

Human rights report



Table of contents

About this report	03	Stakeholder engagement	46
Executive summary	06	Community Forums	51
Human rights risk management	11	Trusted Partners	52
1. Freedom of opinion and expression	12	Case study: Reporting on insights from Syria	54
2. Privacy	13	Case study: Mitigating risks for civic actors in Venezuela	55
3. Equality and non-discrimination	14	Case study: Trusted Partners tackle blasphemy allegations and hostile speech in Pakistan	56
4. Life, liberty and security of person	15	International organizations	58
5. Best interests of the child	16	Transparency & remedy	60
6. Public participation, to vote, and to be elected	16	Annex	64
7. Freedom of association and assembly	17	How human rights are governed and managed at Meta	65
8. Right to health	17	Training Meta employees on human rights	65
Accelerating AI innovation with respect for human rights	19	Links to referenced reports	65
Issues spotlight	25		
2024: The year of elections	25		
Preparing for elections at scale	26		
Managing AI influence risks	26		
Other election integrity efforts	27		
Preparing for the highest-risk elections	29		
Examples from national elections	29		
United States	29		
Mexico	30		
India	31		
European Parliament elections	32		
Child and youth safety	33		
Built-in protections for teens	33		
Fighting sextortion	36		
How we prepare for and respond to crises	37		
Sudan	39		
Middle East	41		
Bangladesh	42		
Georgia	43		
Cybersecurity	44		





About this report

This annual Human Rights Report covers insights and actions from January 1, 2024, through December 31, 2024. We report on Meta services and products including Facebook, Messenger, Instagram, WhatsApp, Threads and Reality Labs.

It builds on Meta's work to respect human rights and reflects progress made on our commitments to the [United Nations Guiding Principles on Business and Human Rights](#) and our [Corporate Human Rights Policy](#). The report shows how we applied these principles across the company in 2024 and provides guidance on where to find more in-depth information.





The content in this report is grounded in our [Comprehensive Human Rights Salient Risk Assessment](#), which was undertaken in 2022. The purpose of the assessment was to identify and prioritize our most significant potential adverse human rights impacts¹ on people who use our products and others who may be affected by our actions. This report outlines these potential salient risks and examples of our actions and mitigations in 2024.²

Human rights continued to be a topic of critical importance to our company and to our stakeholders in 2024. We strive to provide a representative picture of our work across multiple teams and stakeholder engagement around the world.

Policies and progress

In addition to this Human Rights Report, Meta reports annually on policies and progress using the following mechanisms:



Annual Report



Proxy Statement



Responsible Business Practices Report



Transparency Center



Sustainability Report



CDP Climate Change Report



UN Global Compact

This report complements the most recent [Meta Responsible Business Practices Report](#). We [report](#) separately on our efforts to identify and mitigate the risks of modern slavery and human trafficking in our business operations and supply chains. In addition, we comply with mandatory national and European Union reporting, available in our [Transparency Center](#). Links to other Meta disclosures are in the [Annex](#) to this report.

[Go to Annex](#)

¹ The term “adverse human rights impact” is in line with the UN Guiding Principles on Business and Human Rights and means an impact that occurs when an action removes or reduces the ability of an individual to enjoy his or her human rights.

² It does not include the [content policy and other changes](#) we announced in January 2025, when we updated our Hateful Conduct policy, formerly known as our Hate Speech policy, to address concerns about overenforcement and to seek to allow more freedom of expression.



Our Corporate Human Rights Policy applies enterprise-wide. Each Meta service and entity has its own policies and procedures that may have different human rights impacts. This report references actions taken by Meta as a company regarding one or more Meta entities. Statements are not intended to imply that Meta took that same action regarding all entities and/or in all circumstances.³



³ This report's discussion of content moderation and related actions on Facebook and Instagram does not apply to WhatsApp and, unless a policy or action is specified as applying to WhatsApp, it does not apply to WhatsApp. Further, while many actions described in this report apply to Facebook and Instagram, there are intentional distinctions in policies and procedures between the services. If a policy is labeled a "Facebook" policy, it may not apply to Instagram. No statement in this report is intended to create — or should be construed as creating — new obligations (legal or otherwise) regarding the application of a policy or procedure to other services or entities.



Executive summary



This is Meta's fourth annual Human Rights Report. It provides insights into the work Meta did in 2024 to manage human rights risks at scale and live up to our commitments to the [United Nations Guiding Principles on Business and Human Rights](#) (UNGPs).



