人权报告

Meta

见解与行动（2024 年度）

目录

本报告简介

本年度人权报告涵盖了 2024 年 1 月 1 日至 2024 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。

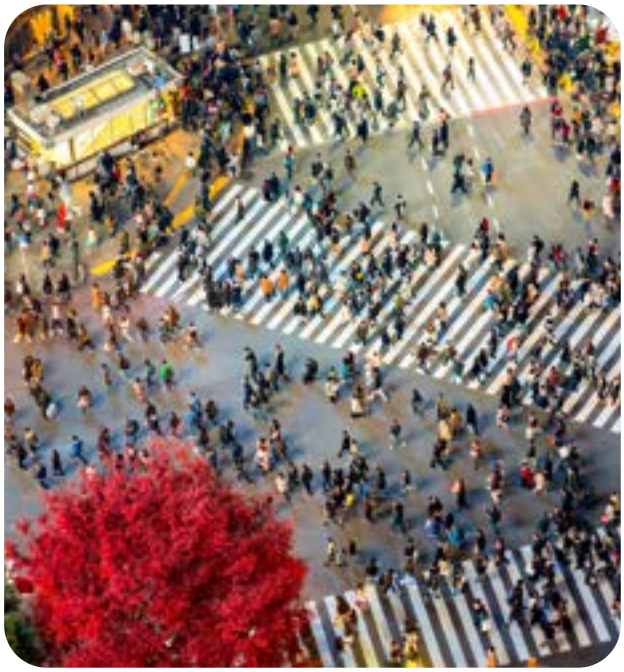
本报告以 Meta 在尊重人权方面的工作为基础，反映了我们在履行《联合国工商企业与人权指导原则》（United Nations Guiding Principles on Business and Human Rights，简称 UNGP）和企业人权在全公司运用这些原则，并介绍了去哪里获取更多深入的信息。

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 2024 年采取的行动和缓解措施。[2]

在 2024 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

本报告是对最新 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。

[2] 本报告不包括我们在 2025 年 1 月公布的内容政策和其他调整，当时我们更新了“仇恨行为”政策（以前称为“仇恨言论”政策），目的是回应对过度执行政策的关切，并力求保障更充分的表达自由。

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。[3]

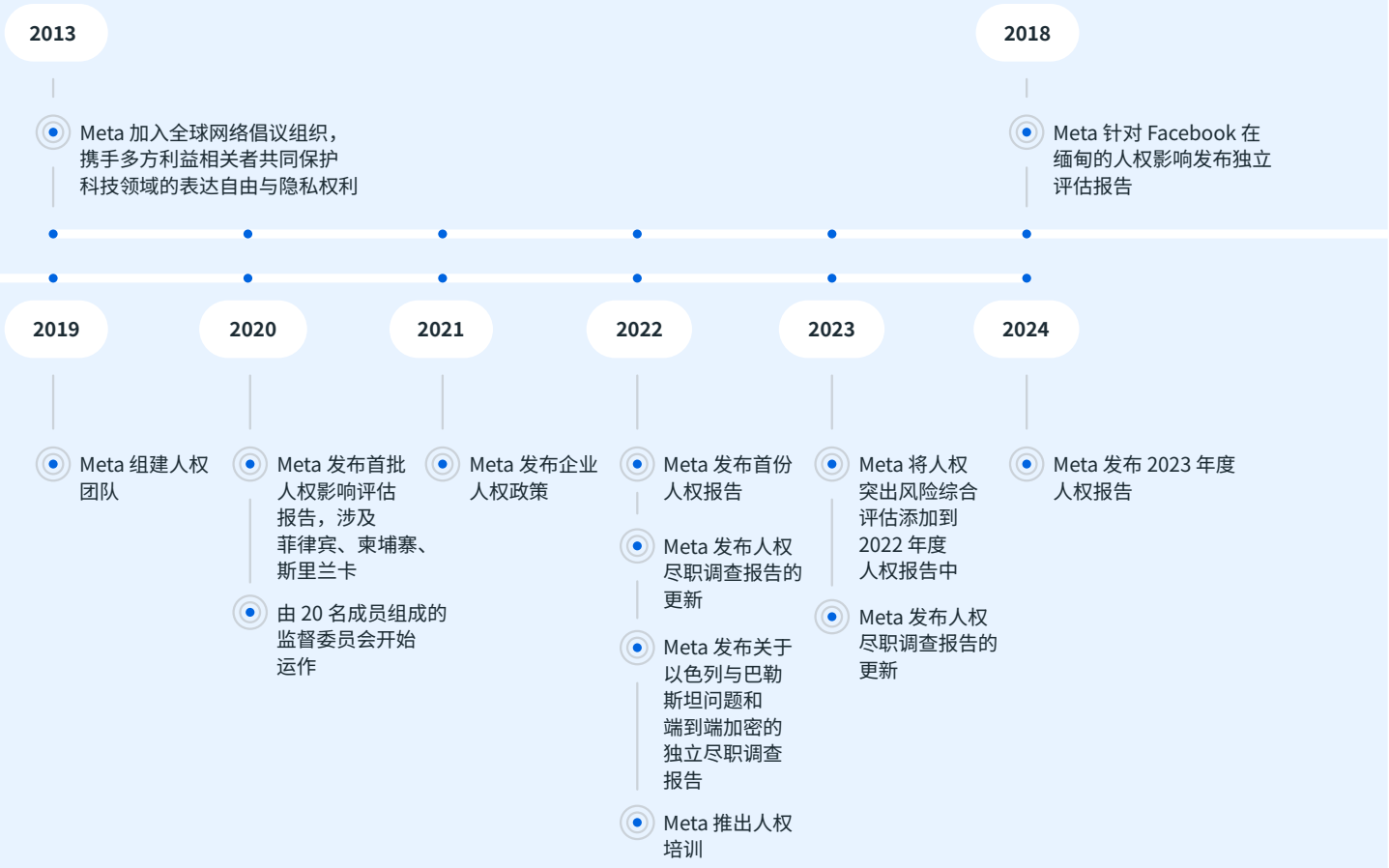
[3] 本报告对 Facebook 和 Instagram 内容审核和相关行动的介绍不适用于 WhatsApp，并且除非指明了某项政策或行动适用于 WhatsApp，否则该政策或行动应被视为不适用于 WhatsApp。此外，虽然本报告中所述的许多行动适用于 Facebook 和 Instagram，但是这两种服务的政策和程序之间存在有意识的区分。如果某项政策被标注为“Facebook”政策，则其不一定适用于 Instagram。本报告中的任何陈述均无意建立与将某项政策或程序应用于其他服务或实体有关的新义务（法律义务或其他性质的义务），也不应被解释为建立了这类新义务。

要点汇总

这是 Meta 的第四份年度人权报告。本报告深入介绍了 Meta 在 2024 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》(UNGP) 的承诺所开展的工作。

人权历程时间表

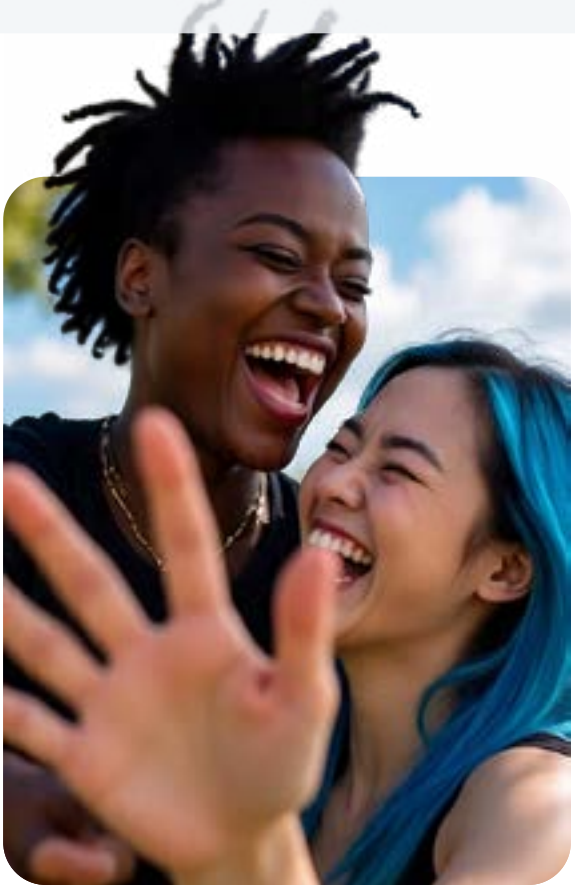
以下图表概述了我们的人权历程，体现了自 2011 年联合国人权理事会通过《联合国工商企业与人权指导原则》以来，我们的相关工作是如何发展演变的。



突出风险评估

阅读更多信息

我们在 2024 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。



加速 AI 创新

2024 年，人工智能 (AI) 加速发展。我们的愿景是普及个性化超级智能，惠及每一个人。

生成式 AI 应用得到越来越广泛的使用，日益改变着我们的沟通、学习、创作和工作方式。我们继续倡导能增强人权的开放式 AI 发展理念。这一理念不仅有助于保障人们的信息获取和自由表达，还能通过提升服务可及性和扩大语言包容性等方式，推动平等和无歧视权利的落实。

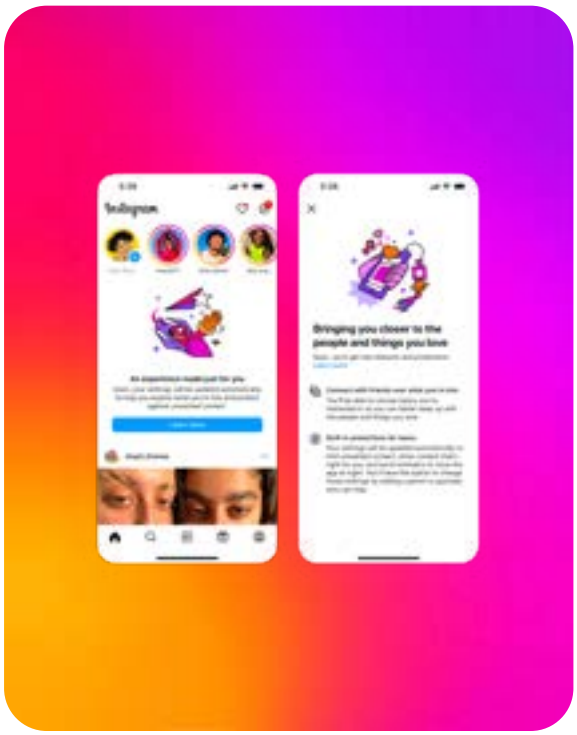
阅读更多信息

2024 年：选举年

2024 年堪称历史上规模最大的选举年。全球超过 70 个国家/地区 (覆盖世界半数以上人口) 举行了全国/地区层面选举。大约有 20 亿人有资格投票。对于举行选举的国家/地区，我们致力于保障其民众的表达自由权、政治进程参与权和信息获取权。

近年来，我们的方案历经数百场选举的打磨，已日趋成熟。这方面的具体工作包括：管理 AI 风险、执行我们与干扰选民投票或人口普查相关的政策、打击恶意行为网络、提高政治广告的透明度，以及为选民提供获取可靠信息的途径。本报告包含来自美国、墨西哥、印度和欧盟地区的例子。

阅读更多信息



儿童和青少年安全

我们持续致力于保护儿童和青少年安全。为此，我们在2024 年采取了多项举措，其中包括推出 Instagram 青少年账户，这是一项面向青少年的新体验，由家长提供指引。青少年账户内置保护功能，可限制谁能与青少年联系以及青少年能看到哪些内容，还能帮助管理青少年使用该应用的时长，同时为青少年探索兴趣爱好提供了新途径。我们的工作平衡了两方面需求：一是支持青少年培养自主意识，二是让家长和监护人能行使和履行其权利和义务。我们在开发这类账户时，遵循了专家的指导建议，以及《联合国儿童权利公约》（UN Convention on the Rights of the Child，简称 UNCRC）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

我们继续将人权原则融入到我们防范与应对危机的方法中。我们的危机政策协议指导我们迅速采用措施来缓解潜在危害。2024 年，我们根据危机政策协议认定了 19 起全球危机局势。在本报告中，我们提供了在孟加拉国、格鲁吉亚、中东以及苏丹开展危机应对工作的例子。

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。2024 年，我们与各类利益相关者交流合作，为公司在多个议题上采取的方案提供参考，这些议题涉及表达、仇恨内容、错误信息和隐私。这些利益相关者涵盖各类人权团体、弱势群体、民间社会成员、学者、智库机构以及监管机构。所沟通的关键主题包括我们在负责任 AI 和诚信选举方面的方案，以及我们在认定危险组织和人物以及暴力事件时采用的信号。

2024 年，我们举行了六次政策交流委员会会议，来自 Meta 的主题专家在这些会议上分享了各种观点，并讨论了社群守则和广告发布守则有可能做出的调整。我们还举办了社群论坛，针对那些存在权衡取舍、尚无定论的议题收集公众意见。这些举措不仅帮助我们改进了产品，预见了新兴技术的潜在风险，也让公司外部的意见在我们的决策流程中拥有了更大话语权。

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。我们还探讨了如何完善相关的上报渠道。在 2024 年的密集选举期和局势动荡加剧的情况下，这些合作伙伴的专业知识和经验尤为重要。在孟加拉国、巴西、科特迪瓦、刚果民主共和国、法国、希腊、印度、印度尼西亚、肯尼亚、伊拉克库尔德地区、墨西哥、尼日利亚、巴基斯坦、塞内加尔、南非、叙利亚和委内瑞拉等国家和地区，他们还提供了相关见解，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

2024 年，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、健康权、平等和无歧视权利等。监督委员会是一个独立机构，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。监督委员会会做出有约束力的裁决，决定是移除还是保留相关内容。根据监督委员会提出的一项建议，对于通过可信合作伙伴计划举报的内容，Meta 评估了对这类内容所做回应的及时性和有效性。

阅读更多信息

管理政府请求

在这一年里，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。2024 年，我们发布了与巴西、德国、印度、伊拉克、以色列、新加坡和土耳其境内的政治言论相关的案例分析。

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。

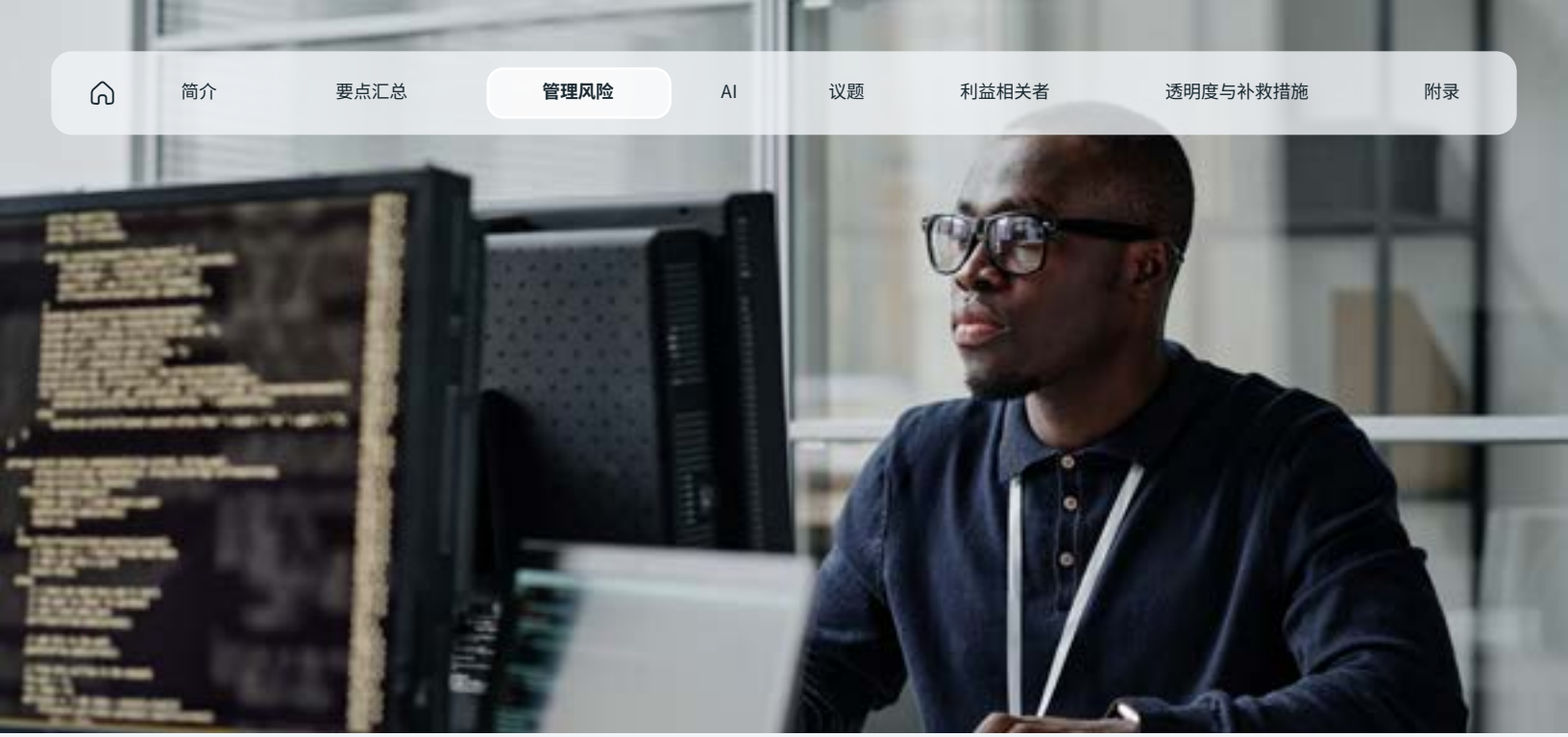
即便已实施相关流程来应对这些风险，仍会残留一定程度的风险，这类风险被称为“剩余”风险。虽然所有风险管理体系中都存在剩余风险，但是由于数字技术动态演进和快速发展的特性，以及高度活跃的第三方行为，与数字技术及其人权影响相关的剩余风险会持续存在。

