Meta

人权报告

洞察与行动
2023 年

目录

关于本报告

本报告是我们的第三份年度人权报告。本报告以 Meta 开展的工作为基础，介绍了 Meta 在履行对《联合国工商企业与人权指导原则》和企业人权政策的承诺上所取得的进展情况。本报告涵盖我们从 2023 年 1 月 1 日至 2023 年 12 月 31 日采取的各项行动。报告涉及的 Meta 服务和产品包括 Facebook、Messenger、Instagram、WhatsApp、Threads 和 Reality Labs。

无论是对于我们的利益相关者还是我们公司而言，人权的话题都至关重要。本报告的内容体现了我们在发布 2022 年人权报告之后收到的利益相关者反馈，并建立在我们的人权突出风险综合评估 (CSRA)（2022 年人权报告对该评估做出了概述）的基础之上。人权突出风险综合评估旨在确定并优先考虑我们可能对人们（包括我们的用户以及可能因为我们整个企业在全球范围内所采取的行动而受到影响的其他人）造成最严重负面人权影响的领域[1]。本报告将重点介绍 Meta 在全球运营中对八个人权突出风险领域的管理方法：

通过人权突出风险综合评估确定的 8 个最重要的潜在人权突出风险领域

人权突出风险领域	在本报告中的章节
意见和表达自由	内容政策、Meta 如何为选举做好应对准备、应对全球危机、以色列和巴勒斯坦：我们的最新行动、儿童和青少年安全、回应政府请求、监督委员会
隐私	内容政策、端到端加密：我们的最新行动、应对安全威胁、儿童和青少年安全、回应政府请求
平等和无歧视	人权背景下的 AI、产品开发、以色列和巴勒斯坦：我们的最新行动
生命、自由和人身安全	Meta 如何为选举做好应对准备、应对全球危机
儿童最大利益	内容政策、儿童和青少年安全、利益相关者的参与
公众参与、投票和竞选	Meta 如何为选举做好应对准备：巴西、巴基斯坦
结社和集会自由	端到端加密：我们的最新行动、应对全球危机：苏丹
健康权	应对全球危机：苏丹

[1] “负面人权影响”一词与《联合国工商企业与人权指导原则》中的含义相同，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。

本报告各章节涉及的人权突出风险领域列于各章节的顶部。

我们通过与一众内外部利益相关者合作，力求介绍我们全球范围内的多个团队所开展的具有代表性的工作。本报告旨在从整体上概述我们公司处理的人权问题，同时会引导读者根据需要探索深度内容。

本报告是对最新 Meta 负责任商业行为报告的补充。对于确定和降低企业运营和供应链中存在的现代奴隶制和人口贩卖风险方面的工作，我们会单独编写报告进行说明。本报告的附录部分提供了其他 Meta 信息披露的链接。

我们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。本报告所提及的行动为 Meta 作为一家公司针对 Meta 旗下一个或多个实体所采取的行动。报告中的陈述无意暗示 Meta 针对所有实体采取了相同的行动。[2]

[2] 例如，WhatsApp 是一款采用端到端加密的消息和通话应用程序，具有自身独有的人权影响领域。本报告对 Facebook 和 Instagram 内容审核和相关行动的介绍不适用于 WhatsApp，并且除非指明了某项政策或行动适用于 WhatsApp，否则该政策或行动应被视为不适用于 WhatsApp。此外，虽然本报告中所述的许多行动适用于 Facebook 和 Instagram，但是这两种服务的政策和程序之间存在有意识的区分。如果某项政策被标注为“Facebook”政策，则其不一定适用于 Instagram。本报告中的任何陈述均无意建立将某项政策或程序应用于其他服务或实体的新义务（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。

执行摘要

我们旨在通过这份人权报告详细阐明 Meta 采用了哪些政策、流程和标准，实时、规模化地管理数十亿人所面临的人权风险。我们希望说明我们如何在日常工作中和每天使用的系统中努力尊重用户和其他社群所享有的权利，包括从我们应对危机、为选举做准备的方式，到我们回应政府请求的方式。

今年的报告涵盖了我们在 2023 日历年所开展的各项工作，并体现了利益相关者就去年的人权报告提出的反馈意见。今年的报告将重点介绍 Meta 在全球运营中对以下八个人权突出风险领域的管理方法：意见和表达自由；隐私；平等和无歧视；生命、自由和人身安全；儿童最大利益；公众参与、投票和竞选；结社和集会自由；以及健康权。

一如既往地，Meta 的各项服务和应用继续为公益事业做出贡献，帮助建设社群、促进创新和调集资源，以支持人道主义救援等。随着各种挑战的涌现，我们以企业人权政策为指导采取行动来应对各种问题，例如在冲突国家/地区使用我们的服务、网络安全和政府过度索取数据等。

在理解和开发人工智能 (AI) 方面，我们仍将继续以人权为核心。我们的生成式 AI 产品允许人们以新颖的方式促进自身的人权，但是我们也认识到，这些系统并非完美，存在继续改进的空间。为了降低潜在风险，我们制定了合理使用政策，并将其纳入了开放式 Llama 2 模型的社群许可协议中，同时我们还提供了安全工具，帮助开发者负责任地进行构建。我们相信，开放式方法可以通过不断的迭代推动创新，而持续的反馈、审查、发展进步和问题缓解，将带来更优质、更安全的产品。

主动与利益相关者接洽是我们开展人权工作的核心，这种方法已纳入我们的企业人权政策中。我们的服务开发、内容政策和内容审核以及社群守则均参考了多方的意见和见解，包括众多民间社会组织、人权维护者、边缘化群体、国际组织、投资者、广告主、用户和可信合作伙伴的意见和见解。这些团体来自各种政治和意识形态领域，为我们提供了多元化的意见和建议。截至 2023 年 1 月，我们的可信合作伙伴网络已纳入来自全球 113 个国家/地区的 400 多个非政府组织、人道主义机构、人权维护者和研究人员。

2023 年是 Meta 加入全球网络倡议 (GNI) 的十周年。我们遵守 GNI 有关科技公司在处理政府请求和限制时应如何尊重用户的表达自由和隐私权的原则，并对此负责。在本报告中，我们概述了 GNI 如何指导我们回应政府请求，包括要求限制内容和访问用户数据的政府请求。

我们在应对 2023 年发生的重大事件和问题时，也会考虑到人权原则。本报告阐述了我们为高风险国家/地区日益加剧的挑战所做的应对准备、我们如何为选举等计划内重大事件做准备，以及我们如何协同应对危机事件。2023 年，Meta 为全世界 130 多场选举做好了应对准备。在本报告中，我们介绍了对苏丹冲突、纳戈尔诺-卡拉巴赫冲突以及中美洲和南美洲移民人权问题的应对措施。参与危机应对工作的人员包括来自内容政策、全球运营、人权和产品领域的一众专家以及其他内部专家，这些专家会根据我们的危机政策协议对危机形势做出评估和回应。我们还阐述了对此前有关以色列和巴勒斯坦的人权尽职调查中所提建议的落实情况、我们就近期冲突所采取的行动，以及端到端加密的最新信息。

我们进一步强化对公开透明和补救措施的承诺，并将其作为人权风险管理的核心宗旨。我们继续与监督委员会开展合作。监督委员会是一项业内领先的创举，其旨在帮助 Meta 解答关于表达自由和网络安全领域的一些最棘手的难题。2023 年是监督委员会具有里程碑意义的一年，这一年监督委员会超越了每年做出 50 项决定的目标，做出的案件决定数量达到 2022 年的三倍多。

此外，我们还针对 Meta 在 Facebook 和 Instagram 上根据政府和法院的法律请求而做出限制的内容，改善了用户的体验。在大多数情况下，我们在通知中告诉用户是哪个国家/地区当局提出了请求而导致内容被限制，以及内容在哪个国家/地区受到了限制。

2023 年，我们向 Lumen 数据库提交了来自奥地利、越南、新加坡、印度和墨西哥的首批移除请求。该独立研究项目由哈佛大学伯克曼互联网与社会研究中心发起，让研究人员能够调查政府和私人行为者就网络内容提出的移除请求。在为全球社群提供分析、报道和倡导互联网用户数字权利的渠道上，该项目向前迈出了新的一步。

序言

随着世界不断向前发展，势必会继续涌现新的复杂挑战。我们在应对这些挑战时，以我们的企业人权政策以及我们对《联合国工商企业与人权指导原则》和联合国全球契约组织 (UN Global Compact) 做出的承诺为指导。我们为此还设立了制衡制度。监督委员会作为一种专业且独立的监督机制，可确保决策的公正与平衡，特别是在表达自由问题上。

随着科技和社会的发展演变，我们的政策和程序也在不断变化。我们会广泛考虑各个利益相关者的意见，重点关注各个维度的多元化，包括观点的多样性。在修订这些政策和程序时，我们会征求员工、专家和人权维护者、边缘化群体和国际组织以及广泛的民间社会团体的反馈意见。

我们在整个企业内使用各种评估工具和协议，帮助识别、预防和减轻可能与 Meta 产品、政策和全球运营相关的人权风险。我们会就往期的评估工作发布报告，分享我们取得的新进展并提供反馈渠道。

2023 年，我们启用了风险管理工具，用于应对危机局势（包括非洲和中东地区的冲突和战争）以及为可预测的世界事件（例如，大量的选举活动）做好应对准备。可见，我们的政策具有足够高的灵活度，可确保我们在平衡各种利益冲突和人权考量因素的同时，能够对每种情形采取适当的、有针对性的应对措施。

十多年来，Meta 一直是人工智能 (AI) 开发领域的先驱。我们知道，进步与责任必须齐头并进。生成式 AI 工具带来了巨大的机遇，我们认为，以尊重人权的方式开发这些技术既是可行的，也是必要的。

一个强大、开放的数字市场不仅利国利民利社会，还对建设社群具有重大意义。如今，我们更需要利用科技紧密连接整个世界。我们将竭尽全力，坚定不移地维护人权。

签署人：

Nick Clegg
全球事务总裁

Jennifer Newstead
首席法务官

人权背景下的 AI

相关的突出风险：
- 意见和表达自由
- 平等和无歧视
- 隐私

2023 年，人工智能 (AI)（尤其是生成式 AI）的快速发展引发了全球关注。鉴于 AI 受到的关注越来越高以及随之而来的潜在机遇和挑战，我们将从人权角度深入探讨我们在 AI 领域所开展的工作。

2023 年，我们发布了开放式 Llama 2 大型语言模型 (LLM)、Meta AI 助手、Emu 图片生成模型和应用内创意工具。

随着成熟的 AI 聊天助手和图片生成工具开始广泛普及，这也引起了公众和监管机构对监督这些技术的重视。例如，欧盟推进了其具有里程碑意义的 AI 法案，美国则在 Meta 和其他 AI 开发者的支持下推出了白宫 AI 自愿承诺书 (White House Voluntary Commitments on AI)。

我们承诺以负责任的方式开发和部署 AI，同时减轻对人权的潜在负面影响。我们在企业人权政策中纳入了对 AI 的规定，同时也认识到《经合组织人工智能原则》(OECD Principles on Artificial Intelligence) 的重要性（该原则已获得二十国集团 (G20) 的广泛采纳和认可）。

我们认为，负责任 AI 可成为推动人权进步的强大工具。

我们认为，负责任 AI 可成为推动人权进步的强大工具。我们的生成式 AI 产品让人们能以新颖的方式行使表达自由权、改善信息和教育资源的获取，以及改进无障碍使用体验。例如，Ray-Ban Meta 智能眼镜的 AI 字幕生成和图像识别功能可为残障人士提供更好的无障碍使用体验。

