

人权报告



目录

本报告简介

要点汇总

人权风险管理

- 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. 儿童最大利益
- 6. 公共事务参与、投票和被选举
- 7. 结社和集会自由
- 8. 健康权

议题聚焦

- 以尊重人权为前提进行创新
 - 通过可穿戴技术提升无障碍水平
 - 持续推进负责任的 AI 创新
- 欺诈和诈骗
 - 打击东南亚犯罪诈骗集团
- 我们防范与应对危机的方法
 - 刚果民主共和国

03

06

09

10

13

14

15

16

16

17

17

17

18

18

19

19

22

22

24

25



利益相关者参与

可信合作伙伴

案例分析：在拉美地区提供有助预防年轻人招募的见解

国际组织

透明度与补救措施

附录

Meta 如何治理和管理人权

对 Meta 员工开展人权培训

参考报告链接

26

27

28

29

30

30

33

34

34

34

34



本报告简介

本年度人权报告涵盖了 2025 年 1 月 1 日至 2025 年 12 月 31 日期间的见解与行动。报告涉及的 Meta 服务和产品包括 Facebook、Messenger、Instagram、WhatsApp、Threads 和 Reality Labs。

本报告以 Meta 在尊重人权方面的工作为基础，反映了我们在履行 [《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）和 [企业人权政策](#) 相关承诺方面的进展。本报告展现了 2025 年我们如何在全公司运用这些原则，并介绍了去哪里获取更多深入的信息。

本报告中的内容基于我们在 2022 年开展的[人权突出风险综合评估](#)。该评估旨在确定并优先考虑我们可能对人们（包括我们产品的用户以及可能因为我们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¹的领域。本报告概述了这些潜在的突出风险，并举例说明了我们在 2025 年采取的行动和缓解措施。

在 2025 年，无论是对于我们公司还是我们的利益相关者而言，人权都依然是一个至关重要的话题。我们在精简报告的同时，继续力求有代表性地展现我们在全球范围内开展的工作，包括旗下多个团队的工作情况和利益相关者参与情况。

政策与进展

除了本人权报告之外，Meta 每年都会通过以下机制发布政策与进展报告：



-  年度报告
-  委托投票说明书
-  负责任业务实践报告
-  政策及信息公示平台
-  可持续发展报告
-  CDP 气候变化报告
-  联合国全球契约

本报告是对 [Meta 负责任业务实践报告](#) 的补充。我们会单独发布[报告](#)，介绍为了识别和降低业务运营和供应链中存在的现代奴隶制和人口贩卖风险，我们做了哪些工作。此外，我们还会遵守各国家/地区和欧盟的强制性报告要求，相关报告可在我们的[政策及信息公示平台](#)查阅。本报告的[附录](#)中提供了其他 Meta 披露信息的链接。

[前往附录](#)

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。



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，可能会对人权产生不同的影响。本报告介绍了 Meta 作为一家公司，针对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 针对所有实体和/或在所有情况下都采取了相同的行动。²



² 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，虽然本报告中所描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。



简介

要点汇总

管理风险

议题

利益相关者

透明度与补救措施

附录

要点汇总



这是 Meta 的第五份年度人权报告。本报告深入介绍了 Meta 在 2025 年为大规模管理人权风险、恪守对[《联合国工商企业与人权指导原则》](#)的承诺所开展的工作。



突出风险评估

我们在 2025 年的优先事项反映了 2022 年[人权突出风险综合评估](#)中确定的突出风险领域：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；公共事务参与、投票和被选举；结社和集会自由；以及健康权。

[→ 阅读更多信息](#)

推动创新

Meta 力求突破人工智能 (AI) 创新的边界，同时将人权考量融入我们的开发流程。我们推出了新模型、安全框架以及[Meta Ray-Ban Display 眼镜](#)和[Meta Neural Band 腕带](#)等产品。这款腕带利用肌肉信号实现 AI 眼镜的免动手控制。我们通过负责任的产品流程开发这些产品，始终将[无障碍](#)纳入考量，并尊重平等和无歧视权利。

[→ 阅读更多信息](#)

打击欺诈和诈骗

我们还加大了打击欺诈和诈骗的力度。我们在[这方面的](#)[工作](#)立足于平台政策的执行，以及与行业、执法部门和当地民间社会组织的跨区域合作。许多诈骗中心依赖人口贩卖和强迫劳动来运作；我们的缓解措施有助于尊重生命、自由和人身安全权，以及关于禁止奴役、苦役和强迫劳动的规定。

[→ 阅读更多信息](#)

儿童和青少年安全

同样，我们持续致力于保护[儿童和青少年安全](#)，其中一项举措是宣布对 Instagram 青少年账户的内容体验进行[全面焕新](#)。此次更新参考了 13+ 电影分级标准和家长的反馈*，以及专家的指导意见和《联合国儿童权利公约》(UN Convention on the Rights of the Child) 中关于符合儿童不同阶段接受能力的原则。

[→ 阅读更多信息](#)

危机应对

我们的 2024 年度人权报告详细阐述了 Meta 如何在[危机政策协议](#)的指导下，将人权原则融入[我们防范与应对危机的方法](#)。这项工作 在 2025 年得以延续。我们运用危机政策协议应对了一系列局势，包括印度（克什米尔）和印度尼西亚的局势，以及马达加斯加、塞尔维亚和其他国家/地区由青年主导的“Z 世代抗议活动”。在本报告中，我们提供了在刚果民主共和国开展危机应对工作的案例分析。

[→ 阅读更多信息](#)



利益相关者参与

我们的[企业人权政策](#)支持我们主动与利益相关者合作，让他们参与进来。2025 年，我们与各类人权团体、弱势群体、民间社会成员、学者、智库机构以及监管机构开展了交流合作。所沟通的关键主题包括 Meta 在 AI 和内容审核方面的方案、我们[1 月政策和产品调整](#)的影响，以及我们打击欺诈和诈骗的工作。

我们继续与世界各地的[可信合作伙伴](#)合作，以便识别趋势，并更好地了解线上内容和行为对本地社群的影响。他们在 20 多个国家和地区提供了相关见解，并识别出了当地的潜在违规内容。

[→ 阅读更多信息](#)

监督委员会

我们还提供了关于[监督委员会](#)工作进展的更新。监督委员会是一个独立机构，负责做出有约束力的裁决和无约束力的建议。在报告期内，监督委员会审议了多起与我们在尊重人权方面的工作有关的案件，涉及的人权议题包括表达自由、青少年的权利和利益、平等和无歧视权利等。

举例来说，根据一项建议，Meta [提高](#)了现有通知的详细程度 — 当政府提出正式举报，声称用户内容违反了当地法律，而 Meta 因此限制了该内容时，会向用户提供更详细的通知。

[→ 阅读更多信息](#)

管理政府请求

在 2025 年全年，我们始终秉持对[全球网络倡议](#)的承诺，力求尊重表达自由和隐私，包括在回应政府提出的内容限制请求时也是如此。此类政府请求一直在增加。除了常规的诚信报告外，我们还发布了与限制政治言论的政府请求相关的[案例分析](#)，这些案例来自奥地利、哥伦比亚、科特迪瓦、德国、印度、伊拉克、意大利、哈萨克斯坦、马来西亚、新西兰、波兰、罗马尼亚、新加坡、南非、坦桑尼亚、土耳其和越南。2025 年，Meta 还接受了第四次 GNI 评估（涵盖 2022 年 4 月至 2024 年 12 月期间）。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”。

[查看案例分析](#)