接下来几页的表格列出了我们的突出人权风险，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 2024 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	示例 — Meta 在 2024 年应对潜在风险的措施
Meta 的内容审核政策及其执行可能会限制表达自由。	我们在制定政策时，继续将表达自由作为指引。2024 年，我们举行了多次政策交流委员会会议，这些会议旨在深入探讨多个领域中表达自由所面临的复杂挑战。
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。这包括发布报告，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）。2023 年度人权报告中详细阐述了我们在回应政府请求时的做法。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。2024 年，我们发布了多项值得关注的透明度案例分析，这些案例分析与政治言论相关，其中包括巴西、德国、印度、伊拉克、以色列、新加坡和土耳其境内的政治言论。
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。	为避免社交媒体和消息功能遭到封锁，我们可能会遵守合法的政府请求，与此同时，我们也会努力恪守我们对全球网络倡议的承诺，做到尊重表达自由。此外，对于无法直接连接到我们应用的用户，我们会继续提供 WhatsApp 代理访问功能。



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	示例 — Meta 在 2024 年应对潜在风险的措施
生成式 AI 模型处理个人数据的方式可能会超出用户的理解或预期。	我们会公开说明 Meta 如何将信息用于生成式 AI 模型和功能，并且制定了内部隐私审核流程，以确保负责任地使用数据（包括用于生成式 AI）。如想了解 2024 年隐私保护工作进展的更新信息，请点击此处和此处。 阅读更多信息
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。	2024 年 10 月，Meta 在 Facebook 和 Instagram 重新引入了人脸识别技术，帮助用户恢复被盗账户，并防范涉及虚假名人代言的诈骗行为。为了平衡潜在的隐私风险与诚信风险，我们为肖像被冒用行骗的公众人物提供了选择权，让他们可自主选择加入或退出该计划。

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	示例 — Meta 在 2024 年应对潜在风险的措施
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。	我们设计和部署了新机制，按方言分派阿拉伯语内容审核任务，从而实现更加高效和精确的审核，这包括在苏丹采取这种做法。新系统可检测阿拉伯语方言，并优先将内容分派给最有可能理解该特定阿拉伯语方言的审核员。
对平等和无歧视有不利影响的内容（例如仇恨行为）	根据相关调研结果、与外部利益相关者的交流以及平台内部核查，我们更新了仇恨行为政策中与攻击“Zionists”（犹太复国主义者）的内容有关的规定。 在训练 AI 模型时，我们测试了训练数据的内容或属性是否有可能增加生成潜在有害内容的风险，例如某个数据集是否涵盖了能代表多个人群的样本。

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 2024 年应对潜在风险的措施

Meta 有关配合实施伤害和宣扬犯罪行为的政策禁止鼓动、组织、宣扬某些犯罪活动或有害行为，或者自曝参与过这类活动或行为。2024 年，我们在该政策中提供了有关战俘的指南，帮助内容审核员更准确地批量移除违规内容，苏丹也纳入了相关规定的适用范围。

我们继续为人权维护者基金提供支持，并重新设计了可信合作伙伴计划，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	示例 — Meta 在 2024 年应对潜在风险的措施
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。	我们推出了 Instagram 青少年账户，这是一种面向青少年的新体验，内置防护功能，并由家长提供指引。 阅读更多信息

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	示例 — Meta 在 2024 年应对潜在风险的措施
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	2024 年，选举相关工作是我们的优先事项。我们针对选举开展了大规模的准备工作（包括为风险最高的选举保驾护航），并帮助选民获取可靠信息，同时采取了其他保障措施。 阅读更多信息

7. 结社和集会自由

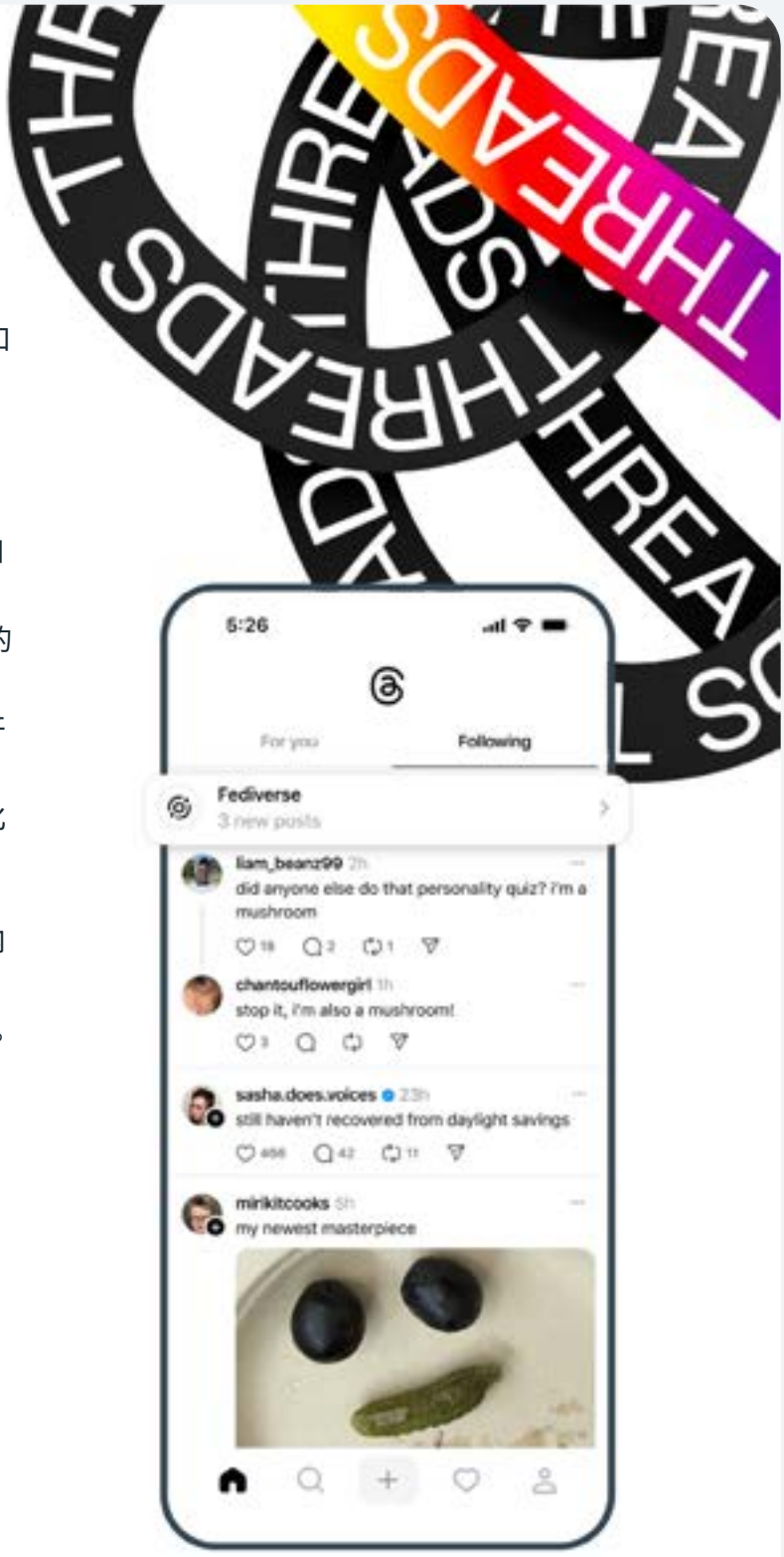
结社和集会自由权对民主至关重要，并且与国际人权法所保障的许多其他权利（包括表达自由权和参与公共事务的权利）相互依存。对 Meta 而言，此权利与我们“让用户发声”以及“联系彼此，建立社群”的核心价值观一脉相承。

示例 — CSRA 中确定的潜在固有突出人权风险	示例 — Meta 在 2024 年应对潜在风险的措施
Meta 旗下平台上的某些内容或合谋造假行为，可能会让一些用户感到无法自由地在 Meta 应用中或线下开展集会活动。	我们部署了危机政策协议，调配资源处理涉及大规模示威活动的违规内容，例如，我们在孟加拉国和格鲁吉亚部署了此协议。 在选举开始前，我们还会提前做好充分准备，降低那些可能阻碍公众集会的违规内容带来的风险，这类内容可能会导致人们因顾虑安全问题，而在选举期间及选举结束后不敢参与集会活动。 Threads 加入了联邦宇宙，这是一个开放性的全球社交媒体服务器网络。此举让用户能够拓展受众社群，并触达新受众。

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	示例 — Meta 在 2024 年应对潜在风险的措施
煽动或意图造成现实伤害的违规内容	我们与 Snap 和 TikTok 携手推出了 Thrive 计划，这是一项全行业信号共享计划，旨在预防自杀及自我伤害相关内容的传播。 我们围绕经监管机构认定存在健康和安全风险的商业内容，举行了政策交流委员会会议。 我们更新了社群守则和广告发布守则，纳入了与已召回商品相关的规定。



新产品和服务

正如 2023 年度人权报告中所述，Meta 在设计和开发产品和服务时，致力于尊重人权。

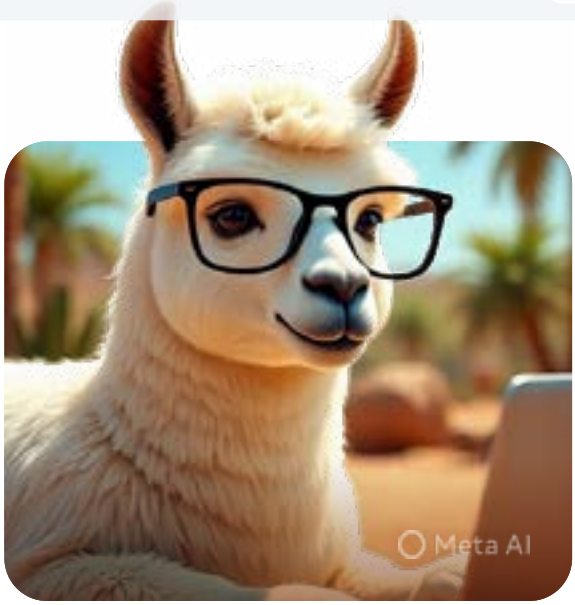
2024 年，Threads 加入了联邦宇宙，这是一个开放性的全球社交媒体服务器网络。如果某个用户开启向联邦宇宙分享内容的功能，那么来自其他平台（例如 Mastodon 或 Flipboard）的用户即便没有 Threads 主页，也能关注该用户的 Threads 内容并与之互动。这让用户可以触达新受众、拓展受众社群、参与所关注话题的公开讨论，从而行使其表达自由权与结社和集会自由权。与此同时，这也有助于打造更加多元化的信息生态系统。

我们还通过帮助中心的专门版块以及隐私中心内新推出的联邦宇宙指南，向 Threads 用户科普“去中心化”和“互操作性”对其隐私有何影响。

以尊重人权为前提，加速 AI 创新

2024 年，人工智能 (AI) 加速发展。生成式 AI 工具和应用得到越来越广泛的使用，日益改变着我们的沟通、学习、创作和工作方式。在 Meta，我们深知 AI 的飞速发展与普及也影响了人权领域，带来了意义重大且往往前所未见的益处与风险。

我们的长期愿景是普及个性化超级智能，惠及每一个人。



2024 年，我们发布了开源的 Llama 3、Llama 3.1、Llama 3.2 和 Llama 3.3 大语言模型 (LLM)。我们还推出了 Meta AI 助手，并将其全面集成到各项 Meta 技术中。Meta AI Studio 作为 AI 人物定制平台闪亮登场；全套生成式 AI 工具帮助广告主拓展业务版图；而 Meta AI 已与 Ray-Ban Meta 眼镜集成。我们继续开展前沿 AI 研究并公开了研究成果，这包括 Movie Gen 模型（可生成视频，并且可根据指令进行精确的视频编辑）和 Video Seal 视频水印模型（用于为 AI 生成的视频添加长效水印），以及其他多项技术突破。

截至 2024 年底，开发者对我们 Llama 开源模型的下载量已超过 6.5 亿次，Meta AI 在全球有近 6 亿月活用户，这让我们的模型成为全球使用最广泛的模型。面对如此庞大的开发者和最终用户群体，我们的责任尤为重大，即必须在尊重人权的前提下打造 AI 技术。

我们的开放理念

我们坚信，为了确保 AI 技术的进步惠及每一个人，开源 AI 是其中的重要一环。正如我们在 2023 年度人权报告中所述，开放理念能给人权带来重要益处。开源 AI 模型：

- 在本质上更有能力应对审查以及对表达自由权的其他限制，因为它们可以在下载后离线运行，从而降低发布后政府可能要求限制输出结果所带来的影响。
- 能更好地实现适配和微调，反映当地的具体情况和特点，这与平等权的原则相契合，同时也提升了技术可及性和语言包容性。
- 让开发者能轻松打造更小、更高效的模型，降低传统上服务欠缺人群的使用壁垒，为经济、社会和文化权利提供支持。
- 任何人都可以仔细检查这类模型，排查潜在风险，这为 AI 安全防护领域的关键研究提供了支持，有助于减轻 AI 可能对人权造成的负面影响。

我们已经见证了这种开放理念带来的切实益处。我们在 2023 年推出了 Llama 影响力补助计划，并在 2024 年继续推进该计划。这项计划与我们 2024 年推出的 Llama 影响力创新奖双管齐下，共同支持和表彰基于我们开源模型、具有积极社会影响的应用案例。

例如，开发者利用 Llama 打造了 Vax-Llama 模型，这是一项供全球医疗服务提供者采用的智能聊天助手服务，旨在提供准确的疫苗接种信息。Llama 还被用于 Llama-Suho 项目，此项目利用韩国相关数据对 Llama 进行微调，从而增强 AI 在韩国背景下的安全防护能力。

以防护为中心

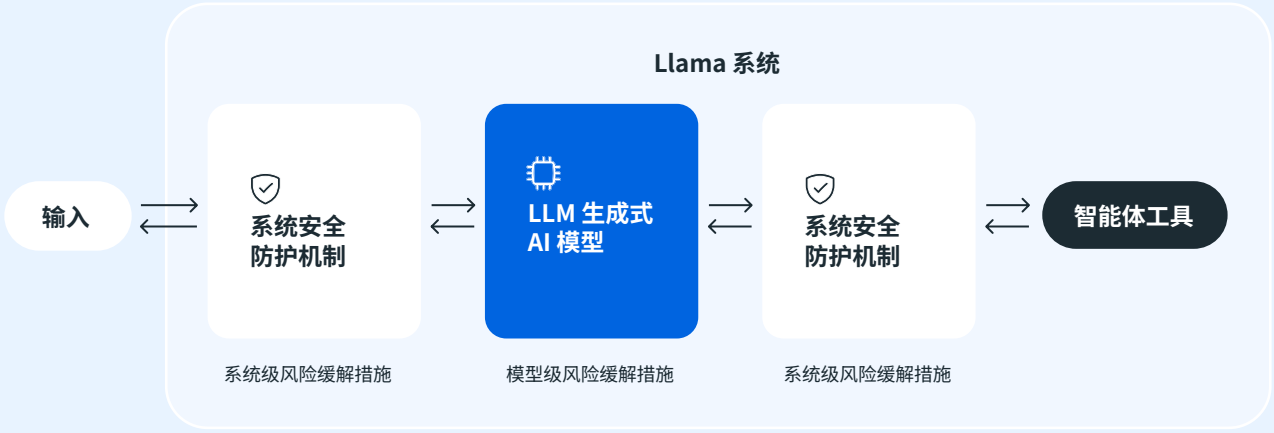
我们始终致力于开发和部署最先进的 AI 产品，同时考虑人权标准，并采取安全防护机制来防范滥用。