我们利用 AI 在越来越多的语言中快速检测和处理涉嫌违反政策的内容和扰人互动，以防止事态恶化升级。在许多 Data for Good（公益数据计划）工作中，我们都会利用 AI 来支持对危机的人道主义响应以及帮助开展公共卫生工作。

采取开放式方法

Meta 致力于以负责任的方法开发和部署生成式 AI 产品和模型。

我们相信，开放式方法可以通过不断的迭代推动创新，而来自社群的持续反馈、审查、发展进步和问题缓解，将带来更优质、更安全的产品。开放式方法还有助于：

- 使受影响的权利所有者能够更好地识别潜在的偏见，从而促进公平并支持各种观点的表达。
- 降低创新壁垒，带来经济效益。
- 让全球社群更容易针对特定语言和背景打造个性化的 AI 工具。

不过，我们也认识到，开放式方法也可能带来潜在的人权风险，包括开发者可能会忽视 Meta 的负责任使用说明，以不安全的方式部署模型。为了降低这些风险，我们开展了安全测试并制定了详细的合理使用政策，这些政策已纳入 Llama 模型的社群许可协议中。我们还随 Purple Llama 项目分享了一份详细的负责任使用指南并提供了安全工具，包括 Llama Guard，帮助开发者以负责任的方式使用这些公开的基础模型进行构建并将其安全地部署到自己的用例中。

我们相信，开放式方法可以通过不断的迭代推动创新，而来自社群的持续反馈、审查、发展进步和问题缓解，将带来更优质、更安全的产品。

安全部署 AI 是整个生态系统的共同责任，正因如此，我们多年来一直在与致力于构建安全、可信 AI 的组织合作。此处略举几例：

- 我们一直在与 MLCommons 及一批全球合作伙伴合作，以有利于开源社群的方式制定责任标准。
- 我们加入了 Partnership on AI 等多方利益相关者倡议，为解决各种问题出谋划策，包括从负责任地部署基础模型，到识别合成内容的最佳方法等等。
- 我们联合成立了 AI 联盟 (AI Alliance)，一个由企业、学术机构、倡导者和政府机构组成的联盟，致力于开发各种工具，建立一个开放、安全的 AI 生态系统。

我们积极与人权利益相关者沟通交流，帮助他们了解我们对 AI 的使用情况并收集他们的反馈意见。在这项工作中，我们还与联合国人权办公室的 B-Tech 项目就其在生成式 AI 方面的工作展开了交流合作，为三份以《联合国工商企业与人权指导原则》为基础的基本文件提供了意见和见解。我们还与联合国促进和保护意见和表达自由权问题特别报告员办公室的一名顾问以及联合国教科文组织 (UNESCO) 交流合作，帮助他们认识 AI 的基本概念和风险，并促进知情讨论，了解人们关切的问题。此外，我们还与斯坦福大学联合举办了一场咨询性社群论坛，其中也探讨了人权问题，目的是帮助 Meta 及其他公司、研究人员和政府做出与生成式 AI 聊天助手相关的决策。这些交流结果突显了人权框架和纳入来自人权来源的信息的重要性。

应对可能有害的生成式 AI 输出结果

我们认识到，生成式 AI 技术可能会产生一些有害输出结果，例如，生成潜在的仇恨、具冒犯性或歧视性内容；加深偏见；提供不准确的信息；以及/或者导致隐私问题。我们也知道，不良行为者可能会滥用我们发布的模型和工具，对他人蓄意造成伤害。我们已经采取了一系列措施，减少我们基础模型（包括 Llama 和 Emu）中的这些风险。我们相信，采取开放、迭代的方法将有助于我们在出现这些问题时实时做出改进。

联合国 B-Tech 项目重点关注多项国际公认权利的多个相关风险领域。我们以这份指南和 Meta 的人权突出风险为指导开展工作。

对于我们的基础模型 Lama 2 和 Emu，这些缓解措施的目的之一是解决重大的潜在人权风险，方法包括解决模型训练中的风险、解决与特定背景相关的问题，以及开展安全评估并做出调整。

随着我们继续开发基础模型，我们公开分享了有关我们负责任方法的详细信息。

除了主动采取防范措施，降低 AI 生成可能有害的输出结果的风险外，一旦发现可能有害的输出结果，我们也会全力予以解决。

图片由 Meta AI 生成

在 10 月 7 日以色列发生恐怖袭击后的几天里，有利益相关者提醒我们注意出现的几个工具问题，所有这些问题很快被我们发现并着手予以解决。其中的主要问题包括：机器翻译在翻译一些用户的 Instagram 主页个性签名时，在译文中添加了“terrorist（恐怖分子）”一词，以及 AI 生成的贴图将巴勒斯坦人描述成暴力分子。

得知这些问题后，我们的工程团队第一时间展开了调查，找出问题的根本原因并完成修复。我们的团队发现，这些问题似乎与模型“幻觉”和所使用的训练数据有关。这两个问题是许多 AI 产品都面临的共同难题，且文献记录颇多（见此处和此处）。

在获知机器翻译产品出现问题后，我们在 90 分钟内对其进行了紧急修复。我们还以最快速度降低了 AI 生成的图片和贴图输出结果与 10 月 7 日恐怖袭击存在的潜在不良关联。例如，我们在模型中加入了已知的姓氏信息，以防止模型出现“幻觉”。此外，我们还对图片生成所使用的基础模型进行了微调，更好地解决了与该冲突事件存在的各种潜在不良关联，并将改进后的模型部署到我们的所有产品中。

详情请参见以色列和巴勒斯坦：我们的最新行动章节。

我们致力于以安全、负责任和人权为核心来开发和部署生成式 AI 产品，在此过程中，我们期待与持不同观点的广大权利所有者持续沟通交流。我们还将采取适当措施，履行我们在该领域所负有的法律义务。

人权风险管理

每一天，我们都积极将人权原则落实到有意义的行动之中。Meta 企业人权政策是我们开展这项工作的宗旨。我们做出的承诺和采取的方法以《联合国工商企业与人权指导原则》以及我们在企业人权政策中列出的国际和地区人权标准为指导。

我们的人权风险管理工作包括但不限于：尽职调查、利益相关者的参与、产品咨询、开展人权培训以及为企业各个方面提供有数据支撑的建议。在本报告中，我们将重点介绍 Meta 开展的各类人权尽职调查。我们在开展尽职调查时，会使用各种评估工具和协议，识别、预防和减轻可能与我们的产品、政策和全球运营相关的风险。

这项工作以及我们的全球网络倡议 (GNI) 承诺为我们看待监管问题提供了指导。

近期的全球趋势表明，监管环境正在发生变化并且会对人权产生影响。举例而言，2023 年，英国颁布了《网络安全法》(Online Safety Act)，欧盟的《数字服务法》(Digital Services Act) 开始生效，新加坡、台湾和乌拉圭等国家/地区也相继出台了其他网络安全法律。

本报告第 15 页上的图表概括说明了我们在人权风险管理上采取的一些行动。如需了解更多详情，还可查阅我们的 2021 年和 2022 年人权报告。

日常尽职调查

在我们的日常尽职调查工作中，许多不同的团队会通力合作，不断将尊重人权融入我们的各项活动中。例如，Meta 的人权和民权专家会持续提供建议，帮助评估和减少内容政策、产品开发、危机和冲突响应以及选举应对准备方面的人权风险。这项工作通常涉及与以下主题相关的事项：仇恨言论、错误信息、可能记录侵犯人权行为的血腥内容、为保护人权维护者而采取的举措以及其他人权相关问题。

截至 2023 年 12 月，平均每天有 31.9 亿人使用我们旗下至少一款应用，鉴于此，我们的产品带来的潜在人权影响会因为时间、地区、内容和受影响社群的不同而存在显著差异。因此，我们会根据 2022 年人权突出风险综合评估 (CSRA)

（这项评估分析了我们的产品对所有国际公认人权的影响）来确定工作的优先次序。这项评估在分析中使用了《联合国工商企业与人权指导原则》提供的标准，确定了八个需优先考虑的最突出风险，如后面图表所示。

对于应该如何就各项人权风险划分行动的优先级，我们往往难以定夺。这种矛盾关系是人权框架所固有的，当我们尝试协调多个且有时相左的 人权时，这种矛盾关系就会凸显出来。我们力求优先考虑最突出的人权风险，即那些有可能对利益相关者造成最严重负面影响的人权风险。

我们在确定应优先考虑的人权领域时，会以《联合国工商企业与人权指导原则》和我们的企业人权政策为指导，另外，我们还会参考围绕多元化的各个维度与利益相关者开展的沟通交流。例如，由于我们的整个用户群体的多样性十分丰富，因此我们有必要确定应优先考虑的人权领域，让青少年等弱势群体能够使用我们的产品和服务。在为青少年提供服务时，Meta 以《联合国儿童权利公约》，尤其是其中的“儿童最大利益”原则为指导。这有时会导致我们需要就应优先考虑哪一种潜在的权利影响做出复杂而棘手的决定，例如，既要保护儿童的安全和健康，又要尊重他们私下独立寻求政治、健康或性别认同等话题相关信息的权利。儿童最大利益框架为家庭中心和家长监督等功能的引入和设计提供了指导。

我们力求优先考虑最突出的人权风险，即那些有可能对利益相关者造成最严重负面影响的人权风险。

我们的人权风险管理方法

人权承诺

我们按照《联合国工商企业与人权指导原则》规定的做法履行尊重人权的承诺，包括：

执行人权政策

开展人权尽职调查并披露相关信息

提供获得救济的方法

坚持实行治理、监督和问责机制

保护人权维护者

优先考虑的权利

人权突出风险综合评估确定的八个优先人权领域：

意见和表达自由

隐私

平等和无歧视

生命、自由和人身安全

儿童最大利益

公众参与、投票和竞选

结社和集会自由

健康权

重要工具和示例

我们采用下列工具和流程来降低风险，本报告将提供一些说明性示例：

内容政策制定	危险组织和人物更新
产品咨询	生成式 AI、Threads
将人权纳入选举应对准备工作	巴基斯坦选举
将人权纳入协调响应工作	优先处理积极响应的国家/地区状况
危机响应	纳戈尔诺-卡拉巴赫冲突和苏丹冲突
GNI 政府请求框架	根据当地法律限制内容和访问用户数据
监督委员会	对提议案件的建议，即国际法中的人质问题
利益相关者的参与	面向家长和青少年的共同设计会议
保护人权维护者	人权维护者基金

相关的突出风险：

- 意见和表达自由
- 隐私
- 平等和无歧视
- 生命、自由和人身安全

内容政策

作为我们基础尽职调查工作的一环，我们的人权专家会为内容政策制定流程提供支持。在此过程中，人权专家会从人权法的角度审核对内容政策的修订提议，以考虑到表达自由权、无歧视权以及其他人权。2023 年，这项工作还包括完善我们的危险组织和人物政策。该项工作于 2023 年分不同阶段开展，2024 年 1 月我们发布了相关的进展更新。

根据“危险组织和人物”政策，我们禁止担负暴力任务或从事暴力活动的组织或个人使用 Meta 的服务。我们所采用的全球统一方法以危险组织类型和分级的详细定义为依据，这些定义已在政策及信息公示平台发布。我们为识别危险组织和人物设立了一套独立的流程，每种情况均由相关领域的专家根据确凿的证据进行审核。我们的定义和认定标准不会因地区或意识形态的不同而有所差异，并且是在咨询世界各地的专家和学者后制定而成的。此外，我们还对美国指定的外国恐怖组织、特别指定的毒枭和特别指定的全球恐怖分子执行这项政策。

