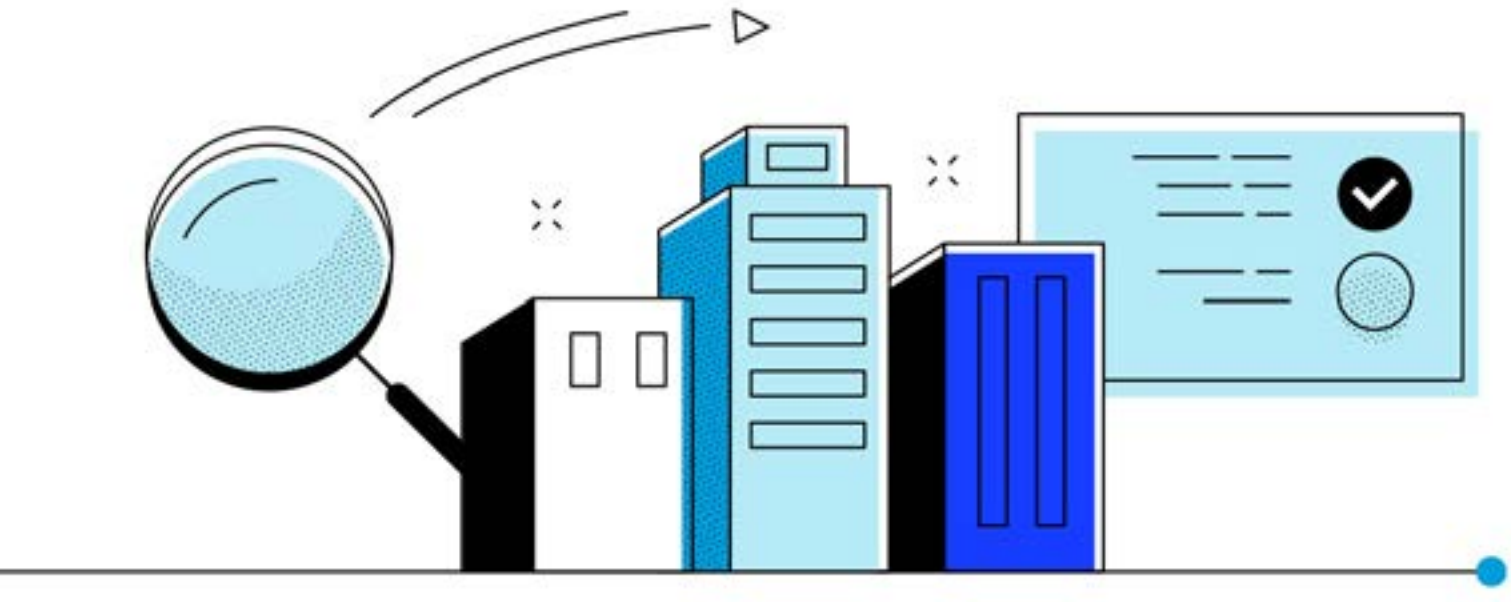
d. 加强包容性研究。探索跨地域、跨语言和跨专题领域的信息生态系统多学科研究，包括信息完整性风险对可持续发展目标的潜在影响，尤其关注研究不足、脆弱和边缘化的环境和社区。制定衡量此类风险和相关危害的严格方法。



对事实核查组织和网络的建议

a. 保持专业标准。恪守专业精神和职业道德，在组织构成和管理、资金来源、所有权和工作实践中坚持独立、无党派和透明的标准。

b. 公开资金来源。采取措施，公开披露资金来源以及与技术公司、媒体机构和民间社会组织等利益攸关方的任何合作。



国家

各国对加强《信息完整性全球原则》负有不可或缺的责任。首先，国家有义务尊重、保护和促进人权，特别是表达自由权，包括寻求、接受和传递信息的权利。

国家因其法律和监管权力、对公共资源的控制以及建立国内和国际联盟的能力等因素，在塑造信息空间方面发挥着核心作用。作为其人权义务的一部分，国家必须保护其领土和/或管辖范围内的人权不受企业侵犯，采取适当措施，通过有效的政策、立法、法规和裁决，预防、调查、惩罚和纠正此类侵权行为。

各国参与信息生态系统的技术和财政能力各不相同。基础设施以及获取技术和财政资源方面的差距造成了数字鸿沟。与此同时，许多大型技术公司虽然几乎实现了全球市场渗透并占据主导地位，但其总部却设在全球北方的少数几个国家。