人权风险管理

[《联合国工商企业与人权指导原则》](#)明确表示，公司应确定他们的负面人权影响，以便有效地预防或缓解这些影响。

鉴于 Meta 业务规模庞大，公司运营可能涉及到各种各样的权利，预见和管理我们的[突出风险](#)至关重要，但这项任务也很复杂。我们需管理两类固有风险：一类源于我们自身行为，另一类源于第三方行为，包括我们平台用户的行为。



下表列出了我们固有的突出人权风险³，这些风险是在我们的 2022 年人权突出风险综合评估（Comprehensive Human Rights Salient Risk Assessment，简称 CSRA）中定义的，该评估的具体情况披露于我们的 [2022 年度人权报告](#) 中。此表举例说明了我们在 2025 年如何应对潜在风险。在本报告后文中，我们将更深入地探讨其中一些例子，以及我们是如何管理与人工智能 (AI)、选举及冲突相关的潜在风险的。

1. 意见和表达自由

[意见和表达自由权](#) 包括寻求、接收和分享各种信息和思想的权利。这是一项基础权利，对于维护人格尊严、个人自主与民主制度至关重要。表达自由是我们使命的核心组成部分，符合我们让每个人都有机会发声的价值观。

示例 — CSRA 中确定的潜在固有突出人权风险

Meta 的内容审核政策及其执行可能会限制表达自由。

示例 — Meta 在 2025 年应对潜在风险的措施

2025 年 1 月，我们更新了 [仇恨行为政策](#)，允许对某些主流热点话题（包括移民和性别）进行 [更多讨论](#)。

我们使用自动化系统来重点处理非法和严重违规行为，例如恐怖主义、儿童性剥削、毒品、欺诈和诈骗，而对于严重程度较低的违规行为，则依靠用户举报后再采取行动。我们还调整了系统，在一般情况下要求具有更大把握才能移除一条内容。2025 年晚些时候，我们在社群守则执行情况报告中汇报了政策执行精确度的进展，该报告属于我们的 [诚信报告](#) 的一部分。

我们终止了 2021 年为减少用户看到的 [公民内容](#) 数量所做的调整，并开始根据个性化信号（例如，点赞某条内容）推荐更多政治内容。

³ 关于“固有”风险与“剩余”风险的区别，请参见 [Meta 2024 年度人权报告](#) 第 11 页。



示例 — CSRA 中确定的潜在固有突出人权风险

Meta 的内容审核政策及其执行可能会限制表达自由。
(续)

示例 — Meta 在 2025 年应对潜在风险的措施

我们在美国终止了第三方事实核查计划，转而开始推行[社群备注](#)计划。在世界其他地方，我们继续依赖[事实核查机构](#)来处理 Facebook、Instagram 和 Threads 中的错误信息，事实核查机构独立于 Meta，且经过了无党派[国际事实核查网络](#) (International Fact-Checking Network) 的认证，在欧洲则经过了[欧洲事实核查标准网络](#) (European Fact-Checking Standards Network) 的认证。

➔ [阅读更多信息](#)

政府对内容的限制过于宽泛

我们致力于履行对[全球网络倡议](#) (GNI) 的承诺。2025 年，Meta 接受了第四次 GNI 评估（涵盖 2022 年 4 月至 2024 年 12 月期间）。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”。我们还发布了报告，说明我们对政府提出的数据请求或内容限制请求所做的[回应](#)（详情请点击[此处](#)和[此处](#)）。在透明度方面，2025 年，我们发布了涉及政治言论且值得关注的[案例分析](#)，这包括来自奥地利、哥伦比亚、科特迪瓦、德国、印度、伊拉克、意大利、哈萨克斯坦、马来西亚、新西兰、波兰、罗马尼亚、新加坡、南非、坦桑尼亚、土耳其和越南的案例。



推出社群备注

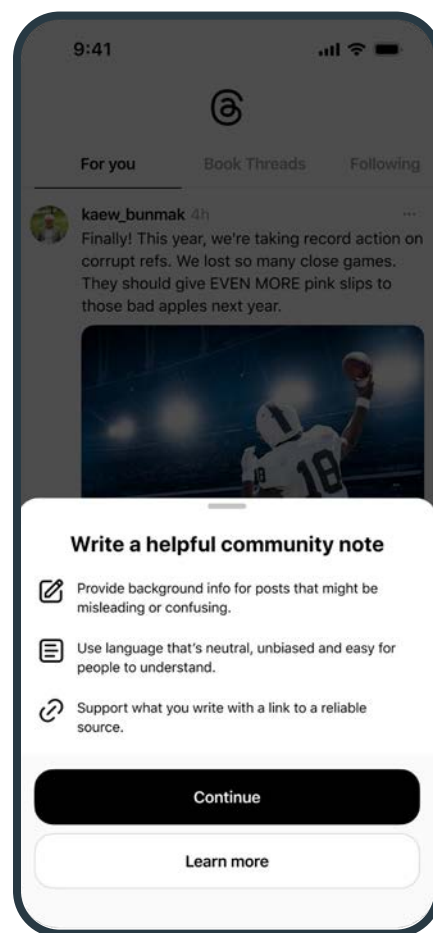
2025 年，Meta 在美国推出了[社群备注](#)。社群备注是一种新工具，供社群为可能存在误导性或令人困惑的内容添加背景信息。我们的方法以人权原则（包括表达自由）为指导，旨在通过相称且能保护言论的干预措施，让人们能更好地获取在用户社群中广受信任的信息。

符合条件的用户（年满 18 周岁、拥有经过安全验证且信誉良好的账户）可撰写和评价备注。只有在达成“有益共识”时，备注才会显示，这意味着有足够多此前意见不一的贡献者现在都认为该备注有帮助。作者姓名会被隐藏，这样可减轻其他成员造成的压力和报复风险，还能鼓励大家积极参与，包括放心地围绕敏感话题撰写备注。

社群备注可以促进表达自由，因为它能为内容添加背景信息（而不是移除合法言论或降低其排名），同时让更多人（而非仅限于少数专业事实核查人员）能参与进来。它还设有防护机制，以帮助防止违规使用。这些机制包括：账户诚信要求、垃圾信息和网址防护，以及对备注执行社群守则。

Meta 针对在决定是否将某个国家/地区排除在社群备注扩展计划之外时，我们应考虑的国家/地区层面因素，包括表达自由水平低、新闻自由缺失以及数字素养水平等，征求了监督委员会的指导意见。

2025 年 11 月，监督委员会接受了[Meta 提出的政策咨询意见请求](#)，以帮助我们制定在美国以外地区扩展社群备注计划的方法。监督委员会于 2026 年 3 月发布了其[建议](#)。





2. 隐私

[隐私权](#)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[企业人权政策](#)中的核心原则之一，就是保护用户安全和隐私。

示例 — CSRA 中确定的潜在固有突出人权风险

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。

示例 — Meta 在 2025 年应对潜在风险的措施

2025 年，我们为 [WhatsApp 中的 Meta AI 开发了隐私化处理技术](#)，该技术基于可信执行环境构建。借助隐私化处理技术，用户可以指示 Meta AI 在安全且私密的云环境中处理其请求。在该环境中，任何其他方都无法访问上述用户的个人消息，即使是 Meta 或 WhatsApp 也不行。该技术目前为 WhatsApp 中的某些可选 Meta AI 功能提供支持，例如总结消息内容和提供写作帮助。当隐私化处理技术完成对请求的响应后，消息不会被存储。请在我们的[隐私保护工作最新进展](#)中阅读更多相关信息。