我们的企业人权政策明确表示，我们的人权承诺同样适用于 AI 领域。

随着 2024 年 4 月 Llama 3 的发布，我们开始强调基于系统的 AI 安全防护方法。基于系统的方法赋予了开发者更大的灵活性，让他们可以针对不同应用场景和受众，采用合适的多层防护。例如，针对某些可能具有冒犯性但合法的言论，我们提供了相关的防护机制，作为可选的系统级风险缓解措施。此外，我们一如既往地在基础模型中嵌入底线防护机制，防止生成儿童剥削内容。

我们认为，这种基于系统的方法有助于在表达自由与其他人权之间实现合理平衡。

AI 安全系统



为实现基于系统的方法，我们开源了三款重要工具（Llama Guard、Prompt Guard 和 Code Shield），开发者可定制这些工具，并结合使用它们或单独使用每款工具来实现防范滥用的机制。

我们的开发者使用指南详细讲解了在各种场景中，如何负责任地部署我们的基础模型和安全系统。我们的合理使用政策继续适用于对我们开源模型的部署。

除了在 2024 年推出上述工具外，我们还采取了多项重要措施，以缓解我们自身部署生成式 AI 所伴随的风险。我们在 2024 年的做法包括：

- 在我们的模型和第一方产品发布前，开展广泛的红队测试，确定并缓解潜在风险，包括涉及潜在负面人权影响的风险。
- 根据独立监督委员会的反馈，更新我们对经编辑影音内容的处理方法，包括为更多类型的视频、音频和图片内容添加“AI 信息”标签和背景信息，并要求创作者披露对 AI 的使用情况。
- 完善我们在测试模型输出是否符合要求时所遵循的内部准则和流程，使其更贴合真实应用场景，并与国际人权标准保持一致。

我们还意识到，AI 安全防护需要跨行业及多方利益相关者的协作。2024 年 2 月，我们与业内同行共同签署了 AI Elections Accord（AI 选举协定），此协定致力于帮助防范具有欺骗性的 AI 生成内容干扰全球选举。2024 年 5 月，我们加入了 Frontier Model Forum（前沿模型论坛），这是一个由行业支持的组织，致力于提升前沿 AI 模型的安全性。

解决不当拒答的问题

不当拒答是指模型收到非违规提示语，却拒绝生成所要求的输出内容，这往往是由模型善意设置的安全防护机制导致的。例如，某个模型可能会不当地拒绝讨论一部包含冒犯性刻板印象或诋毁之词的经典文学作品，也可能会因为设置了防止协助制造化学、生物、放射性、核武器及高威力炸药的安全防护机制，而不当地拒绝回答一道基础的高中化学题。尽管模型安全至关重要，为了限制有害内容的生成，拒答机制必不可少，但是不当拒答会对表达自由、信息获取和其他权利造成负面影响。

从 Llama 3 开始，我们投入了大量精力来解决 Llama 和 Meta AI 不当拒答的问题，并在 2024 年取得了显著进展。

审慎推进国际化布局

2024 年，我们又在 40 多个国家/地区推出了 Meta AI，并增加了对多种语言的支持，包括阿拉伯语、印尼语、菲律宾语、法语、德语、印地语、意大利语、葡萄牙语、西班牙语、泰语以及越南语。

在每个国家/地区推出 Meta AI 及支持每种语言之前，我们均会评估潜在的人权风险，并根据具体情况开展针对性的红队测试，这是减少大语言模型不安全行为的通用做法。

值得注意的是，并非每个提供 Meta AI 服务的国家/地区都在其当地法律中为表达自由提供了有力保障。在 2024 年的国际化工作中，我们制定了一套基于人权的方案，用于回应政府关于限制 Meta AI 输出内容的请求，这套方案以我们的长期政策为基础，并符合我们作为全球网络倡议成员的承诺及我们的企业人权政策。

与利益相关者合作

像 AI 这样技术飞速发展的领域，为利益相关者的合作参与带来了全新挑战。2024 年全年，我们都致力于向利益相关者同步最新信息，并广泛征集有意义的反馈。

具体行动包括：

我们在美国举行了 AI 主题圆桌会议，围绕产品和模型的发布事宜，征集跨学科专家群体的反馈，与会专家来自美国、巴西、布鲁塞尔、约旦、墨西哥以及非洲各地。

我们与斯坦福大学协商民主实验室 (Deliberative Democracy Lab) 合作，在美国、巴西、德国及西班牙举办了多场社群论坛，探讨生成式 AI 智能聊天助手应遵循的原则，并分享了从论坛中获得的相关发现。

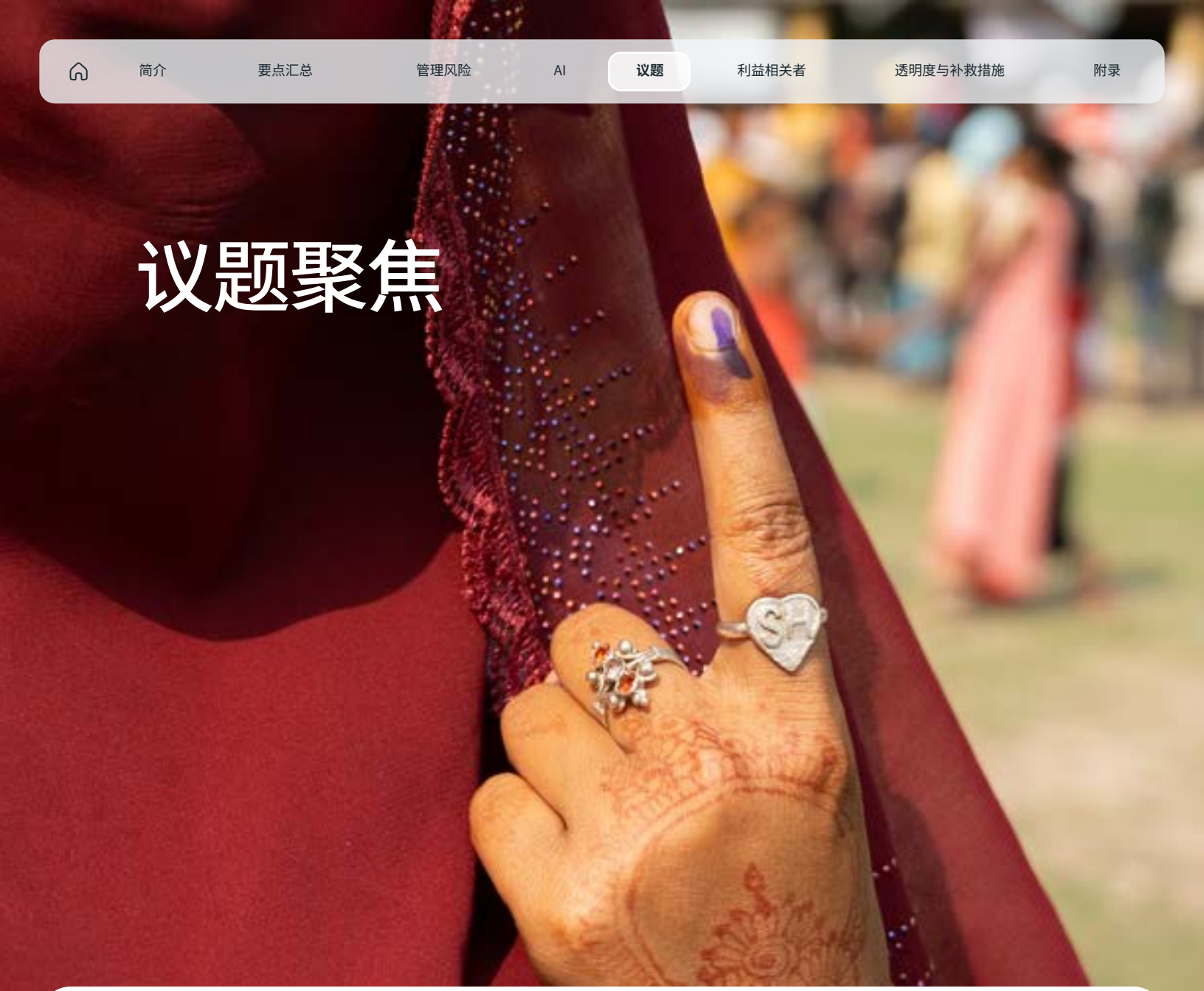
我们举行了一系列 Open Loop 研讨会，旨在探讨如何应对和利用开源 AI 带来的复杂挑战和机遇。在这些研讨会上，来自全球各地的政策制定者、行业领袖、学者以及民间社会代表齐聚一堂，共同制定有效且负责任的 AI 政策。

在第十三届联合国工商业与人权论坛于日内瓦举行之际，我们围绕生成式 AI 产品的人权尽职调查事宜，策划并主导了一场由多方利益相关者参加的互动模拟活动，并在活动中分享了我们的做法，深化了各方对相关风险与挑战的共识。

我们将再接再厉，继续推动 AI 创新，在此过程中，我们仍将致力于与全球的多元化利益相关者进行有效合作与磋商。

阅读更多信息

议题聚焦



2024 年：选举年

2024 年堪称历史上规模最大的选举年。全球超过 70 个国家/地区（覆盖世界半数以上人口）举行了全国/地区层面选举，大约有 20 亿人有资格投票。

Meta 深知保障表达自由权、投票权及公共事务参与权的重要性。选举相关工作是我们全年的重点任务。我们针对各类选举的规模、覆盖面及时间安排做好了充分准备，努力降低用户面临的相关风险，包括 AI 的日益普及带来的潜在风险。

在接下来几页，我们回顾了 2024 年在这方面的工作，并结合欧盟、印度、墨西哥及美国的选举相关情况，提供了说明性摘要。

针对选举开展大规模准备工作

过去几年里，Meta 不断优化调整为选举保驾护航的核心方案。我们在旗下服务覆盖的所有国家/地区部署这一方案，并根据当地需求和风险灵活调整我们的策略。为了做好准备，应对 2024 年选举，我们成立了专门的跨部门团队来协调全公司范围内的相关工作，团队成员包括来自我们情报、数据科学、产品与工程、研究、运营、内容、人权、公共政策及法务团队的专家。在这一年里，我们始终致力于保障人们的表达自由权、投票权和被选举权。

我们的方案涵盖以下具体工作：管理 AI 风险、执行我们与干扰选民投票相关的政策、打击恶意行为网络、保障政治广告的透明度，以及为选民提供获取可靠信息的途径。我们还针对举行选举的国家/地区，评估了分类技术和人工审核员的语言能力是否适配，以支持我们对违反政策的内容采取行动。我们的部分工作重点包括：

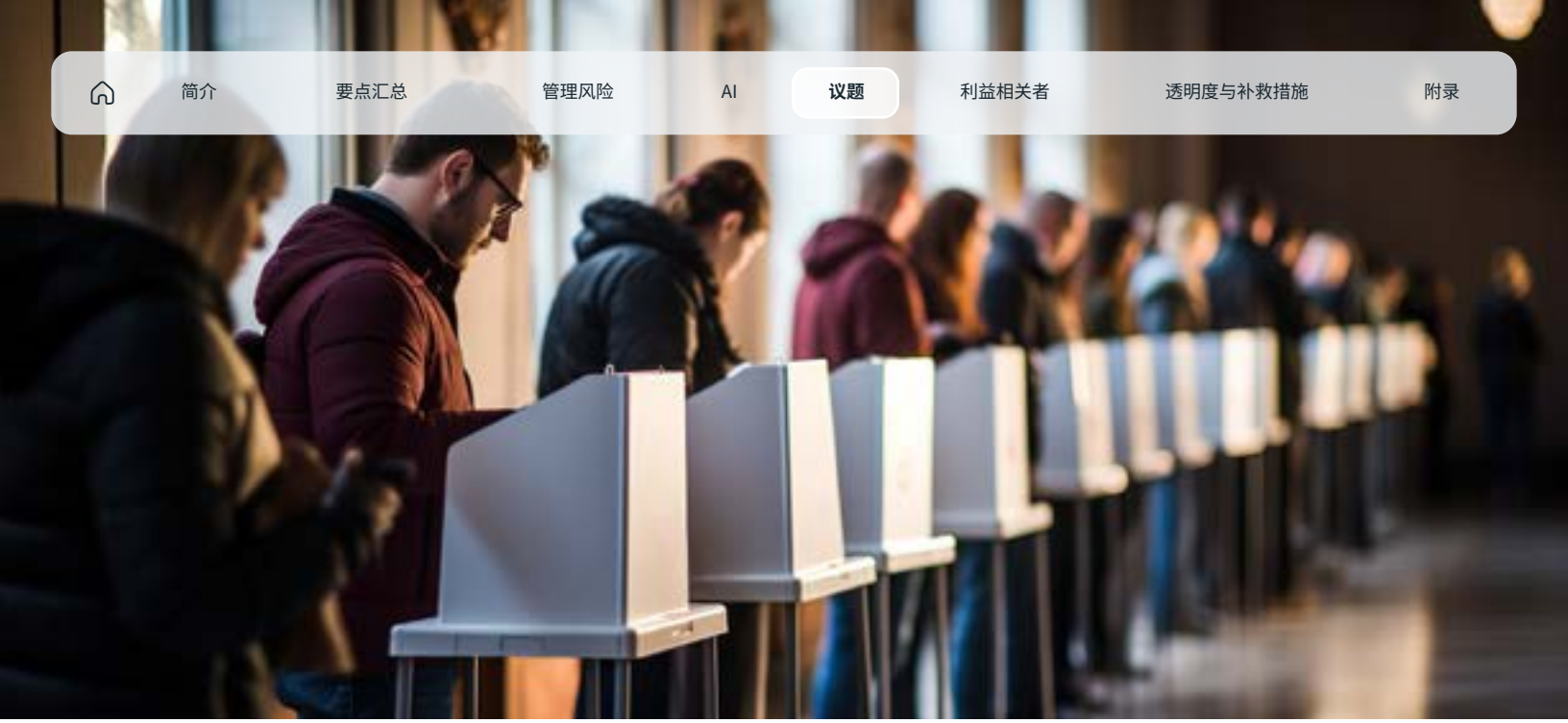
管理 AI 的干扰风险

年初，许多人担忧生成式 AI 可能会威胁选举的公平性，这包括深度伪造内容广泛传播的风险，以及利用 AI 进行虚假信息宣传的风险。对于恶意威胁和 AI 可能对选举造成的干扰风险，我们做好了充分的应对准备并进行了密切监测。根据我们在旗下服务中监测到的情况，这些风险并未大规模显现，其影响程度轻微且范围有限。例如，在一组重要选举进行期间，选举、政治和社会议题相关 AI 内容在经事实核查机构评定的所有错误信息中，占比不到 1%。据显示，我们现有的政策和流程足以降低生成式 AI 内容带来的风险。

我们全年的工作重点是制止各类影响力行动，并依托全球合作来帮助维护选举诚信。

我们密切监测了在利用虚假账户开展的合谋造假活动中，生成式 AI 的潜在违规使用情况。我们发现，使用生成式 AI，仅让这类活动在产出效率和内容生成方面获得了小幅提升。这种小幅提升并未影响我们打击此类影响力行动，因为我们在调查和取缔这类活动时，关注的是其行为本身，而不是其发布的内容（无论该内容是否由 AI 生成）。

我们还与业内同行合作，共同应对生成式 AI 的使用可能带来的威胁。例如，2024 年 2 月，我们与数十家行业领军企业共同签署了 AI Elections Accord（AI 选举协定），承诺携手防止具有欺骗性的 AI 内容干扰 2024 年全球选举进程。后续页面举例说明了特定国家/地区针对 AI 采取的举措。



其他选举诚信举措

除了减轻 AI 可能对选举造成的干扰风险外，我们还致力于为选民提供支持、防止境外势力合谋造假行为、加强候选人安全保障、建立合作关系，以及助力确保广告主信息公开透明。

为选民提供支持

在选举期间，获取可靠信息并负责任地使用网络平台尤为重要。在许多国家/地区，我们通过 Facebook 和 Instagram 的应用内通知，向用户提供选民须知信息与选举日提醒。借助这些功能，用户能够获取来自官方选举机构的权威信息，了解选举日的投票方式、地点与时间。例如，在巴西地方选举期间，Facebook 和 Instagram 上此类通知的互动量累计达到约 970 万次。超过 6,300 万 Facebook 用户和 1.18 亿 Instagram 用户看到了选民注册贴图，点击这些贴图即可跳转获取有关选举和投票的权威信息。

防止境外势力合谋造假行为

我们的安全团队调查和取缔了多个涉及造假账户、公共主页和小组的合谋造假网络。此外，据我们估计，我们的自动化虚假账户检测技术每天阻止了数百万个虚假账户的创建。我们的团队取缔了全球大约 20 起隐蔽的影响力行动，包括在中东、亚洲、欧洲和美国的这类行动。例如，在摩尔多瓦，我们在调查该地区的疑似合谋造假行为时，取缔了一个针对俄语受众的网络。

