Meta

人权报告

见解和行动
2022 年

目录

01 要点汇总

这是 Meta 的第二份年度人权报告，总结了 2022 日历年的情况。它展示了我们如何履行在企业人权政策中做出的承诺，该政策建立在《联合国工商企业与人权指导原则》（United Nations Guiding Principles on Business and Human Rights，简称 UNGP）的基础上。本报告显示了我们如何继续推进尊重人权的工作、如何评估可能与我们服务相关的潜在影响，以及我们正在采取哪些措施来减轻这些风险。

多年来，Meta 的服务和应用帮助草根运动迅速发展、挑战正统观念、倡导权利，并因此改变了世界，2022 年也不例外。随着新的挑战层出不穷，我们的人权工作也在不断发展，以解决诸如在冲突国家/地区使用我们的服务、互联网的开放性面临的威胁以及人工智能的兴起等问题。

我们加强了治理制度，以便在旗下所有产品和服务中推进尊重人权的工作。这包括继续为监督委员会赋权，在 2022 年，监督委员会共发布了 12 项内容审核决定并提出了 91 项建议。为了加强治理，我们的措施还包括培训员工。在 2022 年期间，我们在全公司开展了名为《Bigger than Meta: Human Rights》（人权大过 Meta）的人权培训。

2022 年，Meta 通过了其全球网络倡议评估，评估内容包括对我们政策和流程的审查以及八个具体的说明性案例分析，这些案例分析涉及政府提出的内容移除请求和数据请求。

人权尽职调查是我们企业人权政策的核心。在本报告中，我们披露了人权突出风险综合评估 (CSRA) 的主要发现，并提供了我们落实尽职调查工作所提建议的最新情况。

2022 年，我们与企业社会责任组织（Business for Social Responsibility，简称 BSR）合作开展了 CSRA，这是一家在人权方面拥有专业知识的独立组织。根据 UNGP 标准，该分析确定了八个最优先的潜在突出风险领域。它们分别是：1.) 意见和表达自由；2.) 隐私；3.) 平等和无歧视；4.) 生命、自由和人身安全；5.) 儿童的最大利益；6.) 公众参与、投票和被选举；7.) 结社和集会自由； 以及 8.) 健康权。该评估发现，与 Meta 提供服务的方式和盈利模式相关的潜在风险是多样化而微妙的。

随着新的挑战层出不穷，我们的人权工作也在不断发展，以解决诸如在冲突国家/地区使用我们的服务、互联网的开放性面临的威胁以及人工智能的兴起等问题。

我们还说明了对之前人权尽职调查中所提建议的落实情况，这些调查与菲律宾、以色列和巴勒斯坦以及印度有关。

此外，我们还在继续解决 Meta 2021 年民权审计中提出的建议。2022 年，117 项建议和行动项目中又有 32 项得到落实或正在落实之中，这使得我们的完成率达到了 84%。

为了履行我们的人权责任，我们会倾听外部利益相关者的声音并向他们取经。在我们的企业人权政策中，利益相关者的参与贯穿始终，而且我们致力于与边缘化群体进行更深入的合作交流。我们的服务开发、内容政策和审核以及社群守则均参考了多方的意见和见解，包括民间社会组织、人权维护者、边缘化群体、国际组织、可信合作伙伴、投资者、广告主和用户的意见。

在我们面临最突出人权风险的地区，我们会优先建立合作关系。我们的可信合作伙伴网络包含遍布 113 个国家/地区的超过 400 家非政府、非营利、国家和国际组织，他们会举报有可能违规的内容、帐户和行为，供我们结合语境/背景进行审核。2022 年，我们拓展了可信合作伙伴网络，涵盖了 36 个新的国家/地区。

当地的情况对于了解潜在人权风险至关重要，我们正在支持更多语言，以便能在这方面做得更好。我们意识到，在危机时期，人权尤其会受到威胁。Meta 会评估发生紧迫伤害的风险（甚至早在危机发生之前就会进行此评估），并努力通过具体的政策、服务和业务行动来应对，从而尊重人权。在本报告中，我们回顾了危机政策协议 (Crisis Policy Protocol) 的进展，以确保我们对危机做出的政策响应是有原则和经过合理调整的。我们提供了从我们在俄罗斯、乌克兰和伊朗的工作以及围绕巴西、肯尼亚和尼日利亚选举所做的工作中获得的见解。

我们将再接再厉，以人权为中心来理解和采用快速发展的人工智能 (AI) 技术。AI 模型可以支持人们行使自己的权利，但是也可能表现出有问题的偏见或歧视性影响，并生成有问题的内容。我们致力于以开放和合作的方式解决这些问题，并通过各种努力来解决社交媒体和 AI 带来的一些最棘手的问题，对此我们深感自豪。

我们充分认识到所肩负的责任，努力大规模推进我们的人权原则，以造福我们的用户和整个社会，赋予人们建设社群的力量。

02 关于本进展报告

Meta[1] 的第二份年度人权报告反映了我们在履行人权承诺方面的进展，这些承诺基于《联合国工商企业与人权指导原则》和我们的企业人权政策。报告中介绍了我们如何解决我们的人权影响。这包括回顾我们最近开展的突出风险评估、从人权尽职调查中获得的见解以及我们采取的应对行动。

我们的第一份年度人权报告发布于 2022 年 7 月，涵盖了我们从 2020 年 1 月 1 日至 2021 年 12 月 31 日获得的认识和进展。

本报告是我们的第二份年度人权报告，涵盖了我们从 2022 年 1 月 1 日至 2022 年 12 月 31 日获得的认识和进展。它立足于第一份基础性的报告。报告中涉及的 Meta 服务包括 Facebook、Instagram、WhatsApp、Messenger 和 Reality Labs。

本报告受 Meta 的第一份负责任业务行为报告 (Responsible Business Practices Report) 的补充，后者符合全球报告倡议组织（Global Reporting Initiative，简称 GRI）和可持续发展会计准则委员会（Sustainable Accounting Standards Board，简称 SASB）的标准。正如 Meta 2023 年负责任业务行为报告中所述，在 Meta 最新的优先主题评估中，人权作为优先主题得分很高。

Meta 的企业人权政策适用于整个企业。Meta 旗下各服务和实体有各自的政策和程序，有时会对人权产生不同的影响。本报告提到了 Meta 作为公司就 Meta 旗下一个或多个实体采取的行动。报告中的陈述无意暗示 Meta 就所有实体采取了相同的行动。[2]

在编写本报告时，我们与内部和外部的许多利益相关者进行了交流，力求在简洁与全面之间取得平衡。

[1] 2021 年 10 月 28 日，Facebook, Inc. 更名为 Meta Platforms, Inc.。为了保持一致性，本报告使用“Meta”来指代该公司，不论背景时间是在 2021 年 10 月 28 日之前还是之后。凡是提及“Facebook”，只是指相关的社交媒体服务和应用，并非指整个公司。

[2] 例如，WhatsApp 是一款端到端的加密消息和通话应用，它具有独特的人权触点。本报告对 Facebook 和 Instagram 内容审核和相关行动的讨论不适用于 WhatsApp，而且除非指明了某项政策或行动适用于 WhatsApp，否则该政策或行动不适用于 WhatsApp。此外，虽然本报告中所述的许多行动适用于 Instagram 和 Facebook，但是这两种服务的政策和程序之间存在有意的区分。如果某项政策被标注为“Facebook”政策，它不一定适用于 Instagram。本报告中的任何陈述均无意建立与将某项政策或程序应用于其他服务或实体有关的新义务（法律义务或其他性质的义务），也不应被解释为建立了这类新义务。

03 前言

Meta 的使命是赋予人们建设社群的力量，让世界联系更紧密。我们的服务和应用帮助草根运动蓬勃发展，挑战既定的权威和正统观念。不管是为了伊朗的抗议者、乌干达的 LGBTQIA+ 活动人士、基辅的记者还是使用我们服务的每位用户，我们都要努力尊重人权。

每天都有数十亿人使用我们的应用，我们力求履行在如此庞大的规模下运营所伴随的责任。随着技术和政治的发展日新月异，了解我们的服务影响人权的方式并解决这些问题是一项不断变化的挑战。十几年来，Meta 一直是 AI 研发领域的领导者。随着生成式 AI 等强大新工具的快速发展，我们不断调整方法，以履行我们的人权责任，并对新出现的风险保持警惕。

我们面临的人权问题错综复杂，因国家和地区而异，并受到世界各地政府法规变化的影响。例如，欧盟于 2022 年颁布的《数字服务法》建立了适用于 Meta 的重要尽职调查和报告义务，但是该法明确与该地区的基本权利框架挂钩。这一切发生在互联网的开放性和我们用户的基本权利受到不断变化的威胁的背景下，这种威胁的源头在于专制互联网模式的蔓延，这使得公民与全球互联网的其余部分隔绝开来。

“每月有超过 35 亿人使用 Meta 旗下应用 — Facebook、Instagram、WhatsApp 和 Messenger。这意味着地球上有三分之一到一半的人都在使用它们。借助这些应用，人们可以打破地域限制，在一个紧密联系的世界中畅游，接触到各种各样的人物、思想、新闻和社群，并开展商业活动。这种联结规模在人类历史上是前所未有的。”

Nick Clegg，全球事务总裁

本报告分享了我们在所有计划、服务和政策中推进尊重人权工作的进展。我们会继续探索新的途径来将人权融入到我们的服务和实践中。这包括开展以下方面的工作：推进民主、确保建设开放和可互操作的元宇宙，以及找到方法来帮助确保 AI 公平地为不同社群服务。

我们的所有努力都基于我们根据企业人权政策和《联合国工商企业与人权指导原则》(UNGP) 做出的承诺。我们并非单打独斗。监督委员会作为一个独立而专业的仲裁机构，发挥着行业领先的关键作用，他们会审核内容决定，并提供政策咨询意见和建议。根据其章程，监督委员会特别关注用于保护自由表达权的人权准则。