我们对 CrushAI 等“脱衣” (nudify) 应用背后的实体提起了[诉讼](#)，这些应用允许用户在未经他人同意的情况下，利用 AI 生成裸体或露骨色情图像。我们进一步[宣布](#)了为打击此类应用而采取的额外措施。这项工作补充了我们长期以来的努力，即通过[StopNCII.org](#)和[美国国家失踪与受虐儿童中心的 Take It Down 服务](#)等工具，帮助防止私密图像（无论是真实的还是 AI 生成的）在网上传播。



3. 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歧视。为了尊重此权利，我们的举措包括禁止平台上出现仇恨行为，“仇恨行为”的定义可参见我们的[政策](#)。

示例 — CSRA 中确定的潜在固有突出人权风险

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。

示例 — Meta 在 2025 年应对潜在风险的措施

我们[推出](#)并开源了 [Meta Omnilingual Automatic Speech Recognition](#)（Meta 全语种自动语音识别系统），这是一套为超过 1,600 种语言（包括 500 种低资源语言）提供自动语音识别 (ASR) 能力的模型。对于此前缺少技术支持的语言，该技术可以将其口语内容转换为文本，从而开辟了新的教育机会，并进一步增强了表达自由，让人们能够以自己最熟悉的语言获取知识和分享内容。

对平等和无歧视有不利影响的内容（例如仇恨行为）

在 2025 年 1 月宣布[政策和产品方面的调整](#)之前，我们评估了这些调整的风险，并考虑了相关缓解措施（例如其他政策和用户举报的可用性），以处理此类内容。在发生冲突或危机的国家/地区，我们可能会在某些政策领域（包括仇恨行为违规）默认采用主动检测。例如，在部署危机政策协议后，我们在利比亚采用了主动检测。



4. 生命、自由和人身安全

[生命、自由和人身安全权](#)涉及不受人身伤害和人身限制的自由。对 Meta 来说，尊重此人权意味着减轻内容可能引起伤害的风险，这包括以下几方面的风险：暴力和人口贩卖、受国家支持的网络威胁，以及参与或倡导暴力活动或仇恨行为的非国家团体。

示例 — CSRA 中确定的潜在固有突出人权风险

存在以下行为的不良行为者：

- 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、活动人士和其他弱势群体

示例 — Meta 在 2025 年应对潜在风险的措施

通过深入研究以及与合作伙伴的磋商，Meta 发现苏丹存在童兵招募风险加剧的可能性，并针对该地区定制了阿拉伯语、法语和英语的青少年安全提示信息。2025 年 11 月，Meta 发起了两项互补的宣传活动：一项是在 Facebook 和 Instagram 上向约 100 万名苏丹青少年（13-18 岁）投放快速推广提示；另一项是向约 40,000 名与[战略性网络瓦解行动](#)所涉网络有关联的青少年，精准推送互动式自定义消息弹窗。这些消息提醒青少年注意招募风险，并引导他们获取外部安全资源。

[→ 阅读更多信息](#)

我们继续为人权维护者基金以及我们的[可信合作伙伴计划](#)提供支持，以完善针对人权维护者及其他弱势群体的应急响应机制。

5. 儿童最大利益

《联合国儿童权利公约》(UNCRC) 规定：关于儿童的一切行动，“均应以儿童的最大利益为一种首要考虑”。Meta 的[儿童最大利益框架](#)符合 UNCRC 的基本价值观。保护青少年儿童的网络安全是 Meta 的首要任务。我们为青少年、家长和监护人提供内置保护功能的工具，帮助保护青少年的安全，同时为青少年行使表达自由权和信息获取权提供空间。

示例 — CSRA 中确定的潜在固有突出人权风险

青少年儿童可能会接触到扰人、不当的内容，或者遭到诱骗侵害。

示例 — Meta 在 2025 年应对潜在风险的措施

在[青少年账户](#)的基础上，我们增加了[新的保护措施](#)，并推出了[更新的内容设置](#)，此举参考了 13+ 电影分级标准和家长的反馈。这意味着 Instagram 的目标是让青少年默认看到与适龄电影中类似的内容，同时考虑到有些家长希望为孩子设置更严格的体验，我们为此推出了更严格的新设置。我们还宣布推出新的控制选项，帮助家长管理[青少年在我们应用中与 AI 互动的方式](#)。此外，我们对“脱衣”应用背后的公司[采取了法律行动](#)，并[与学校合作应对网络欺凌](#)。

6. 公共事务参与、投票和被选举

在自由公正的选举中享有[公共事务参与权](#)、[投票权和被选举权](#)是民主的基石。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。无论是在选举前、选举期间还是选举后，我们都致力于在网上为选举保驾护航。

示例 — CSRA 中确定的潜在固有突出人权风险

违规内容可能会对公共事务参与、投票或竞选公职产生负面影响。这种负面影响可能源于多种行为，包括但不限于：

- 不良行为者有组织地干预选举
- 针对候选人的威胁，包括现实伤害与暴力风险
- 阻碍他人投票的个人行为、垃圾信息增加、境外势力合谋造假行为或者违规内容举报所反映出的行为

示例 — Meta 在 2025 年应对潜在风险的措施

我们立足于[2024 年度人权报告](#)中详述的规模化、基于风险的选举保障方案，继续推进选举保障工作。

2025 年，我们的选举准备工作涵盖澳大利亚、喀麦隆、智利、科特迪瓦、厄瓜多尔、埃及、德国、伊拉克、日本、缅甸、菲律宾、波兰、罗马尼亚和坦桑尼亚等国家/地区。

7. 结社和集会自由

[结社和集会自由权](#)对民主至关重要，并且与国际人权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）相互依存。对 Meta 而言，此权利与我们“让用户发声”以及“联系彼此，建立社群”的核心价值观一脉相承。

示例 — CSRA 中确定的潜在固有突出人权风险

Meta 平台上某些类型的内容，可能会让一些用户感到无法自由地在 Meta 应用上或线下建立联结。

示例 — Meta 在 2025 年应对潜在风险的措施

我们部署了[危机政策协议](#)，以调配资源用于处理与大规模示威活动（包括在伊朗、马达加斯加、尼泊尔和塞尔维亚等地发生的“Z 世代抗议活动”）相关的违规内容。

8. 健康权

[健康权](#)是指“人人有权享有能达到的最高的体质和心理健康的标准”。Meta 通过多种方式来尊重此权利，包括让人们更容易获取健康信息，使有类似健康问题的人能够相互联系，并使人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

煽动或意图造成现实伤害的违规内容

示例 — Meta 在 2025 年应对潜在风险的措施

我们宣布了[在 AI 工具中为青少年提供的保护机制](#)，包括从工具设计着手，让 AI 能够安全地回应与自我伤害、自杀和饮食失调相关的提示词。我们还[宣布推出](#)新的控制选项，包括让家长能了解孩子与 AI 所聊的话题。

在美国，我们进一步加大了应对校园网络欺凌的力度，为教师、教育工作者和管理人员提供了[优先举报机制](#)，用于举报青少年安全问题。

[简介](#)[要点汇总](#)[管理风险](#)[议题](#)[利益相关者](#)[透明度与补救措施](#)[附录](#)