Human rights timeline

The following chart outlines our human rights journey and how our work has evolved since the adoption of the UNGPs by the UN Human Rights Council in 2011.

2013

- Meta joins the Global Network Initiative, a multistakeholder collaboration to protect freedom of expression and privacy in tech

2018

- Meta releases Independent Assessment of the Human Rights Impact of Facebook in Myanmar

2019

- Meta establishes its human rights team

2020

- Meta publishes first Human Rights Impact Assessments — Philippines, Cambodia, Sri Lanka
- 20-member Oversight Board begins operating

2021

- Meta launches Corporate Human Rights Policy

2022

- Meta issues first Human Rights Report
- Meta publishes updates on Human Rights Due Diligence reports
- Meta publishes independent Israel and Palestine and End-to-End Encryption Due Diligence
- Meta launches human rights training

2023

- Meta adds Comprehensive Human Rights Salient Risk Assessment to Human Rights Report covering 2022
- Meta publishes updates on Human Rights Due Diligence reports

2024

- Meta publishes Human Rights Report covering 2023

Salient Risk Assessment

→ [Read more](#)

Our priorities in 2024 reflected the salient risks that were identified in the 2022 [Comprehensive Human Rights Salient Risk Assessment](#): freedom of opinion and expression; privacy; equality and non-discrimination; life, liberty and security of person; best interests of the child; public participation, to vote, and to be elected; freedom of association and assembly; and right to health.



Accelerating AI innovation

Advancements in artificial intelligence (AI) accelerated in 2024. Our vision is to build personal superintelligence, and make it widely available so all can benefit.

Generative AI apps became more widely used and increasingly transformed how we communicate, learn, create and work. We continued to promote an open approach to AI that can enhance human rights. This approach helped enable people's access to information and freedom of expression, as well as advance the rights to equality and non-discrimination, including by improving accessibility and expanding language inclusivity.

[→ Read more](#)

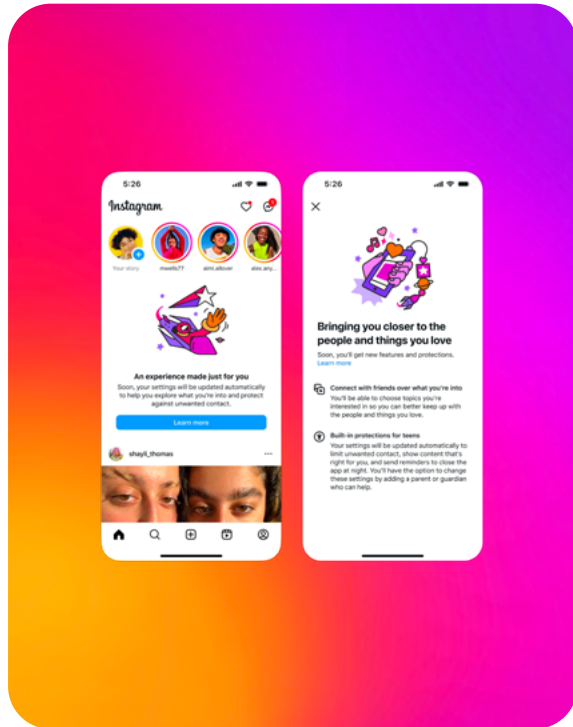


2024: The year of elections

2024 was the [largest election year in history](#). More than 70 countries, home to more than half the world's population, held national elections. Roughly 2 billion people were eligible to vote. We focused on enabling the rights of freedom of expression, participation in political processes and access to information of those in countries undergoing elections.

[Our approach](#) matured over hundreds of elections in recent years. It involved efforts to manage AI risks, enforce our [policies on voter or census interference](#), disrupt adversarial networks, increase transparency of political advertising and connect voters to reliable information. This report includes examples from the U.S., Mexico, India and the European Union.

[→ Read more](#)



Child and youth safety

Our commitment to [child and youth safety](#) continued. Among other initiatives, our work in 2024 included the launch of [Instagram Teen Accounts](#), a new experience for teens, guided by parents. Teen Accounts have built-in protections that limit who can contact teens and the content teens see, and ways to help manage the time teens spend on the app, while providing new ways for them to explore their interests. Our efforts balanced supporting the promotion of youth autonomy with the rights and duties of parents and guardians. We developed them in line with expert guidance and the principle of the evolving capacities of the child outlined in the [UN Convention on the Rights of the Child](#).