我们的工作还会借鉴专门的多边组织和联合国组织以及多方利益相关者联盟的原则和实践。我们是联合国全球契约组织的成员，并支持其原则。而且我们继续参与联合国人权事务高级专员办事处的 B-Tech 项目，该项目致力于打造与在科技领域实施 UNGP 相关的权威指南和资源。作为全球网络倡议组织（Global Network Initiative，简称 GNI）的成员，我们致力于践行其原则和实施准则。2022 年，我们通过了 GNI 每三年一次的评估。

我们还在不断探索新的治理形式。例如，我们试行了社群论坛，将来自世界各地的不同群体汇聚一堂，大家在此共同讨论棘手问题，并分享对诸多建议的看法。

本报告旨在深入揭示 Meta 在 2022 年为了识别、减轻和防范人权风险所做的工作。与往常一样，我们还有更多工作要做，我们期待在未来几年更多地报告我们的进展。

签名：

Nick Clegg
全球事务总裁

Jennifer Newstead
首席法务官

04 Meta 如何管理和治理人权工作

清晰的管理和治理结构使我们能够在所有计划、服务和政策中推进尊重人权的工作。Meta 人权团队负责指导企业人权政策的落实，这项工作受全球事务总裁和首席法务官的监督。

正如我们的第一份年度人权报告所述，该团队的任务包括：促使企业人权政策融入到现有和正在制定的政策、计划和服务中；开展尽职调查；以及为有关企业人权政策的员工培训提供支持。该政策指导各个团队打造尊重人权的产品，应对新出现的危机，并迅速灵活地大规模落实人权。

我们的企业人权政策要求我们定期向董事会报告重要的人权问题。董事会下属的审计与风险监督委员会负责监督公司面临的各种风险，包括与人权有关的风险，以及管理层为监测或减轻这些风险而采取的措施。该委员会定期听取与人权团队的现有工作和正在着手开展的工作相关的情况通报。

培训 Meta 员工

在 Meta，如何开发与开发什么同样重要。通过我们的人权培训，员工能更好地了解其责任以及履行这些责任所需的知识和技能。

《Bigger than Meta: Human Rights》（人权大过 Meta）是我们在 2022 年开展的培训，强调了 Meta 的服务、政策和业务决策对人权的潜在和实际现实影响。该培训力求在我们的日常工作中促进人权观念，鼓励尊重人权，让使用我们服务的所有用户受益。

这一培训补充了 Meta 的民权培训，后者于 2022 年 7 月开展，以无歧视、公正和公平原则为中心。

申诉和获得补救

在 Meta，我们努力提供途径来供利益相关者报告问题，供 Meta 审核这些问题，以及供 Meta 提供符合 UNGP 第 31 条的补救措施。Meta 提供多种申诉途径，这些途径在行为规范以及平台和应用上的帮助中心里有述，包括向同类首创的监督委员会提出申诉的流程。

我们在移除内容后，会通知用户，并清楚说明就内容移除决定提出申诉的途径。与移除决定和政策执行的其他方面有关的数据会通过政策及信息公示平台发布。如需更多信息，请参见我们的第一份年度人权报告。

监督委员会

监督委员会发挥着独一无二、行业领先的作用，旨在尊重自由表达权。该委员会为公众提供了向外部机构提出额外申诉的机制。2022 年，Meta 批准向监督委员会信托基金提供为期三年、价值 $1.5 亿美元的新资助，此举强调了 Meta 对监督委员会的承诺。这让监督委员会能继续作为独立机构进行运作。

监督委员会负责制定完全独立的内容审核决定以及与内容政策、服务和运营有关的建议。2022 年，监督委员会发布了 12 项审理决定，其中有 9 项推翻了 Meta 的决定。这些审理决定涵盖来自许多国家/地区的内容，包括哥伦比亚、埃塞俄比亚、阿富汗、克罗地亚、苏丹、印度和美国。所涉及的问题包括在有关军事政变的内容中描绘血腥暴力、在新闻报道中提及塔利班，以及描述对未成年人的性暴力等等。这些审理决定都为平台用户提供了获得补救措施的途径。

除了其审理决定外，监督委员会在 2022 年提出了 91 项建议，其中有 89 项建议处于考虑执行、正在执行或已经完全落实的状态。

2022 年监督委员会建议

监督委员会建议	Meta 正在评估和/或执行的建议	已经完全落实的建议
91 项	75 项	14 项

全球网络倡议组织

自 2013 年以来，Meta 一直是全球网络倡议组织 (GNI) 的成员。GNI 是一个不断发展壮大的联盟，由互联网和电信公司以及各种民间社会团体组成，其中包括来自世界各地的人权和新闻自由团体。我们致力于践行 GNI 有关表达自由和隐私的原则（下称 GNI 原则），这些原则基于国际上认可的法律和标准。2021 年，我们在企业人权政策中重申了这些承诺。

每两到三年，所有 GNI 成员公司都要针对其原则执行情况开展独立评估。

2022 年，Meta 通过了其 GNI 评估，其中包括对我们政策和流程的审查以及八个具体的说明性案例分析，这些案例分析涉及政府提出的内容移除请求和数据请求。GNI 董事会一致认为 Meta“在执行 GNI 原则方面付出了真诚的努力，并随着时间推移而不断改进”，这一条是评估成员公司的标准。GNI 将发布一份单独的公开评估报告，其中涵盖 2021–2022 周期的所有成员公司评估。

05 我们如何识别和解决突出的人权风险

2022 年，我们开展了人权突出风险综合评估 (CSRA)，这是一项基础的计划工具，用于帮助指导未来的风险管理。CSRA 的目的是在全球范围内确定 Meta 整个企业最突出的固有人权风险及其优先顺序。最突出的固有人权风险是指 Meta 有可能对人们（包括用户和可能受 Meta 行动影响的其他人）造成的最严重的负面人权影响[3]。“固有”风险是指在执行流程或缓解措施之前便已存在的风险，与之相对的是“剩余”风险，这是指在执行了用于解决固有风险的流程后依然存在的风险。

该评估是如何开展的：企业社会责任组织 (BSR) 开展了桌面研究，并与内部和外部利益相关者广泛交流合作，这包括与来自多个地区和各种领域的民间社会团体、联合国代表和投资者举行了十多场会议。该评估还借鉴了从先前开展的人权尽职调查和利益相关者参与中获得的见解。该评估依据 UNGP 和 UNGP 报告框架中对“严重性”（范围、规模、可补救性）和“可能性”的定义来确定风险的优先顺序：

[3] “负面人权影响”一词与 UNGP 中的意思一致，是指当某项行动剥夺或削弱个人享受其人权的能力时产生的影响。除非另有明确说明，否则本报告不应被解释为暗示任何特定个人或群体遭受了负面影响。

1. **范围：**潜在负面人权影响可能会影响多少人？
2. **规模：**对受影响者来说，潜在负面人权影响的严重程度如何？
3. **可补救性/不可逆性：**补救措施能否使有可能受影响者恢复至与受影响前同样或同等的状况？
4. **发生的可能性：**潜在负面人权影响发生的可能性有多大？

根据 UNGP，CSRA 特别关注很可能成为弱势或边缘人士的个人的权利、需求和挑战。

我们获得的信息：该分析确定了全球业务中八个最优先的人权突出风险领域，如下所示：

1. **意见和表达自由**
2. **隐私**
3. **平等和无歧视**
4. **生命、自由和人身安全**
5. **儿童的最大利益**
6. **公众参与、投票和被选举**
7. **结社和集会自由**
8. **健康权**

我们的突出风险

CSRA 使用 UNGP 标准和 UNGP 报告框架，分析了整个全球性企业在国际公认的所有人权领域的固有突出风险，并确定了八个风险最突出的优先领域。下文探讨了这些风险，并配有说明性示例来揭示一些潜在的风险因素和 Meta 采取的一些防范和缓解措施。

意见和表达自由

在 Meta，我们坚信意见和表达自由权至关重要，它是我们努力保护的核心权利。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 解决潜在风险的措施
互联网中断和对社交媒体的封锁阻碍了人们行使表达自由权，并切断了他们接收和发送重要新闻和信息的途径	我们最近推出了一项功能，让人们能在互联网连接中断或者对 WhatsApp 的访问被阻止时通过代理服务器连接到 WhatsApp。
政府对内容的限制过于宽泛	我们根据 GNI 原则来评估政府提出的内容移除请求 (TDR) 是否合法有效，并且我们的执行情况会定期受到独立评估。 我们实行广泛的操作控制措施来审核内容移除请求的有效性，并会在政策及信息公示平台中提供与政府提出的内容移除请求有关的信息。 我们会根据国际人权准则定期修订内容政策，并会就此征询不同利益相关者的意见。
Meta 执行的内容审核可能会限制表达自由	AI 模型可以预测一条内容是仇恨言论还是暴力和血腥内容。我们的政策执行技术是单独的系统，负责决定是否应采取措施，例如删除内容、对内容降级或将内容发送给人工审核团队以供进一步审核。通过执行我们的政策，我们力求减轻仇恨言论、暴力与煽动暴力以及其他人权危害的风险，同时维护表达自由。

隐私

“正确保护隐私权需要我们全公司持续共同的努力，也是 Meta 的每个人为推动我们的使命而需肩负的责任。”

Michel Protti，产品首席隐私官

隐私权是 Meta 使命的核心，也是实现其他人权（例如表达自由、集会与结社自由以及宗教信仰自由）的必要条件。

我们致力于改进和调整我们的做法和政策，以便预测和应对我们的服务和应用中存在的与用户隐私相关的挑战。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 解决潜在风险的措施
政府提出过于宽泛或不必要的用户数据请求	我们通过专门的执法响应团队来保护用户免受不合法或过于宽泛的政府数据请求的影响。 我们会回绝不符合国际公认标准（包括企业人权政策中所述的标准）的政府请求，并在必要时在法院对这些请求提出质疑。 我们会响应合法有效的请求，只要我们有充分的理由认为法律要求我们这么做。此外，我们也会在法律未强制要求的情况下响应某些合法请求，但前提是我们有充分的理由认为这种做法： • 符合某位用户可能位于的另一个司法管辖区的法律要求， • 会影响该司法管辖区的用户，且 • 符合国际公认的标准，包括我们的企业人权政策。 当我们履行政府提出的数据请求时，我们会尽力在政府请求获取某用户数据时通知该用户，并针对收到的政府请求发布透明度报告。