议题聚焦

以尊重人权为前提进行创新

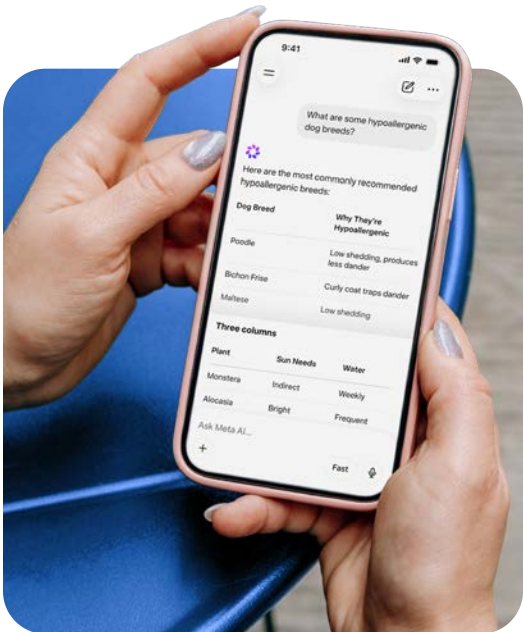
可穿戴技术和人工智能 (AI) 的创新为推动人权和支持负责任的风险缓解提供了重要机遇。

通过可穿戴技术提升无障碍水平

2025 年，我们[推出了 Meta Ray-Ban Display 眼镜和 Meta Neural Band 腕带](#) — 这是下一代 AI 眼镜与腕带的组合。这款腕带利用[表面肌电图 \(sEMG\)](#) 技术，将肌肉细微运动产生的电信号转化为指令，实现免动手控制。

我们在开发该腕带的原型时，始终将[无障碍](#)纳入考量，并尊重平等和无歧视权利。基于一项涉及近 200,000 名参与者的研究，该腕带为因脊髓损伤、中风或其他致残事件而导致行动受限的人士，以及有震颤症状或手指少于五根的人士提供了[新的可能性](#)。为保护隐私，将肌电信号转化为数字指令的机器学习过程[在设备本地进行处理](#)。

我们的技术搭配 AI 眼镜，可以[为残障人士提供支持](#)，让他们能够更独立地应对日常生活。免动手通话、实时翻译和实时字幕等功能打破了沟通障碍，而我们还与 [Be My Eyes](#) 携手打造了“[呼叫志愿者](#)”功能，让盲人和低视力人士能够联系志愿者，获取周围环境的实时描述。在构建这些工具时，我们直接与[美国盲人退伍军人协会](#)、美国退伍军人事务部的[盲人康复中心](#)等残障领域的利益相关者交流合作，以制定培训指南并支持无障碍应用场景。



持续推进负责任的 AI 创新

我们发布了[独立的 Meta AI 助手](#)，同时推出了更新的开源大语言模型 [Llama 4](#)。我们还推出了[隐私化处理](#)，这是一项[利用可信执行环境的新技术](#)，可在 WhatsApp 中启用 Meta AI 助手，同时采用加密通信等方法保护隐私。详情请参见我们此前的尽职调查分析，其中讨论了[端到端加密在人权方面的积极作用](#)。

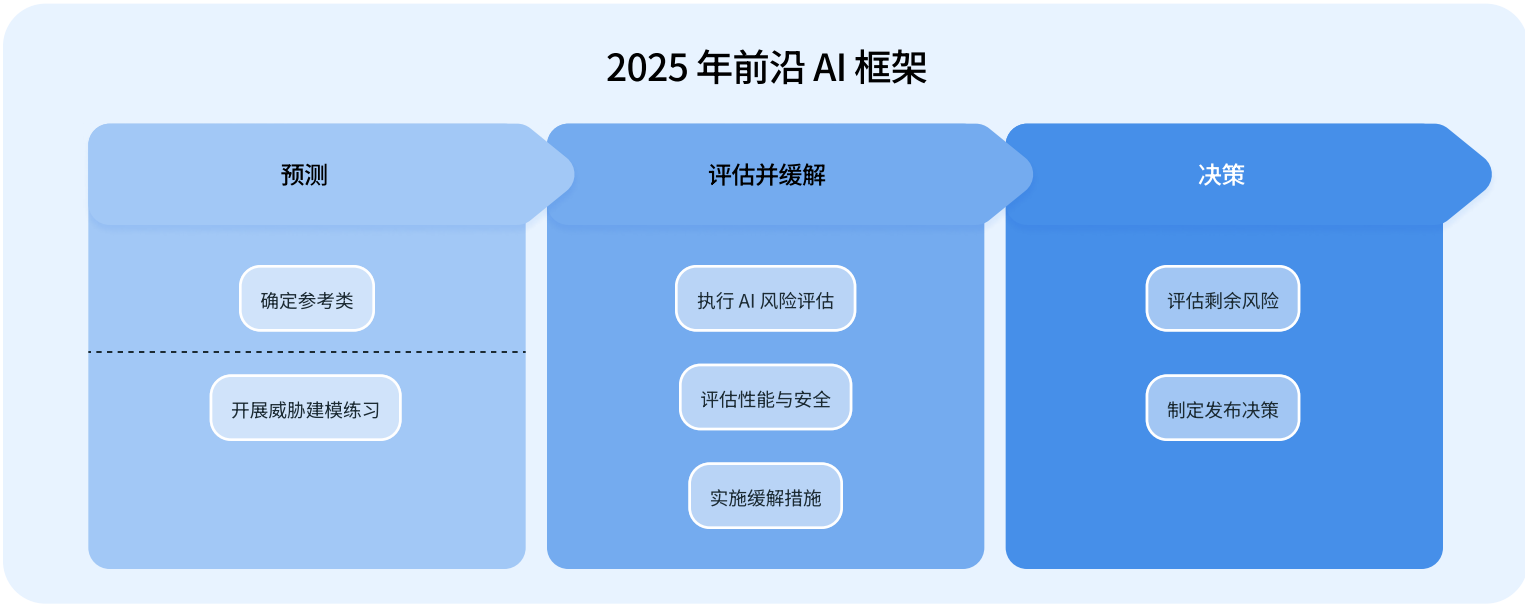
为降低我们的 AI 产品和模型被滥用及输出潜在有害内容的风险，Llama 4 的开发工作包括训练前数据过滤、训练后缓解措施和系统级缓解措施。这些内容在我们[关于 AI 安全防护的开发者使用指南](#)中有介绍。

我们还发布了[前沿 AI 框架](#) (Frontier AI Framework)，重点关注网络安全威胁与化学和生物武器风险领域中最关键的风险。我们的框架概述了我们在开发前沿 AI 系统时，用于预测和缓解风险的一系列流程，包括：

识别要防范的灾难性后果	我们的框架识别了与网络、化学和生物风险相关的潜在灾难性后果，这是我们全力防范的情况。该框架的重点是评估这些灾难性后果是否因技术进步而成为可能，如果是，则找出缓解这些风险的方法。
威胁建模练习	我们开展了威胁建模练习（必要时与外部专家合作），以便预测不同行为者可能如何设法滥用前沿 AI 来产生这些灾难性后果。这些练习为我们以后果为导向的方法打下了基础。
确定风险阈值	我们根据模型在多大程度上可能助长威胁场景来定义风险阈值。我们已制定相关流程，将风险控制在可接受的水平内，包括采取缓解措施。

我们还继续支持打造广泛的生态系统，以推动负责任的 AI 开发和部署。为支持开发者构建安全的应用，我们加大了对[系统级安全防护](#)的投入，包括推出 [Llama Guard 4](#)（可跨文本和图片模态实现统一防护），以及 [Llama Prompt Guard 2](#)（可更好地检测越狱和提示词注入）。

我们的工作参考了与民间社会、国际组织、学术界和公众围绕 AI 的人权影响开展的广泛交流与磋商。这包括与斯坦福大学协商民主实验室合作举办的[社群论坛](#)，其中汇集了与 AI 智能体及其治理的未来发展相关的多元化观点。