根据监督委员会和企业社会责任组织 (BSR) 提供的以色列和巴勒斯坦人权尽职调查建议，我们对“危险组织和人物”政策做出了下列更新：

- 我们更新了政策条文，在继续履行我们法律义务的前提下，允许用户发布更广泛的社会和政治言论，包括有关选举、冲突解决、灾难和人道主义救援的言论。
- 我们收到反馈称，我们此前对“赞扬危险组织或人物”的定义过于宽泛，因此我们更新了政策，使这项定义变得更加细致和适度。我们现在禁止“美化”危险组织和人物的暴力和仇恨行为，“美化”的定义比“赞扬”的范围更窄。
- 我们简化了对危险行为者的评估和分类方法，根据他们造成的现实世界伤害和暴力的分级进行评估和分类。
- 我们更新了除名流程，提供更加详细、全面的标准，仅当危险组织或人物满足这些标准后，我们才会考虑将其从名单中移除。这项更新可确保我们指定的危险组织或人物符合不断变化的形势。

2023 年 9 月，我们还修订了暴力与煽动暴力政策，完善了有关高强度暴力的措辞。2023 年 3 月，我们根据监督委员会的建议修订了欺凌和骚扰政策，纳入了对“公众人物”的定义。2023 年 6 月，我们根据监督委员会对苏丹的血腥暴力视频一案提出的建议，举行了有关侵犯人权背景下暴力和血腥内容的政策交流委员会 (Policy Forum) 会议。我们审视了暴力和血腥内容政策是否在尊重受害者的隐私和尊严、表达自由和社群福祉之间取得了适当的平衡。我们的分析表明，这项政策已适当平衡这三个方面，因此没有修改政策。

为提高透明度，具体的政策变更会纳入到我们的社群守则中，并且可在相关更改日志和政策交流委员会的公开会议记录中查阅。

此外，我们还于 2023 年更新了处罚制度。新系统规定的限制期更少，这样做是为了帮助确保对违反政策的行为采取相称的应对措施。对于屡次违规的用户，在给予充分的警告和解释，帮助其理解我们移除其内容的原因后，我们仍然会对该用户实施账户限制，通常是从第七次违规开始时执行。对于更严重的违规行为，例如发布包括恐怖主义、儿童剥削、人口贩卖、宣扬自杀、性剥削、销售非医疗药物或宣扬危险组织和人物的内容，我们仍然将立即采取处理措施。

产品开发

相关的突出风险：
- 意见和表达自由
- 隐私
- 平等和无歧视

我们的工程师致力于创造性地解决现实世界中的问题。人权是我们建立负责任创新实践的指导原则，并且我们致力于对新产品开展尽职调查。

2023 年，我们推出了许多令人振奋的新产品，例如 7 月推出的 Threads、9 月推出的 Ray-Ban Meta 智能眼镜和 WhatsApp 频道，以及 10 月推出的混合现实头戴设备 Quest 3。在所有这些产品的开发过程中，我们都考虑到了人权问题。

此外，我们还在继续建设元宇宙，包括利用元宇宙帮助提高教育水平、消除偏见和打击仇恨行为。例如，我们推出了多款虚拟现实体验，帮助用户构建社群，包括 MLK: Now is the Time、Inside the Mosque（共两集）和 ABLE。

我们的团队在快节奏的迭代环境中构建、测试、完善和部署产品和服务，包括基于 AI 的产品和服务。我们在构建过程中注重隐私保护，并根据我们动态变化的产品开发流程的特殊需求，制定了产品风险缓解方法。在开发过程中的早期，产品团队就会根据人权原则，指导进行负责任的创新。这有助于团队预测和减轻对个人、社群和社会造成的潜在危害。例如，我们的民权和人权专家可以提供建议或开展快速冲刺行动，以确定产品发布的最高风险国家/地区。

例如，在 Threads 产品的开发过程中，我们的人权专家就利用了现有的产品和政策审核流程来评估这些产品的潜在人权影响。他们与 Threads 团队合作实施缓解措施，包括应对政府可能提出的内容审查请求。

我们还在继续开发 Project Height，此框架可供产品团队用来评估产品发布过程中出现的公民权利问题。这使产品开发团队能够在产品开发过程中考虑到民权问题。此框架是对我们评估产品安全时所用其他风险流程的补充。

Meta Quest 推出的 MLK: Now is the Time

我们如何为重大事件和全球危机做好应对准备

相关的突出风险：
- 生命、自由和人身安全
- 公众参与、投票和竞选
- 意见和表达自由
- 平等和无歧视
- 隐私
- 儿童最大利益

我们在全公司范围内努力将人权原则纳入 Meta 的诚信工作中[3]，这包括我们如何为高风险国家/地区日益加剧的挑战和计划内的重大事件（例如选举）做好应对准备，以及我们对危机事件的协调响应。这种跨多个专业团队的协调工作模式，有助于我们预测并高效应对任何地方出现的紧张局势升级情形。

Meta 如何为选举做好应对准备

相关的突出风险：
- 公众参与、投票和竞选
- 意见和表达自由
- 生命、自由和人身安全

2023 年，许多国家/地区举行了选举，同时，我们也需要为 2024 年的更多选举做好应对准备。2023 年，Meta 为全球 130 多场选举做好了应对准备，包括巴基斯坦、阿根廷、土耳其和尼日利亚的选举。我们将在下文分享有关巴西和巴基斯坦两场相关选举的洞察信息以及我们采取的行动。

我们设有专门的团队负责推动整个 Meta 公司内的应对准备工作，争取在全球选举活动到来前将保护措施落实到位。这些工作可能包括：打击恶意威胁行为的高级安全保障行动、更新政策和流程，以便从我们平台上移除潜在有害内容、我们业界领先的全球事实核查网络，以及政治和社会议题类广告的信息透明度。

我们还会不断评估在选举等重大事件期间发生紧迫伤害的风险，以便采取有针对性、有时限的政策和产品应对措施，帮助保护人们的安全。

虽然我们的大部分工作都是积极主动的，甚至在选举日期之前提前很久就开启启动，但是我们也会为高风险事件做好应对准备。在这些情况下，Meta 可以采取多种方法来应对各种情况，例如，更改或限制产品功能、引入消息流量限制以及限制内容的传播。

我们识别潜在有害趋势的一些方法包括：从公开报道中收集信息、审核我们的可信合作伙伴提出的建议、持续观察内容趋势、开展人权尽职调查，以及审核我们的情报专家提供的评估结论。这些信息，外加现有的社群守则和政策执行系统，有助于我们确定可以使用哪些类型的产品和政策缓解措施来预防高风险选举期间的违规行为。这些产品和措施包括我们的危机政策协议以及我们为应对风险加剧的情形而可能采取的独立系统调整措施。我们还提供透明度工具，例如我们的广告资料库。

[3] “诚信”是一个 Meta 内部的专业术语。信任和安全、计算机和账户安全、减少不良体验、相关隐私问题等由 Meta 内部一个团队网络负责处理，他们通常将之称为“诚信”工作。除了开展其他许多工作外，这些团队还会开发各种工具，用于预防伤害、审核平台内容和执行我们的政策。

巴西

早在巴西 2022 年总统选举的一年前，我们就开始启动相关的应对准备工作。选举结束后我们的工作仍在继续，正因如此，我们才能够迅速应对 2023 年 1 月 8 日对巴西国会、最高法院和其他公共建筑发起的袭击事件。Meta 还根据危机政策协议将这场选举后发生的动乱指定为危机事件，以帮助公司评估如何降低内容风险。

我们为 2022 年巴西选举所做的应对准备工作还包括绘制选举期间和选举之后的风险情景图。我们考虑了有关表达自由和其他人权的国际标准。我们还对产品做出了变更并调整了政策，以保护巴西选举的诚信。这些工作包括与最高选举法院合作，向民众提供有关投票的可靠信息。

在 2023 年 1 月 8 日的袭击事件发生之前、期间和之后，我们均部署了一系列工具和方法来打击潜在的仇恨言论、煽动暴力和错误信息。在选举前不久，我们将巴西指定为临时高风险地区，这让我们能够移除呼吁携带武器或强闯政府大楼的内容，详见监督委员会对 “巴西将军的演讲” 一案的决定。

从 2022 年 8 月 16 日竞选活动开始到 2023 年 1 月底，Facebook 和 Instagram 上因违反我们在巴西的暴力与煽动暴力政策而被移除的内容分别超过 100 万条和 96 万条。这些内容包括呼吁军队进行干预的帖子。

在应对巴西选举活动的这五个月内，我们移除的违反仇恨言论政策的内容数量如下：

57 万+

52 万+

另外，我们在巴西移除的违反欺凌和骚扰政策的内容数量如下：

38 万+

63 万+

性别暴力

在选举期间，针对女性候选人、记者和人权维护者的性别骚扰和暴力威胁往往会加剧。为了降低这些风险，同时也作为我们致力于加强保护旗下平台上的女性并支持她们获取公共服务这一更广泛工作的一部分，我们在巴西采取了以下行动：

- 与 Ministry of Women（妇女部）合作，在巴西的 WhatsApp 上推出一个官方频道，作为与监察员联络的另一种方式，让女性能够向监察员提交有关针对女性的暴力行为的投诉或请求获取有关此类行为的信息。用户可以通过此频道获取相关法律信息，找到专门的妇女援助服务机构的地址，以及直接与工作人员交谈。
- 发布葡萄牙语版的在线指南，保护女性免受网络暴力的侵害。
- 与行业团体、民间社会组织和监管机构（包括 Ministry of Women）就解决性别暴力和骚扰问题的联合倡议展开合作。

巴基斯坦

虽然巴基斯坦选举是在 2024 年举行，但我们针对巴基斯坦选举的应对准备工作却始于 2022 年。这些工作包括全公司范围内在产品、政策和运营方面的努力，以避免和减轻与使用我们平台有关的人权风险。

其他行动还包括开发各种信号，专门用于识别与权利问题相关的内容和高优先级的升级问题。根据 GNI 原则，我们在政策及信息公示平台发布了一份实时的案例分析，介绍了 2023 年 12 月从巴基斯坦政府收到的内容移除请求，但要求移除的内容并未违反我们的社群守则或当地法律。

利益相关者的参与至关重要。选举前，我们与地区人权利益相关者举行了关于选举诚信工作的情况通报会。我们还与选举委员会沟通合作，解释我们对政府行动请求和人权的处理方法。在选举前的一段时间里，互联网被切断，于是，Meta 与利益相关者分享了信息，告诉他们如何设置代理服务器以接入 WhatsApp。

作为多方利益相关者团体（其中包括 GNI 和亚洲互联网联盟）的成员，Meta 参与了就选举前就互联网中断和侵权立法对权利的影响发表的关切声明。

在选举前几个月，我们还采取措施，在我们的平台上加强对人权维护者和其他弱势个人的保护。这些措施包括根据对违反社群守则的虚假指控，为内容移除提供保护。我们的做法具有包容性，符合我们的企业人权政策，该政策采用了联合国《人权维护者宣言》中对人权维护者的广泛定义。

在积极响应的国家 / 地区做好应对挑战的准备

积极响应的国家/地区分类是在专业团队和我们的人权专家的指导下，通过循证流程确定的，并且在评估现实世界伤害和暴力风险最高的国家/地区及划分优先次序上，参考了我们的人权尽职调查流程。一旦某个国家/地区被确定为积极响应的国家/地区，将触发进一步的风险缓解措施。这些措施可能包括强化的风险监测和缓解工作以及增加投资，例如，加强对冲突相关语言的内容审核，以移除违反政策的内容，以及提供定制化的产品支持。由于风险错综复杂，我们还联合可信合作伙伴和第三方事实核查机构，支持我们在积极响应的国家/地区所开展的工作。

应对全球危机

我们积极应对世界各地的突发危机，包括国际冲突、恐怖袭击以及环境灾难等非暴力危机。参与危机应对工作的人员包括来自内容政策、全球运营、人权和产品领域的一众专家以及其他内部专家，这些专家会根据我们的危机政策协议对危机形势做出评估和回应。正如我们在 2022 年人权报告中所述，危机政策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和人权原则，迅速使用有针对性或特殊的政策手段来减轻潜在伤害。