为确保所有国家都能为信息生态系统做出贡献并从中受益，需要采取紧急和持续的举措，提高各国扩大数字连通性的能力，积极预防可能出现的“人工智能鸿沟”，并在尊重人权的同时，加强各国充分应对信息空间风险的能力。最终，这些努力将加强信息的完整性，促进人权，并有助于实现可持续发展目标。

建议

a. 尊重、保护和促进人权。根据国际人权标准和准则，尊重、保护和促进人权，尤其是表达和意见自由权，包括信息权。确保为解决信息完整性的各种因素而实施的法规或其他措施符合适用的国际法，包括国际人权法，公民社会能够充分参与，成为加强人权和建立信任的更广泛努力的一部分。确保对表达自由权的限制是例外情况，在实施限制时，必须符合国际人权法的要求， 即，由法律规定，为保护他人的权利或名誉，或国家安全、公共秩序、公共卫生或道德所必需，并符合适当性原则。确保限制措施不会在实践中扼杀表达自由。采用并有效执行符合国际法（包括国际人权法）的个人数据隐私保护措施。

b. 维护完整性。不在国内或跨国开展或赞助蓄意传播虚假信息或利用仇恨言论的信息活动。不以任何形式关闭或限制互联网。维护并执行联合国安理会相关决议，包括有关保护联合国和平行动免受影响授权执行的信息生态系统完整性风险的决议。

c. 保护民众。重申并加倍努力，确保在法律上和实践中保护处于弱势和边缘化境况的群体并赋予他们权力，这些群体往往是在线和离线信息空间的目标，如妇女或 LGBTQI+ 个人或少数族裔或宗教群体，同时应解决儿童的特殊需求和权利问题。遵守国际人权法规定的义务，依法禁止构成煽动歧视、敌意或暴力的战争宣传或鼓吹民族、种族或宗教仇恨的行为。

d. 提供获取信息的渠道。以所有人都能理解和使用的所有语言和形式，一视同仁地提供及时获取公开信息的渠道——包括新闻媒体，同时促进服务不足的社区获取信息。确保在危机情况下获取可靠、准确的信息。采用符合道德规范的可信传播方式，积极主动地与社区互动，建立对公共机构的信任。

e. 确保媒体自由。确保、保护和促进自由、可行、独立和多元化的媒体环境，采取有力措施保护记者、媒体工作者和事实核查人员，特别关注妇女以及弱势和边缘化群体成员，使其免受一切形式的歧视、虐待、骚扰和暴力威胁。在法律和政策上尊重和保护数字内容创作者和公民记者的权利。

f. 保护研究人员和民间社会。尊重学术自由，保护学者和民间社会免受恐吓、骚扰或报复行动。

g. 提供透明度。对技术公司和媒体机构提出的要求和数据请求提供充分的透明度。采取措施解决不透明和欺骗性的游说策略，以及技术公司与决策者之间破坏信息完整性的利益冲突，如不道德的雇佣行为和经济激励。

h. 加强全球团结、能力建设和发展援助。参与国家间的合作和伙伴关系，支持能力建设，以加强信息的完整性，提高对信息空间风险的抵御能力，特别是在发展中国家。以完全透明的方式分配财政资源，用于数字、信息和媒体扫盲和宣传计划的培训和能力建设，包括以所有语言开展人工智能技术方面的培训和能力建设。支持发展中国家在国家主导下努力建设社会抵御信息生态系统完整性风

险的能力，开展强有力的媒体和信息扫盲培训，支持公益媒体，包括提供专门和充足的发展援助。支持包括图书馆在内的公共机构的工作，改善扫盲培训和资源的获取。

i. 促进政治参与。 在整个选举过程中，保护所有选举利益攸关方获得准确、及时的信息。采取措施促进包容性政治参与和领导力，维护妇女在公共生活中的权利，包括保护她们免受一切形式的歧视、虐待、骚扰和暴力威胁。

j. 优先开展包容性的公益研究。 优先考虑、投资和支持遵守道德标准的独立研究，以及与信息完整性有关的跨学科审查，包括考虑到人工智能技术新出现的和尚未知晓的能力和影响。支持跨地域、跨语言和跨专题领域的研究，包括信息生态系统完整性风险对可持续发展目标的潜在影响，尤其关注服务不足、研究不足和面临风险的环境和社区。促进和宣传对研究成果的开放式获取，以便在国家内部和国家之间公平共享信息。