为内容审核工作打造的开源健康工具

审核员若想准确地执行我们的政策，还必须关注自己的[抗风险能力](#)和[健康状况](#)。

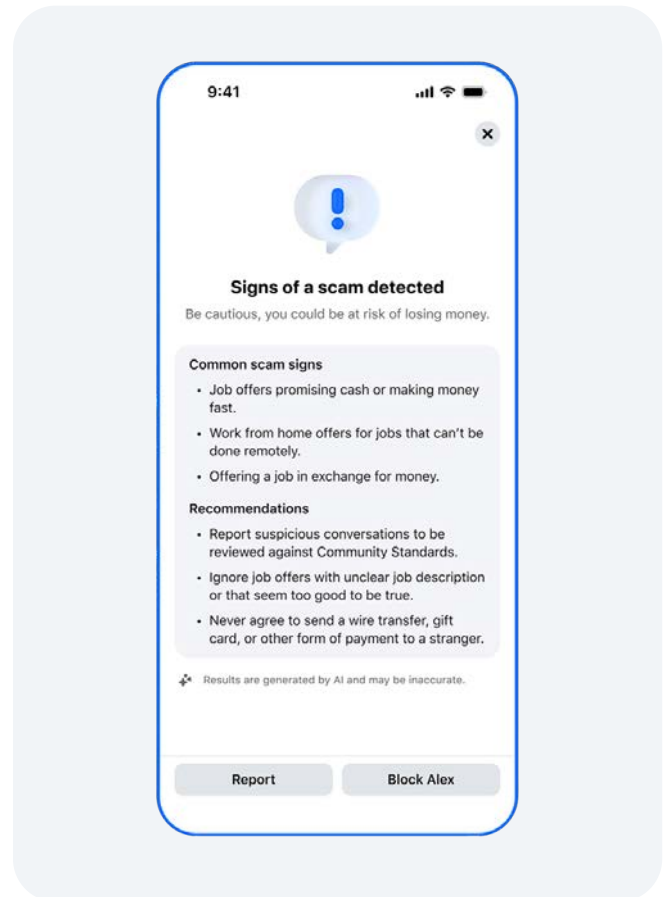
2025 年，我们开源了一套内容审核滤镜，旨在为审核员的身心健康提供支持，让其他开发者也能从我们在此领域的经验中受益。这些可自定义的图片和视频滤镜具有模糊处理、灰度/棕褐色转换、细节减少和基于 AI 的预警画面等功能，有助于减少看到血腥或令人不适的内容时对心理健康可能造成的负面影响。

欺诈和诈骗

欺诈和诈骗网络在过去十年中大幅增长，对人权造成了负面影响。许多诈骗中心依赖人口贩卖和强迫劳动来运作，诈骗受害者可能遭受重大的经济损失和其他伤害。2025 年，我们[移除](#)了超过 1.59 亿条违反我们政策的诈骗广告，还[封禁](#)了 Facebook 和 Instagram 上与犯罪诈骗中心相关的 1,090 万个账户。

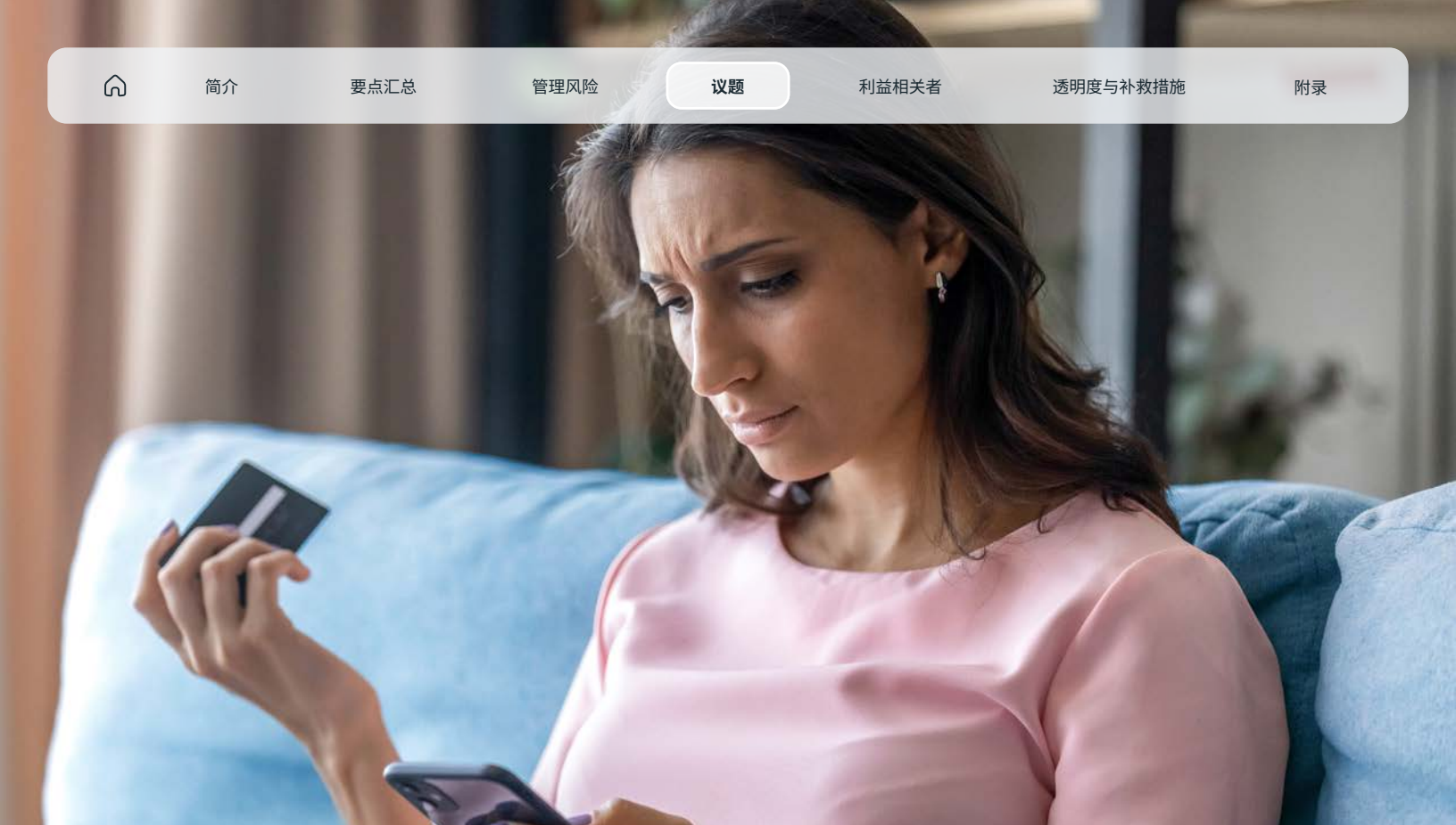
我们通过[平台防御措施](#)优先保障安全，例如利用[人脸识别](#)打击冒充行为，在 Messenger 上推出[诈骗检测警告](#)，以及在 WhatsApp 上推出[屏幕共享警告](#)。为了打击诈骗分子，我们的[工作](#)还包括提起诉讼和支持执法部门追究其责任。我们一方面要保障社群的安全与福祉，另一方面又要采取相称且尊重人权的风险缓解措施，两者兼顾并非易事。在采取这些行动时，我们始终考虑隐私保护，并遵循国际人权标准，包括[全球网络倡议原则](#)。