→ [Read more](#)

Crisis response

We continued to integrate human rights principles into [how we prepare for and respond to crises](#). Our [Crisis Policy Protocol](#) guides our expedited use of levers to mitigate potential harm. In 2024, we designated 19 global situations under the Crisis Policy Protocol. In this report, we provide examples of our crisis response in [Bangladesh](#), [Georgia](#), the [Middle East](#) and [Sudan](#).

→ [Read more](#)





Stakeholder engagement

Our [Corporate Human Rights Policy](#) supports our proactive engagement with stakeholders. In 2024, we connected with a broad range of stakeholders to inform the company's approach to issues related to expression, hateful content, misinformation and privacy. These stakeholders included a diverse range of human rights groups, vulnerable communities, civil society members, academics, think tanks and regulators. Key topics included our approach to responsible AI and election integrity, as well as our designation signals for dangerous organizations and individuals, and violent events.

In 2024, we conducted six [Policy Forums](#), where subject matter experts from Meta share varying viewpoints and discuss potential changes to Community Standards and Advertising Standards. We also held [Community Forums](#) to leverage public input on issues where there were competing tradeoffs and no clear answers. These helped us improve products and anticipate potential risks of emerging technologies, and enabled voices outside the company to have a greater say in our decision-making.

We continued to engage our [Trusted Partners](#) around the world to help identify trends, and better understand the impact of online content and behavior on local communities. We also explored how to strengthen relevant escalation channels. Their expertise was particularly valuable during 2024's

intense election cycle, and in situations of heightened unrest. They also provided insights and identified potentially violating content in Bangladesh, Brazil, Côte d'Ivoire, Democratic Republic of Congo, France, Greece, India, Indonesia, Kenya, Kurdistan-Iraq, Mexico, Nigeria, Pakistan, Senegal, South Africa, Syria and Venezuela, among other countries and regions.

[→ Read more](#)

Oversight Board

In 2024, the [Oversight Board](#) considered cases concerning our efforts to respect human rights, including freedom of expression, the right to health, and the right to equality and non-discrimination, among other topics. The Oversight Board is an independent body that reviews cases referred by Meta or appealed by individuals on Facebook, Instagram or Threads who disagree with our content moderation decisions. It provides binding rulings on whether to remove or leave up content. In response to a recommendation by the Oversight Board, Meta evaluated the [timeliness and effectiveness](#) of responses to content reported through the Trusted Partner program.

[→ Read more](#)

Managing government requests

Throughout the year, we continued to be guided by our commitment to the [Global Network Initiative](#) to respect freedom of expression and privacy, including when responding to government requests to restrict content. In 2024, we published [case studies](#) related to political speech in Brazil, Germany, India, Iraq, Israel, Singapore and Türkiye.

[View case studies](#)



Human rights risk management

The [United Nations Guiding Principles on Business and Human Rights](#) make it clear that companies should identify their adverse human rights impacts in order to effectively prevent or mitigate them.

Given the scale of Meta's operations and the range of rights it could implicate, anticipating and managing our [salient risks](#) is important but complex. We manage two types of inherent risks: those stemming from our own activities and those stemming from the activities of third parties, including people who use our platforms.

Once processes are implemented to address these risks, some level of risk will always remain. This is known as "residual" risk. While residual risks exist in all risk management systems, those associated with digital technologies and their impact on human rights persist due to the dynamic and rapidly evolving nature of digital technologies and the high degree of third-party activity.



The table on the following pages lays out our salient human rights risks as defined in our 2022 Comprehensive Human Rights Salient Risk Assessment (CSRA), which we disclosed in our [2022 Human Rights Report](#). This table provides illustrative examples of how we addressed the potential risks in 2024. Later in this report, we take a deeper dive into some of these examples and how we managed potential risks in relation to artificial intelligence (AI), elections and conflicts.

1. Freedom of opinion and expression

The [right to freedom of opinion and expression](#) includes the right to seek, receive and share information and ideas of all kinds. It is a foundational right, essential for the protection of human dignity, individual autonomy and democracy. Freedom of expression is a core part of our mission, consistent with our value of giving everyone a voice.

Examples of potential inherent salient human rights risks identified in CSRA

Examples of how Meta addressed the potential risks in 2024

Meta's content moderation policies and enforcement may limit freedom of expression.

We continued to develop our policies with freedom of expression as our north star. In 2024, we conducted several [Policy Forums](#) that sought to develop a nuanced appreciation of freedom of expression challenges in a range of areas.