候选人安全

Meta 还为民选官员、候选人及其工作人员的账户提供了更强的防护，以抵御针对这些账户的入侵、冒充和骚扰行为。我们为候选人开展了多场安全培训，培训中介绍了可用于在我们平台上应对骚扰问题的指南。我们还发布了科普内容，以便这些内容能被所有选举参与者广泛获取。

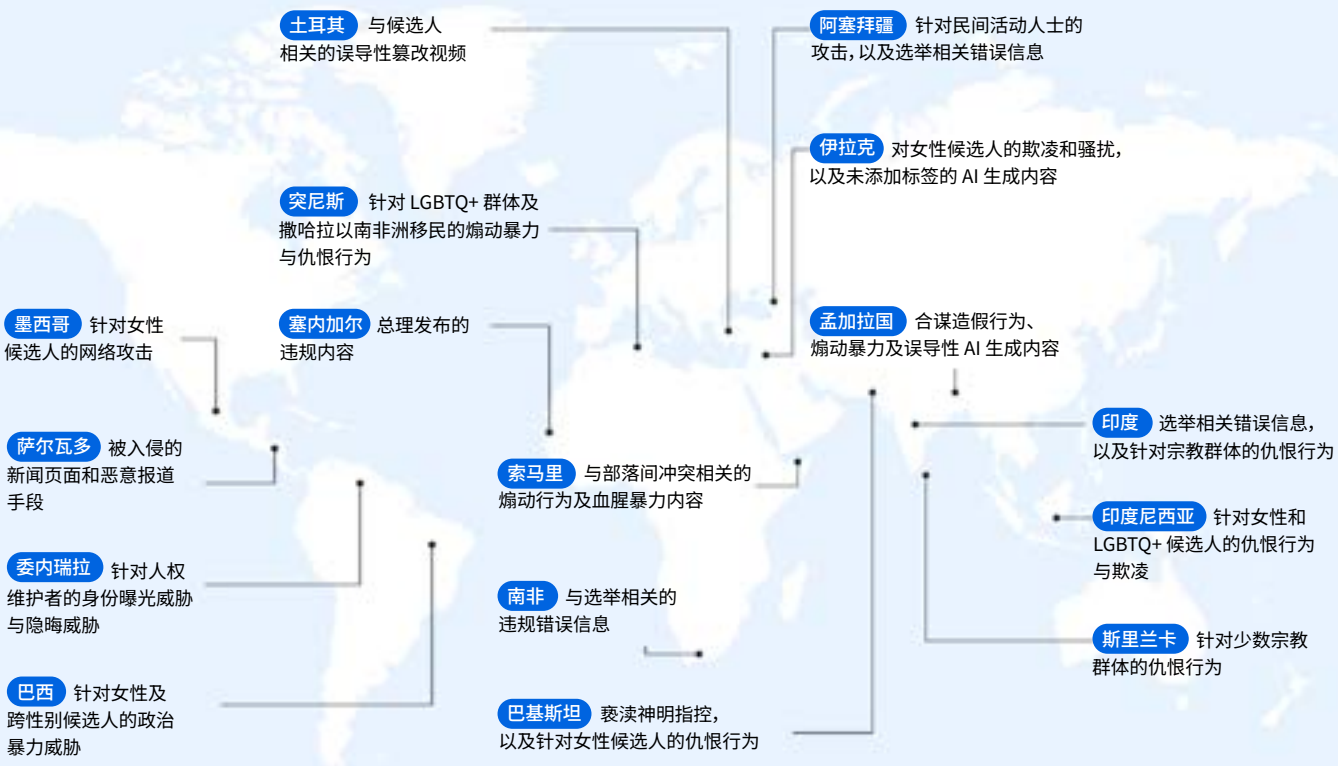
外联与合作

我们开展了外联工作，与政府部门及执法机构建立了沟通渠道，方便他们举报可能违反我们社群守则或当地法律的内容。我们还与民间社会团体、事实核查机构和其他科技公司合作，帮助我们发现和制止新兴威胁，以及虚假信息的传播。

广告信息公示

我们继续为社会议题、选举和政治类广告提供业内领先的信息公示力度。在我们提供上述广告的大多数市场，广告主必须完成授权流程并在其内容中添加“赞助方”标签，才能投放这类广告。该标签可能包含对这条广告负责的组织或个人的相关信息，尽管具体要求可能视国家/地区而异。这些广告随后会存入我们公开的广告资料库。2024 年，我们新增了规定，要求广告主在某些情况下，披露使用 AI 或其他数字技术制作或更改社会议题、选举或政治类广告的情况。

2024 年，可信合作伙伴为 25 个国家/地区的选举诚信工作提供了支持



为风险最高的选举做好准备

我们认为一些选举的风险较高，要做好其保障工作，需要进行更多准备、投入额外资源并专门定制方案。我们会考虑多种因素，例如选举类型、我们用户在该国家/地区人口中的占比、政治暴力风险、弱势群体受攻击的情况以及我们自身的运营能力。我们所做的额外工作包括设置专门的监测机制和临时风险应对措施，它们可以针对不同国家/地区和语言进行设计和应用。

我们在全球设立了多个选举运营中心，以便监测和迅速应对所出现的问题，包括高风险选举中的问题。您可以在网上详细了解我们为巴西、法国、印度、印度尼西亚、墨西哥、巴基斯坦、南非、英国、美国和欧洲议会选举开展的保障工作。

全国/地区层面选举案例

接下来几页选取了不同国家/地区的四个简要案例，帮助说明我们在 2024 年如何管理选举风险。在每种情形下，我们都至少提前一年便开始准备。

美国

为做好迎接美国选举的准备，我们的工作重点包括帮助用户获取可靠的选民须知信息，打击境外势力合谋造假行为，以及确保广告主信息公开透明。

选民须知信息	在 2024 年美国大选期间，Facebook 和 Instagram 动态版块顶部的提醒累计获得了超过 10 亿次展示。这些提醒包含选民登记、邮寄投票、提前到场投票及选举日投票的相关信息，点击次数累计超过 2,000 万次。用户点击提醒后，即可跳转至政府官方网站了解详情。
境外势力合谋造假行为	我们针对网上与选举相关的境外势力合谋造假行为做好了充分准备，包括进一步扩大对俄罗斯官方媒体机构的政策执行范围，同时我们继续打击规模最大、最顽固的隐蔽影响力行动之一，即代号为“Doppelganger”（幽灵分身）的行动。10 月至 11 月，“Doppelganger”行动针对美国发起了多次干预尝试，但其中大多数均在内容曝光前被我们主动拦截。
广告投放限制期	在大选竞选活动的最后一周，我们禁止投放新的政治、选举和社会议题类广告 — 这一举措自 2020 年起持续推行。此次实行此限制期的理由与往年一致：我们认识到，在选举的最后几天，如果允许投放广告的话，人们可能没有足够的时间对广告中的新说法提出质疑。

墨西哥

2024 年是墨西哥历史上规模最大的选举年，有大约 90,000 名候选人竞选超过 20,000 个公职。竞选活动期间的暴力活动也达到历史新高。至少有 37 名候选人遇害，记录在案的非致命袭击事件超过 828 起。竞选公职的女性人数创下墨西哥历届选举之最，而女性候选人遭受性别暴力和遇害的比例也居高不下。

我们采取了与其他高风险环境中类似的举措，并且受益于 Meta 当地专家的支持。在选举前和选举期间，我们移除违规内容的力度比平时更大。这些违规内容包括干扰选民投票或兜售选票的内容、仇恨内容，以及 Facebook 和 Instagram 上针对女性候选人的性别骚扰与暴力威胁。

为帮助防止选举中断并降低现实伤害风险，我们的工作围绕以下几方面展开：候选人安全、提供易用的选民须知信息，以及培养公众的媒体素养。

候选人安全	我们将逾 3,000 名候选人（包括所有联邦级候选人和州长候选人）纳入了交叉检查计划，以防止政策执行失误，同时还为相应候选人的账户开启了高级安全保护。这包括监测潜在的入侵威胁。我们与非营利组织和媒体机构合作，推出了名为“Vote Against Violence”（投出反暴力的一票）的宣传活动，旨在遏制网络上的性别暴力行为。此宣传活动在我们平台上触达了超过 120 万用户，且在其他渠道中得到了进一步传播。当局发现针对候选人的暴力行为或暴力威胁后，会向我们发送内容移除请求。
选民须知信息	我们与墨西哥国家选举机构（National Electoral Institute，简称 INE）合作，在 WhatsApp 推出了智能聊天助手“Inés”，为选民提供帮助。该智能聊天助手可回答与选举流程有关的问题，例如投票地点和方式、如何办理选民证，以及海外墨西哥选民的投票程序。在投票日，我们在 Facebook 和 Instagram 发送了提醒，并在这两款应用中推出了投票主题贴图，鼓励人们投票。
媒体素养	为帮助防止虚假信息的传播，我们与 INE 和民间社会组织 Movilizatorio 合作，推出了“Soy Digital”（“We Think Digital”）宣传活动。该宣传活动提供了通俗易懂的学习模块与资源，帮助人们树立数字公民意识、提升信息素养能力，包括了解如何保持网络安全。该宣传活动触达了超过 1,500 万人。此外，我们还培训了 300 名选区级别的负责人，随后他们又对数千名投票站工作人员进行了媒体素养培训。

印度

为迎接 2024 年印度大选，Meta 提前 18 个月开始做准备。我们的工作重点是确保平台诚信并推动选民教育。我们采取灵活的方案，能够为长达 60 天的选举期提供持续支持，在此期间，投出的选票超过 6.4 亿张。我们的准备工作包括：

选民教育与意识提升	印度选举委员会 (Election Commission of India) 通过其 Facebook 公共主页发起了投票提醒通知，触达了 1.45 亿用户。印度选举委员会还部署了 WhatsApp 应用程序编程接口 (API)，用于开展投票提醒宣传活动，其消息触达了大约 4 亿用户。
确保平台诚信	我们采取了措施，防止平台被违规使用。内容审核员对 Facebook、Instagram 和 Threads 内容进行了审核，涵盖超过 20 种印度本土语言的内容和英语内容。我们移除了虚假账户，并履行了我们在 2019 年与其他社交媒体公司共同签署的《自愿道德准则》下的承诺。
打击错误信息	我们与跨行业的错误信息打击联盟（Misinformation Combat Alliance，简称 MCA）合作，在 WhatsApp 上推出了专门的事实核查求助热线，以打击 AI 生成的虚假信息。我们与 MCA 携手推出了这一 WhatsApp 求助热线，依托此热线，MCA 设立了全球首个 Deepfakes Analysis Unit（深度伪造分析小组），用于评估用户怀疑可能是深度伪造内容的任何音频或视频内容。我们还与 MCA 合作，为数百家执法机构提供了打击深度伪造的培训。

欧洲议会选举

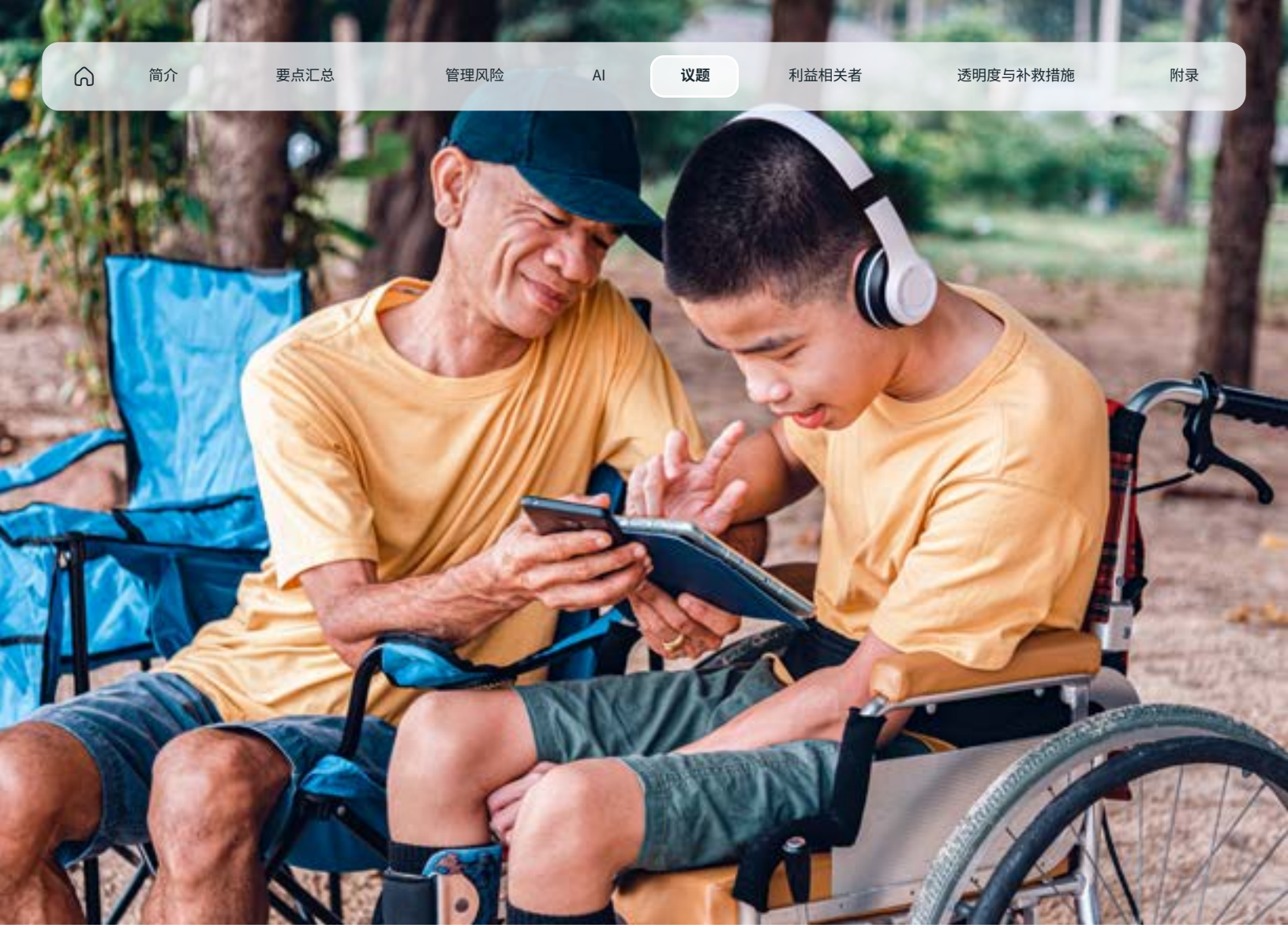
Meta 在为欧洲议会选举做准备时，借鉴了从此前全球各地选举中学到的重要经验，同时遵循了《数字服务法》确立的监管框架，并履行了我们在欧盟 (EU)《反虚假信息行为准则》下的承诺。

针对欧盟地区，我们的选举保障措施聚焦于以下方面：

普及选举信息，推动公民参与 	我们通过应用内的“选民须知信息单元”和“选举日信息”，为用户提供可靠的选举信息，并引导他们查阅选举流程相关信息。在 Facebook 和 Instagram，用户与这些通知的互动次数分别超过了 4,100 万次和 5,800 万次。
打击影响力行动 	在打击合谋造假行为的工作中，我们重点关注与欧洲议会选举相关的具体威胁。我们封禁了多个以欧盟为目标的网络，这包括多次对源自俄罗斯的网络“Doppelganger”采取处置措施。
打击错误信息 	我们与欧洲事实核查标准网络合作，共同打击由 AI 生成和经数字手段编辑过的影音内容，并携手开展媒体素养宣传活动，以提升公众对相关风险的认知。
应对与生成式 AI 技术滥用相关的风险 	得益于我们为治理生成式 AI 内容所采取的政策和措施，欧洲议会选举期间及前后，在欧盟地区的 Facebook 和 Instagram 上，有近 6,000 条社会议题、选举或政治类广告以及超过 570 万条内容都添加了 AI 相关披露标签，这进一步提高了信息透明度。