Meta 应用中的内容或行为可能会对隐私和数据保护权利产生负面影响（例如人肉搜索、数据抓取）	Meta 制定了强大的 Facebook 社群守则和 Instagram 社群守则来约束用户对我们应用的使用。我们禁止发布分享、提供或索取个人身份识别信息或其他隐私信息，因而可能引发人身伤害或财务损失的内容。一旦发现违反我们政策的内容，我们会立即将其移除。 “数据抓取”行为违反了我们的服务条款，这是指未经我们许可，使用自动化技术访问或收集来自 Meta 服务和应用的数据。我们有专门的团队、技术措施和流程，专注于检测、调查和阻止未经授权的数据抓取尝试行为。
Meta 或第三方可能会以对隐私和数据保护权利有负面影响的方式，将敏感或非常私人的用户数据（包括与受保护特征相关的数据）用于投放定向广告。	我们收集和处理的任何信息都取决于我们提供的服务。为了帮助我们执行这项任务，我们正在开发隐私保护强化技术 (PET)，以便最大限度减少数据收集和使用。我们的团队致力于打造隐私保护强化技术，供 Meta 各团队使用，这些技术侧重于一些重要领域，例如在收集数据时去除身份识别信息，以及让团队能在我们的应用和服务中实现端到端加密 (E2EE)。 在咨询民权专家、政策制定者和其他利益相关者的意见后，我们推出了一些调整，以改善用户对广告的控制，并且我们移除了一些与人们可能会觉得敏感的主题相关的广告受众定位选项。 我们与美国司法部合作开发和推出了差异降低系统 (VRS)，这项新技术有助于在我们的应用中以更公平的方式传播某些广告。我们在美国针对住房广告推出了差异降低系统，而且会将其应用范围扩展至就业和信贷广告。

CSRA 确定了与定向广告相关的多个风险途径，同时强调了这些风险的微妙性质。例如，虽然基于广告的业务模式可能会给隐私权带来风险，但是我们有可能制定缓解措施来解决这些风险，同时依然保留让世界各地的用户汇聚一堂和表达想法的广泛能力。同样，广告也是中小企业开拓商机的关键手段，如果没有广告，他们在如何触达潜在顾客这个问题上将束手无策。

平等和无歧视

平等和无歧视权利旨在平等地保护所有人不受任何歧视。虽然随着科技的普及，上网越来越容易，网民的范围也越来越广，但是并非所有群体在网上都有相同的体验，尤其是边缘化的个人和群体。

我们努力通过 Facebook 和 Instagram 社群守则为所有用户创造更加包容和平等的网络环境，这些守则禁止仇恨言论、欺凌和骚扰。我们的人权与民权团队共同努力，力求尊重个人的平等权利，并在我们的所有技术中促进公平。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 解决潜在风险的措施
对平等和无歧视有不利影响的内容（例如仇恨言论）	Facebook 社群守则（包括与基于受保护特征的仇恨言论有关的政策）概述了 Facebook 允许和禁止发布的内容。Instagram 社群守则概述了 Instagram 允许和禁止发布的内容。对于违反我们政策的内容，我们会采取处理措施。广告发布守则针对我们允许和禁止发布的广告内容类型以及我们禁止的受众定位提供了政策详情和指南。 我们继续打造负责任 AI 系统，以减少偏见和提高公平性，所涉及的领域包括广告投放，以及减少生成式 AI 应用中有害用语的流行。为了帮助研究人员衡量我们 AI 模型的公平性，我们向前迈出了重要一步。缺少能代表各种人群和经历的多元化数据，可能会导致利用 AI 得到的结果反映出有问题的刻板印象，或者无法平等地为每个人服务。因此我们编制了 Casual Conversations v2，这是一个基于用户同意且公开可用的数据集，让研究人员可以更有效地评估某些类型的 AI 模型在公平性和可靠性方面的表现，从而提高模型的包容性。

审核某些语言和方言可能比审核其他的语言或方言更具挑战性。

我们结合使用 AI 和人工内容审核员来帮助移除我们应用中违反政策的内容。

我们使用人工翻译和机器翻译来加强内容审核。我们不断改进举报流程以及我们所支持的语言使用的完整性分类技术。

仇恨言论

2022 年，我们从 Facebook 移除了大约 5,000 万条仇恨言论内容，其中有 91% 是我们主动发现并移除的。我们从 Instagram 移除了 1,600 万条仇恨言论内容，其中有 92% 是我们主动发现并移除的。从用户浏览的 Facebook 和 Instagram 内容来看，仇恨言论的占比不到 0.02%。

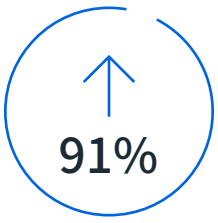
如想详细了解 Meta 民权团队在平等和无歧视方面的工作进展，请查看此进展报告和此更新。

无障碍

随着科技的进步，不让残疾人士掉队十分重要。我们坚持不懈地改进我们的工作，确保在旗下所有服务和应用中提供无障碍体验。让残疾人士能无障碍地使用这些服务和应用是我们企业人权政策中的一项重要承诺，该政策将《残疾人权利国际公约》列为一项核心的国际标准。

我们的行动： 拓展跨行业努力，打造更加无障碍的体验

不论是通过辅助技术、康复或训练解决方案，还是通过能提高认识、开阔眼界的体验，虚拟、增强和混合现实技术以各种方式为残疾人士带来了特别的希望。Meta 继续与 XR 协会 (XR Association) 开展项目合作，以提高对无障碍的认识和加强跨行业努力，例如与 XR Access 携手推出以无障碍为重点的新网站 xraccessibility.github.io。最近的创新包括由 Meta 提供支持的 Ray-Ban Stories 智能眼镜，这款智能眼镜可以在 WhatsApp 和 Messenger 实现免持消息交流和通话，还可以使用语音命令和触控功能来截图和录屏。



2022 年从 Facebook 移除的 5,000 万条仇恨言论内容中，有 91% 是 Meta 主动发现并移除的

除了在我们服务中嵌入功能外，Meta 还制定和分享了技术建议，帮助开发人员创建无障碍的虚拟现实应用。我们还与 XR 协会合作，与残疾人社群和行业组织一起制定和推出了面向开发人员的 XR 无障碍指南。

生命、自由和人身安全

生命、自由和人身安全权涉及不受人身限制的自由以及保护身心免受伤害的权利。对 Meta 来说，为了尊重此人权，要做的工作包括解决内容可能引起伤害的风险，这包括涉及以下方面的内容：人口贩卖、受国家支持的网络威胁以及参与或倡导暴力或仇恨的非国家团体。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 解决潜在风险的措施
不良行为者： • 利用 Meta 服务和应用来配合实施网络或现实伤害 • 违规使用服务和应用进行网络攻击或网络钓鱼 • 威胁和骚扰人权维护者 (HRD)、活动人士和其他弱势群体	我们的安全政策旨在识别和防御对抗性威胁。Meta 拥有保护人权维护者和新闻工作者的工具和资源，这些工具和资源是依据人权维护者的直接反馈来打造的。
	我们对更多人开放了 Facebook Protect 计划，尤其是人权维护者，这让他们能更好地控制自己的帐户，包括设置允许谁查看他们的帖子。 我们隐私中心里的安全指南汇总了各种资源，帮助人们管理他们的隐私设置，以免受到网络监视或者未经授权或强制的访问。该安全指南还能帮助受到欺凌和骚扰者寻求支持。我们就一系列问题与专家和人权维护者合作，以创建政策、工具和资源，这涉及数字素养、儿童保护等方面以及给家长的建议。
	Meta 大力打击专制政府、恐怖组织或其他试图监视政权批评者、反对派人士和人权维护者的不良行为者违规使用我们服务的行为。我们在发现此类活动时，会采取行动不让他们的网域基础设施在我们的服务中得到分享，并在可行和适当的情况下，通知我们认为这些恶意活动想要攻击的对象。

我们尊重生命、自由和人身安全权的更多方式

以下举措强调了我们用于帮助面临风险的群体和个人的方法。

保护人权维护者和新闻工作者

对于人权维护者、新闻工作者和其他弱势群体，我们为其开发了资源和工具，它们可在 Meta 安全中心里找到。我们的记者安全指南有助于确保记者利用工具和资源来管理他们的网络安全，从而保护其消息来源、联系人、登录信息和个人信息的安全。我们还通过 Meta Journalism Project（Meta 媒体人项目）来赞助多项计划，为国际新闻社群提供支持。

2022 年，我们采取了这些行动：

- 加大力度保护活动人士和新闻工作者免受暴力威胁，包括移除“暴露”活动人士身份的内容，只要这类内容有可能危及活动人士的安全。
- 保护在专制国家或冲突环境中被逮捕或拘留的人权维护者的帐户，防止当地执法部门、安全部门和冲突行为者未经授权访问这些帐户。
- 推进 Meta 面向亚太地区的人权维护者基金及新闻工作者安全倡议 (Human Rights Defender Fund and Journalist Safety Initiative for Asia Pacific)，通过这项举措，国际人权团体“公民权利维护者组织”(Civil Rights Defenders) 为 500 多名人权维护者（主要在南亚和东南亚）提供了紧急撤离、临时安置、安全培训和法律援助服务以及数字安全措施。公民权利维护者组织还开展了安全评估，促进了安全措施的落实，并向这些地区的人权维护者和组织提供了其他与安全有关的支持。
- 加强与人权维护者的交流合作，以便就地区和全球范围内人权维护者所面临威胁的影响提供持续反馈，并对人权维护者进行与隐私中心里的安全指南有关的培训。
- 为新闻工作者和人权维护者推出免费的在线数字安全培训课程，该课程由国际记者中心 (International Center for Journalists) 与 Border Center for Journalists and Bloggers 合作创建和管理。

2022 年，我们将隐私中心里的安全指南呈交给了联合国人权维护者处境特别报告员办公室。

人口贩卖和剥削

为了预防和阻止伤害，我们会移除对剥削（包括人口贩卖）行为进行协调或推动的内容。[4]