我们的人权突出风险综合评估可帮助我们在冲突局势中对不同的潜在人权风险排列优先次序，例如，对自由和人身安全权、表达自由权以及我们社群的安全和福祉排列优先次序。我们还可能会调用工具箱中的一些临时措施，帮助保护人们的安全以及降低可能将我们平台用于进一步加剧网络和现实世界中紧张局势的风险。为遵守《联合国工商企业与人权指导原则》关于在冲突期间加强人权尽职调查的指导，我们可能会考虑公共国际法的相关发展变化，包括国际人道法，也被称为武装冲突法。下文列出了我们这项工作的一些例子。我们将在下一小节介绍我们在以色列和巴勒斯坦人权尽职调查及相关冲突方面的最新工作进展情况。

苏丹

2023 年 4 月，苏丹武装部队 (SAF) 和快速支援部队 (RSF) 之间爆发战斗，这场冲突造成民众大规模流离失所、粮食不足、医疗资源匮乏以及其他需要人道主义援助的挑战。根据 Meta 危机政策协议，苏丹被评估为风险最高的国家。我们以危机政策协议为指导采取了行动，以尊重人权、将伤害风险降至最低，以及协助提供人道主义援助。

2023 年 8 月，我们根据危险组织和人物政策将 RSF 指定为危险组织，该政策禁止担负暴力任务或从事暴力活动的组织或个人使用我们旗下平台。这有助于遏制潜在有害内容的传播。这项决定是在与联合国机构磋商后，根据人权尽职调查研究和主要非政府组织关于严重侵犯人权行为的报告所做出的。该决定经过了监督委员会的审核和批准。我们还将苏丹指定为临时高风险地区，这有助于移除呼吁平民拿起武器或将武器带到该国特定地点的内容。我们与第三方事实核查机构合作，在我们旗下平台上揭穿、标记并减少传播与冲突相关的错误信息。我们还移除了经可信合作伙伴评估可能会诱发紧迫人身伤害或暴力风险的错误信息。

相关的突出风险：
- 生命、自由和人身安全
- 意见和表达自由
- 结社和集会自由
- 隐私
- 健康权

由于苏丹首都喀土穆发生大规模流离失所现象，我们的许多利益相关者（包括政府官员、民间社会团体、媒体和人权维护者）都很难或无法取得联系。在这种情况下，我们需要寻找新的团体并与之合作，以支持人道主义工作的开展并帮助识别和阻止潜在有害内容和错误信息的传播。我们与苏丹国内和主要散居在国外的活跃民间社会团体进行了接洽，包括医生、媒体、民间记者和妇女团体。我们向他们概述了我们的内容政策，并为他们提供了数字安全培训和 Meta 危机应对工具。这包括举办关于如何使用 WhatsApp 社群的讲座，在后来协调疏散行动和提供人道主义援助中，这些社群都派上了用场。

尽管安全局势不断恶化，我们的可信合作伙伴仍继续报告与暴力与煽动暴力、错误信息、仇恨言论以及欺凌和骚扰有关的内容。可信合作伙伴的评估有助于我们针对两种内容执行“错误信息和伤害”政策：一种是使用断章取义的图片煽动紧张局势的内容，另一种是针对在苏丹开展工作的人道主义机构提出指控的内容。

与我们处理其他冲突的方法类似，我们移除了一些违规的血腥内容，但没有对账户应用违规记分（违规记分是对违规行为的一种惩罚，受到多次违规记分会导致账户限制升级），以避免过度惩罚或限制那些试图提高人们对冲突所造成影响的认知的用户。这种处理措施受到了 2022 年监督委员会对苏丹血腥暴力视频一案所提建议的启发。该案件促使我们在 2023 年重新审视了有关分享暴力和血腥内容以提高人们对侵犯人权行为的认知的政策，并对我们如何应用“允许保留具新闻价值内容”的政策做出了更详细的解释。

由于缺乏医疗基础设施，苏丹国内的药品供应严重短缺，因此我们临时修改了政策，使用户能够请求获取或捐赠药品，而此类行为通常是我们的管制商品政策所禁止的。此外，我们还与苏丹医疗专业委员会 (Sudan Medical Specialization Board, SMSB) 合作推出了 SMSB 苏丹诊所 (SMSB Sudan Clinic)，这是一款使用 WhatsApp API 构建的远程医疗应用。此应用通过我们的 WhatsApp 社会影响力费用减免计划 (WhatsApp Social Impact Fee Waiver Program) 提供，该计划旨在为政府和非营利性合作伙伴减免 WhatsApp 消息使用资费。

我们还需要平衡人口移动和安全信息的传播，既要帮助民众迁移，又要确保他们的安全，同时又不能助长向中非共和国、乍得、埃及、埃塞俄比亚、利比亚和南苏丹等邻国的非正常移民。这是因为，随着流离失所人口各种需求的上升，他们遭受剥削的风险也随之增加，这将导致这些地区出现不良行为者。我们采取了行动识别不良行为者，并移除了可能给人们带来风险的内容，包括偷运人口的内容。

增强抵御力：数字安全和内容政策培训

2023 年：

6 节 培训课

300 个 受益人

冲突结束后：

2 节 培训课

70 个 受益人

纳戈尔诺-卡拉巴赫

2023 年 9 月 19-20 日，阿塞拜疆在自封的阿尔扎赫共和国（纳戈尔诺-卡拉巴赫争议地区中的一片区域）发起军事进攻。这次进攻导致亚美尼亚人逃离该地区，引发了一场难民危机。

我们一个包括人权专家在内的跨职能团队一直在监测这场危机局势，并迅速调集资源和执行我们的危机政策协议。随着仇恨言论、欺凌、血腥内容以及提及危险组织和人物的内容的传播度不断增加，我们具有相关语言技能和背景知识的内部专家对这些内容采取了处理措施。我们还与当地的联合国机构进行了接触，以便更好地了解危机的变化形势。Meta 为冲突双方的民间行为者提供了额外的保护，以减少冒充、骚扰和黑客入侵的风险。

与其他冲突一样，我们必须决定如何在“保护战俘的隐私和尊严”与“记录人权侵犯行为”之间取得适当的平衡。Meta 社群守则禁止发布可能泄露战俘身份或位置的信息。这项规定符合《联合国工商企业与人权指导原则》规定的公司人权责任。这些责任在武装冲突期间尤为重要，而且必须以国际人道法的规定为指导。一方面，泄露战俘的身份和位置会危及他们的安全、隐私和尊严及其家人的安全。

另一方面，这种曝光可以为公众舆论提供依据，并提高人们对潜在虐待行为（包括违反国际人权法和国际人道法的行为）的认知。此外，这种信息披露还有助于推动权利保护和问责行动。鉴于战俘图像在社交媒体上的传播规模大、速度快，解决这些相互冲突的利益变得非常棘手。

相关的突出风险：

- 意见和表达自由
- 生命、自由和人身安全
- 隐私

2023 年，监督委员会维持了我们的一项决定，即保留包含一段可认出战俘身份的视频的 Facebook 帖子，并为视频添加“该内容可能令人不适”的警告画面。监督委员会同意 Meta 的观点，认为在这种情况下，这些帖子可以为公众舆论提供依据，并提高人们对潜在虐待行为（包括违反国际人权和国际人道法的行为）的认知，这“超过了战俘面临的安全和尊严风险”。

保留证据

我们支持为所有国际罪行的受害者伸张正义。多年来，我们一直致力于探索在尊重权利的前提下保留和披露证据的方案，并与民间社会团体、学术界以及国际检察专家和机构开展了广泛的磋商。

正如我们在回应监督委员会的建议时所述，我们制定了一种方法，允许国际法院和联合国授权的问责机制（例如实况调查团和调查委员会）向我们提出请求，以延长保留与其正在开展的调查有关的数据。这项工作已基本完成，其有别于我们回应执法部门保留请求的长期政策。

我们向若干联合国授权机制和特别报告员介绍了我们的方案，并概述了向 Meta 提出延长保留请求的流程。我们将仔细审核收到的所有请求，确保其符合我们的政策和适用法律。

这是一个全新的领域，没有既定的或经过检验的最佳做法可供参考，并且这项工作在法律、隐私和政策方面仍有许多重要的内在考虑因素。我们预计将在监督委员会半年更新报告和年度人权报告中详细介绍我们在这方面的最新工作情况。

相关的突出风险：

- 生命、自由和人身安全
- 隐私

中美洲和南美洲的移民

人口偷运可能导致想要离开原籍地且通常是为了追求更美好生活的弱势个人遭到剥削。我们采取了行动，移除助长或促成剥削（包括人口贩卖和人口偷运）的内容。

2023 年，我们发现了北美洲、中美洲和南美洲偷运移民方面存在潜在的人权风险，包括达连隘口沿线，达连隘口是横跨哥伦比亚北部和巴拿马南部的一片危险莫测且森林茂密的丛林地带。在这片区域，非法商业活动、危险组织的线上活动以及关于移民和难民的错误信息的风险都更高，尤其是因为其中一些人试图前往美墨边境。

为了应对这些风险，我们执行了社群守则和商业交易政策，禁止剥削行为，包括偷运人口。我们与民间社会团队和政府当局合作，共同应对错误信息，并继续监测和执行针对违规自然、付费和商业内容的政策。我们还做出调整，让第三方事实核查机构更容易找到与美墨边境相关的内容并对其做出评定，因为我们认识到，面对重大事件，响应速度尤为重要。我们会利用关键词检测功能将相关的内容集合到一起，方便事实核查机构查找。事实核查机构会评定我们平台上的内容并针对错误信息发布英语和西班牙语的文章。

我们还通过 We Think Digital 等诚信和培训计划，为移民和难民群体以及非政府组织提供支持。We Think Digital 计划侧重于培养可识别错误信息、欺诈和诈骗内容的数字技能。

人权影响评估

开展人权尽职调查的方法在不断演变。我们在确定如何改进方法和强化行之有效的措施时，会考虑到潜在的风险并从过往的工作中汲取经验教训。

我们在人权突出风险综合评估的指导下，对特定国家/地区、应用、服务、硬件或战略举措开展了人权影响评估。这些评估旨在帮助我们预测潜在的影响，尤其是在为新产品和新功能做准备时，例如端到端加密。人权影响评估涉及利益相关者的参与，这是人权尽职调查的一个基本要素。在我们开展人权影响评估的过程中，每项建议在全公司范围内的实施往往都需要多个团队的协作并经历多个工作流。

我们过去已经分享过人权影响评估的摘要和建议，这些信息均列于附录中。为回应针对菲律宾、印度、端到端加密以及以色列和巴勒斯坦问题与 Meta 平台相关的潜在人权风险评估，我们采取了相关行动，下面我们将分享有关这些行动的更新。

以色列和巴勒斯坦：我们的最新行动

就在我们于 2023 年 9 月发布了以色列和巴勒斯坦人权尽职调查更新后不久，哈马斯于 10 月 7 日对以色列发动了恐怖袭击，作为回击，以色列随即对加沙采取了军事行动，该地区的其他行为者也卷入了这场冲突并使得事态愈发升级。我们知道，我们对该地区持续冲突局势的回应深深影响了该地区及世界各地的人民。

袭击发生后，Meta 立即将这一暴力事件指定为我们危机政策协议中的最高级别暴力事件，并立即实施了危机应对措施，包括组建一个全天候跨职能专门团队。我们的行动以核心人权原则为宗旨。我们的企业人权政策以《联合国工商企业与人权指导原则》为基础，旨在优先考虑和降低最突出的人权风险。我们还将国际人道法作为重要参考。