k. 促进扫盲。 通过有针对性的媒体和信息扫盲活动，从儿童早期就将数字技能与正规和非正规教育课程完美结合起来，培养批判性的、知情的公共讨论。积极提高公众（包括儿童）对在线权利、数字信息环境如何运作以及个人数据如何使用的理解和认识，同时考虑到不同年龄和背景的人们在社会、文化和语言方面的具体需求。优先考虑处于弱势和边缘化境况的个人和群体的扫盲需求，包括妇女、儿童、青 年、老年人、残疾人和即将上网的数十亿人。围绕与人工智能技术相关的具体问题开展扫盲工作，并不断更新扫盲工作，以反映新的和正在出现的技术和挑战。

l. 增强儿童、家长、监护人和教育工作者的能力。 为儿童、家长、监护人和教育工作者提供关于安全和负责任的数字行为、浏览网络媒体以及了解儿童表达和信息自由权利的持续资源。让各方参与制定媒体和数字扫盲准则和倡议，以获得更安全的在线体验，同时利用青年的数字能力。

对所有政治行为者的建议

参与并影响政治进程的个人、团体和实体

a. 维护选举的公正性。避免并公开谴责破坏选民资格、投票、计票和结果等信息完整性的行为。

b. 保护包容。公开谴责针对候选人和公职人员的虐待和骚扰行为，尤其是针对妇女和弱势及边缘化群体成员的虐待和骚扰行为，并采取措施加以解决。

c. 提供透明度。保持传播的透明度，包括广告的资金来源和数据驱动的目标定位技术的使用。

联合国

信息完整性全球原则》适用于联合国及其国际公务员。通过遵守《全球原则》，本组织为全球社会负责任地管理信息完整性树立了令人信服的榜样。加大工作力度，加强信息生态系统的完整性，将有助于推进本组织的使命，即确保和平、促进可持续发展以及促进和保护所有人的人权。

联合国将

a. 加大努力。加大努力，加强信息的完整性，包括通过针对具体情况的研究、监测、风险评估、社区参与以及跨越不同背景和语言的联盟建设。将信息完整性纳入计划和行动，以加强预防、缓解和应对，并确定新出现的机遇和挑战。

b. 支持能力建设倡议。协助各国的能力建设，提供技能发展倡议，包括对青年的培训，以帮助加强信息的完整性。

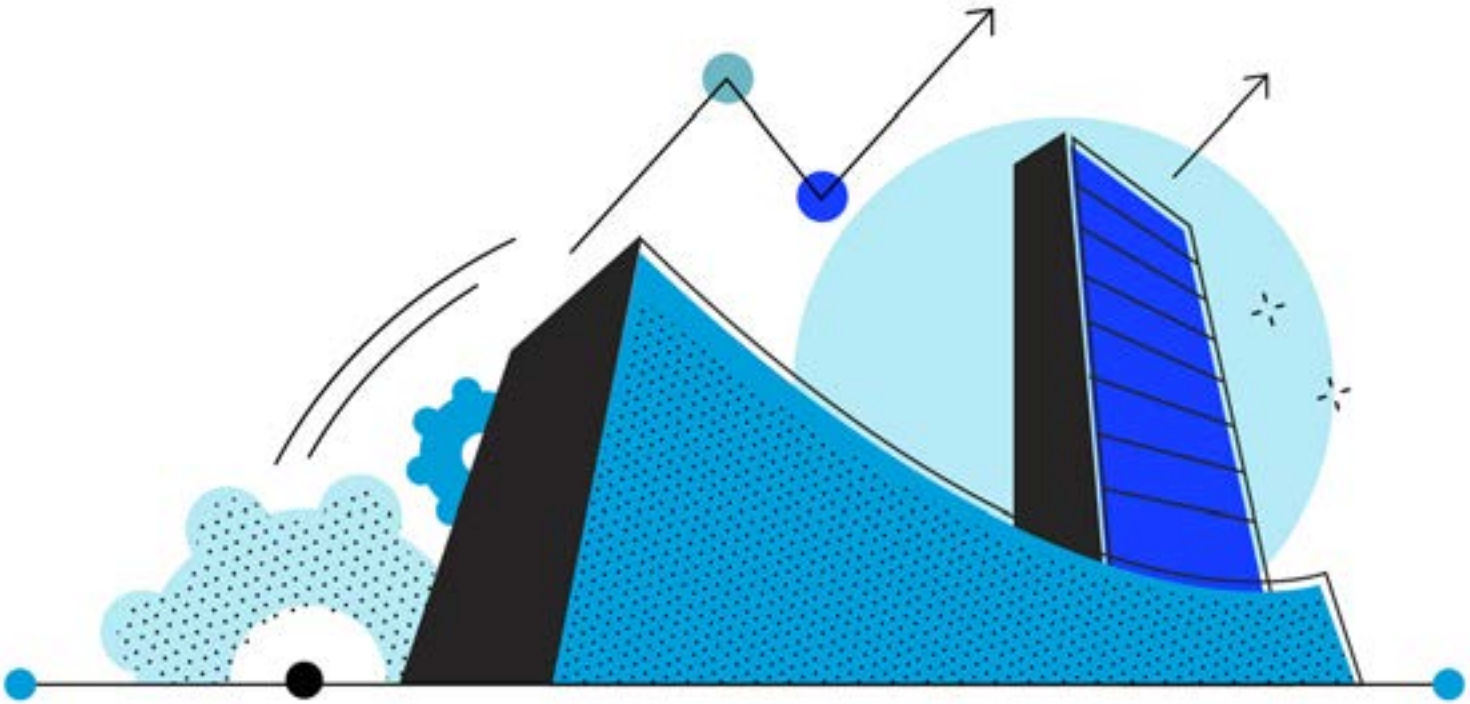
c. 开展宣传。在全球、各国和各社区推广和宣传《全球原则》，特别关注得不到充分服务的环境以及处于弱势和边缘化境况的群体。积极促进社会和谐，加强社区抵御信息完整性风险的能力，支持实现可持续 发展目标的努力。