2022 年，我们更新了有关剥削的政策，禁止发布索求人口偷运服务的内容。我们还在 Facebook 提供了信息资源，介绍与偷运者接触的风险、潜在剥削的迹象以及寻求合法移民（包括庇护）的途径。我们在打造这些资源时咨询了专家的意见，包括咨询国际移民组织 (International Organization on Migration)。

[4] 如需有关人口贩卖和剥削的更多信息，请查看我们第一份年度人权报告的第 32 页。

我们仍然允许发布索求或分享以下方面信息的内容：个人安全和过境，以及如何离开一个国家/地区或寻求庇护。我们允许发布旨在谴责人口贩卖或提高人口贩卖问题认知的内容。

印度比哈尔邦为被贩卖人口设立的帮助热线

人口贩卖活动在印度比哈尔邦尤为猖獗。为了帮助受害者，Meta 支持开发了一条位于 WhatsApp 的帮助热线，用于辅助比哈尔邦劳工部的追踪系统。用户可以向特定 WhatsApp 号码举报童工事件，该热线以英语、印地语和摩揭陀语提供服务。这条帮助热线还用于监控该邦康复童工取得的进展。

雇佣监视

在雇佣监视行业的帮助下，一些政府和不良行为者能够通过违规间谍软件来攻击和恐吓活动人士、持不同政见者、新闻工作者和研究人员，从而达到压制声音的目的。这类间谍软件无视人权、公民自由和法治，在人们不知情的情况下对其进行攻击。

雇佣监视行为违规使用我们的服务，将这些服务作为在用户的移动设备上植入恶意软件的载体。我们已经破坏并报告了许多这类行动。2022 年，我们携手行业合作伙伴，提出了预防和减轻这类威胁的原则和做法。在《网络安全技术协议》(Cybersecurity Tech Accord) 的支持下，我们在民主峰会上发表了这些原则。在 2022 年 12 月的政策建议中，我们为其他利益相关者（包括政府、雇佣监视行业本身以及民间社会）提供了路线图，以共同促进在线权利和自由。

未经同意的私密图像分享 (NCII)

未经他人同意，在线上或线下分享其私密图片或视频的行为侵犯了此人的隐私权和安全权。图像往往会通过多个平台和应用传播。未经同意的私密图像分享经常被称为“报复性色情”，因为分享私密内容可能是在关系破裂后对某人的一种“报复”行为。

Meta 与非营利组织和其他公司合作，打造了一款免费工具 StopNCII.org，它让潜在受害者能够标记图像，以便加入 StopNCII.org 计划的公司可以检测和移除这些图像，防止它们在网上分享。只需一个操作，个人就可以申请移除多个平台和应用上未经同意分享的私密图像。StopNCII.org 建立在 Facebook 和 Instagram 开发的技术基础上，提供 22 种语言版本，支持 Facebook、Instagram 和加入该计划的其他公司的平台，包括 TikTok 和 Bumble。

儿童的最大利益

儿童的最大利益赋予了儿童权利，要求人们在对儿童有影响的所有决定和行动中考虑儿童的最大利益，包括避免儿童遭到人口贩卖。Meta 的“儿童的最大利益”框架符合《联合国儿童权利公约》(UN Convention on the Rights of the Child) 的基本价值观。

我们希望儿童拥有安全、积极的网络体验。保护儿童在网上的安全是 Meta 的首要任务，我们为青少年提供工具，让他们在网上度过更有意义的时间，包括提供更多方法来帮助他们保持专注和设置界限，例如休息提醒、专家建议、勿扰模式、切换主题提示，以及隐藏动态版块帖子“赞”数的选项。在打造这些技术时，我们采用的方法立足于研究发现以及来自家长、青少年及心理健康和儿童心理学专家的直接反馈。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 解决潜在风险的措施
儿童接触到扰人内容或不当内容	我们的政策规定，只有年满 13 岁者才能在 Facebook 和 Instagram 创建帐户。我们的政策是为了帮助年满 13 岁的用户获得安全的体验，在制定这些政策时，我们会咨询儿童安全专家的意见。我们会为某些内容（例如，某些类型的血腥暴力内容）设置年龄限制；我们努力避免向未成年人推荐某些非违规内容（例如有关烟酒的内容）；而且我们有敏感内容控制选项，该选项默认对未成年人采用最严格的设置。我们禁止发布性化未成年人的内容，而且我们一旦得知明显的儿童性剥削行为，会向美国国家失踪与受虐儿童援助中心（National Center for Missing and Exploited Children，简称 NCMEC）报告。 我们还针对未成年人制定了更严格的广告政策（例如，除了禁止烟酒等管制商品或服务的广告外，我们也不允许向未成年人投放减肥产品广告）。 我们的政策禁止发布鼓励自杀、自残、饮食失调行为或者欺凌和骚扰的内容，而且 Facebook 和 Instagram 力求阻止这类内容。我们发现这类内容后会将其移除，而且我们会继续改进用于检测和移除这类内容的技术。

对儿童的性剥削或虐待	我们安全中心的儿童保护版块重点介绍了我们用于保护儿童安全的全面方法，这包括禁止儿童性剥削的政策，以及用于预防、检测、移除和举报违规行为的先进技术。 我们打造了 30 多种工具来为儿童、青少年及其家长提供支持，其中包括在 Instagram 的家庭中心里提供的监护工具，这些工具可以帮助家长及其小孩一起度过在网上的时光。 在 Facebook 和 Instagram，我们使用成熟的检测技术来检测和移除剥削儿童的图片和视频以及阻止这类内容的分享。我们会向 NCMEC 和世界各地的有关当局报告我们网站上出现的明显儿童剥削事件。2022 年，我们支持 NCMEC 开发了 Take It Down 平台，此平台面向担心自己的私密图像在网络平台和应用中传播的未成年人。

公众参与、投票和被选举

在自由公正的选举中享有公众参与、投票和被选举的权利是民主的基石。此权利对于法治、社会包容、经济发展和促进所有人权也至关重要。

在我们的服务和应用中保护选举诚信是我们的首要任务之一。Meta 在选举前后和期间会投入大量精力和资源来在网上为选举保驾护航。我们的目标是让民众有机会发声，帮助他们参与公民进程，同时打击仇恨言论、干涉和境外干预势力。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 解决潜在风险的措施
对公众参与、投票或竞选公职有不利影响的内容	我们禁止发布企图干涉或压制投票的内容。我们努力识别和移除这类内容，不论该内容来自于谁。 对于政治类广告和公共主页，我们提供业界领先的透明度，因此用户可以看到谁在试图影响投票。在一些市场，我们禁止投放劝阻投票、过早宣称选举胜利、企图否定选举合法性或者与卫生部门有关安全投票的规定不一致的广告。 仇恨言论可能会影响选民。我们的社群守则涵盖有关骚扰、暴力与煽动暴力以及仇恨言论的政策，并禁止针对族裔或宗教等特征对他人发起攻击。我们一旦发现违反这些规定的内容，会立即将其移除。
试图阻止人们投票的行为；垃圾信息、境外势力合谋造假行为或违规内容举报的数量出现增长	对于因企图违反我们的政策或造成现实伤害而被 Meta 封禁的任何组织，我们会主动搜索来自这些组织的内容。 2022 年，我们更新了 Facebook Protect，该计划会为候选人、他们的竞选活动和民选官员提供安全工具和额外保护。
有组织地干预选举的不良行为者	我们先进的安全行动通过多种方式来支持选举安全，包括打击操纵性竞选宣传活动、识别新出现的威胁，以及部署业界领先的全球事实核查网络。 我们使用 AI 来识别虚假帐户，并在这些帐户被举报之前移除其中绝大部分帐户。自 2017 年以来，我们移除了超过 150 个合谋造假网络。

与我们在全国大选方面的工作有关的更多信息

对于更有可能引发违规内容的选举，我们组建了一个跨职能的专家团队，他们在选举诚信、错误信息、安全、人权和网络安全方面具有专业知识，并具备该市场的相关语言专业知识。该团队实时监控并应对新出现的风险。他们还会监控其他社交网络和传统媒体上的新闻报道和选举相关活动。这些工作为我们提供了全面的视角，有助于我们追踪什么样的内容可能会热传。

2022 年，我们建立并加强了多个团队，继续为世界各地的选举做好准备，这包括巴西、肯尼亚和尼日利亚的选举，以及美国的中期选举。我们投入了资源、与利益相关者和专家交流、建立了合作关系并改进了技术。

2022 年美国中期选举

在这些选举的筹备阶段，我们与州和地方选举官员合作，在动态版块内发送投票提醒，共发出超过 8,000 万条与选举相关的通知。我们打击了国内外的选举干预活动，揭露并瓦解了数十个试图干预选举的网络。我们封禁了一千多项军事化的社会运动，并从我们的应用中移除了数以万计与“匿名者 Q”组织 (QAnon) 相关的公共主页、小组和帐户。我们还移除了原来的 Stop the Steal 小组，并封禁了超过 270 个白人至上主义组织。

巴西

巴西于 2022 年 10 月初举行了大选，并在当月晚些时候举行了第二轮总统选举。这些选举是在经济危机中举行的，当时人们对否定选举合法性言论的担忧也甚嚣尘上。我们在选举诚信方面的工作是持续性的，而且我们提前一年就在紧锣密鼓地加强工作，以便迎接 2022 年巴西大选。为迎接选举，我们重点关注四个关键领域：防止干预、与选举当局合作、打击错误信息和虚假信息，以及提高广告的透明度。

我们启动了选举行动中心，让巴西和世界各地的诚信专家能够实时识别我们服务和应用中的潜在威胁。从 8 月到 10 月，我们借助人工审核和 AI，主动检测并移除了巴西用户发布的大量违规 Facebook 和 Instagram 内容，包括超过 31 万条违反暴力与煽动暴力政策的内容，以及超过 29 万条违反仇恨言论政策的内容。此外，对于我们的技术判定有可能违反这些政策的内容，我们还减少了其传播，以防止它们迅速扩散。我们在 Facebook 和 Instagram 发布了选举日提醒，以及更新选民身份证的提醒。