详情请参阅我们的政策及信息公示平台。

前往政策及信息公示平台



儿童和青少年安全

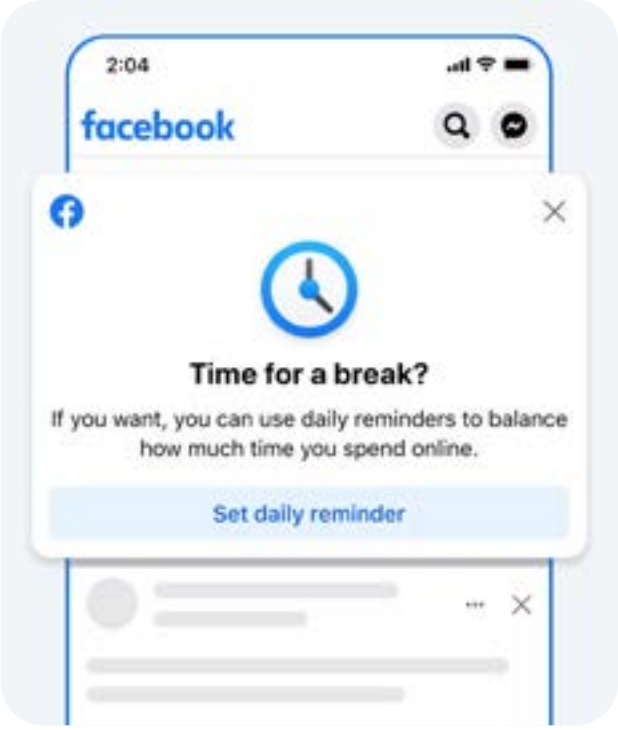
保护青少年儿童的网络安全是 Meta 的首要任务。我们为青少年及其家长提供内置保护功能以及各种工具，帮助保护青少年在我们应用和服务中的安全。

内置青少年保护功能

保护青少年的网络安全，需要全球多方利益相关者携手合作，其中包括家长、儿童问题专家、学者、业界同行、政府、民间社会组织等各界力量。我们始终致力于协助保护青少年的安全，同时为他们创造空间，让他们能在家长的指引下，自由表达自我并获取各类信息。

多年来，我们开发了超过 50 项工具与资源，用于为青少年及其家长和监护人提供支持，同时，我们花费了十多年来制定政策并开发技术，以治理违反我们规定的内容和行为。

2024 年，我们更新了政策和产品设计，为青少年打造差异更明显的独特体验。在我们儿童最大利益框架的指导下，这些更新继续帮助青少年看到适龄内容。在现有 Instagram 保护功能的基础上，我们又在美国、英国、加拿大和澳大利亚推出了 Instagram 青少年账户，并计划后续向全球推广。重新设计的青少年账户内置多种保护功能，可限制谁能与青少年联系以及青少年能看到哪些内容，还能帮助管理青少年使用该应用的时长。这些更新还提供了新途径，供青少年在家长指引下探索兴趣爱好。这项新的 Instagram 青少年账户体验符合专家的指导建议，以及《联合国儿童权利公约》中关于符合儿童不同阶段接受能力的原则。



我们开发了家长监护面板并在全球推出，供使用我们监护工具的家长和监护人一站式查看和管理孩子的账户。借助此面板，家长和监护人可以通过自己的账户设置控制措施，以查看和管理孩子接触到的任何扰人联系或不当内容，并限制孩子的屏幕使用时间。

我们还围绕智慧用屏计划，在美国开展了一系列研讨会，帮助家长了解如何与家人展开有关安全使用电子设备的对话，同时让家长进一步了解 Meta 提供的家长监护工具，以便能为孩子设置最合适的界限和保护措施。

“Meta 新推出的 Instagram 青少年账户意义重大，既能赋予家长指引青少年的能力，又不会剥夺大龄青少年的自主性。在青少年保护的道路上，这些新设置加上增强的安全隐私工具与建议，共同构成了向前迈出的一大步。”

— ConnectSafely 首席执行官 Larry Magid

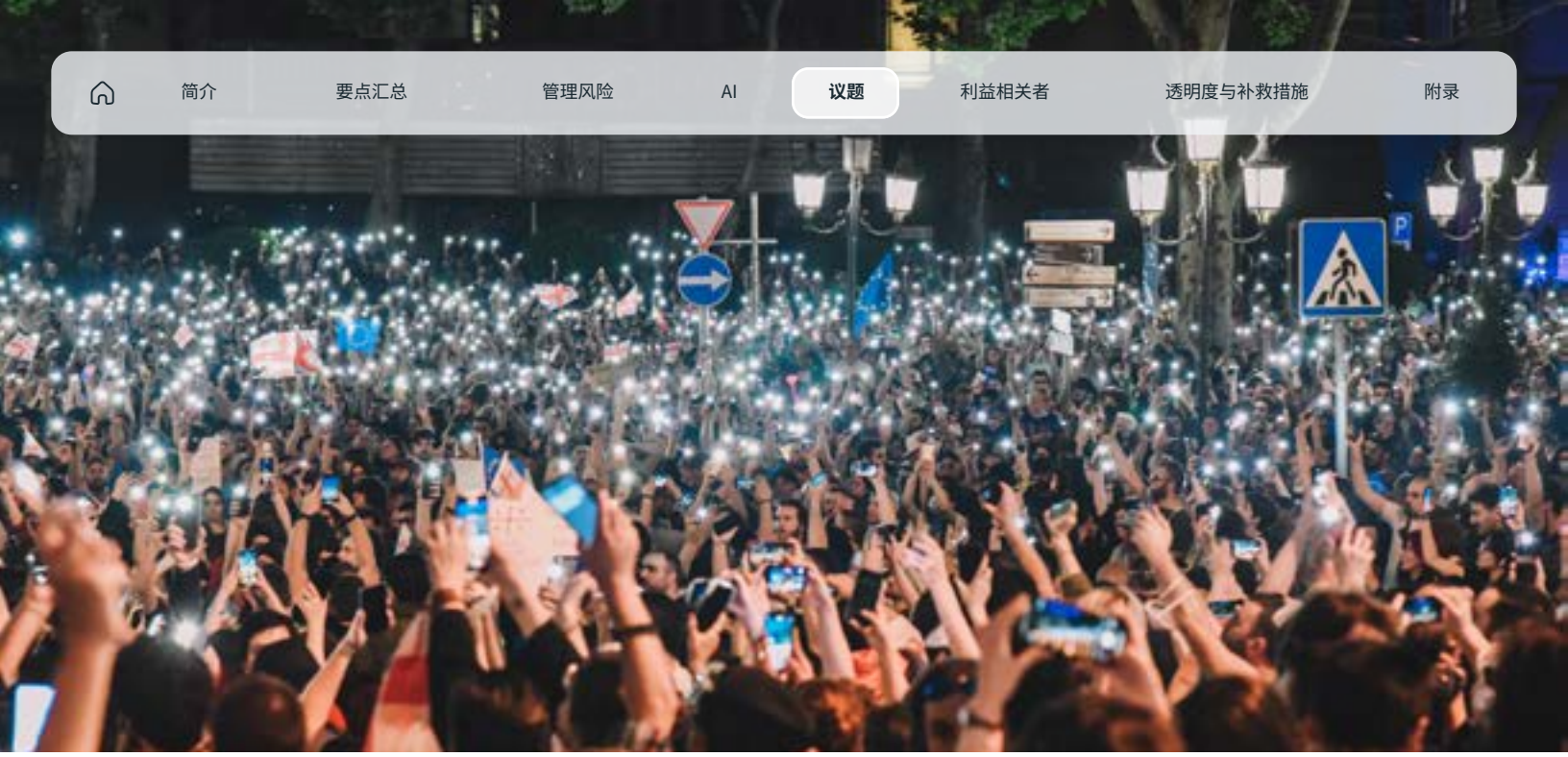
打击性勒索行为

帮助保护青少年儿童免受恶意用户的伤害，始终是 Meta 的首要任务。2024 年，我们继续与美国国家失踪与受虐儿童援助中心 (NCMEC) 合作，将 Take It Down 计划扩展到更多国家/地区，并使其支持更多语言，让更多青少年能够重新掌控自己的私密图像。我们开发了新工具来防范性勒索，并让潜在诈骗分子和犯罪分子更难找到青少年并与之互动。我们还发起了一项教育宣传活动，借鉴 NCMEC 和 Thorn 的专业见解，帮助青少年识别性勒索骗局，也为家长提供方法，支持孩子规避此类骗局。该宣传活动会引导青少年和家长访问这方面的专家建议，这些建议由 Thorn 制定并经过 Meta 的调整，可以为任何需要获取性勒索相关支持和信息的人提供帮助。

我们是 Lantern 计划的创始成员，这是由技术联盟 (Tech Coalition) 运营的一项计划，支持我们与其他科技公司共享信号，识别违反成员公司儿童安全政策的账户和行为。我们向 Lantern 计划共享了性勒索的专属信号，以便能依托行业内的这一重要合作，努力遏制各平台上的性勒索骗局。2024 年，参与 Lantern 计划的公司数量增长至两倍，总数达到 26 家。

请点击此处查看完整列表，了解我们为了支持青少年和家长而提供的工具、功能和资源。

↗ 阅读更多信息



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。2024 年，我们在多个国家和地区发起并协调了危机应对行动，这包括孟加拉国、格鲁吉亚、肯尼亚、新喀里多尼亚、尼日利亚、韩国、英国、委内瑞拉等。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。

在危机期间，危机政策协议是我们使用的一种关键工具。该协议指引我们迅速采用以下层面的措施来缓解潜在危害：

- 政策层面，例如向审核员提供额外指导意见。具体示例：针对违反我们暴力和血腥内容政策的某些行为，提供暂缓违规记分的指导意见，以免过度惩罚或限制那些试图让公众了解冲突影响的用户。
- 产品层面，例如调整产品体验。具体示例：调整产品设置，仅允许好友和家人评论帖子。
- 人力层面，包括调配资源来重点处理特定问题。

借助危机政策协议，我们可以对有可能引发平台内风险的局势开展线下评估。一旦认定了危机局势，我们会开展评估来识别平台内风险，并确定是否需要采取任何额外措施。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。

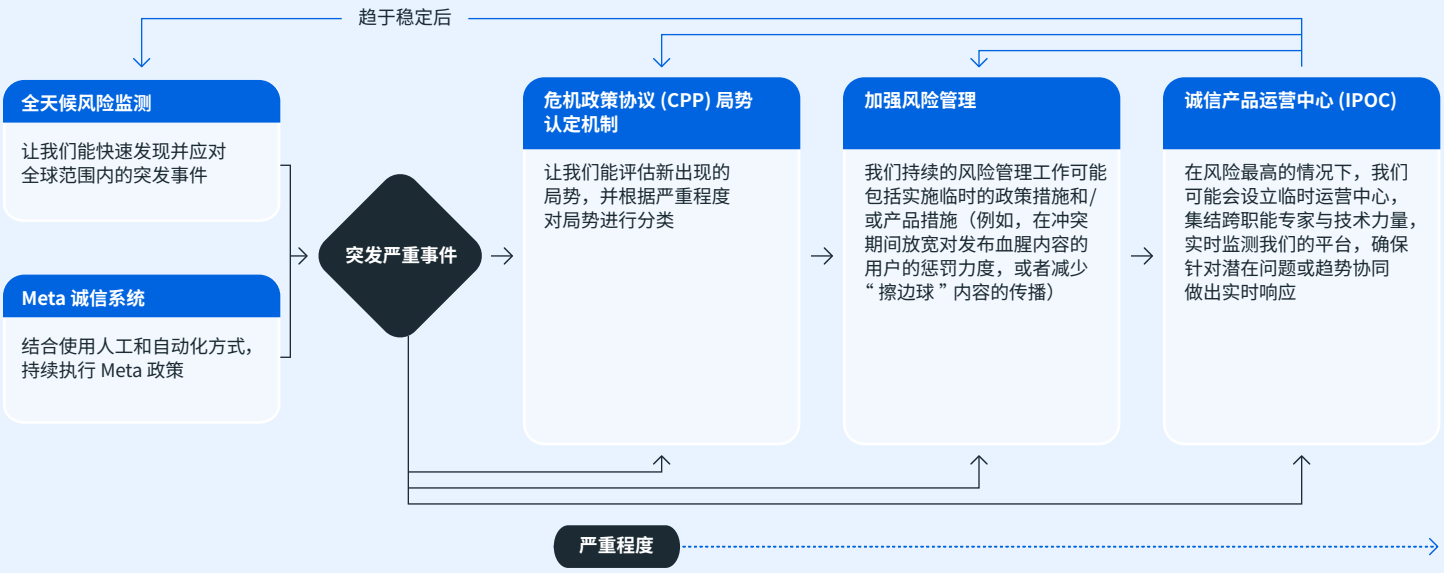
我们将在下一页阐述我们防范与应对危机和冲突的方法。我们还会提供案例，展示我们如何使用危机政策协议，这些案例也体现了我们的工作所涉及的广泛地域。

防范与应对危机和冲突[4]

不论是危机政策协议，还是我们针对有风险的国家/地区开展的工作，都是我们用于预防、发现和缓解风险的关键工具。我们的产品、政策和运营团队会评估不断变化的实地情况，为制定有效、相称的应对措施提供依据。

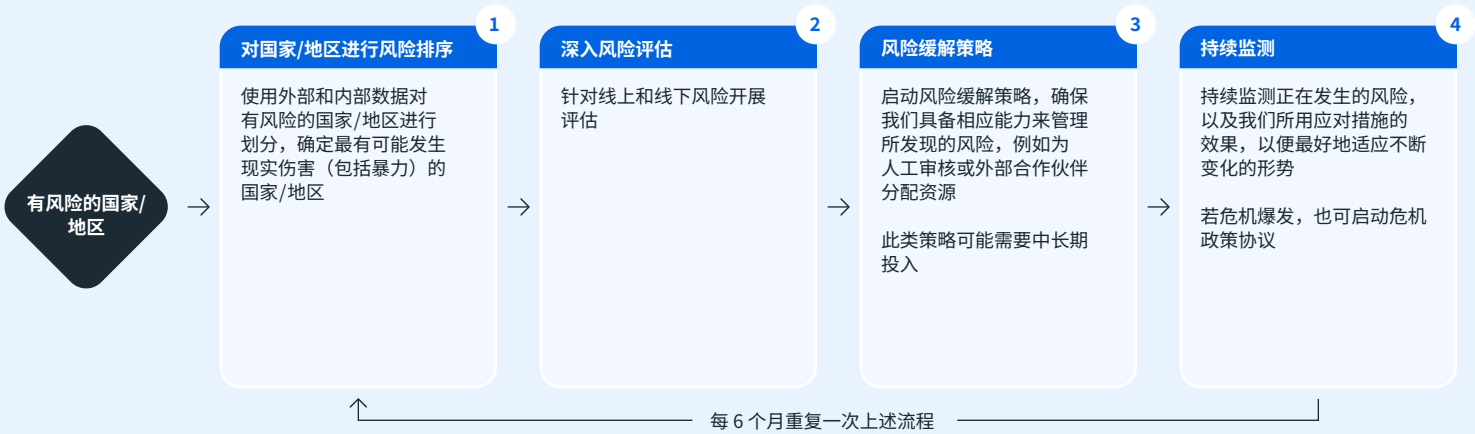
应急响应

我们如何快速应对突发严重事件？



长期措施

我们如何采取长期措施缓解冲突风险？



[4] 我们的危机应对工作涵盖全球范围内的许多情况，包括冲突、族群内部暴力、内乱、大规模抗议、环境灾害，以及恐怖袭击或其他犯罪袭击等。

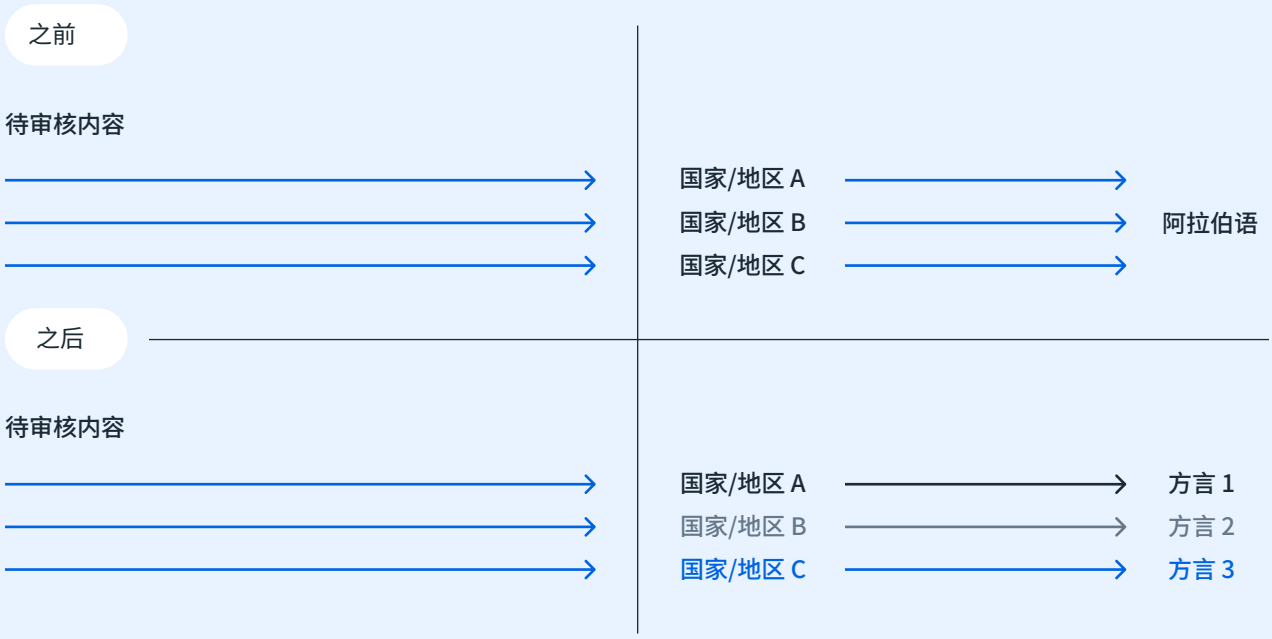
苏丹

2024 年，苏丹武装部队 (SAF) 与快速支援部队 (RSF) 之间的冲突进一步升级，加剧了该国的动荡局势与人道主义危机。在苏丹，违规内容的数量较冲突前有所增长，且在全年居于高位，这包括涉及暴力与煽动暴力、配合实施伤害、剥削以及危险组织和人物的内容。