我们最初于 2023 年 10 月 13 日以英语、阿拉伯语和希伯来语发布了一篇博文，详细说明了我们做出的回应，并于 10 月 18 日、12 月 5 日和 12 月 8 日提供了更多进展更新。此后，我们继续完善所采取的措施，以应对不断变化的冲突形势，包括加沙持续的人道主义危机和仍然被哈马斯扣押的人质。

相关的突出风险：

- 生命、自由和人身安全
- 意见和表达自由
- 平等和无歧视
- 隐私

在冲突局势中，平衡“我们社群的安全和福祉”和“允许在我们的平台上发声”这两者尤为棘手，而当冲突涉及受美国制裁的实体（例如哈马斯、真主党和巴勒斯坦伊斯兰圣战组织，根据我们的政策，这些组织均被指定为危险组织）时，要取得这种平衡就更具挑战性。

我们为回应以色列和巴勒斯坦人权尽职调查而开展的工作为我们制定行动方法提供了指导。例如，我们能够在内部系统中为阿拉伯语内容分配更合适的审核人员，以实现更高的审核准确度。我们还为希伯来语市场增加了内容审核资源。如需了解详情，请访问 Meta 2024 年以色列和巴勒斯坦人权尽职调查更新。

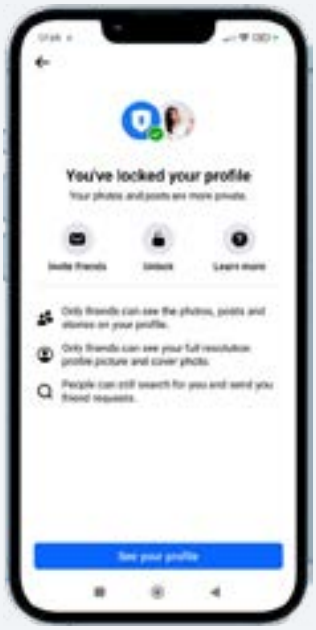
我们对这场冲突采取的应对举措远不止尽职调查中概述的措施，并且这些应对举措同样以我们的危机政策协议为指导。我们采取了一系列有时限的临时产品和政策措施，以降低突出的人权风险。我们知道，在某些情况下，这些措施（例如，降低自动执行政策的门槛），会在无意中限制对世界重大事件的讨论，从而影响表达自由权。这便是我们在发现风险时，采取有时限并与风险严重程度相称的措施的原因。

临时政策措施

我们最初依据“危险组织和人物”(DOI) 政策采取了行动，并扩大了“暴力与煽动暴力”政策的应用范围，移除描述可认出身份的人质被绑架或关押的内容，即使这些内容是为了表达谴责或提高人们对人质所面临处境的认知。我们这样做是为了按照国际人道法的标准保护受害者的隐私和尊严，并根据 DOI 政策防止哈马斯的宣传内容出现在我们的平台上。

随着战争的持续，我们发现用户开始分享人质内容，目的是反驳一种新出现的言论，该言论否认 10 月 7 日发生的恐怖袭击。为应对这种情况，我们采取了更加细致的方式，允许用户出于提高认知或谴责袭击行为的目的分享描述人质被绑架的内容。如果用户分享内容的意图不明确，则我们会以安全为重，采用更加谨慎的方法，继续移除内容。

我们还做出了其他旨在保护表达自由的调整。对于某些政策领域，例如某些类型的暴力和血腥内容，我们移除了违规内容，但没有对账户应用违规记分（违规记分是对违规行为的一种惩罚，受到多次违规记分会导致账户限制升级），以避免过度惩罚或限制那些试图提高人们对冲突所造成影响的认知的用户。



临时产品措施

我们采取了一系列临时产品措施，这些措施旨在帮助确保民众的安全和减轻突出的人权风险，我们描述冲突应对措施的博文对此进行了概述。这些措施包括调整自动处理内容的置信度阈值、在搜索中屏蔽某些话题标签、调整产品以解决不当和有问题的评论，以及锁定个人主页工具等安全措施。

我们还临时降低了“擦边球”内容或潜在违规内容（例如，描述血腥暴力画面的图片或视频）可能被取消推荐资格的门槛。这一措施适用于关系网络外的内容，此类内容是指某用户尚未选择关注的人发布的内容，这些内容可能出现在用户的动态、搜索、发现和 Reels 等版块。

除了临时政策措施，我们还采取了其他旨在保护表达自由的调整。为了应对我们产品的使用量大幅飙升的情形，我们临时调整了一些自动流量限制措施，这些措施旨在防止根据我们的政策被定义为垃圾信息的行为，例如高频率发帖 。调整后，我们对此类内容的处理更加宽松，从而减少对合法用户造成限制的风险。

另外，我们关于可推荐性和降级的宏观政策（不是我们危机应对行动的一部分）也可能会影响冲突相关内容（例如血腥和暴力内容）的可见度。

Meta 经常开展调查，以评估我们内容审核系统的表现及其对内容执行政策时是否过度或不足。展望未来，我们确定了需要持续改进的领域，包括优先按方言审核阿拉伯语内容，以及加强对图像和视频内容库的审核和监督，以降低过度执行政策的潜在风险以及对表达自由的相关影响。这是一项长期持续的工作。

如需详细了解我们的应对行动，请访问此处和此处，参阅我们对民间社会组织所提质询的回复函件。有关“2024 年以色列和巴勒斯坦人权尽职调查更新”的更多详情，请参阅此处。

相关的突出风险：

- 隐私
- 意见和表达自由
- 结社和集会自由
- 平等和无歧视

端到端加密：我们的最新行动

2023 年 12 月，我们开始为 Facebook 和 Messenger 上的所有个人聊天对话和通话推出默认端到端加密。我们于 2019 年开始致力于将默认端到端加密推广至所有消息应用，而上述行动标志着我们在这方面迈出了具有里程碑意义的一步。我们也在继续为 Instagram Direct 开发端到端加密功能。WhatsApp 自 2016 年起就一直提供默认的端到端加密。

端到端加密可确保除您本人和您的消息发送对象外，没有其他人能够看到你们之间的消息，从而可以增强安全性和隐私保护。此功能提供了一层额外的安全屏障，可以保护您与亲友之间的消息和通话，这意味着，包括 Meta 在内的任何一方都无法看到或听到你们之间互发的消息或通话内容。举报端到端加密对话时，该对话中近期的消息将从您的设备安全地发送给我们的帮助团队。

在扩大端到端加密的应用范围上，人权框架一直是指导我们开展工作的关键。端到端加密可直接促进隐私权的保护，而隐私权又能促进其他各种人权的实现，包括表达自由、结社自由、意见自由、行动和人身安全。但另一方面，个人违规使用任何消息服务（包括加密功能）以伤害他人的风险始终存在。

2022 年，我们分享了委托企业社会责任组织 (BSR) 开展的独立人权影响评估的完整结果，这项评估旨在调查我们扩大默认端到端加密功能应用范围的计划带来的潜在人权利益和风险。该评估向 Meta 提出了 45 项建议，这些建议旨在最大限度提升潜在利益并减轻潜在的负面影响。自我们初步回应这些建议以来，我们一直在继续努力履行我们做出的承诺并公布了大量最新进展情况，包括在 2023 年 1 月和 8 月发布的最新情况报告，以及一份详细的安全白皮书。

问题聚焦

相关的突出风险：

- 隐私
- 生命、自由和人身安全

应对安全威胁

在过去六年中，Meta 发布了多份公开报告，介绍了我们检测和打击平台上的安全威胁所开展的各项工作。在公民实验室 (Citizen Lab) 和 45 个民间社会团体向民主峰会 (Summit for Democracy) 提出“禁止销售间谍软件，直至防止侵犯人权的保障措施落实到位”的提议上，我们的阻断后分析和调查为他们提供了部分依据。随后，美国总统乔·拜登签署了一项行政命令，禁止美国联邦实体使用间谍软件助长侵犯人权的行为。

我们在 2023 年减轻了一些与雇佣监视行业相关的威胁。雇佣监视公司包括为各种目的提供监视和跟踪个人或组织的服务的企业。专制政府、犯罪组织和其他不良行为者滥用雇佣监视服务，试图监视政权批评者、反对派人士、记者和人权维护者，压制他们的表达自由，这些现象引发了对侵犯隐私的担忧。

2023年，我们在 Facebook 和 Instagram 上识别并移除了六个独立的账户网络，这些账户与来自意大利、西班牙和阿联酋的八家雇佣监视公司存在关联。他们惯用的一些伎俩包括：在互联网上实施社交工程攻击和网络钓鱼，以及使用冒充抗议者、记者和年轻女性的角色，诱骗受害者提供电子邮箱和电话号码并点击恶意链接。

> “私人间谍软件行业的崛起使获得先进监视能力的途径变得更加广泛，同时还隐去了购买和利用这些服务的客户的身份。这种无差别、不透明的攻击行为对人权维护者构成了重大威胁。正因如此，我们才致力于与民间社会合作伙伴和整个行业合作，追究非法间谍软件公司的责任”。
>
> David Agranovich，Meta 威胁阻断总监

相关的突出风险：
- 儿童最大利益
- 隐私
- 意见和表达自由

儿童和青少年安全

儿童和青少年的网络安全[4]是整个互联网面临的一项共同挑战，需要整个行业、政府和民间社会的共同努力与协作，因为我们既要帮助保护用户的安全，又要为用户行使一系列人权（例如，表达自由权和信息获取权）保留余地。

十年来，我们一直致力于解决这一难题，并继续以保护权利为本设计我们的服务和制定内容政策。我们的“儿童最大利益框架”参考了联合国《儿童权利公约》、英国《适龄设计准则》(UK Age Appropriate Design Code)、爱尔兰数据保护委员会的《儿童基本原则》(Children's Fundamentals) 以及法国全国信息与自由保护委员会 (CNIL) 的《未成年人建议书》(Recommendation on Minor) 等法规和指南。

儿童剥削

保护儿童安全一直是我们的首要任务。我们力求通过执行政策和开发预防工具来帮助防止伤害的发生。我们一直在努力让用户更容易找到儿童剥削内容的举报工具，在这项工作的推动下，2023 年第 1 季度青年用户在 Messenger 和 Instagram 上向我们提交的举报比 2022 年第 1 季度增加了 75%。我们还引入了更多创新方法，主动查找、移除或限制可能违反我们儿童安全相关政策的账户。

例如，我们推出了主动检测技术，如果账户（以及与之关联的任何账户）在我们旗下平台上从事的活动和互动行为表现出对儿童存在性兴趣，我们便会自动检测并停用这些账户。由于采用了这种方法，从 2023 年 8 月 1 日至 2023 年 12 月 31 日，我们检测并自动移除了 90,000 多个账户。我们力争在出现成年用户账户与青少年进行或关于青少年的扰人互动，或试图与青少年进行潜在不安全接触的最初迹象时，便主动检测并限制成年用户的账户。

[4] “青少年”的定义没有统一的标准。联合国《儿童权利公约》将“儿童”定义为 18 岁以下的任何人（第 1 条），也包括 13 岁以上的青年。在本报告中，“青少年”和“青年”可以互换。

作为 Take It Down（该工具旨在主动预防青少年的私密图片在网络上传播）的创始成员，我们与美国国家失踪与被剥削儿童中心 (NCMEC) 合作，将该工具扩展至更多语言，并鼓励更多行业参与者加入，这项举措让能够访问该平台的青年用户人数增加了数百万。

我们也是 Lantern 计划的创始成员之一，这是技术联盟 (Tech Coalition) 于 2023 年 11 月推出的一项新计划，旨在为科技公司提供一种渠道，在各网络平台上分享有关“猎手”账户和行为的信号。企业社会责任组织对该计划开展了人权影响评估。