欺诈和诈骗是全球性挑战。信号共享、防御者社群之间的协作以及全社会参与，对于有效开展工作至关重要。2025 年全年，Meta 积极参与了多项信号共享合作，例如加入[全球信号交换](#) (Global Signal Exchange) 项目、[全球反诈联盟](#) (Global Anti-Scam Alliance) 和[欺诈情报互惠交换计划](#) (Fraud Intelligence Reciprocal Exchange)。我们还[倡导](#)公布未能采取充分行动的国家/地区的名单，并推动实施国际处罚措施；同时注意这些举措对人权的影响，避免对弱势群体（包括被贩卖并被迫从事诈骗活动的人员）产生非预期后果。



打击东南亚犯罪诈骗集团

过去十年中，产业化、复杂的诈骗网络和园区大量[涌现](#)。在[亚太地区](#)，这些园区雇佣了数十万人，且几乎一直逍遥事外。犯罪诈骗窝点通过欺骗性招聘信息引诱求职者，并强迫他们从事网络诈骗工作，通常还伴有身体虐待的威胁 — 这触犯了国际人权标准中关于禁止奴役、苦役和强迫劳动的规定。



2025 年 1 月至 10 月期间，我们以[违反欺诈、诈骗和欺骗行为](#)政策以及[危险组织和人物](#)政策为由，[移除了](#)超过 6,400 个 Facebook 账户和公共主页。其中一个被打击的网络源自柬埔寨，该网络[冒充](#)全球执法机构，谎称能提供损失追回服务。这项打击工作需要跨境合作，并密切监控地缘政治变化（例如缅甸境内及柬埔寨北部边境沿线的冲突）的影响。我们在 2026 年上半年[打击恶意威胁的报告](#)中更详细地讨论了这些工作。



我们 2025 年的打击工作在参与[联合打击行动周](#)试点计划时达到高潮。在这项于曼谷举行的计划中，Meta 与美国联邦调查局 (FBI)、泰国皇家警察和其他执法机构开展了合作。该试点计划[使得](#) 59,000 个账户、公共主页和小组从 Meta 平台上移除，并签发了六份逮捕令。



除打击行动外，我们还更新了[社群守则有关剥削的政策](#)，以覆盖更多潜在犯罪诈骗集团的招募内容。更新后的政策限制在以下地区发布招聘帖：经执法部门和当地非政府组织确认，存在劳动剥削高风险的地区。

Meta 还与联合国毒品和犯罪问题办公室合作开展了 [#TrappedInScamCrime \(#身陷诈骗犯罪\)](#) 宣传活动，以帮助打击与东南亚诈骗中心相关的、旨在强迫他人从事犯罪活动的在线招募和人口贩卖行为。该宣传活动结合幸存者的亲身经历来加强预防宣传，帮助人们识别欺骗性招聘信息并获取支持资源。

我们进一步与东南亚国家联盟合作，制定了应对网络诈骗的[原则](#)。这项工作是我们更广泛的[反人口贩卖工作](#)的补充。

我们防范与应对危机的方法

我们会提前防范并及时应对全球范围内的许多危机，包括冲突、族群内部暴力、内乱、大规模抗议、环境灾害，以及恐怖袭击和枪击事件等。2025 年，我们在喀麦隆、印度（克什米尔）、印度尼西亚、利比亚、马达加斯加、尼泊尔、塞尔维亚、坦桑尼亚和泰国-柬埔寨等国家和地区发起并协调了危机应对行动。针对乌克兰、苏丹和中东地区的冲突，我们根据[危机政策协议](#)，继续推进对已认定危机局势的处置工作。我们在[此前的人权报告](#)中阐述了这些已认定的危机局势。

在危机期间，危机政策协议是我们依赖的关键工具。该协议指引我们迅速采取措施来缓解潜在危害，具体内容详见我们的[2024 年度人权报告](#)。我们会根据所观察到的风险，采取特定类型的应对措施。这些措施会参考过往危机干预经验、人权原则以及武装冲突法。

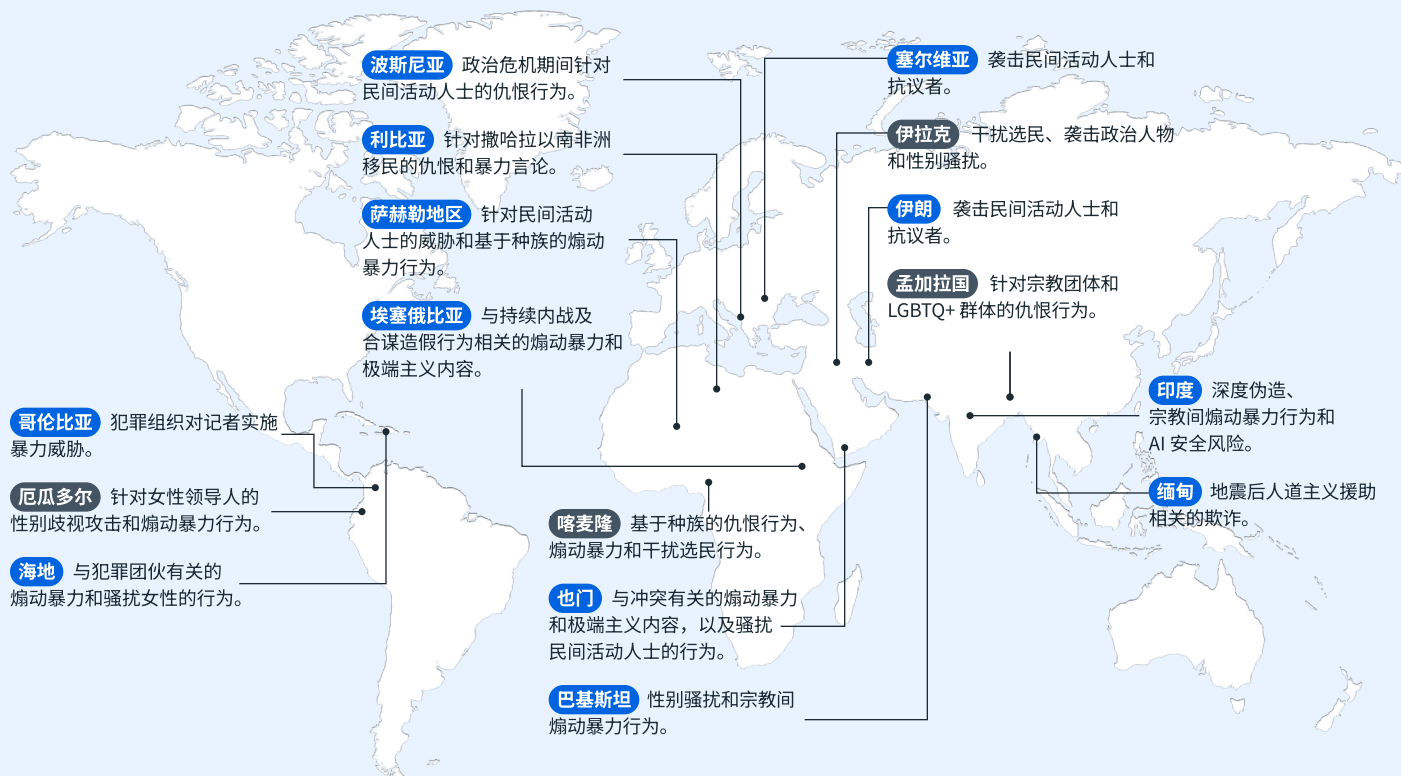
在下一页，我们以前报告中未涉及的一个地区为例，阐述我们针对危机、受冲突影响地区和高风险地区所采取的防范与应对措施。