为减少违规内容的传播，我们以危机政策协议为指导，立足于我们 2023 年采取的措施来进一步加强应对。随着冲突持续，我们实施了临时措施并制定了长期缓解方案，以应对违规内容数量居高不下的风险。

其中一项长期方案是设计、开发并上线了一套系统，该系统可识别特定的阿拉伯语方言，并将相关内容优先分派给更可能精通该方言且了解当地背景的审核员。此前的系统将阿拉伯语视为单一语言，并将相关内容分派给负责审核阿拉伯语内容的审核员。新系统则可识别具体使用的阿拉伯语方言，并将相关内容分派给最可能理解该方言的审核员。对苏丹地区而言，这一改变使数量更多的内容能得到更精确的审核，进而减少政策执行中的失误。此项工作参考并借鉴了以色列和巴勒斯坦人权尽职调查的相关成果。

按方言分派审核任务



2024 年，冲突双方都越来越多地在网上曝光战俘的身份信息。这种身份曝光行为不仅增加了战俘遭受现实伤害的风险，还违背了《关于战俘待遇的日内瓦公约》中关于保护战俘尊严与安全的规定。根据监督委员会于 2023 年提出的一项建议（监督委员会于 2024 年在“苏丹快速支援部队俘虏视频”案中重申了此建议），我们也认识到，部分与战俘相关的内容可能具有公共利益价值，例如提升公众对潜在人权侵犯行为的认知，或是帮助找到失踪的战俘。因此，在有关配合实施伤害和宣扬犯罪行为的政策中，Meta 向内容审核员提供了有关战俘内容的指南，以便他们能更好地批量处理该地区的潜在违规内容。

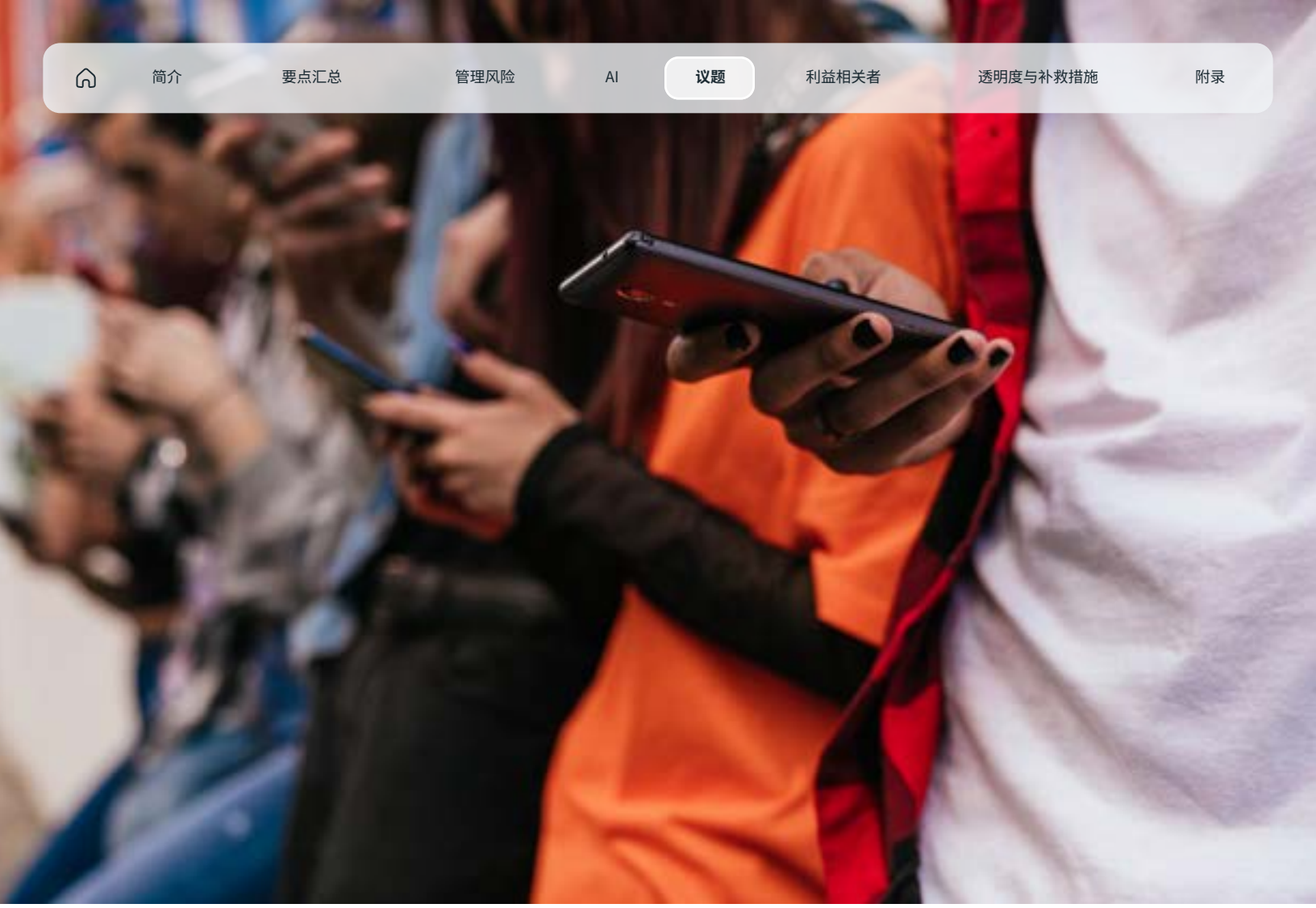
阅读 2024 年上半年监督委员会报告

查看监督委员会案件

可信合作伙伴发挥了关键作用：他们针对当地局势发展以及与冲突相关的潜在违规内容，提供了至关重要的见解。这些见解不仅有助于我们执行相关 Meta 政策，包括仇恨行为政策、欺凌和骚扰政策及剥削政策，还帮助我们在错误信息和伤害政策中，认定了经过预先审核的潜在有害言论，从而最终为营造更安全的网络环境做出了贡献。根据我们的剥削政策，我们得以识别出与童兵图像和童兵招募相关的潜在风险，并移除这类内容并减少其传播。

我们还面向人权维护者、新闻工作者、本土组织和侨民组织开展了培训，帮助他们向包括移民和难民在内的苏丹用户提供指导。这些培训重点关注内容政策、数字安全以及如何提升他们在 Meta 旗下平台上的影响力。

武装冲突往往会引发大规模的流离失所，因此在苏丹，我们重点识别潜在的剥削行为，包括人口偷运与贩卖、对妇女和女童的性剥削，以及强迫婚姻等。我们移除了超过 19,100 篇提供人口偷运服务的小组帖子，并移除了美化强迫婚姻的内容。与侨民群体保持交流合作仍是一项重要策略，这有助于我们掌握内容趋势，还能为国家层面的监测工作提供清晰指引。为此，我们开展了 30 多场专项交流活动，为平台用户的安全保驾护航，这包括探讨如何识别与仇恨言论和人口偷运相关的新兴字词和短语。这些见解使我们能够更有效地发现并处理违反我们政策的内容。



中东

中东冲突仍是 Meta 的关注重点。2024 年，我们重点关注由以色列与加沙之间的持续暴力事件引发的风险，当时战火蔓延到了整个地区，而且该地区内的其他势力进一步介入，导致冲突升级。我们一方面努力确保用户在我们平台上的表达自由，另一方面也致力于防止煽动恐怖主义、暴力行为及其他现实危害的内容在平台上传播。

我们的核心策略与 2023 年保持一致，主要包括：根据危险组织和人物 (DOI) 政策，继续将 2023 年 10 月 7 日哈马斯发动的袭击事件认定为恐怖袭击，并按照我们的政策处理违规内容。同时，我们终止了 2023 年实施的临时产品调整。

在 10 月 7 日的恐怖袭击发生后，Meta 立即将这一暴力事件及随后引发的冲突认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，并采取临时的产品措施和政策措施。我们参考了《联合国工商企业与人权指导原则》（这是我们企业人权政策的基石），并结合 2022 年的尽职调查工作，来指导我们的应对策略。有关此次危机应对的详情，可参阅 2023 年度人权报告和新闻中心帖子。

2024 年全年，我们持续与以色列、中东阿拉伯国家乃至全球的政府、民间社会组织和其他相关方沟通合作，以展现我们在相关事务上的透明度与积极响应态度。我们还对多起监督委员会案件做出了回应。

此外，我们继续实施 2022 年人权尽职调查报告中提出的建议。我们报告了 2023 年 6 月 30 日至 2024 年 6 月 30 日期间 Meta 的工作进展，包括针对希伯来语投入更多内容审核资源，以及更新我们的危险组织和人物政策，允许在某些情况下，进行更多有关社会和政治话题的讨论。此前有反馈指出，我们的危险组织和人物政策经常会波及新闻报道、对时事的中立讨论，甚至是对恐怖组织和仇恨团体的谴责等内容，此次更新就是为了回应上述反馈。值得注意的是，任何赞扬或支持危险组织或人物、其暴力行为或使命的内容，仍然禁止发布。

在 2024 年 6 月 30 日至 2025 年 6 月 30 日期间，我们上线了一套系统，该系统可检测阿拉伯语方言，并优先将内容分派给最有可能理解该特定阿拉伯语方言的审核员。我们还优化了可信合作伙伴上报渠道，从而提升了对上报问题的快速响应能力。如要详细了解我们在此期间的工作进展，可参阅 2025 年 12 月最终更新：以色列和巴勒斯坦人权尽职调查。

↗ 2023 年度工作进展　↗ 2024 年度工作进展

孟加拉国

我们积极为 2024 年 1 月的选举做准备，这些准备工作帮助我们在报告期内预见并应对了多重风险。我们的目标是在保障用户安全的同时，支持他们行使投票权并表达自我。通过这些选举相关的准备工作，我们得以在年中针对学生抗议活动、暴力镇压事件以及随后的政府更迭做出响应。

鉴于动荡局势的严重性，我们启动了危机政策协议。我们主动识别各类风险，包括仇恨言论、针对少数宗教群体的煽动行为、错误信息以及合谋造假行为。我们采取了一系列风险缓解措施，包括实施临时高风险地区政策、与可信合作伙伴和第三方核查网络合作，以及对人权维护者的账户启用增强的保护机制。

我们采取的其他措施包括：

- 建立精确的检测信号，以便识别可实时处理的违规内容激增情况，例如血腥暴力和仇恨行为相关内容。
- 运用包括 AI 检测在内的工具与技术，来识别和处理违规内容及关键词搜索。
- 在错误信息和伤害政策中，认定了更多经过预先审核的潜在有害言论。

对于与抗议活动相关的内容，若政府提出的内容移除请求不符合国际人权标准，我们未予执行。这种做法符合我们作为全球网络倡议成员的承诺及我们的企业人权政策。

格鲁吉亚

2024 年，我们两度为格鲁吉亚启动了危机政策协议。2024 年 3 月，在一系列反对《境外影响力透明法》(Law on Transparency of Foreign Influence) 提案的大规模示威活动发生后，我们首度启动该协议。2024 年 12 月，在全国选举结束后，当地再次发生大规模示威活动，同时警方和其他安全部队的暴力行为也不断升级，我们随即再度启动该协议。

启动危机政策协议后，我们的团队得以加强风险缓解工作，处理违规内容激增与人身暴力风险增加的问题，并帮助保护人权维护者的安全。我们审核了诋毁字词列表（历来被用于攻击特定群体的字词），以识别并管理我们平台上的仇恨内容。我们移除了旨在操纵公众舆论或传播潜在有害内容的虚假账户。此外，我们还捣毁了一个以格鲁吉亚为目标的合谋造假行为 (CIB) 网络，并移除了其他造假账户。

在整个危机期间，我们与民间社会组织、事实核查机构和可信合作伙伴开展了合作，他们协助我们掌握事态发展情况，并促进了我们与格鲁吉亚更广泛的民间社会和反对派之间的信息交流。围绕以反对派群体为目标的违规内容，可信合作伙伴提供了关键信号与见解。我们还与可信合作伙伴沟通，帮助他们更好地理解哪些内容在我们社群守则下构成违规。此外，我们联系了民间社会合作伙伴，以识别面临风险的人权维护者，帮助确保他们的账户受到增强的保护。

网络安全

我们的安全政策对于保障用户的表达自由权、信息获取权和隐私权等权利至关重要。我们继续在全公司范围内开展协同工作，以识别和抵御针对平台的恶意威胁，包括影响力行动、网络间谍活动、监视行为，以及欺诈和诈骗活动。我们安全工作中的重要任务之一，就是打击从事恶意活动的恶意行为网络。

2024 年，在中东、亚洲、欧洲和美国，我们取缔了 20 个合谋造假行为网络，原因是它们违反了我们的合谋造假行为政策。这类网络通过使用虚假账户或误导性手段，试图操纵公众舆论，从而达成特定战略目标。我们采取的措施包括：持续监测并打击曾取缔的网络试图在我们平台上卷土重来的行为；通过威胁报告公开分享信息；将调查所得见解融入到我们的检测系统和产品设计中，进一步提升其抗风险能力。

我们继续检测并取缔针对和/或伪装成特定族群或宗教群体的合谋造假行为网络。2024 年的众多案例中，有一例是源自孟加拉国的网络，它因合谋造假行为被我们取缔。该网络以孟加拉国本土受众为目标，使用虚假账户来发布内容和管理公共主页。它以虚构的新闻机构身份作为伪装，并盗用现有新闻组织的名称，传播反对孟加拉国民族主义党的内容，同时支持执政党。该网络的行动与孟加拉国一家非营利组织及人民联盟党相关人员存在关联。

另一例涉及源自中国的网络，该网络以全球锡克教社群为目标，利用被盗账户和虚假账户伪装成锡克教徒，鼓吹名为“Operation K”（K 行动）的虚构行动主义运动，号召在新西兰和澳大利亚等地举行支持锡克教的抗议活动。该网络利用 AI 生成的图片和帖子来发布英语和印地语内容，具体涉及旁遮普地区洪灾、全球锡克教社群、卡利斯坦独立运动、哈迪普・辛格・尼贾尔遇刺事件，以及对印度政府的批评。

阅读更多信息

在打击间谍软件公司的政策执行工作中，我们终止并取缔了 Paragon Solutions 的活动，这是一家间谍软件供应商，以 WhatsApp 上的众多用户为目标，其中包括新闻工作者和民间社会成员。我们联系了可能受影响的 WhatsApp 用户，并为他们提供了与自我保护相关的资源。我们还向这些用户提供了多伦多大学公民实验室的相关信息，该实验室可为民间社会成员提供更多资源。2024 年，我们成为《帕尔摩备忘录》(Pall Mall Memorandum) 的创始签署方之一，这是一项旨在遏制间谍软件滥用的多国合作倡议。

2024 年 12 月，美国一名联邦法官裁定 NSO 集团负有法律责任，原因是该集团违反了美国州法与联邦法律，且违反了 WhatsApp 服务条款。这是首次有间谍软件公司依据美国法律被判负有责任。2019 年，Meta 和 WhatsApp 对 NSO 集团提起本案诉讼，背景是该集团曾在未经授权的情况下访问 WhatsApp 服务器，以便在 1,400 多名 WhatsApp 用户的移动设备上安装“飞马”(Pegasus) 间谍软件。这些用户包括新闻工作者、人权活动人士和政治异见人士等。

20

取缔合谋造假行为网络

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。

2024 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负责任人工智能 (AI) 和诚信选举方面的方案，以及我们在认定危险组织和人物以及暴力事件时采用的信号。

例如，为了评估我们与“Zionist”（犹太复国主义者）一词相关的政策是否恰当，我们与全球来自民间社会和学术界的 145 位利益相关者进行了磋商。与会者包括政治科学家、历史学家、法学学者、数字权利和民权组织、表达自由倡导者和人权专家。我们还与其他利益相关者互动，包括来自我们可信合作伙伴计划的非政府组织，以及代表不同观点的各类侨民群体。