“防止儿童剥削是我们行业目前面临的最重要挑战之一。网络“猎手”是指利用各种应用和网站以青少年为目标进行犯罪的顽固犯罪分子。他们还会试探每个平台的防御能力并迅速找到应对伎俩。因此，我们现在也要一如既往地努力保持领先地位。除了开发能从根本上铲除网络“猎手”的技术外，我们还聘请了专门负责在线儿童安全的专家，并与业内同行和执法部门共享信息。”

Antigone Davis，Meta 全球安全负责人

青少年安全和健康

我们致力于为青少年提供安全、适龄的网络体验，同时尊重他们的数字权利。我们采取的方法侧重于从源头上预防伤害，为用户提供掌控自身体验的方式，以及对违反我们政策的行为迅速做出反应。我们开发了许多工具、功能和资源，为青年用户及其家长提供支持。这是一项长期持续的工作。

2023 年，我们继续在各项服务中拓展家长监护工具相关的工作。我们通过 Trust, Transparency and Control Labs (TTC Labs)（这是一场跨行业运动，目的是让用户能够掌控自己的隐私。），与家长、青少年、监护人和专家举行了共同设计的会议，并与青少年顾问合作，了解青少年的行为，为产品设计提供依据。我们还在 Meta Quest 2 和 3 上面向 10-12 岁的青春期前儿童推出了家长代管账户。

我们定期咨询青少年发展、心理学和心理健康方面的专家，以确保为我们平台中的青少年打造安全且适龄的体验，包括不断了解哪些类型的内容更适合青年用户。我们还会咨询世界各地的安全合作伙伴、青少年团体和青少年顾问以及联合国儿童基金会 (UNICEF)，向他们介绍我们所做的工作并鼓励他们参与进来。

听取专家的观点后，我们制定了新的保护措施，在我们的应用中为青少年用户打造更加适龄的体验，例如，隐藏更多搜索结果、在 Instagram 上提示青年用户更新隐私设置，以及在 Facebook 和 Instagram 上自动为青年用户开启最严格的内容控制设置，等等。

利益相关者的参与

我们通过 TTC Labs 举行了共同设计的会议，让用户、我们的社群和专家参与进来，倾听青年人的需求和关切，为我们的产品设计提供参考，帮助为青少年创造积极向上的网络体验。

87 名	青少年
71 名	监护人
221 名	专家
13 个	国家/地区

利益相关者的参与

与广泛的利益相关者积极合作，是 Meta 开展人权工作的核心。这为我们的产品和内容政策制定流程吸纳了外部知识、多元化的观点和反馈意见，并有助于加强问责制和提高透明度。与我们合作的利益相关者包括国际组织、各种民间社会团体和学术机构、边缘化群体、代表性不足的群体，以及青少年、家长和看护人等用户。我们从问题的各个角度出发收集世界各个地区的意见和观点，确保取得良好的平衡。

此外，我们继续与可信合作伙伴合作，以识别各种相关趋势，更好地了解我们的升级处理渠道所面临的挑战，以及探讨如何为民间社会合作伙伴改进这些机制的效果。

与广泛的利益相关者积极合作，是 Meta 开展人权工作的核心。

边缘化群体和人权维护者

我们认识到，与来自边缘化社群的利益相关者开展有意义的交流合作非常重要，同时考虑到区域内甚至国家内部的差异也很重要，这些差异都是我们希望努力去了解的方面。

我们努力在一系列问题上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 取得的相关最新进展。他们的见解为我们的内容政策制定以及促进“自由表达”和“安全”的政策执行等提供了参考。

在世界各地，人权维护者利用数字平台组织各种运动，他们经常因为在网上开展旨在促进和保护人权的行动而遭遇恐吓、压制和面临法律挑战。这种情况在边缘化群体中尤为明显，例如女性和 LGBTQIA+ 群体。

在包容性框架的指导下，我们通过举行圆桌会议、研讨会和一对一会议，了解相关的内容政策问题、找出政策上的漏洞以及寻求参与政策制定的机会。举例而言，我们与少数群体权利团体 (Minority Rights Group) 合作，在被认为存在武装冲突或社会暴力风险的国家/地区举行了八次全球会议。

我们还与超过 250 名来自各种背景的利益相关者开展合作，包括女性群体、LGBTQIA+ 社群成员、宗教少数群体、少数族裔和土著团体，收集有关这些群体所面临政策挑战的宝贵见解。我们与中东和北非地区的 LGBTQIA+ 组织和人权维护者举行了超过 19 次磋商和研讨会，根据从中收集的反馈意见，我们与约旦 Open Source Association (JOSA) 合作，为人权维护者发布了一个数字安全工具包 (Digital Security Toolkit)。这份指南提供阿拉伯语和英语两个版本，确定了数字安全方面的领先做法，并提供了诸多安全功能、技巧和应对措施，用于加强中东和北非地区活动人士和人权维护者的网络安全。我们还在撒哈拉以南非洲地区发起了一场 LGBTQIA+ 具影响力人士宣传活动，该活动旨在提高 LGBTQIA+ 活动人士对安全资源的意识，突显了我们对保护该地区高风险用户的承诺。

在为期两年以亚太地区为重点的试点项目取得成功之后，我们于 2023 年扩展了与 Civil Rights Defenders 合作运作的人权维护者基金 (Human Rights Defender Fund) 的覆盖范围，为全球人权维护者提供支持并追加了 $50 万美元的资金用于 2024 年的开支。

AfricanDefenders

我们以亚太地区人权维护者基金的经验和原则为蓝本，与 AfricanDefenders 合作发起了非洲人权维护者基金，该基金旨在支持那些因开展支持人权的在线活动而成为骚扰、迫害和/或起诉目标的人权维护者。这些人权维护者包括倡导自身权利的边缘化群体成员，例如活动人士和民间记者、非暴力政治活动家、女性人权维护者和 LGBTQIA+ 群体。该基金拨付的小额资金将用于在紧急情况下提供支持，以及用于购置新设备和安全技术、安排临时安置以及获取紧急法律支援和安全救助。

可信合作伙伴

截至 2023 年 1 月，我们的可信合作伙伴网络已纳入来自全球 113 个国家/地区的 400 多个非政府组织、人道主义机构、人权维护者和研究人员。除了有广泛而多元化的利益相关者持续参与之外，可信合作伙伴也为我们带来了丰富的知识和经验，为我们的内容审核工作以及制定有效、透明的政策提供参考。可信合作伙伴还可通过升级处理渠道报告潜在的有害内容和账户安全问题，对于这些问题，我们会根据情况采取相应措施，以保障用户的安全。

2023 年，来自可信合作伙伴的报告帮助我们移除了 49,600 篇违反我们政策的内容。可信合作伙伴还帮助我们揭露并移除了在美国、格鲁吉亚、布基纳法索和多哥从事合谋造假行为的六个欺诈性网络。此外，我们还对南亚的三个独立网络间谍行动采取了处理措施。这些网络意图操纵公众舆论，以实现其战略目标，这可能对人权构成威胁。

2023 年，我们全面审核了民间社会合作伙伴向我们提交紧急内容和账户安全相关问题的升级处理渠道。这项工作的重要一环是组织召开了一次全球峰会，与会者包括来自 Meta 的政策、运营和法律专家，以及来自亚太、中东、欧洲、拉美和非洲 13 个国家/地区的 15 个民间社会合作伙伴。通过这次交流，我们得以全面评估我们为民间社会提供的升级处理渠道，并为未来制定了共同的愿景：改进报告的优先级排序，加强应急响应和洞察情报收集流程。

案例分析：海地出现新的危害

2023 年 7 月 7 日是海地总统约韦内尔-莫伊兹遇刺的两周年纪念日。该国的局势依然高度紧张，帮派暴力的死灰复燃进一步加剧了这种紧张局势。我们可信合作伙伴报告的内容涉及多个违规领域，包括暴力与煽动暴力、危险组织和人物、欺凌和骚扰以及血腥暴力。他们还提醒我们，我们的平台上出现了一个新的民团运动 — Bwa Kale。这些洞察情报让我们的运营和调查团队能够利用关键词，主动识别潜在的违规内容，有效降低现实世界中的伤害。

案例分析：埃塞俄比亚报告激增

鉴于埃塞俄比亚紧张的社会政治局势和持续不断的暴力事件，我们每月都会与可信合作伙伴举行会议，讨论相关的内容趋势，以及加强通过可信合作伙伴渠道报告高危内容的做法。每两个月，我们还与埃塞俄比亚和散居国外的民间社会组织开展一次埃塞俄比亚对话系列活动。我们在埃塞俄比亚的合作伙伴持续向我们报告内容，并对错误信息和隐晦威胁报告开展危害评估。通过与这些合作伙伴合作，我们能够更好地识别关键问题，包括断章取义的图片、传播错误/虚假信息的虚假账户、利用自身影响力传播错误信息的已认证账户、仇恨言论、合谋骚扰和恶意暴露个人信息、最初由公众人物创造的带有政治色彩的措辞，以及危险组织和人物。

案例分析：缓解孟加拉国宗教内部冲突

2023 年 3 月，孟加拉国穆斯林少数民族逊尼派穆斯林和阿哈默底亚派穆斯林之间爆发了宗教内部骚乱。我们在孟加拉国的可信合作伙伴报告的内容包括针对阿哈默底亚派穆斯林群体的有害错误信息、仇恨言论、暴力与煽动暴力内容。通过利用以前指定的、预先审核过的有害言论，我们能够迅速对错误信息报告执行政策。可信合作伙伴还提供了重要的本地信号，这让我们能够根据“仅限升级处理”政策移除内容（例如隐晦的威胁）并深入了解本地仇恨言论措辞。

国际组织和多方利益相关者组织

随着科技的进步及其对人权影响的增加，我们继续与广泛的国际政府机构交流合作，其中包括：

- 联合国人权事务高级专员办事处
- 联合国教科文组织
- 防止灭绝种族罪行问题及保护责任办公室
- 秘书长技术事务特使办公室
- 联合国儿童基金会
- 联合国难民署
- 互联网治理论坛
- 联合国基金会
- 经合组织
- 世界经济论坛
- 其他区域性组织

我们是众多多方利益相关者倡议的成员，例如全球互联网反恐论坛 (Global Internet Forum to Counter Terrorism)、全球网络倡议 (Global Network Initiative) 和网络自由联盟咨询网络 (Freedom Online Coalition Advisory Network)。我们还积极参与行业合作，为政府政策提供意见和支持。

我们参加了第二届民主峰会，并且为响应峰会呼吁私营企业积极推进民主进程的号召，我们还公开分享了自己的承诺。在峰会上，我们因共同牵头制定“应对‘网络水军’带来的日益严重的威胁”的行业准则而受到了表彰，并在倡导私营企业支持增进民主反审查技术以打击专制政权的行动呼吁上发表了自己的意见。此外，我们在实现安全、私密的连接，以及打击政府滥用数字技术（特别是通过代理使用 WhatsApp 和使用雇佣监视）方面的工作也得到了认可。

我们参加了与秘书长技术事务特使办公室就制定全球数字契约 (Global Digital Compact) 进行的磋商，以支持在互联网治理和保护网络人权（包括隐私权和表达自由权）方面采取全球统一方法的必要性。我们还参与了联合国信息完整性行为守则的磋商，并提交了一份详细的意见书。

我们继续与联合国人权事务高级专员办事处的官员、独立专家和特别报告员定期会晤，讨论全球和具体国家/地区出现的问题，并积极参与 B-Tech 项目。在《世界人权宣言》(Universal Declaration of Human Rights) 发表 75 周年之际，我们做出了人权承诺。这包括对人权尽职调查和信息披露以及与民间社会组织和受影响社群沟通交流等方面的新承诺。