从 2022 年 1 月 1 日至 2022 年 10 月 2 日，有超过 3,000 万人点击了 Facebook 上有关巴西选举的帖子中添加的选举信息标签，并跳转至高等选举法院（Tribunal Superior Electoral，简称 TSE）网站上的官方信息。超过 470 万人在 WhatsApp 订阅了 TSE 的智能聊天助手，以获取权威信息。为了打击针对参政女性的网络暴力，我们在 TSE 法院和巴西妇女民主网络 (Women's Democracy Network) 的支持下发布了一份指南。

为了打击错误信息，我们与独立的事实核查机构开展了合作（这项合作仍在继续），核实 Facebook 和 Instagram 上葡萄牙语的帖子、Reels、快拍、视频、图片和评论的真实性。2022 年，我们在巴西的事实核查计划的合作伙伴数量从 4 个增加到 6 个。

为了打击虚假信息，我们的内部团队发现并移除了多个违反我们合谋造假行为政策的网络。

为了提高广告的透明度，在巴西，我们将透明度政策的适用范围从政治广告扩展到 Facebook 和 Instagram 上有关经济、安全和教育等社会议题的广告。我们拒绝了从国内外来源提交的 135,000 条在地区定位中包含巴西的广告。

请点击此处和此处，详细了解我们围绕巴西选举开展的工作。

肯尼亚

肯尼亚于 2022 年 8 月举行了大选。以前的选举流程因侵犯人权而留下了污点。在 2022 年选举之前，我们与主要的人权维护者 (HRD)、联合国机构、新闻工作者和民间社会成员合作，采取了一系列行动来加强我们的安全和诚信措施，包括扩展我们的可信合作伙伴计划。

与对待类似选举一样，我们成立了一个具有语言专业知识的跨职能团队来领导我们围绕该选举开展的工作。我们的工作以弱势群体为重点。我们将女性领导者纳入了 Facebook Protect 计划，以帮助保护经常成为攻击目标的帐户，我们还采用了更强大的安全保护措施（包括监控潜在的帐户盗用威胁），并在 Facebook 和 Instagram 开展了骚扰清理行动。我们还开发了检测和行动系统，帮助保护新闻工作者和人权维护者。

我们为人权维护者和独立女性新闻工作者举办了一系列培训。我们还通过移除旨在压制投票的内容，努力打击错误信息在平台内外的传播。这包括移除断章取义的言论或虚假的暴力行为指控。我们的行动在一定程度上参考了当地合作伙伴的建议。

我们加大了对合作的肯尼亚第三方核查机构的投资，例如 AFP、Africa Check 和 PesaCheck。我们还向全肯尼亚的当地新闻编辑室提供资助，以加强当地语言的事实核查工作。为了推动有关错误信息和仇恨言论的教育，我们以多种语言在平台上和通过当地广播开展了宣传活动，并推出了两项支持服务，分别是 Trending Event（热门事件）和 CrowdTangle Live Display（CrowdTangle 实时视图），以便帮助第三方核查机构识别和处理可能的错误信息。

结社和集会自由

结社和集会自由权对民主至关重要，也是国际法所保障的许多其他权利（包括表达自由权和参与公共事务的权利）的基础。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 解决潜在风险的措施
Meta 旗下平台上的某些内容可能会让一些用户感到无法自由地在 Meta 应用中聚集	仇恨言论、欺凌和其他形式的骚扰可能会影响个人的集会和结社权。我们在收到受害者的举报后会移除这类内容，因为它们让人们无法在 Meta 应用和现实中感到安全和受尊重。Meta 针对欺凌和骚扰以及暴力与煽动暴力（包括通过不受欢迎的公共主页、小组和活动发布的针对个人的内容）制定了广泛的政策。 我们力求消除对群体的骚扰，因为它侵犯了许多权利，包括集会权。这包括将目标对准有较高风险受到现实伤害的个人，包括人权维护者、未成年人、暴力事件或悲剧的受害者，以及选举期间高风险国家/地区的反对派人士。
政府对内容施加的限制过于宽泛，限制了集会和结社权	WhatsApp 的端到端加密功能可保护隐私权，因此让用户能够行使其集会和结社权。 我们根据全球网络倡议组织的原则来评估政府提出的内容移除请求 (TDR) 是否合法有效，并且我们的执行情况会定期受到独立评估。 我们实行广泛的操作控制措施来审核内容移除请求的有效性，并会在政策及信息公示平台中提供与政府提出的内容移除请求有关的信息。我们会根据国际人权准则定期修订内容政策，并会就此征询不同利益相关者的意见。

Meta 尊重供应商员工组织工会的权利。供应商员工的工会组织不会改变我们与该供应商合作或聘用该供应商的决定。我们认为，与我们合作的公司不应反对或限制其员工成立工会的权利，这一点非常重要。

健康权

世界卫生组织指出：“享受最高而能获致之健康标准，为人人基本权利之一。不因种族，宗教，政治信仰，经济或社会情境各异，而分轩轾。”

Meta 通过多种方式来尊重健康权，包括让人们更容易获取可信的健康信息，使有类似健康问题的人能够相互联系，并使他们能够就自己的身心健康做出明智的决定。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 解决潜在风险的措施
Meta 旗下平台上的医疗错误信息或虚假信息	2020 年，Meta 采用了一项政策，目的是移除符合下列情况的健康卫生错误信息： • 突发公共卫生事件期间的健康卫生错误信息； • 权威的国际卫生组织或地方卫生部门告知我们某言论不实时；以及 • 上述卫生组织或部门告知我们，某言论可能直接造成紧迫的人身伤害风险时。 Meta 继续使用 AI 工具来扩展事实核查机构的工作，以检测虚假和误导性的健康卫生信息，手段包括根据图片的视觉内容对图片进行分类。这使我们能够禁止某些医疗产品的违规广告和商业交易帖。
煽动或意图鼓励现实伤害行为的违规内容	Meta 的安全中心提供自杀预防的相关资源和其他信息，能帮助到可能有自杀念头的人。 我们的政策禁止发布鼓励自杀、自残、饮食失调行为或者欺凌和骚扰的内容，而且 Facebook 和 Instagram 力求阻止这类内容。我们发现这类内容后会立即将其移除，而且我们会继续改进用于检测和移除这类内容的技术。

06 尽职调查的最新情况

我们根据 UNGP 第 17 和 21 条，通过企业人权政策中规定的人权尽职调查来努力评估人权影响。这包括采取适当行动，落实调查结果，监督落实情况，以及每年报告相关见解和行动。

2022 年，我们发布了对以色列和巴勒斯坦、印度以及端到端加密进行人权影响评估 (HRIA) 和人权尽职调查的结果。我们继续落实这些评估和先前的评估中提出的建议。我们还推进了对元宇宙的潜在人权影响的尽职调查工作。

菲律宾

2021 年 12 月，针对我们在菲律宾委托第三方开展的独立人权影响评估，我们发布了其结果和我们做出的回应。此次人权影响评估就以下方面提出了建议：煽动暴力、监视、网络性剥削、贩卖人口、极端分子活动以及企业问责。

我们随后的一些行动包括：

- 在 2022 年菲律宾大选期间，我们与选举委员会、选举监督机构、独立核查机构和民间社会组织合作，打造了新的服务，并制定了更强有力的政策。这些工作拓展了我们多方面的能力，包括移除违规内容和网络，让更多人能获得可信的选举信息，提高数字素养和促进公民参与，以及提高政治广告的透明度。
- 我们在菲律宾开放了 Facebook 广告资料库功能，并要求投放选举、政治和特定类别社会议题广告的广告主完成我们的广告授权流程，并为广告添加“赞助方”信息。
- 我们对社群守则进行了多项更新，包括扩大对新闻工作者和人权维护者 (HRD) 等公众人物的保护。
- 我们移除了更多类型的违规内容，我们的政策现在针对性别骚扰为人们提供更强有力的保护。我们推出了针对大规模骚扰和控评的新政策，并移除了针对“有较高风险受到现实伤害的个人”发起的大规模合谋骚扰攻击。这包括对持不同政见者的攻击。
- 我们与国际记者中心和 Border Center for Journalists and Bloggers 合作，推出了一项免费的数字安全保障计划，目的是帮助新闻工作者和人权维护者保护他们的数字资产，并打击网络骚扰。Meta 的新闻工作者安全中心集中了我们应用中提供的所有资源和工具。

更多详情可在 2023 年人权落实情况更新中找到。

以色列和巴勒斯坦

2021 年 5 月期间，企业社会责任组织 (BSR) 在以色列和巴勒斯坦针对我们政策和流程的影响开展了独立的人权尽职调查 (HRDD)，我们发布了该组织有关这项调查的报告。这项调查是为了回应监督委员会的以下请求：通过独立审核来确定 Meta 对阿拉伯语和希伯来语内容的审核，包括对自动化工具的使用，是否存在任何偏见。

该报告揭示了在受冲突影响的地区，整个行业在内容审核方面长期面临的挑战，并指出有必要保护表达自由，同时降低网络服务被用来传播仇恨或煽动暴力的风险。该报告还强调，由于社会和历史的演变、各种快速发展的暴力事件以及恐怖组织的行动和活动，管理与冲突相关的问题非常复杂。BSR 没有发现 Meta 在这些问题上有意偏倚的证据。

在报告期间，我们随后采取的一些行动包括：

政策

- 与利益相关者广泛交流后，我们修订了危险组织和人物 (DOI) 政策中对“赞扬”一词的定义，这些利益相关者包括来自世界各地的学者、民间社会行为者，以及反恐、人权和表达自由领域的其他专家。

政策执行

- 经过广泛的内部评估，我们认为，在我们的系统中为各种阿拉伯语方言创建专门的审核分派路线，将有助于更精确地审核高严重性的阿拉伯语内容。我们目前正在评估如何开发相关的审核分派机制，高效地分派阿拉伯语方言内容的审核任务，从而提高阿拉伯语内容审核的准确性，并更好地防止政策执行不力和执行过度的问题。
- 我们已经针对以下事项开展了分析：构建一项针对方言的阿拉伯语分类技术，用于检测使用该语言的任何内容。根据这些结果以及语言学家和语言模型专家的意见，我们将在系统中添加扩展的语言识别功能，该功能可识别使用不同阿拉伯语方言的内容。
- 我们致力于更新我们的分类技术，以便定期提高准确性和表现。
- 我们推出了一项希伯来语分类技术，它可以主动检测并处理违规的希伯来语敌意言论内容。