图例

危机应对

选举

2025 年危机应对与选举诚信方面的工作重点聚焦





刚果民主共和国

2025 年，刚果民主共和国 (DRC) 东部的战斗加剧，M23 武装组织（据报道该组织受卢旺达[支持](#)）在北基伍省夺取了领土，导致平民大规模流离失所以及[严重的人权侵权行为](#)。可信合作伙伴向我们发出警报，指出该国东部暴力情况愈演愈烈，族群间的暴力风险也不断上升，这促使 Meta 根据危机政策协议认定了该危机局势。

自 2021 年底冲突再度爆发以来，由 Meta 专家组成的跨职能团队（涵盖政策、运营、安全和公共政策领域）一直在监测刚果民主共和国的线下局势，并根据我们的[错误信息和伤害](#)政策、[暴力与煽动暴力](#)政策、[配合实施伤害和宣扬犯罪行为](#)政策以及[危险组织和人物](#)政策，对违规内容进行处置。我们还部署了缓解措施以应对潜在的线上内容风险，包括在“错误信息和伤害”政策中认定经过预先审核的潜在有害言论，以及实施我们的[临时高风险地区政策](#)。此外，我们此前已针对“卢旺达语群体、卢旺达人、巴尼亚穆伦格族人或图西族人”，认定了具有身份曝光风险的相关群体，以帮助预防现实伤害风险。

随着 2025 年冲突升级，我们评估并缓解了相关内容风险。我们提供了政策说明和指导；审核了关键词，以确保有效的检测和精确的政策执行；并且应用了临时高风险地区政策。

Meta 还咨询了可信合作伙伴，以便在可能要认定新的经过预先审核的有害言论时参考他们的意见。“经过预先审核的有害言论”是指“错误信息和伤害”政策中认定的虚假或误导性言论，它们可能直接导致即将发生的人身伤害风险。利用可信合作伙伴的见解，我们能够更有效地检测和移除违反政策的内容，例如可能激化当地对立的内容。

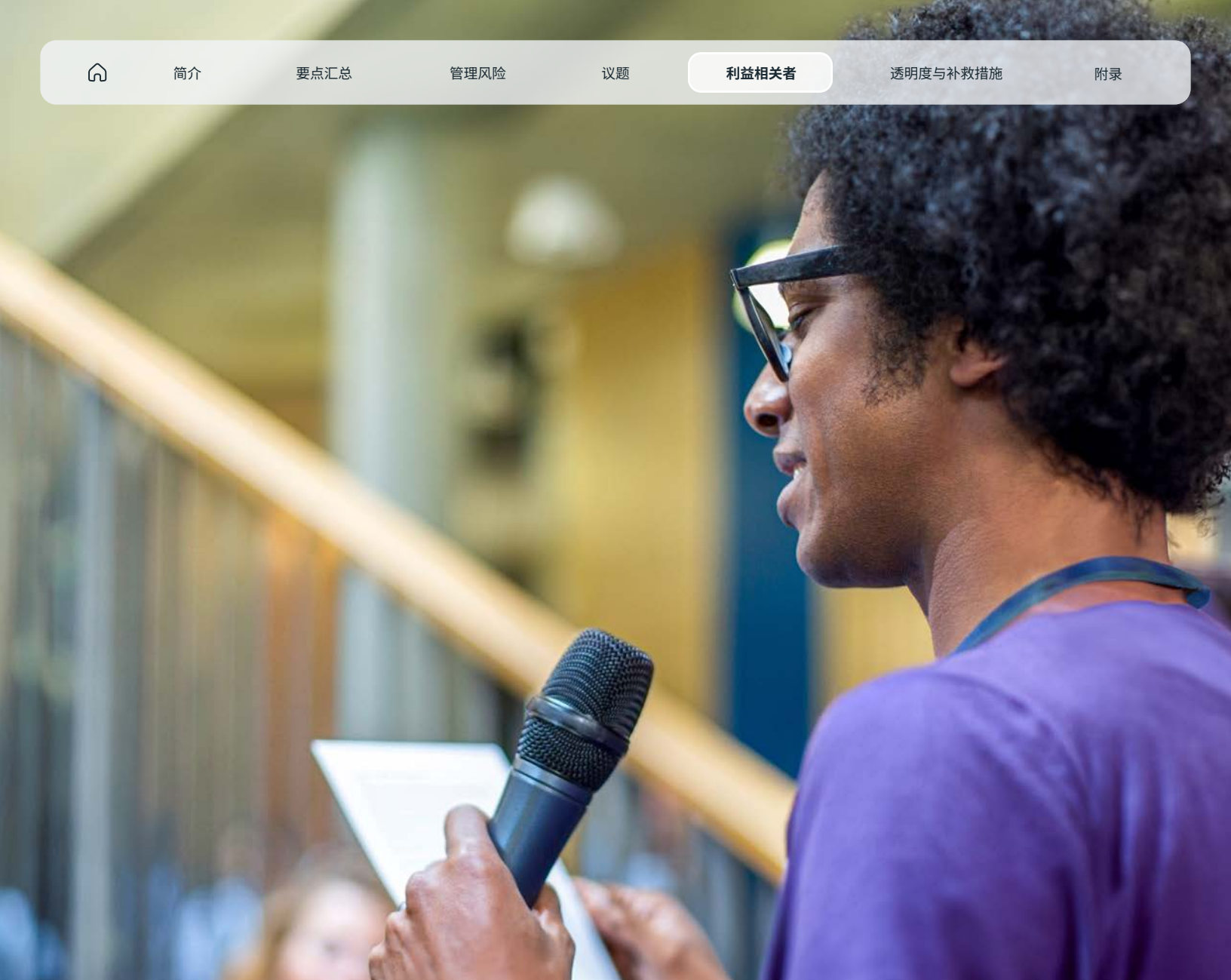
利益相关者参与

我们有规划地主动与全球用户社群[交流合作](#)，这有助于塑造 Meta 的政策，也是我们人权风险管理工作中的核心环节。

2025 年，我们与各类利益相关者广泛开展交流合作，包括民间社会成员、学者、智库机构、人权专家和监管机构。所沟通的关键政策问题包括我们在人工智能 (AI) 和内容审核方面的方法、2025 年 1 月推出的产品和政策调整（包括[社群备注](#)）的影响，以及欺诈和诈骗问题。

我们与来自 65 个国家/地区的超过 615 位外部利益相关者进行了交流，为多个领域的政策制定工作提供参考，这些领域包括危险组织和人物，以及健康与养生方面的广告发布政策。我们还与数百位其他相关方进行了磋商，内容涉及风险预测、趋势分析、政策执行和危害预防等工作。

此外，我们就一系列表达自由和监管议题与民间社会进行了交流，包括在阿根廷和巴西，以及在马来西亚举行的[亚太数字权利大会](#) (Digital Rights Asia Pacific Assembly Conference) 上展开相关对话。我们还面向民间社会成员开展了关于我们政策的培训。例如，在北非，我们面向民间社会团体开展了关于[错误信息和伤害](#)政策以及[仇恨行为](#)政策的培训，帮助他们识别可能违反这些政策的内容。这类培训还帮助人权维护者和弱势群体掌握了保护其数字身份的工具。



可信合作伙伴

在识别严重违反社群守则的行为方面，可信合作伙伴是我们的重要盟友。我们继续与[可信合作伙伴](#)合作，以便识别趋势，更好地了解线上内容和行为对本地社群的影响，并探讨如何改进民间社会的上报渠道。

2025 年，在动荡加剧或举行选举的国家和地区，可信合作伙伴提供了相关见解并识别了有害内容。这些国家和地区包括阿富汗、柬埔寨、科特迪瓦、刚果民主共和国、印度、印度尼西亚、以色列、肯尼亚、黎巴嫩、马达加斯加、墨西哥、尼日利亚、巴勒斯坦、索马里、苏丹、叙利亚、泰国、坦桑尼亚、乌克兰和委内瑞拉等。