Meta 参与了联合国教科文组织 (UNESCO) 的“可信任的互联网”(Internet for Trust) 合作项目，出席了 2 月在巴黎举行的峰会，并参与了多方利益相关者流程，该流程旨在为数字平台法规制定以人权为本的监管准则。

我们与世界经济论坛合作，参与了一个广泛的项目，该项目旨在就新兴技术相关政策达成共识。我们帮助制定并签署了《全球数字安全原则：将国际人权理念引入数字领域》(Global Principles on Digital Safety: Translating International Human Rights for the Digital Context)，该文件建立了一个框架，旨在帮助各国政府和在线服务提供商采用多方利益相关者模式来促进数字安全。我们还是 AI 治理联盟 (AI Governance Alliance) 和“定义和构建元宇宙”倡议 (Defining and Building the Metaverse) 的成员，该倡议旨在为开发一个符合道德标准和负责任的元宇宙提供指导。我们在去年加入了 AI 治理联盟，该联盟旨在为 AI 治理建立全球标准，以道德和多元化的承诺为指导，积极推动科技进步，造福社会。

“我们单凭一己之力无法做到。要想这个项目真正发挥影响力，我们必须建立这些战略参与和战略伙伴关系”。

Lene Wendland，联合国人权事务高级专员办事处，工商企业和人权部门负责人

我们开始与联合国麻醉药品委员会 (UN Commission on Narcotic Drugs) 合作成立预防联盟 (The Prevent Alliance)，这是 Meta、Snap、美国政府和联合国毒品和犯罪问题办公室之间的一项公私合作倡议，旨在防止滥用数字平台从事与合成药物非医疗用途相关的非法有害活动。

我们向全球难民论坛 (Global Refugee Forum) 做出了承诺，支持难民、寻求庇护者和国内流离失所者的权利。

Meta 的 Data for Good 计划

我们的 Data for Good 计划继续与世界各地的人道主义机构、非营利组织、研究人员和政府合作，为边缘化群体提供支持。例如，该计划继续支持 IMPACT Initiatives（影响力倡议）等非营利组织的工作，帮助重新安置因乌克兰战争而流离失所的人们。该计划还为包括国际移民组织在内的国际机构提供支持，帮助他们改进全球相关指标，以支援全球移民和因自然灾害而流离失所的人们。此外， Meta 的 Data for Good 计划还通过 Development Data Partnership 等大规模合作项目，为包括世界银行在内的发展性机构从事的工作做出了重大贡献。举例而言，由 Data for Good 开发的 AI 驱动工具（例如，Relative Wealth Index [相对财富指数]），帮助世界银行更好地了解了空气污染对中低收入国家/地区特困人口的影响。

公开透明和补救措施

我们承诺做到公开透明并提供补救措施，并将其作为人权风险管理的核心宗旨。这两点均为《联合国工商企业与人权指导原则》的核心要素。我们会定期发布透明度报告，并且我们的年度人权报告也会介绍我们如何行使 Facebook 社群守则和 Instagram 社群守则、我们对政府请求的回应，以及我们如何保护知识产权。我们还在努力不断提高社群守则的可访问性，截至 2023 年年底，社群守则已发布 90 种语言版本。我们相信，监督委员会将继续提供社交媒体行业中独有的补救途径。

回应政府请求

相关的突出风险：

- 意见和表达自由
- 隐私
- 公正审判权
- 有效补救权

我们不会仅仅因为政府实体的要求而对内容采取行动或披露用户数据。作为全球网络倡议 (GNI) 的成员，我们承诺遵守基于国际公认人权标准而制定的 GNI《言论自由和隐私原则》（简称“GNI 原则”）。这项承诺非常重要，因为这些原则可指导我们如何回应政府请求，并帮助我们尽量减少对用户表达自由和隐私权的负面影响。

近年来，政府机构提出的请求在不断变化。政府行为者要求我们采取一系列广泛的行动，而不仅仅是限制在某个国家/地区访问某条内容。自 2020 年以来，政府请求的数量一直在不断上涨，Meta 根据当地法律做出限制的内容数量也在增加。

加入全球网络倡议十周年

“GNI 原则是我们管理政府提出的内容移除请求的‘指路明灯’。根据当地法律提出的内容限制请求往往会在人权方面带来非常现实、影响深远的难题。在当地法律与我们的人权承诺发生冲突时，我们会以 GNI、国际公认的人权标准和我们的企业人权政策作为指导”。

Siobhán Cummiskey，Meta 内容政策总监

GNI 是一个由多方利益相关者组成的倡议组织，其成员承诺维护重要人权条约《公民权利和政治权利国际公约》(ICCPR) 中规定的表达自由权和隐私权。GNI 为负责任的公司决策制定了一个全球标准，以促进和推动整个科技生态系统的表达自由权和隐私权。在《联合国工商企业与人权指导原则》的背景下，GNI 原则成为了政府保护人权的义务与企业尊重人权的责任之间的纽带。

2023 年是 Meta 加入 GNI 和承诺保护人权的十周年。我们遵守 GNI 有关科技公司在处理政府请求和限制时应如何尊重用户的表达自由和隐私权的原则，并对此负责。我们会定期开展独立评估，衡量我们对这一承诺的履行情况。

政府的请求既包括内容限制，也包括要求访问用户数据。这些请求分别涉及表达自由权和隐私权。当我们收到来自政府实体或法院的内容限制报告或用户数据访问请求时，我们会根据统一的全球流程对这两种请求进行审核（请见此处和此处）。

- **有时，政府官员会在开展官方调查的过程中，要求我们提供使用我们平台的用户的数据。**对此，我们将继续遵守 GNI 原则来回应政府提出的各种形式的请求。大部分此类信息索取请求都与刑事案件有关。在许多案件中，这些政府请求会要求获得用户的基本信息，例如姓名、注册日期和网龄。其他请求还包括获得 IP 地址记录或账户内容。我们采用严格的准则处理所有政府数据索取请求。我们会报告收到的请求数量、请求的用户/账户数量，以及我们生成数据的请求百分比。自 2016 年以来，我们一直在公开报告这些数据以及我们在回应请求时所依据的政策。

- **有时，政府当局会要求移除涉嫌违反当地法律的内容。**我们也可能会收到法院要求移除内容的命令。如果我们认定相关内容并未违反我们的政策，则我们可能会在根据我们的企业人权政策和我们作为 GNI 成员做出的承诺，对该内容开展仔细的法律和人权评估后，限制在声称内容违法的国家/地区对该内容的访问。如果法律没有禁止，我们也会通知受影响的用户。我们会考虑是否可以采取任何缓解措施，例如采取措施减轻请求造成的影响，包括在表示反对的前提下遵守请求（在某些情况下，这会导致此类请求被撤销），或由 Meta 对移除命令提出法律上诉、在我们的政策及信息公示平台发布案例分析以实时公开相关信息，以及/或者将请求副本提交至 Lumen 数据库。如果我们认为政府请求或法院命令不具备法律效力、范围过宽或违背国际人权标准，则我们可以要求澄清、上诉或不采取行动。

- **有时，我们会收到政府和法院的法律请求或命令，要求我们根据当地法律规定采取新的行动，**包括为内容添加更正声明（例如，说明当地政府认为该内容是错误信息）、限制用户访问特定功能、恢复之前因违反我们的政策而被移除的内容、要求我们在某个国家/地区内大规模自动限制内容，以及在全球范围内限制内容，尽管相关内容仅违反了一个国家/地区的当地法律。我们尊重业务运营地区的法律法规，但是如果当地法律义务与我们的人权和透明度承诺发生冲突，我们会力求尽最大可能尊重国际公认的人权原则。我们可以要求澄清、提出质疑或不采取行动。

- 在某些情况下，法律可能会禁止我们公布内容移除请求或其中包含的某些信息。在这种情况下，我们会力争在不违反法律义务的前提下，尽可能多地公布有关该命令及其存在的信息。

我们如何解读内容限制报告

我们会收到来自政府、监管机构、法院以及非政府实体和公众人士的报告，称内容涉嫌违反当地法律。我们自 2013 年起每半年发布一次内容限制报告，详细列举我们根据当地法律限制访问 Facebook 和 Instagram 内容的具体实例。

- 在该报告的特定国家数据部分，我们会详细介绍每六个月根据当地法律限制的内容数量、性质和类型。自 2023 年起，该报告提供的信息还包括：在某些国家/地区内，我们有义务根据当地法律的规定在该国家/地区内大规模自动限制内容，或在一些国家/地区内为内容添加更正声明。

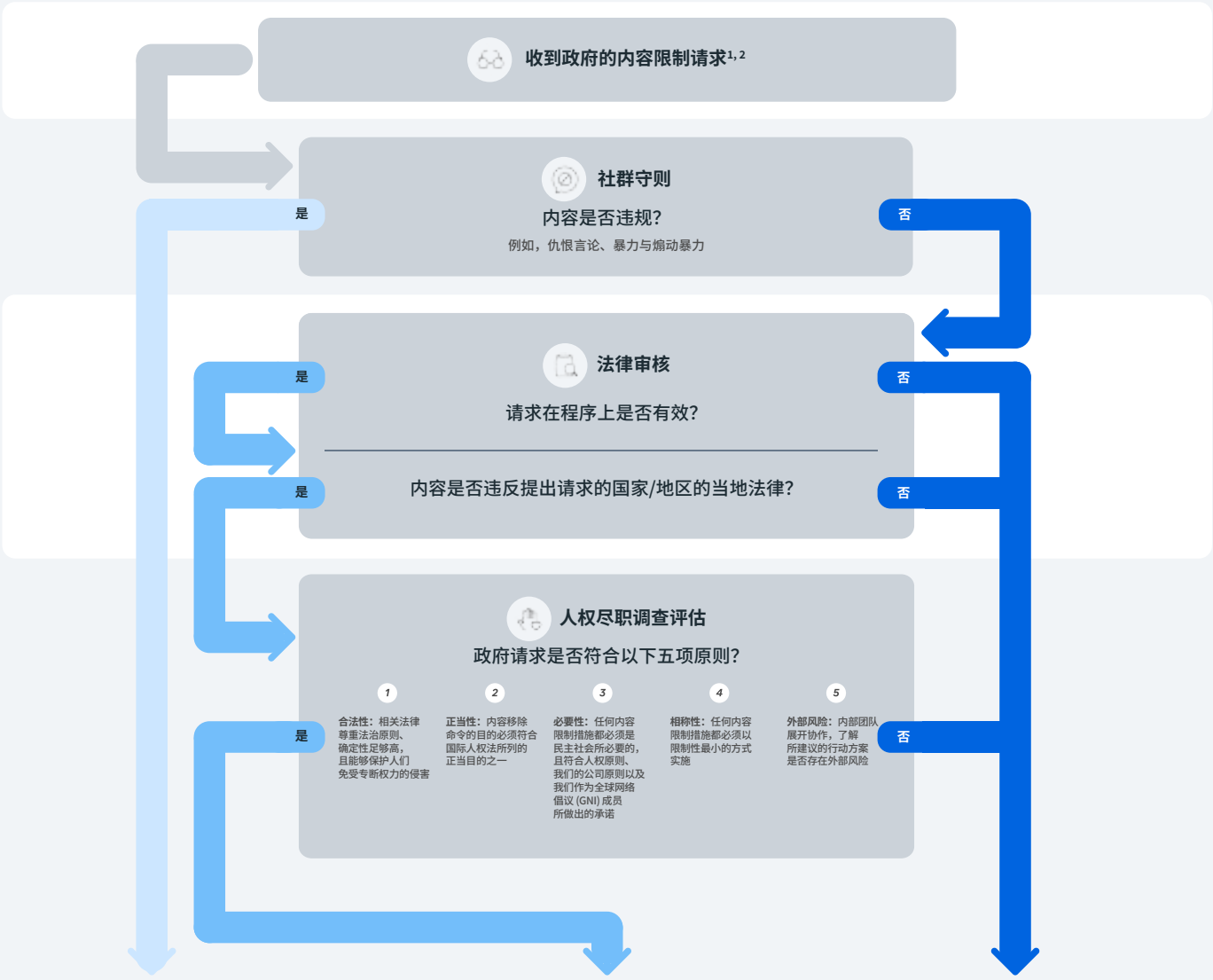
- 在该报告的全球限制部分，我们会详细介绍我们偶尔因域外法律要求而被迫在全球范围内限制内容可见性的情况，尽管相关内容仅违反了特定国家/地区的当地法律。