如需详细了解建议落实情况的最新进展，请参见 Meta 更新：以色列和巴勒斯坦尽职调查最新情况。

印度

在 Meta 的第一份年度人权报告中，我们详细总结了印度人权影响评估的情况，该评估的设计参考了丹麦人权研究所 (Danish Institute for Human Rights) 有关数字活动人权影响评估的指南中的报告和评估部分。

根据此研究所和其他人权专家的指南，我们认为这种披露形式可减轻 UNGP 第 21(c) 条所述的安全风险。在本报告中，我们详细介绍了我们已经采取或打算采取的行动。我们将继续研究该评估的发现，并将其建议作为基准来确定和指导相关行动。2023 年及之后，我们将定期追踪我们对建议的落实情况。

自那时起，我们取得了以下进展：

- **利益相关者参与：**为了回应与 Meta 在印度全面深化利益相关者参与的潜力相关的见解，Meta 正在完善与民间社会交流合作和进行外联的方法，包括通过一项具体的计划来促进印度民间社会和社群组织的参与。
- **透明度报告：**我们关于印度的报告变得更加详细。除了报告政府的用户数据收集情况外，我们继续在每月提供详细的印度报告，其中包括以下方面的信息：针对 Facebook 和 Instagram 上的违规内容采取的行动，从印度用户那里收到的申诉，以及从最近根据 IT 规则成立的申诉上诉委员会收到的命令。我们继续完善我们的系统，以便针对因政府请求而移除的内容收集相关指标数据并进行公开报告，为了回应监督委员会的建议，这项工作反映在我们向其提交的报告中（2022 年第 4 季度和 2023 年第 1 季度）。
- **拓展合作：**我们承诺拓展与印度民间社会的合作并取得了良好进展，这包括数字素养、女性和儿童安全以及打击极端主义等各方面的合作。2023 年，女性安全中心变身为 Meta 安全中心重新亮相，并翻译成了阿萨姆语、孟加拉语、古吉拉特语、印地语、马拉地语、旁遮普语、泰米尔语、泰卢固语、卡纳达语和马拉雅拉姆语。我们扩大了印度组织对 Resiliency Initiative（抵御力倡议）的参与，并且正准备推出 Search Redirect Program（搜索跳转计划），让使用极端主义和暴力相关搜索词的用户跳转访问相关资源、教育内容和外展团体。
- **内容审核：**为了回应与进一步防止针对弱势群体的歧视性和仇恨内容的政策相关的见解，Meta 开发了一个在评估潜在仇恨言论时使用的原型测试，用于指导《拉巴特行动计划》的运用。我们已将《拉巴特行动计划》中的原则改编为有效的内容政策工具，包括基于升级处理的框架，用于评估攻击某些概念（而不是攻击人）的言论和涉及国家威胁使用武力的内容。在印度，对于所有类别的已认定危险组织和个人，我们都增加了其包含的已认定团体数量，以便确保我们的平台不会被不良行为者违规使用。

此外，Meta 在确保选举诚信方面的经验也为我们准备迎接即将到来的邦选举和大选提供了参考信息。这些工作包括启动我们的选举行动中心，确保内容审核员以 20 种印度语言提供支持，将我们的独立核查合作机构数量从 7 个增加到 11 个（目前覆盖 15 种印度语言），提高政治广告的透明度，以及与选举当局和民间社会密切合作。

07 利益相关者参与：外部其他方如何加强我们的工作

与世界各地的外部利益相关者交流合作有助于我们履行人权责任，并为实现问责和透明创造了重要手段。特别是，我们认识到，与来自代表性不足群体和边缘化群体的利益相关者进行有意义的交流合作非常重要。在我们的企业人权政策和相关工作中，利益相关者的参与贯穿始终。我们努力听取和寻求人权专家、活动人士、学者和其他人士的建议，并向他们介绍 Meta 的最新进展。他们的见解为多个领域提供了参考信息，例如我们内容政策的制定和执行，以及我们的算法排名和推荐方法。

2022 年，Meta 加强并扩展了我们与外部利益相关者的持续交流合作，这涉及以下团体。

利益相关者	咨询与合作方面的努力
民间社会组织	为社群守则审核以及选举等重要领域的诚信工作提供支持
人权维护者	采取措施保护帐户安全，最大限度减少错误信息和违规内容的传播
弱势或边缘化群体	为制定内容政策和打造安全服务建立反馈机制
国际组织	分享对数字人权政策的看法，并在重点国家/地区改善沟通
可信合作伙伴	这项全球计划旨在推动多方面的改进，包括识别和标记违规内容，贡献有助于我们执行政策的当地知识，以及打造安全服务
投资者和广告主	就 Meta 的人权工作（包括人权尽职调查）进行磋商并听取情况介绍
用户	就有争议的问题和分散决策提出意见

我们的工作包括：

在尼日利亚帮助建立安全的网络空间，支持选举诚信

在 2023 年尼日利亚大选之前，我们与各种各样的利益相关者交流合作。这些交流合作聚焦于我们的选举政策，以及为了维护我们应用中信息的完整性和准确性所做的工作。我们收集了有关违规内容趋势的洞察资讯，用于帮助我们改进违规内容的检测以及维护信息完整性。

我们的行动：

- 面向女性政治家和女性公众人物、人权维护者 (HRD) 和活动人士等弱势群体开展了数字安全培训
- 在尼日利亚举行了危险组织和人物 (DOI) 政策圆桌会议，并在伦敦与尼日利亚侨民举行了两场圆桌会议
- 通过研讨会和咨询与可信合作伙伴合作，帮助他们识别和上报仇恨言论、错误信息、暴力与煽动暴力、干预投票、欺凌和骚扰等性质的内容

让边缘化和代表性不足的群体参与内容政策的制定

我们希望在制定内容政策时考虑边缘化群体和代表性不足群体的意见，并在这方面取得了进展。我们制定了一个包容性框架，确保在制定政策时考虑多元化利益相关者的意见，并使用这个框架来指导社群守则的制定。

我们的行动：

- **征询有关社群守则的反馈：**在欧洲、中东和非洲，来自在我们应用中代表性不足的宗教、土著和种族群体的领导人参加了八场圆桌会议，讨论我们的社群守则在 Meta 旗下服务和应用中创造包容环境的效果。其中包括在肯尼亚大选后举行的圆桌会议以及分析波兰移民危机的圆桌会议。
- **加强与 LGBTQIA+ 群体的关系：**针对 Meta 内容政策（特别是有关仇恨言论、欺凌和骚扰的政策）对用户的影响，我们在欧洲、中东和非洲与团体和代表进行了更多交流合作。这包括来自英国、西班牙、德国、肯尼亚、加纳、纳米比亚、乌干达、埃及、苏丹、津巴布韦、南非、冈比亚和坦桑尼亚的 LGBTQIA+ 人权维护者、民间社会组织、学者和活动人士。
- **拓展与加勒比民间社会的联系：**我们听取了与潜在违规内容的传播有关的担忧，并引导大家就数字权利、网络安全和保障、女性安全以及 LGBTQIA+ 权利等问题展开讨论。来自海地、特立尼达和多巴哥、牙买加、巴拿马、多米尼加共和国和开曼群岛的专家和民间社会团体参加了首场此类圆桌会议。
- **与宗教群体开展更广泛的交流合作：**我们扩展了政策交流合作的对象范围，以包括英国的穆斯林女性、萨赫勒地区的穆斯林群体、英国的锡克教群体，以及巴哈伊教群体（其代表来自加拿大、美国、智利、巴西、德国、西班牙、瑞士、南非、阿联酋、埃及和突尼斯）。
- **制作了一份有关吸引残疾人互动的指南：**我们与从事残疾人与科技政策交叉领域工作的残疾人权利组织合作，为我们的内容政策团队制作了一份指南。

土著群体对非医用药物的看法

我们与北美、加蓬、喀麦隆和津巴布韦的利益相关者进行了交流合作，以便在制定有关在宗教或传统活动中使用非医用药物的政策时考虑他们的意见。我们探讨了如何在以下两者之间取得平衡：一是，与具有传统和宗教用途的物质有关的表达自由；二是，死藤水等非医用药物的潜在推广带来的安全风险。我们咨询了范围广泛的学者和民间社会组织，包括专注于健康、医学、人类学、临床心理学、药物政策、人种学、民族鸟类学、宗教和迷幻药物研究的学者，以及法律专家、监管机构、表达自由倡导者、医疗保健专业人士、传统/宗教治疗师协会、社群领袖和土著团体。我们根据收到的意见，修订了我们在积极讨论将非医用药物用于传统或宗教用途这一问题时所用的方法。

与国际组织交流合作

科技的快速发展和影响促使许多国际组织制定政策，解决数字世界中的人权问题。我们经常与各种国际政府机构合作，包括人权事务高级专员办事处、联合国教科文组织、Office on Genocide Protection and the Responsibility to Protect（防止灭绝种族罪行和保护责任问题办公室）、联合国儿童基金会、世界经济论坛、互联网治理论坛、联合国难民署、经济合作与发展组织、联合国基金会以及各种地区组织。我们提供了与我们的社群守则、它们的执行情况以及我们如何努力保护人权维护者有关的深入见解。2022 年，我们在国家层面（包括在埃塞俄比亚、肯尼亚、乌克兰、海地和阿富汗）加强了与联合国的合作。

我们的行动包括：

- 联合国儿童基金会与 Meta 密切合作，确保为乌克兰人提供关键信息，同时提供广告抵用金，以支持在全球主要市场、乌克兰以及诸如波兰和罗马尼亚等邻国传递筹款和倡导信息。
- 我们定期与联合国人权事务高级专员办事处的官员会晤，讨论全球和具体国家/地区的问题，并积极参与 B-Tech 项目。
- 我们为联合国驻埃塞俄比亚、尼日利亚和海地的国家工作队提供了社群守则和报告机制方面的培训。