借助可信合作伙伴计划，我们移除了逾 6 万条违反政策的内容。可信合作伙伴在拉美地区的工作成效示例见下一页。

案例分析

在拉美地区提供有助预防年轻人招募的见解



在拉美地区，可信合作伙伴在内容举报方面发挥了关键作用，这包括举报美化、支持[危险组织和人物](#)以及充当其代言人的内容。他们也就预防年轻人招募的举措提供了反馈。

在巴西，可信合作伙伴举报了 Instagram 上描绘某个已认定犯罪组织的成员的内容，包括展示违禁枪支、暴力行为和毒品的内容。在哥伦比亚，可信合作伙伴举报了代表犯罪团伙的主页，背景是这些团伙对瓜维亚雷省的新闻工作者进行了威胁和暴力袭击。在海地，被举报的内容美化了已认定的帮派关联人员的犯罪行为。这些见解为我们的多项工作提供了支持，包括移除违规内容、减轻平台上由危险组织和人物带来的风险，以及降低人身袭击风险。

Meta 与可信合作伙伴和区域专家合作，在拉美地区发起了面向年轻人的信息宣传活动。这些宣传活动面向墨西哥 80,000 名 13-25 岁的用户，以及哥伦比亚、海地和巴西的另外 55,000 名用户，旨在告知用户哪些内容在平台上属于违规、向他们宣传与危险组织和人物接触的风险，并支持他们在网上自行做出更明智的选择。该宣传活动让账户停用和删除好友的行为出现了显著增长。



国际组织

Meta 在这一年里继续与联合国系统各机构保持合作。我们参加了第 80 届联合国大会 (UNGA) 的政策讨论，议题包括全球数字和 AI 治理、AI 在[国家安全](#)领域的应用，以及弥合 AI 能力差距。在联合国大会高级别周期间，我们还参加了联合国“人工智能治理全球对话”机制的启动会议。

Meta 继续与[人权事务高级专员办事处](#)保持密切合作，包括参与 [B-Tech 项目](#)。我们回应了提交意见的请求，并与[非洲人权和人民权利委员会](#)以及联合国人权特别程序下的独立人权专家进行了磋商，这些专家包括促进和保护意见和表达自由权问题特别报告员。我们还与其他联合国机构以及人道主义组织进行了交流合作，以便为我们的危机应对工作提供参考。

此外，我们参加了 2025 年[互联网治理论坛](#)、[联合国工商业与人权论坛](#)以及 [RightsX Summit](#)，所讨论的议题包括对人权维护者有正面意义的 AI 应用场景等。我们继续参与了其他多个多方利益相关者论坛，包括[全球互联网反恐论坛](#)、[RightsCon 大会](#)和 [Tech Against Trafficking](#)，并在这些论坛上与利益相关者展开了积极交流。

透明度与 补救措施



[监督委员会](#)作为独立机构，旨在帮助我们解决关于网络表达自由的一些最棘手难题，包括哪些内容应该删除，哪些内容可以保留，以及相关原因。监督委员会负责审理 Meta 提出的案件，或者 Facebook、Instagram 或 Threads 上反对我们内容审核决定的用户提出的申诉案件，并会做出有约束力的裁决，决定是移除还是保留相关内容。监督委员会还会提供建议来改进我们的内容审核实践，并应请求提供政策咨询意见。



2025 年，监督委员会根据国际人权框架，审议了多起与我们执行内容政策有关的案件，涉及的人权议题包括表达自由、儿童的权利、平等和无歧视权利等。

监督委员会的审理决定示例：



监督委员会推翻了 Meta 不为一段可能经编辑的音频片段添加标签的决定，该片段涉及两名伊拉克库尔德政客密谋操纵议会选举。监督委员会认为，为符合 Meta 的人权责任，Meta 应该为本案中的内容以及该音频片段的所有实例添加“高风险”标签。



监督委员会多数成员维持了 Meta 保留两篇帖子的决定，其中一篇帖子中的视频涉及一名跨性别女性因使用女厕所而被质问，另一篇帖子中的视频涉及一名赢得田径比赛的跨性别运动员。

根据监督委员会建议所采取行动的示例：



根据伊朗女子街头遭遇对峙一案中的建议，同时为了契合我们减少过度执行的努力，我们更新了方法，转向更依赖举报的执行方式。这有助于降低对可能没有威胁性的比喻性言论或玩笑执行政策的风险。



鉴于监督委员会建议 Meta 提供与内容决定相关的更多详情，Meta 提高了现有通知的详细程度 — 当政府提出正式举报，声称用户内容违反了当地法律，而 Meta 因此限制了该内容时，会向用户提供更详细的通知。

详情请参见 [Meta 关于监督委员会的半年度报告](#)和我们的[监督委员会影响力页面](#)。

影响：通过用户通知提高透明度

+9%

2024 年第 1 季度 –
2025 年第 1 季度

在 Facebook 上根据社群守则
采取符合条件的处理措施时的
通知率增幅

+16%

2024 年第 1 季度 –
2025 年第 1 季度

在 Instagram 上根据社群守则
采取符合条件的处理措施时的
通知率增幅

~100%

2025 年第 1 季度

在 Instagram 和 Facebook,
内容因违反当地法律而被
限制时的通知率



我们还为监督委员会在表达自由和系统性风险评估方面的思想领导力提供了支持，以便监管机构能够借鉴监督委员会的经验。在这一年里，监督委员会参加了联合国大会的政策讨论，并与非洲人权和人民权利委员会等利益相关者进行了交流合作。

***免责声明：** 社交媒体与电影之间存在很多差异。我们在更新内容设置时并未与美国电影协会 (MPA) 合作，他们没有对 Instagram 上的任何内容进行分级，也没有以任何方式认可或批准我们的内容设置。不过，我们参考了家长已经熟悉的 MPA 公开指南以及家长的反馈。我们内容审核系统所用的标准与电影分级委员会不同，因此体验可能不完全相同。



简介

要点汇总

管理风险

议题

利益相关者

透明度与补救措施

附录

附录



Meta 如何治理和管理人权

我们的人权专家负责指导[企业人权政策](#)的执行，这项工作受首席全球事务官和首席法务官的监督。

人权专家的任务包括：推动企业人权政策整合工作，将该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；为面向员工的企业人权政策培训提供支持。企业人权政策为打造尊重人权的產品、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。

对 Meta 员工开展人权培训

在 Meta，我们的开发原则与开发成果本身同样重要。我们的人权培训强调了在现实生活中，我们的服务、政策和业务决策对人权的潜在和实际影响。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。

我们于 2022 年推出了《Bigger than Meta: Human Rights》（人权大过 Meta）培训，并在 2025 年全年持续开展这项培训。我们的隐私培训也为人权培训目标提供了支持。该隐私培训重点提升我们全员的隐私保护能力，避免个人（尤其是边缘化群体）因个人数据处理而受到伤害。

参考报告链接

[负责任业务实践报告](#)

[可持续发展报告](#)

[2024 年度人权报告](#)、[2023 年度人权报告](#)、[2022 年度人权报告](#)、[2021 年度人权报告](#)

[2026 年反奴役及反人口贩卖声明](#)

[2025 年度冲突矿产报告](#)

[Meta 透明度报告](#)

[监管报告和其他透明度报告](#)

[打击恶意威胁的报告](#)

[Meta 关于监督委员会的半年度报告](#)