2024 年，我们与撒哈拉以南非洲、中东和北非地区的当地民间社会组织联合成立了发声与表达工作组，以了解这些组织对沙特阿拉伯、约旦、尼日利亚和塞内加尔等国立法提案的关切。在这些会议中，我们探讨了如何在做到以下两点的前提下，保障用户对我们平台的访问：一是遵守当地法律的内容限制要求，二是履行我们在全球网络倡议下关于捍卫表达自由和用户隐私的承诺。

我们还试点推出了“人权委员会合作计划”，邀请埃塞俄比亚、加纳、肯尼亚、尼日利亚和南非的国家人权机构参与，重点探讨 Meta 如何处理潜在有害内容以及配合网络内容监管。

此外，我们在埃塞俄比亚、巴勒斯坦、索马里、苏丹和突尼斯举行了冲突应对研讨会。对于举行选举的国家/地区，我们会为当地的人权维护者和新闻工作者提供培训，帮助他们掌握保护数字身份的工具。

通过我们的 Open Loop India 计划和 Open Loop Sprint 工作，我们与其他公司、政策制定者和 AI 专家合作，围绕利益相关者参与在 AI 生命周期和价值链中的作用提出了见解。

我们在利益相关者参与方面的做法

集思广益，广泛汲取不同观点和专业看法：与全球各地的主题专家交流，发掘重要见解，获得多元化的全球视角与本地洞见。

公开透明：与外部利益相关者深入探讨相关挑战及改进措施。

建立反馈闭环：展示我们的政策如何随时间发展变化。

建立信任：让各方认可我们政策及其执行的正当性。

464

来自 **34** 个国家/地区的 464 位利益相关者为 **6** 次政策交流委员会专题会议做出了贡献。

121

121 位利益相关者为 Meta 的其他政策制定工作做出了贡献。

100+

100 多位利益相关者参与了选举相关工作通报会，我们还制作了 **7** 期选举专题简报。

290+

超过 290 名新闻工作者、人权维护者和活动人士接受了培训。

Meta 政策制定闭环流程

持续审核

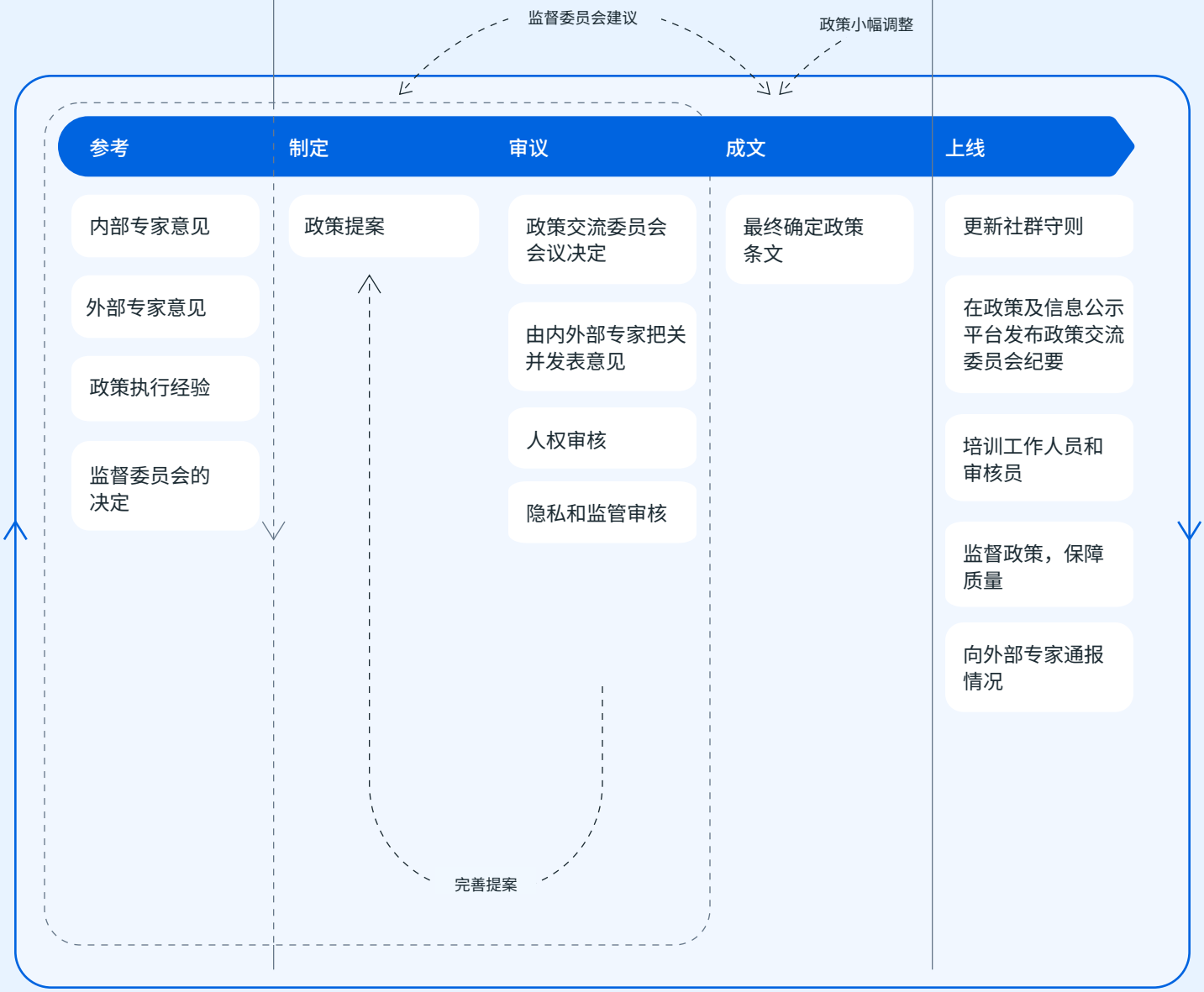
我们参考从各种来源获取的经验意见，持续审核我们的政策。

制定

提案的制定需遵循严格的流程，确保提案符合原则、具备可操作性且清晰易懂。

上线

更新政策执行系统，让政策在我们的服务中“上线”。



政策交流委员会会议

我们致力于制定尊重人权、包容多元化视角的政策，让多种观点和信仰都能被倾听和体现。Meta 的政策交流委员会会议是定期举办的会议，主题专家会在会上讨论社群守则和广告发布守则有可能做出的调整。这些会议涉及提出新政策或修订现有政策，并遵循一套政策制定流程。该流程包括与全球利益相关者开展广泛交流合作，以及对内外部研究成果进行评估。

我们在 2024 年举行了六次政策交流委员会会议，分别涉及以下主题：

1. 将“Zionist”（犹太复国主义者）作为指代词来实施仇恨行为
2. 暴力犯罪事件
3. 可能带来健康与安全风险的商业内容
4. 移除敏感图像
5. 饮食失调相关内容
6. 哀悼已认定危险人物的内容

社群论坛

Meta 的社群论坛植根于协商式治理理念，旨在针对那些存在权衡取舍、尚无定论的议题收集公众意见。这种做法让公司外部的意见能在我们的决策流程中拥有更大话语权，并让我们能够预见公众舆论未来的演变趋势。

2024 年，Meta 与斯坦福大学协商民主实验室合作举办了一场社群论坛，重点探讨用户希望 AI 智能体开发遵循哪些原则。该论坛吸引了来自印度、尼日利亚、沙特阿拉伯、南非和土耳其的约 1,000 名参与者。点击此处可获取详细报告。

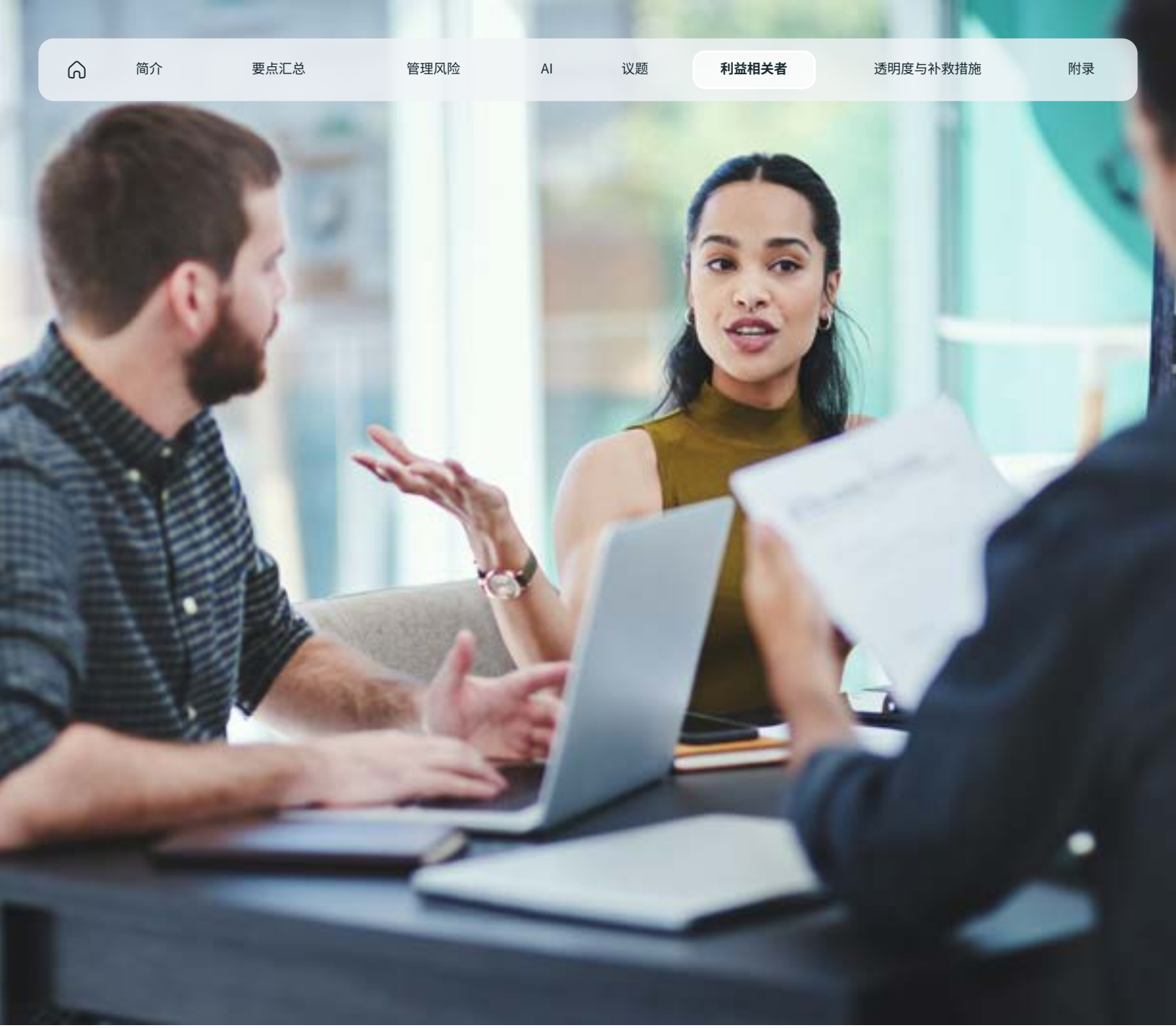
在论坛举办期间，参与者不仅能直接听取主题专家的见解，还可相互交流探讨，并向 Meta 提出宝贵反馈。借助这种协商式方法，参与者可以深入探讨个性化体验中存在的固有矛盾，权衡个性化体验的价值与数据收集和存储等取舍因素。

我们在用户控制权和个性化体验方面的做法

对于 AI 智能体，从论坛中获得的相关发现为我们在用户控制权和个性化体验方面的做法提供了依据。这些发现包括：

- 参与者支持 AI 智能体通过记住之前的对话来提供个性化体验，尤其是在信息透明并赋予用户控制权的情况下。
- 相比标准化的 AI 智能体，参与者更支持结合文化/地域背景进行定制的 AI 智能体。
- 参与者青睐能回应情感信号的拟人化 AI 智能体。

此外，我们启动了一个试点项目，旨在针对“具有文化相关性的 AI 模型应具备哪些特质”征询公众看法，并根据这些反馈建立偏好数据集，同时会将数据开源供开发者使用。该项目的成果将是一系列即取即用的数据集，让 Llama 大语言模型在不同文化背景下更具相关性和实用性。



可信合作伙伴

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组织的上报渠道。

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，在 2024 年这个选举年，他们的作用尤为突出。在局势动荡加剧的国家和地区，这些合作伙伴提供了相关见解，并识别出了当地的有害内容。这包括孟加拉国、巴西、科特迪瓦、刚果民主共和国、法国、希腊、印度、印度尼西亚、肯尼亚、伊拉克库尔德地区、墨西哥、尼日利亚、巴基斯坦、塞内加尔、南非、叙利亚和委内瑞拉等国家和地区。

2024 年，借助可信合作伙伴计划，我们移除了逾 10 万条违反政策的内容。

可信合作伙伴提供了选举相关内容趋势方面的见解，这些见解旨在为选举诚信工作提供参考，帮助我们检测并移除违规内容，以及识别平台上面临高风险的用户，以便落实额外保护措施。可信合作伙伴可以有效识别以下情况：针对边缘化群体的敌意言论激增、对新闻工作者和人权维护者的攻击行为，以及 AI 内容违规使用问题。

为管理仇恨行为风险，我们移除了所认定的诋毁性措辞。我们与可信合作伙伴合作，更好地了解这类措辞的使用语境，以便能更准确地执行我们的政策。

我们咨询了 20 个国家/地区的 40 余家可信合作伙伴，为涉及以下方面的政策制定和产品开发流程提供依据：移除敏感图像、剥削、认定危险组织和人物的信号、将“Zionist”（犹太复国主义者）作为指代词来实施仇恨行为、AI 智能聊天助手等等。

为回应监督委员会的建议，对于通过可信合作伙伴计划举报的内容，Meta 评估了对这类内容所做回应的及时性和有效性。在 2022 年第 2 季度至 2024 年第 4 季度的两年间，对于通过可信合作伙伴计划举报的内容，Meta 的响应速度显著加快。

2024 年，Meta 通过加大培训投入、简化政策执行系统和开发新工具等措施，有效提升了举报量和审核效率。

全球成效

在全球范围内，可信合作伙伴计划在 2022 年第 2 季度收到了超过 **11,800** 条举报内容，这一数字在 2024 年第 2 季度增长至超过 **49,200** 条，约为前者的**四倍**。

在全球范围内，可信合作伙伴渠道两年间的成效增长情况

2022 年第 2 季度至 2024 年第 4 季度

+ 4 倍

通过可信合作伙伴计划收到的举报内容数量增至 4 倍

+ 12 点

上报后 5 天内解决的案件占比提升 12 点

+ 15%

从处理时间（天数）中值来看，效率提升 15%

+ 15 倍

接受政策复审的举报内容数量增至 15 倍

接下来几页举例说明了可信合作伙伴在巴基斯坦、叙利亚和委内瑞拉的工作成效。

案例分析

报告叙利亚局势，提供见解分析

在 2024 年 12 月阿萨德政权倒台后，可信合作伙伴在报告和分析当地局势发展、提供有关当地内容趋势的见解以及上报严重违规行为方面发挥了关键作用。

凭借本地专业知识，可信合作伙伴报告的信息为 Meta 的危机应对工作提供了依据，让我们能够更及时、更高效地执行政策并降低风险。对于阿拉维派、基督教徒、库尔德人等宗教少数派和少数族裔的身份曝光风险，以及声称这些群体与倒台政权有关联的言论，可信合作伙伴均表示关切；同时，可信合作伙伴还警示了前叙利亚军队内部不同极端派系抬头的趋势。这些见解为我们的风险缓解工作提供了支持，这包括降低我们平台上由危险组织和人物带来的风险，以及降低基于个人特征的线下袭击风险。