- 我们与秘书长技术问题特使办公室展开了交流，讨论了《全球数字契约》，并强调该契约需要支持全球统一的互联网治理方法，促进数据的自由流动。这对于在网上保护人权至关重要，包括隐私权和表达自由权。
- 在联合国大会高级别周那段时间，我们与联合国高级官员和外交官举行会议，讨论人权及相关问题。
- 我们与联合国人权理事会的多位特别报告员进行了交流，其中包括专注于人权维护者处境、表达自由、阿富汗和缅甸等问题的特别报告员，以便分享信息并更好地了解他们有哪些与 Meta 服务所扮演的角色相关的担忧。
- 我们继续回应来自联合国缅甸独立调查机制的数据请求，以协助问责工作。

依靠可信合作伙伴来大规模识别和上报问题

在世界上我们面临最突出人权挑战的地区，我们会优先建立合作关系。

我们的可信合作伙伴网络包含遍布 113 个国家/地区的超过 400 家非政府、非营利、国家和国际组织，他们会举报有可能违规的内容、帐户和行为，供我们结合语境/背景进行审核。我们努力了解我们的服务对当地的影响，而这些可信合作伙伴是我们在这方面的重要盟友。他们帮助我们跟上新兴趋势，还能标记我们可能漏掉的违规内容，并改进错误信息和伤害等领域的政策和执行。他们的举报提供了宝贵的见解，有助于为制定政策提供参考信息，并帮助保护我们用户的安全。

我们的行动：

- 2022 年，我们拓展了可信合作伙伴网络，涵盖了 36 个新的国家/地区，包括也门、乌克兰、埃塞俄比亚、马里、肯尼亚、尼泊尔、海地和柬埔寨。
- 在合作伙伴的提醒下，我们注意到了一些我们原本可能会忽略的事态发展和风险。例如：
  - 可信合作伙伴举报称，在也门，我们应用中发生的性勒索事件激增，因此我们展开了调查，并最终移除了大量 Facebook 用户、企业帐户和 Instagram 帐户。
  - 在柬埔寨，可信合作伙伴标记了一些为东南亚呼叫中心招聘工作人员的 Facebook 帖子，这些帖子包含人口贩卖和劳动剥削的关键迹象。超过 30 篇帖子被移除。
  - 在巴基斯坦，可信合作伙伴标记了有关《跨性别人士法》(Transgender Persons Act) 的错误信息以及将个人置于风险之中的针对性内容。我们移除了该内容，并采取了措施确保我们会捕捉到今后出现的类似违规内容。

通过社群论坛进行创新

Meta 继续探索如何将硅谷之外的更多声音融入到决策当中。我们试行了基于审议民主的社群论坛，“审议民主”是全球政府和组织都在使用的一种决策方式。它使代表性人群汇聚一堂，就特别复杂或有争议的问题进行有条理的对话，并做出决定。

我们的行动：
2022 年，我们采用了基于审议式民调 (Deliberative Polling®) 的集体审议模式，“审议式民调”是斯坦福大学审议民主实验室 (Stanford Deliberative Democracy Lab) 创建的一种方法。我们与该实验室以及位于英国的数据分析团体“行为洞察团队”(Behavioural Insights Team) 合作，确保该流程是公平而独立的。

此论坛是迄今为止规模最大的审议式民调论坛，有来自 32 个国家/地区的 6,300 多人参加。该论坛侧重于解决社交虚拟空间中的欺凌和骚扰问题。

参与者收到了经过独立审查且与该主题相关的教育材料，在多个小组会议中展开了审议，并有机会向来自世界各地的独立专家提问。这些专家包括网络安全专家、人权律师、反欺凌/骚扰活动人士、沉浸式虚拟现实专家，以及 Twitter、TikTok 和 Second Life 的前品牌安全负责人，还有联合国和世界经济论坛咨询委员会成员和国际新闻自由奖获得者。

在这些会议之后，参与者接受了民意调查，表明是否支持他们所讨论的提案。这些调查旨在为虚拟体验的服务和应用开发提供参考信息。我们致力于创建计划来促进更具包容性的决策，让更多人在我们的应用和技术开发过程中拥有发言权，此活动就是这项大计中的一环。

总的来说，参与者强烈支持这样的观点：创作者和平台所有者都有责任监督他们所创造的世界，不论是公开空间还是仅对成员开放的空间。

参与者支持使用自动语言检测、视频捕捉以及可见版主和隐身版主等工具。相比仅对成员开放的空间，在公开空间中使用这些工具的呼声更高。

如需更详细的结果，请参见斯坦福大学的结果报告。

08 语言

语言是攸关 Meta 人权影响的最重要因素之一。人权突出风险综合评估中的利益相关者咨询意见部分强调了语言的重要性。2022 年，我们提升了能力以支持更多语言，让更多人能在网上行使其权利，包括获得补救。

了解当地的情况对于了解潜在人权风险至关重要。因此 Meta 的语言能力包括人工翻译和机器翻译。语言支持包括语言检测和基于语言的分类技术、内容审核、用户界面的翻译以及其他内容，例如帮助中心或社群守则。

我们一直在努力增加对更多语言的支持。从 2021 年到 2022 年底，我们的团队又以 30 种额外语言为 Facebook 用户界面提供专业支持。这份列表包括在互联网上代表性不足的语言，以及难以进行语言学溯源和翻译的语言，例如爪哇语、绍纳语、基隆迪语、僧伽罗语、尼泊尔语和阿萨姆语。

此外，截至 2022 年底，我们的社群守则已有 77 种语言版本。

No Language Left Behind

Meta 的 No Language Left Behind（意为“一种语言也不能少”，简称 NLLB）是同类首创的突破性 AI 项目，其使用的开源模型能够直接在 200 种语言之间提供经过评估的优质翻译，这包括阿斯图里亚斯语和卢干达语等较少使用的语言。该项目旨在让人们有机会以自己的母语访问和分享网络内容，从而能与任何地方的任何人交流，不论其语言偏好如何。

NLLB-200 模型背后的技术现在可通过维基媒体基金会的内容翻译工具获得，该技术正在支持维基百科编者将信息翻译成他们的母语和首选语言。这有助于以更多语言向世界各地的维基百科读者传播更多知识。

09 问题聚焦

在危机期间，人权往往面临特别的威胁。在这种情况下，Meta 会评估平台内外发生紧迫伤害的风险，并通过具体的政策、服务和业务行动来应对，从而尊重人权。

2022 年，Meta 加强了整个组织的应对能力，在任何国家/地区的紧张局势升级期间（包括在冲突、选举或内乱期间）为民众提供支持。为此，我们组建了一支跨职能团队，重点关注如何将支持和资源用在刀刃上，高效地预测世界任何地方紧张局势升级的时刻并做出妥善应对。这项工作是围绕全球选举、积极响应的国家/地区以及全球准备情况来组织的，以便预测将来的关键事件并做好应对准备。我们制定和完善了一系列完整性标准，让用户能够举报违规内容，这些标准通过各个国家/地区独具特色的语言专业知识来体现。

危机政策协议

2022 年 8 月，根据监督委员会的建议，我们推出了内部危机政策协议 (CPP)，以帮助确保我们对危机做出的政策响应是有原则和经过合理调整的。

危机政策协议是一个动态的框架，使我们能够识别紧急危机情况并评估其相对严重性。该协议指导我们根据观察到的风险，借鉴过去的危机干预措施，迅速使用有针对性或特殊的政策手段来减轻潜在伤害。因此，借助危机政策协议，我们既能在全球采取一致的危机应对措施，又能适应迅速变化的当地情况，做到两者兼顾。为了制定危机政策协议，我们咨询了 50 多位人权、冲突预防、仇恨言论、人道主义响应和国家安全领域的全球外部专家。此外，我们还开展了原创调研。

此协议旨在补充和加强全公司现有的危机应对工作。危机政策协议有助确保 Meta 采用全面、跨学科的政策方法，推动服务、政策和运营团队不断改进危机应对工作。

在我们的危机应对活动中，值得一提的例子包括：

俄罗斯和乌克兰

俄罗斯在 2022 年 2 月入侵乌克兰，给 Meta 带来了重大人权挑战。影响公众舆论的宣传和虚假信息在网上泛滥，监视、互联网服务的中断和宣传违规内容的行为也出现激增。我们需要加快内容审核工作，同时兼顾表达自由，以帮助乌克兰公民协调人道主义救援工作，并表达他们对入侵军队的抵抗。

在俄罗斯，Facebook 和 Instagram 都被封禁了，并且 Meta Platforms, Inc. 被当局列入了极端组织名单。然而，WhatsApp 仍然可用。

尽职调查：
自战争爆发以来，我们的人权尽职调查包括解释相关的国际人道主义法和人权标准，帮助我们的团队应对挑战。这些工作包括审核揭露战俘身份的视频、宣扬仇恨言论或错误信息的视频，以及传播受俄罗斯政府控制的、否认俄罗斯入侵乌克兰的宣传内容的视频。

我们在俄罗斯和乌克兰与人权维护者 (HRD)、活动人士、新闻工作者、民间社会团体和人权团体进行了接触，以了解他们面临的问题。我们还与联合国和其他国际机构进行了交流，以加深我们对相关问题的理解，并进一步了解战争时期的内容审核最佳实践。

我们的行动：
Meta 采取了一系列措施来应对挑战，包括：

- **建立一个特别行动中心**，配备来自全公司的专家，包括俄语和乌克兰语专家，该中心全天候运作，以实时监控和应对瞬息万变的战争形势。
- **加大力度处理违规内容**，例如暴力与煽动暴力、仇恨言论、合谋造假行为、错误信息和虚假信息。这包括在该地区拓展我们对俄语和乌克兰语内容的第三方事实核查能力，并向乌克兰的事实核查合作机构提供额外的财政支持。我们还牵手五个新的可信合作伙伴，与他们召开了聆听会，以提高我们对内容相关风险的认识，并制定适当的举报机制。