- 该报告的案例分析部分会详细介绍我们从政府和法院收到的一部分内容移除请求。自 2023 年起，我们侧重于只发布具有较高公共利益价值的案例分析。

- 该报告的“我们如何评估关于内容违反当地法律的报告”部分包含以下信息：我们对政府移除请求的审核流程，以及我们根据国际人权标准为减轻内容限制所造成影响而采取的缓解措施。

2024 年 5 月，我们发布了 2023 年下半年根据当地法律对内容做出限制的相关数据（请见此处）。

政府请求的生命周期



最终决定

我们可能会采取一些缓解措施，例如提起诉讼或对外分享政府请求限制内容的相关信息，以降低这些请求造成的影响。

根据我们的政策采取行动

当我们发现违反 Facebook 社群守则或 Instagram 社群守则的内容时，我们将对其采取适当的处理措施（例如，移除内容或设置年龄限制等）。

在内容违法的司法管辖区限制对内容的访问

在我们根据举报对违反当地法律的内容做出限制时，我们会将这一情况告知发布内容的用户，并且当用户尝试在其所在国家/地区查看因政府请求限制而被限制的内容时，我们也会告知用户这一情况。我们也会通知相关政府机构，我们已应其请求采取行动。

未采取行动或需要更多信息

我们可能会联系相关政府机构，要求他们提供有关请求的更多信息。

[1] 此图表描述了我们对政府提出的 Facebook 和 Instagram 内容限制请求的审核和回应流程，详见我们的政策及信息公示平台。如需查阅详细的说明，请访问政策及信息公示平台（点击此处的链接）。广告和其他 Meta 产品可能在此方面存在一些差异。

[2] 在紧急情况下，我们可能会被迫偏离上述生命周期。有时候，某个国家/地区的法律可能会要求我们根据当地法律的规定，在特定国家/地区大规模自动限制对内容的访问，但这种情况并不常见。在这种情况下，我们仍将以我们做出的全球网络倡议承诺和企业人权政策为指导行事。

有关我们加入 Lumen 的更新

2022 年 3 月，我们承诺加入 Lumen，这是一个由哈佛大学伯克曼互联网与社会研究中心发起的独立研究项目。该项目让研究人员能够调查政府和私人行为者就网络内容提出的移除请求。2023 年 11 月，我们向 Lumen 提交了来自奥地利、越南、新加坡、印度和墨西哥的首批移除请求，这些请求可在 Lumen 数据库中查阅。此举将进一步促进全球社群分析、报道和倡导互联网用户数字权利的工作。请参阅我们的内容限制报告，了解我们在公开共享政府移除请求的相关信息上所遵循的指导原则。

改进用户通知

为了进一步提高我们的信息透明度，并遵守有效补救权和公正审判权的标准，我们于 2023 年在 Facebook 和 Instagram 上改进了有关 Meta 根据政府和法院提出的法律请求而限制内容的用户通知。在大多数情况下，我们会在通知中告诉用户是哪个国家/地区当局提出的请求导致其内容受到限制，以及内容在哪个国家/地区受到限制。这一点不适用于因违反我们政策而被移除的内容。

相关的突出风险：

- 有效补救权
- 公正审判权
- 意见和表达自由
- 生命、自由和人身安全
- 公众参与、投票和竞选
- 平等和无歧视
- 儿童最大利益
- 隐私

监督委员会

监督委员会是一项业内领先的创举，其旨在帮助 Meta 解答关于表达自由和网络安全领域的一些最棘手的难题，包括哪些内容应该移除，哪些内容可以保留，以及相关原因。监督委员会是一个独立机构，可以审理 Meta 直接提出的案件，也可以审理 Facebook、Instagram 或 Threads 上反对我们内容审核决定的用户提出的案件。除了就是否应保留还是移除内容发布具有约束力的决定外，监督委员会还会发布有助于我们改进内容审核的建议，以及应我们的请求提供政策咨询意见。监督委员会以全球人权标准为指导，为 Meta 提供重要的专业见解和建议，帮助确保有关我们政策和产品的决策符合用户的最大利益。

2023 年是监督委员会具有里程碑意义的一年，这一年监督委员会超越了每年做出 50 项决定的目标，做出的案件决定数量达到 2022 年的三倍多。

监督委员会对其细则做出了重要调整，这使得“在特殊情况下，包括内容可能带来紧急的现实影响时”，可采用加急审理流程。此流程可在 30 天内发布加急内容审理决定。监督委员会对与以色列-哈马斯冲突有关的两个案件首次使用了加急审理流程。

监督委员会的决定涉及许多国家/地区各种不同的问题，包括伊朗抗议口号“Death to Khamenei”（哈梅内伊去死）、柬埔寨首相洪森 (Hun Sen)、性别认同和裸露内容、亚美尼亚战俘、巴西选举和古巴妇女抗议等。除发布内容决定外，监督委员会还在 2023 年向 Meta 提出了 60 项建议。监督委员会还发布了有关移除新冠疫情错误信息的政策咨询意见。请参阅监督委员会的透明度报告，了解更多详情（请见此处、此处和此处）。

允许保留具新闻价值的内容

作为对监督委员会所提建议的回应，我们进一步提升了透明度，说明在某内容可能违反 Facebook 社群守则或 Instagram 社群守则但展示该内容又符合公众利益时，我们会如何以及在何种情况下对此内容应用“允许保留具新闻价值内容”的规定。

- 从 2022 年 6 月 1 日到 2023 年 6 月 1 日，我们记录了 69 条符合“允许保留具新闻价值内容”规定的内容。
- 其中有 9 条（占比约 13%）是由政界人士发布的帖子。
- 在这 69 条符合“允许保留具新闻价值内容”规定的内容中，我们总共记录了 17 条可“批量”适用该规定的内容，换句话说，对于这些内容，适用“允许保留具新闻价值内容”的规定的内容不止一条。

我们的交叉检查计划

2023 年也是我们交叉检查计划的一个重要转折点。这一年年初，我们对监督委员会针对该计划的技术成熟度、操作严谨性、公平性、治理和透明度提出的 33 项建议做出了初步回应。我们采取了行动，通过与来自九个不同地区的合作伙伴举行反馈收集会议，加强了可信合作伙伴和民间社会组织对交叉检查治理发展工作的参与。

2023 年，我们共完成了 122 项建议中 61 项建议的执行工作，并继续努力落实其他许多建议（详见我们的监督委员会半年最新情况报告）。我们已经执行的建议涵盖我们的运营、政策和产品领域，推动了我们的整个公司和全球社群做出广泛而有意义的改进。

2023 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*	已完全执行的建议*
66 条	69 条	61 条
（2022 年有 91 条）	（2022 年有 75 条）	（2022 年有 14 条）

* 一些正在评估和/或执行的建议或已完全执行的建议包括前几年提出的建议（详见我们的 2022 年人权报告）。

展望未来

正如本报告所述，2023 年，我们见证了许多快速涌现的发展变化，也发现了许多将人权纳入决策的机会。我们力求做到这一点。然而，社会政治宏观环境并不稳定，许多信源表明，表达自由和民主原则正在遭到侵蚀。虽然人工智能 (AI) 技术已诞生数十年之久，但直到 2023 年，公众对 AI 带来的机遇和风险的认识才开始呈现爆炸式增长。

快速发展的监管环境为人权的推进带来了巨大的希望，同时也伴随着挑战：一方面，这可能会推动企业在应用符合主要人权原则（包括《联合国工商企业与人权指导原则》）的人权风险管理举措的方式和范畴上取得重大创新；另一方面，对合规的要求也会削弱企业对创新的追求和限制创新所需的空间。愈发明显的一点是，在长期的规模化人权风险管理工作中，需要持续使用符合《联合国工商企业与人权指导原则》的数据和工具，并将这些数据和工具纳入各项广泛的公司流程中。

> “我们发现，企业和政府有很多机会可以将人权考量纳入日常决策中”。
>
> John Ruggie，《联合国工商企业与人权指导原则》作者

Meta 努力在人权风险管理的方法上进行变革创新，包括在与利益相关者的合作上。体验式学习、模拟练习和案例分析等做法似乎大有可为，但尚未成为主流，我们将努力加强在这些领域所开展的工作。正如我们在本报告中所述，我们还认识到，齐心协力和变革创新对于解决与生成式 AI 相关的潜在人权风险至关重要。

我们知道，未来任重而道远。我们希望以我们所开展的工作和总结的经验教训为基础，为科技行业、商界和人权团体在认知和良好惯例的演变发展上做出一份贡献。

对于推进科技行业的人权事业，我们现在正处在一个挑战和希望并存的时刻。我们所有人，包括企业、民间社会、政府、联合国体系、投资者和监管机构，都需要努力培养新的技能、掌握新的知识以及开创新的合作方式，才能在塑造、共享和了解如何开展规模化的人权风险管理上取得一番成就。

放眼未来，澎湃的创造力浪潮正席卷而来，要想驾驭这股浪潮，我们需要勠力同心、无惧困难、勇于尝试以及富有企业家精神的果敢抉择。值得庆幸的是，人权运动从不缺乏创新的勇气。让我们一起乘风破浪！

附录

Meta 如何治理和管理人权

清晰的管理和治理结构使我们能够在各项计划、服务和政策中推进尊重人权的工作。我们的人权专家负责指导企业人权政策的执行，这项工作受全球事务总裁和首席法务官的监督。

人权专家的任务包括：促使将企业人权政策融入现有和正在制定的政策、计划和服务中；开展尽职调查；以及支持为企业人权政策相关员工提供培训。企业人权政策为打造尊重人权的产品，应对新出现的危机以及迅速灵活地大规模落实人权提供了指导。

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专家的现有工作和正在着手开展的工作相关的情况通报。

对 Meta 员工开展人权培训

在 Meta，开发的方式与开发的内容具有同等重要地位。通过参加人权培训，员工能更好地了解自己肩负的责任以及履行责任所需的知识和技能。

我们于 2022 年推出了《Bigger than Meta: Human Rights》（人权大过 Meta）培训，并在 2023 年全年持续开展这项培训。该培训强调了我们的服务、政策和业务决策对人权的潜在和实际现实影响。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。这项培训与我们的民权培训相辅相成，后者以无歧视、公正和公平原则为中心。

我们的人权培训目标还得到了年度强制性隐私培训的支持。该隐私培训重点旨在培养我们共同具备的能力，保护个人（尤其是边缘化群体）免受因处理个人数据而造成的伤害。这项培训有助于保护人们的隐私权和数据保护权。

引用报告的访问链接

- 2024 年可持续发展报告
- 2024 年负责任商业行为报告
- 2022 年人权报告、2021 年人权报告
- 2023 年反奴役及反人口贩卖报告
- 2024 年反奴役及反人口贩卖报告
- 2023 年冲突矿物报告
- 2023 年民权工作进展报告
- Meta 透明度报告
- 监管报告和其他透明度报告
- 人权影响评估：端到端加密、菲律宾、缅甸、印度尼西亚、柬埔寨、印度、斯里兰卡以及以色列和巴勒斯坦

Meta

humanrights.fb.com/annual-human-rights-report