d. 提高专门能力。在联合国秘书处设立一个中央单位，负责制定创新的、细致的方法，以应对影响联合国任务交付和实质性优先事项的信息生态系统完整性所面临的风险，并根据需要与其他能力进行协调，为整个联合国系统提供服务。

e. 制定灵活的传播战略。利用创新、循证、灵活和量身定制的传播战略，利用数字和离线信息空间促进共同利益，更好地满足联合国所服务的所有人的需求。

f. 提供多语种资源。建立多语种在线信息完整性资源中心，共享适用于不同情况的研究、指导和最佳做法，以支持全球、区域和国家层面的倡议。

g. 支持多方利益攸关方的行动计划。支持区域和国家多方利益攸关方行动计划和联盟，利用现有机制，并利用本组织在国际能力建设和协调方面的专门知识和经验。



下一步工作

在信息生态系统完整性面临的不断升级的风险，以及人工智能技术迅速发展的背景下，加强信息完整性的紧迫性不容忽视。鉴于全世界都在探索数字时代的复杂性，《全球原则》为保护和促进信息完整性提供了一个全面、统一的行动框架，并期待在未来峰会上找到多边解决方案。

为此，敦促利益攸关方

- 公开承诺、采纳并积极宣传《联合国信息完整性全球原则》，将其作为立即行动的框架。
- 利用“全球原则”组建并积极参与广泛的跨部门信息完整性联盟，汇集来自民间社会、学术界、媒体、政府和国际私营部门的各种专业知识和方法，包括能力建设方面的专业知识和方法，并确保青年的充分和有意义的参与（如通过专门的青年咨询小组）。
- 在区域、国家和地方层面合作制定多方利益攸关方行动计划，让社区参与进来，支持基层倡议并从中学习，确保青年充分而有意义的参与。

通过拥护《联合国信息完整性全球原则》，所有部门的利益攸关方都能表现出团结一致的精神，共同开辟一条通往重振信息生态系统的道路，为所有人提供信任、知识和个人选择。

附录

资源

1. 联合国秘书长的“我们的共同议程--政策简报 8：数字平台的信息完整性”（2023 年）
https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-zh.pdf

2. 联合国教科文组织数字平台治理准则（2023 年）
https://unesdoc.unesco.org/ark:/48223/pf0000388080

3. 联合国秘书长的报告：“打击虚假信息，促进和保护人权和基本自由”（2022 年）
https://daccess-ods.un.org/access.nsf/Get?OpenAgent&DS=A/77/287&Lang=C

4. 联合国教科文组织《人工智能伦理问题建议书》（2021 年）
https://unesdoc.unesco.org/ark:/48223/pf0000381137_chi

5. 联合国关于仇恨言论的战略和行动计划（2019 年）
https://www.ohchr.org/zh/documents/outcome-documents/rabat-plan-action

6. 关于禁止构成煽动歧视、敌意或暴力的鼓吹民族、种族或宗教仇恨言论的拉巴特行动计划（2012 年）
https://www.ohchr.org/zh/documents/outcome-documents/rabat-plan-action

7.《联合国工商企业与人权指导原则》（2011 年）
https://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_CH.pdf