- **在全球范围内对受俄罗斯政府控制的媒体机构的 Facebook 公共主页和 Instagram 帐户内容进行降级处理**，使这些内容更难在我们的应用中找到。我们还开始对符合以下情况的 Facebook 和 Instagram 帖子和快拍添加标签和进行降级处理：包含指向受俄罗斯政府控制的媒体网站的链接。在欧盟和乌克兰，我们在其各自政府的坚持下封锁了所有俄罗斯官方媒体。
- **推出旨在保护人权维护者和其他弱势群体的安全功能**，包括让人们能锁定自己的 Facebook 个人主页，取消查看和搜索好友名单的功能，并在 Messenger 添加工具，帮助人们免受攻击。
- **帮助确保人权维护者**和独立媒体的帐户受到更多保护，免遭骚扰或冒充，并通过我们的交叉检查计划确保他们的声音不会受到不当限制。这项工作仍在进行中。
- 我们的 Data for Good（公益数据计划）团队与可信合作伙伴分享可保护隐私的数据集，由此**帮助预测难民潮**。
- 临时将社群互助平台用作 Facebook 上的资源集中地，**帮助乌克兰人**和该地区的其他人找到来自当地联合国机构和红十字会的可靠信息。

伊朗

2022 年 9 月，22 岁的玛莎 · 阿米尼 (Mahsa Amini) 因头巾佩戴“不规范”而被伊朗道德警察逮捕。她在遭警方拘留期间死亡，此事在伊朗和世界各地引发了广泛的抗议，而且抗议活动仍在继续。为应对抗议活动，伊朗当局大力压制言论自由和集会自由，并限制民众使用互联网以及 Instagram 和 WhatsApp 等伊朗人广泛使用的应用。

Meta 迅速行动，采取了一系列措施，帮助人们连接到我们的应用并在应用中保持安全。这包括成立一个专门的波斯语团队，重点解决抗议活动引发的问题。此团队确保我们正确地运用我们的政策并保护用户的安全。我们将继续密切关注事态发展。

我们的行动：帮助抗议者传播信息

伊朗人广泛使用 Instagram 来通报抗议活动的情况和揭示当地事件的真相。在抗议活动的头五个月，与伊朗抗议活动有关的话题标签在 Instagram 的使用次数超过 1.6 亿次。抗议者与国际媒体分享了他们在 Instagram 的视频，其中许多媒体无法从伊朗进行报道。

- **帮助人们保持连接：**2022 年，我们在伊朗推出了 Instagram Lite，帮助人们在带宽减少的情况下访问 Instagram。此外，WhatsApp 推出了代理服务器功能，让人们能在互联网中断或被封锁的情况下进行连接，这对于伊朗民众尤为重要。
- **保护活动人士和人权维护者：**在抗议活动期间，活动人士和新闻工作者的社交媒体帐户经常成为欺凌、骚扰、冒充和黑客攻击的目标。我们与伊朗活动人士合作，携手保护他们的帐户，并帮助他们提高对人权侵犯行为的认识。

10 我们在人权和负责任产品开发方面的方法

元宇宙

2021 年，Meta 创建了 XR 计划和研究基金 (XR Programs and Research Fund)，这是一项为期两年、金额为 $5,000 万美元的投资，用于与行业合作伙伴、民权和人权团体、政府、非营利组织和学术机构合作开展计划和外部研究。这些投资重点关注 Meta 必须采取正确做法的领域，以实现元宇宙的潜在益处。下面列出了四个优先政策领域，所有提案必须与 Meta 在其中一个或多个领域的工作保持一致：

- 经济机遇：我们如何给用户提供更多选择，以及保持数字经济的蓬勃发展
- 隐私保护：我们如何在产品中以有意义的方式做到信息透明，并为用户提供控制权
- 安全诚信：我们如何保障用户在我们平台上的安全，并提供各种工具，供用户在看到或遇到不良内容时采取行动或寻求帮助
- 公平包容：我们如何确保这些技术采用包容性设计，且人人都可使用

这项举措的合作伙伴包括大学和研究机构、国际组织、民权团体和非政府组织，这些非政府组织来自世界各地代表性不足的群体，或与这些群体有合作。

人工智能

我们认为，以负责任的方式开发和部署的人工智能 (AI) 技术可成为推动人权的强大工具。AI 在帮助我们解决本报告中讨论的许多潜在负面人权影响方面发挥着重要作用。例如，我们已经借 AI 之力在越来越多的语言中扩展了对问题内容的检测和审核，提高了对潜在扰人联系和互动的识别能力，通过自动生成字幕让内容更加易懂，并加强了我们的一系列 Data for Good（公益数据计划）工作。

与此同时，AI 的发展和使用也引发了新的人权风险。除其他潜在风险外，AI 模型可能会表现出有问题的偏见或歧视性影响，并生成有问题的内容，或者是过度执行政策，从而对表达自由产生不利影响。

我们致力于以开放和合作的方式解决这些问题。例如，若想评估某项服务、应用或流程是否公平对待所有群体，人口统计数据至关重要。然而，收集这些数据是敏感行为，会引发人们对如何保护个人隐私的重要担忧。

我们的方法以责任和人权为中心。AI 已纳入到我们的企业人权政策中。我们也认识到了《经合组织人工智能原则》(OECD Principles on Artificial Intelligence) 的重要性，这些原则明确提到了人权问题，并得到了二十国集团 (G20) 的广泛采纳和认可。我们希望促进各个学科领域的人员和 AI 技术的受众开展更加协作和透明的对话，讨论这些关键问题的未来走向。

我们创建了专门的跨学科负责任 AI (RAI) 团队，旨在确保 Meta AI 技术能造福人类和社会。该团队的成员包括伦理学家、社会和政治科学家、政策和权利专家、研究人员和工程师。负责任 AI 团队与人权团队密切合作，帮助确保以负责任的方式设计和使用我们的机器学习系统。

我们还开创了一种 AI 系统说明卡方法，以标准化方式提供透明信息，说明 AI 系统（而不仅仅是单个模型）如何运作，我们首先重点对 Instagram 动态排名试行了此方法。

负责任 AI 的五个要素

Meta 的负责任 AI 建立在五个要素的基础上：

- 隐私与安全
- 公平与包容
- 稳健与可靠
- 透明与可控
- 问责与治理

我们在解决其中许多问题方面取得了进展，但仍然任重而道远。2022 年，我们致力于：

- **扩展语言支持**
  - 我们的开源 No Language Left Behind 项目利用 AI 直接在 200 种语言之间提供优质翻译。
  - 我们的大规模多语言语音（Massively Multilingual Speech，简称 MMS）识别模型支持 1,107 种语言的语音转文字和文字转语音，并支持超过 4,000 种语言的语言识别。一些受支持的语言（例如 Tatuyo）只有几百名使用者，而且其中大多数语言以前并不存在语音技术。我们公开分享了这些模型和代码。
  - 维基百科与 Meta AI 团队合作，使用 Meta 的开源 AI 程序 Sphere 对引文进行事实核查。
- **解决偏见**
  - 为了确保 AI 系统公平对待每个人，我们引入并开源了多个数据集和模型，它们负责以保护隐私的方式解决这方面的不足。
  - 我们训练了一个用于减少文本中的人口统计偏见的 AI 模型，这是一种人口统计文本扰动器，有助于打破自然语言编程 (NLP) 数据中存在的刻板印象关联。
  - 我们引入并开源了两个新的数据集，用于帮助衡量 NLP 模型中的公平性并减少潜在偏见。这些数据集更全面地体现了不同的人口统计维度，包括与性别认同、年龄、种族和残疾人士相关的术语，以衡量这些模型中的公平性。
- **更好地了解有问题的内容关联**
  - 我们组建了一支跨学科团队，其中包含来自我们的民权、人权、工程、负责任 AI、AI 研究、政策和产品团队的人员，以便更好地了解我们的一些端到端系统中有问题的内容关联。我们努力采用技术缓解措施来减少使用 AI 模型的应用中出现有问题的内容关联的几率。

- **评估公平性**
  - 为了解决 AI 公平性问题，我们的方法之一是创建和分发更多样化的数据集。我们也通过 Casual Conversations v2 项目，设计了一个基于用户同意的大型数据库，用于衡量算法偏见和稳健性。
  - Meta AI 研究团队开始探索新的公开公平性指标，用于定量评估计算机视觉模型中三种有据可查的潜在伤害和偏见类型。这些公平性指标补充了现有的负责任 AI 开发方法，例如数据和模型文档。这些指标经过专门设计，可随着研究的进展和新方法的出现而调整和发展。

我们也在 SEER (SElf-SupERvised) 上取得了进展，这是 Meta AI 研究团队开发的自监督计算机视觉模型，专注于改善不同图像集的结果，而无需像传统计算机视觉训练那样进行仔细的数据搜集整理和标注。

11 前景展望：未来的考量

人权标准和原则仍然是人类对普世价值观的最佳诠释，也是建立合法治理制度的最佳基础。过去几年来，互联网治理模式取得了长足进步。正如本报告所述，我们在整个 2022 年都在努力继续推进我们的治理模式，并履行我们的企业人权政策和我们根据《联合国工商企业与人权指导原则》做出的承诺。

这项工作仍在继续，我们也将继续履行承诺，努力取得进展。我们期待实施本报告中讨论的许多先进的问责机制。例如，我们在继续开展工作，让人们能通过社群论坛和集会参与治理措施。

我们还将再接再厉，以人权为中心来理解和采用快速发展的人工智能 (AI) 科学技术。虽然前方挑战重重，任重道远，但我们能通过各种努力来解决社交媒体和 AI 技术带来的一些最棘手的问题，对此我们深感自豪。

我们将继续依靠人权尽职调查来更好地确定使用我们服务的用户所面临的最重要的人权风险，并制定政策来应对这些风险。

我们在一个错综复杂而支离破碎的监管环境中运营。我们许多人每天都在使用的无国界且基本上自由的互联网正在受到威胁。我们必须与政府、联合国和民间社会合作，确保不断发展的监管框架建立在人权原则和价值观的基础上，这一点至关重要。我们充分认识到所肩负的人权责任，并将努力在我们的日常工作中推进这些原则，以造福我们的用户和整个社会。