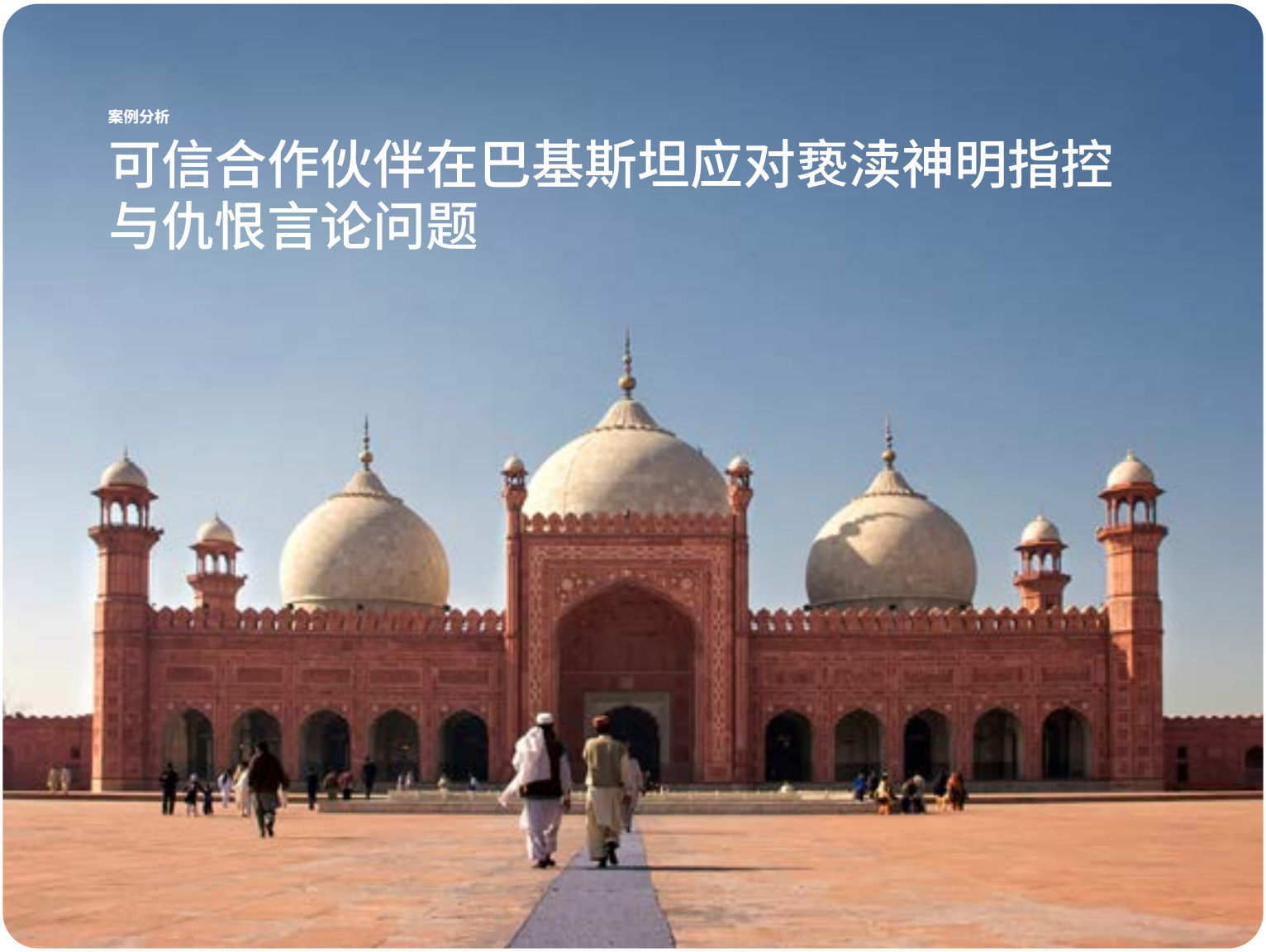
案例分析

降低公民行为者面临的风险

在 2024 年 7 月 28 日委内瑞拉大选前夕，我们与可信合作伙伴开展了合作，并与民间社会组织建立了新的合作关系，以便防范选举相关风险，并提高潜在违规内容的举报率。

大选结束后，委内瑞拉爆发抗议活动，政府随后采取镇压措施，包括大规模拘留抗议者，以及针对性逮捕政治反对派人士。我们的可信合作伙伴提供了有关当地局势发展的关键见解。他们举报了有害内容，包括针对抗议者和反对派支持者的隐晦威胁与身份曝光信息，这类内容令上述人士面临遭受任意拘留和人身伤害的风险。此外，可信合作伙伴还报告了新闻工作者、反对派成员和人权维护者等公民行为者的账户受到的攻击。

这些见解为开展主动检测提供了依据，帮助我们为这些账户启用高级安全保护功能，并通过实施交叉检查机制来防止政策执行失误。在高压环境下，这些措施为新闻报道工作和公民参与提供了支持。



案例分析

可信合作伙伴在巴基斯坦应对亵渎神明指控与仇恨言论问题

在巴基斯坦，可信合作伙伴在向我们通报针对边缘化群体（包括宗教少数派和性别少数群体）的潜在有害内容方面，发挥了重要作用。

在 2024 年 2 月巴基斯坦大选期间，可信合作伙伴重点举报了两类与选举相关的内容：一是针对政治候选人的仇恨言论，二是可能构成煽动行为的亵渎神明指控。在巴基斯坦，亵渎神明指控可能引发法律诉讼和人身暴力事件。

根据这些举报，我们得以根据有关配合实施伤害和宣扬犯罪行为的政策，移除与亵渎神明指控相关的内容。

此外，在其他关键时期，例如爆发教派暴力事件时，可信合作伙伴也为我们提供了信号和见解。他们的工作使我们能够迅速做出反应，移除平台上的违规内容，并加强内容检测和政策执行力度。

与巴基斯坦利益相关者交流合作

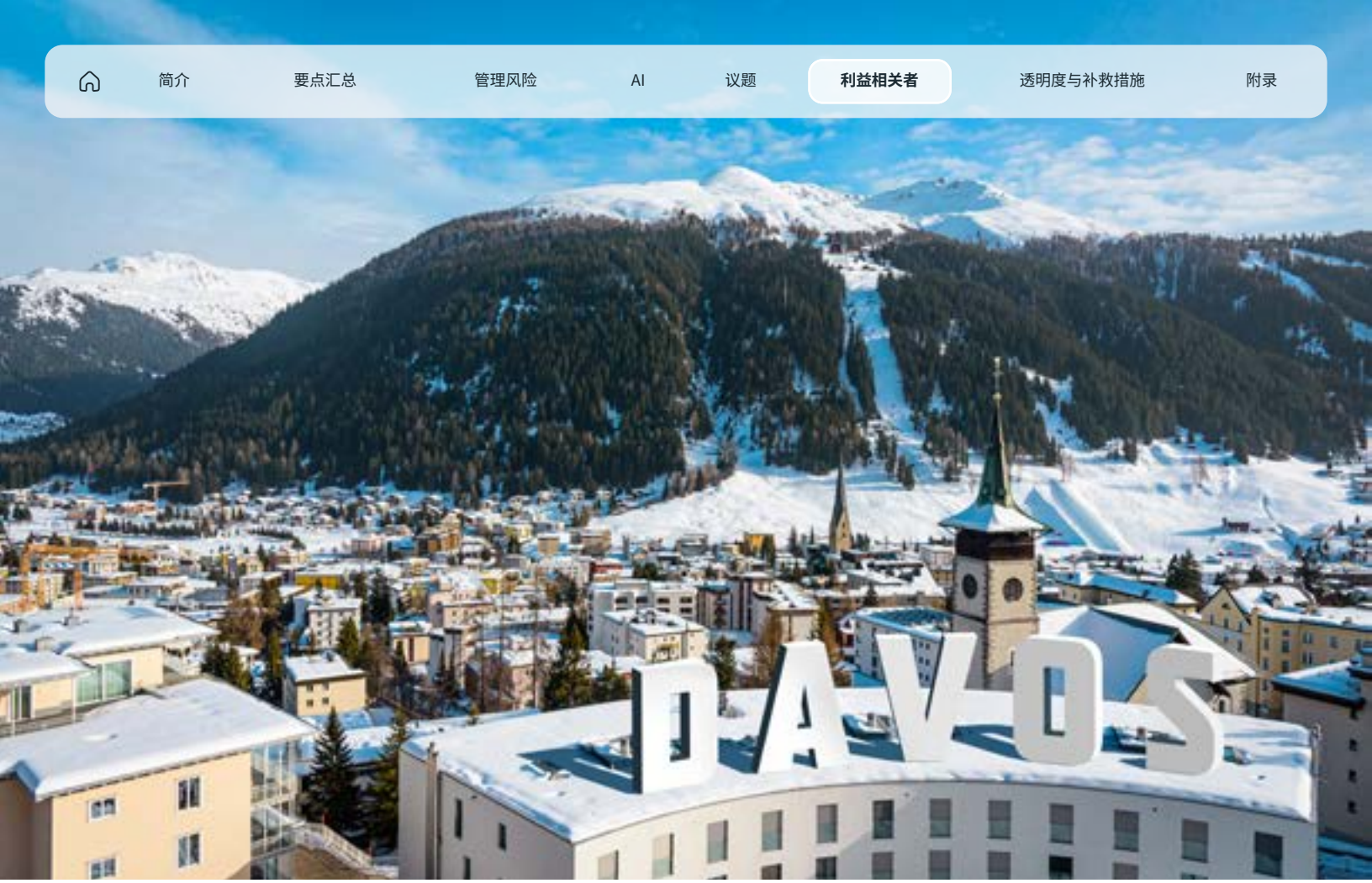
作为人权尽职调查工作的一环，Meta 在巴基斯坦与包括政府和非政府组织在内的多个利益相关者开展了交流合作，重点探讨如何平衡安全与表达自由。主要活动包括：

与巴基斯坦人权部、国家儿童权利委员会、国家人权委员会和数字权利基金会联合主办青少年网络安全圆桌讨论会。会上我们讨论了青少年账户，以及为巴基斯坦用户推出的乌尔都语 Take It Down 门户。

与多元化的人权维护者群体开展交流，收集有关互联网中断的见解，并探讨在倡导工作方面可能的合作途径。围绕政府设立的“防火墙”以及虚拟专用网络 (VPN) 白名单机制所产生的影响，这些人权维护者提供了宝贵见解。

与民间社会组织举行圆桌讨论会，深入讨论 Meta 的关键人权承诺，以及人权团队的工作内容。这包括探讨如何在无需中断互联网或限制社交媒体平台（包括我们旗下应用）的前提下，应对各种情况。

在每场活动中，每位与会者都谈到了针对用户发起的、恶意且无端的亵渎神明指控所造成的影响。当与会者了解到 Meta 有关身份曝光风险的政策，以及 Meta 为了保护受针对用户的安全而持续开展的工作后，他们均感到安心。



国际组织

2024 年，联合国成员国协商并通过了《全球数字契约》(GDC)，为数字技术与 AI 的全球治理设立了综合框架。我们与联合国成员国、联合国各机构及行业联盟携手，共同敲定了《全球数字契约》的最终文本。我们的工作旨在支持表达自由，同时为所有人创造一个更安全、更包容和更开放的数字化未来。

在这一年里，Meta 还以其他方式参与了联合国系统的各项活动。我们的工作包括协助联合国儿童基金会制定《Digital Technologies, Child Rights and Well-Being》（数字技术、儿童权利与福祉）文件，为科技行业履行尽职调查义务提供指引；同时与联合国教科文组织合作，开展数字平台虚假信息治理工作。我们还构建了一个翻译界面，为联合国教科文组织提供支持。该翻译界面基于“Meta 一言不漏”（Meta No Language Left Behind，简称 Meta NLLB）AI 模型，支持 200 种语言的优质翻译，其中包括阿斯图里亚斯语、卢干达语、毛利语、斯瓦希里语和乌尔都语等边缘化语言，助力促进语言多样性和信息获取。

Meta 继续与联合国人权事务高级专员办事处（下称“人权高专办”）保持密切合作。我们定期与人权高专办工作人员会面，并积极参与 B-Tech 项目，该项目为科技行业实施《联合国工商企业与人权指导原则》提供了权威指导和资源。我们还参与了该项目下设的“Community of Practice”（实践社群），这是一个与其他科技公司开展保密对话的空间。我们也积极参与了关于 AI 与人权标准的持续讨论。此外，我们参加了 2024 年联合国工商业与人权论坛，并在“网络仇恨言论”和“保护新闻自由”专题讨论会上发言。

Meta 在未来峰会及第 79 届联合国大会期间，参与了一系列边会活动中的政策讨论，议题涵盖 AI 在全球治理中的作用、数字创作者赋能、侨民主导的经济创新，以及网络犯罪与内容法律对表达自由的影响等。我们还参与了关于保护人权维护者、利用社交媒体在人道主义危机中传递救生信息等议题的讨论。

此外，我们与联合国人权特别机制下的独立人权专家开展了磋商，其中包括促进和保护意见和表达自由权问题特别报告员，以及人权维护者处境问题特别报告员等。

在这一年里，Meta 与七国集团 (G7)、二十国集团 (G20)、联合国教科文组织 (UNESCO) 及经济合作与发展组织 (OECD) 合作，推进 AI 包容性与治理方面的专项工作。我们还与相关政府围绕信息完整性的重要意义展开了对话。我们继续参与世界经济论坛“全球数字安全联盟”的工作，这些工作促成了《干预之路：数字安全措施有效实施路线图》的发布。

此外，我们积极参与了多个多方利益相关者论坛，并在论坛会议上与各界利益相关者展开交流，这些论坛包括 Eradicate Hate Summit（根除仇恨峰会）、非洲互联网自由论坛 (FIFAfrica)、全球互联网反恐论坛、互联网治理论坛 (IGF)、RightsCon、Tech Against Trafficking 大会以及联合国妇女地位委员会会议。

作为全球网络倡议成员以及数字信任与安全伙伴关系 (Digital Trust and Safety Partnership) 的参与者，Meta 出席了“European Rights and Risks:Stakeholder Engagement Forum”（欧洲权利与风险：利益相关者参与论坛），此次参会为我们依据《数字服务法》开展系统性风险评估提供了指导。

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。

了解监督委员会影响的渠道

我们会定期报告 Meta 提交给监督委员会的案件以及监督委员会建议的最新实施进展，2024 年，上述报告的频率从每季度一次调整为半年一次。此外，我们还上线了一个政策及信息公示平台专题页面，专门追踪监督委员会的各项建议产生的影响。除上述页面外，我们还设有监督委员会建议专题页面，其中列出了监督委员会就相关案件提出的建议、我们的承诺水平以及建议实施状态。

2024 年与监督委员会建议相关的行动

监督委员会提出的建议	Meta 正在评估和/或实施的建议[5]	已实施的建议[5]
48	70	41
（2023 年为 66 项）	（2023 年为 69 项）	（2023 年为 61 项）

2024 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、健康权、平等和无歧视权利等。下面举例说明了我们根据监督委员会在 2024 年的审理决定采取的行动。详情请参见 Meta 关于监督委员会的半年度报告。

监督委员会在 2024 年的审理决定示例：

监督委员会推翻了 Meta 移除三篇 Facebook 帖子的决定，这些帖子显示了 2024 年 3 月莫斯科恐怖袭击事件的视频。监督委员会要求 Meta 恢复这些帖子，并为帖子添加警告画面，将其“标记为令人不适的内容”。Meta 政策禁止描绘对可见受害者进行已认定袭击的画面。监督委员会认为，尽管这些帖子违反了该政策，但移除这些帖子不符合 Meta 的人权责任。

监督委员会维持了 Meta 移除一条巴基斯坦政客演讲视频的决定，因为此内容有引发现实伤害的风险。该视频配文称这名政客“crossing all limits of faithfulness”（逾越了一切忠诚底线），并使用了“kufr”（不信道）一词来暗示该政客亵渎神明。

[5] 一些正在评估和/或实施的建议或已完全实施的建议包括前几年提出的建议（详见我们的 2023 年度人权报告）。

根据监督委员会的建议采取的行动示例：

根据有关 AI 的一系列建议（比如说，请参见此处），我们调整了处理 AI 生成内容的方式，包括更新相关标签和政策，例如错误信息政策。

根据监督委员会有关内容政策的建议，Meta 修改了危险组织和人物政策，针对所有使用“shaheed”一词的语言，允许用户发布使用该词的内容，除非该内容伴有暴力信号或存在其他违反我们政策的情况（例如，美化已认定的危险人物）。

2024 年，监督委员会还尝试针对紧急案件缩短审理时间。例如，委内瑞拉在 7 月总统大选后爆发了暴力事件，当时我们将两条涉及“Colectivos”（集体组织）的内容提交给监督委员会进行加急审理。“Colectivos”（集体组织）是一个统称，用于指代与政府关系密切的非正规武装团伙或准军事性质组织。这些案件的审理时间缩短至 14 天。

我们还与监督委员会合作，与非洲、拉丁美洲、中东和土耳其等地区的监管机构和民间社会组织互动，提升各方对监督委员会职权范围及案件甄选流程的认识。

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，这项工作受全球事务总裁（现称“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。2024 年，人权事务总监向董事会下属的审计与风险监督委员会进行了汇报。

2024 年，Meta 在公司的第三方风险管理计划中，增设了人权风险管理专项模块。这项控制措施彰显了我们的承诺：持续改进我们的人权风险管理，努力与第三方开展负责任、尊重人权的合作。

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 2024 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

2025 年负责任业务实践报告

2025 年可持续发展报告

2023 年度人权报告、2022 年度人权报告、2021 年度人权报告

2024 年反奴役及反人口贩卖报告

2024 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

先前发布的人权影响评估：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