Overbroad government limits on content

We strive to follow our [Global Network Initiative](#) (GNI) commitments. This included reporting on our [responses](#) to government requests for data or content restrictions ([here](#) and [here](#)). Our approach to responding to government requests is detailed in our [2023 Human Rights Report](#). In cases where we believe that requests from the government or orders from courts are not legally valid, are overly broad or are inconsistent with international human rights standards, we may request clarification, appeal or take no action. In 2024, noteworthy transparency [case studies](#) involving political speech included those in Brazil, Germany, India, Iraq, Israel, Singapore and Türkiye.

Internet disruptions and blocks to social media prevent people from exercising their right to freedom of expression and cut them off from receiving and sending vital news and information.

In order to prevent blocks to social media and messaging, we may comply with lawful government requests while seeking to honor our GNI commitments to respect freedom of expression. We also continue to provide the [WhatsApp by proxy](#) feature for people who cannot connect to our apps directly.



2. Privacy

The [right to privacy](#) is a necessary condition for the realization of other human rights, such as freedom of expression, freedom of assembly and association, and freedom of belief and religion. One of the core principles outlined in our [Corporate Human Rights Policy](#) is to keep people safe and protect privacy.

Examples of potential inherent salient human rights risks identified in CSRA

Generative AI models may involve processing of personal data in ways that people do not expect or understand.

Examples of how Meta addressed the potential risks in 2024

We are transparent about how Meta [uses information](#) for generative AI models and features, and we have an internal [Privacy Review process](#) for responsible data use, including generative AI. An update on our privacy progress in 2024 is available [here](#) and [here](#).

[→](#) Read more

Content or behavior on Meta apps may adversely impact privacy and data protection rights.

In October 2024, Meta [reintroduced facial recognition technology](#) on Facebook and Instagram to help people recover compromised accounts and prevent scams involving fake celebrity endorsements. To balance potential privacy with integrity risks, we give public figures, whose likenesses are being abused to scam others, the option to participate in or opt out of the program.



3. Equality and non-discrimination

The [right to equality and non-discrimination](#) provides for equal protection against discrimination. As part of respecting this right, we don't allow hateful conduct on our platforms, as defined in our [policy](#).



Examples of potential inherent salient human rights risks identified in CSRA

Some languages and dialects may be more challenging to moderate than others.

Content adversely impacting equality and non-discrimination (e.g., hateful conduct)

Examples of how Meta addressed the potential risks in 2024

We designed and deployed new mechanisms to route Arabic content by dialect for more efficient and precise moderation, including in [Sudan](#). The new system detects and prioritizes directing content to moderators who are most likely to understand that particular Arabic dialect.

Based on research, [external engagements](#) and investigation on our platforms, we updated our Hateful Conduct policy regarding [content attacking "Zionists."](#)

During training of our AI models, we tested training data for content or properties that could increase the risk of generating potentially harmful content, such as whether a dataset was representative across multiple demographics.



4. Life, liberty and security of person

The [right to life, liberty and security of person](#) concerns freedom from physical harm and confinement. For Meta, respecting this human right includes mitigating the risk that content may provoke harm, including risks of violence and human trafficking, state-sponsored online threats, and non-state groups engaged in, or advocating for, violence or hate.

Examples of potential inherent salient human rights risks identified in CSRA

Bad actors who:

- Exploit Meta services and apps to coordinate online or offline harm
- Misuse services and apps for cyber-attacks or phishing
- Threaten and harass human rights defenders, activists and other vulnerable groups

Examples of how Meta addressed the potential risks in 2024

Meta's [Coordinating Harm and Promoting Crime policy](#) prohibits facilitating, organizing, promoting or admitting to certain criminal or harmful activities. In 2024, we provided guidance on prisoners of war in the policy so that content reviewers were better able to remove violating content at scale, including in [Sudan](#).

We continued supporting the Human Rights Defender Fund and redesigned our [Trusted Partner program](#) to improve emergency response for human rights defenders and other vulnerable people.



5. Best interests of the child

The UN Convention on the Rights of the Child (UNCRC) states that in all actions concerning children, “the best interests of the child shall be a primary consideration.” Meta’s [Best Interests of the Child Framework](#) aligns with the fundamental values of the UNCRC. Online child protection is a top priority for Meta. We offer tools for teens, parents and guardians with built-in protections to help keep them safe while providing space for them to exercise their right to freedom of expression and access to information.

Examples of potential inherent salient human rights risks identified in CSRA

Children can be exposed to unwanted, inappropriate content or predatory behavior.

Examples of how Meta addressed the potential risks in 2024

We launched [Instagram Teen Accounts](#), a new experience for teens with built-in safeguards, guided by parents.

[→](#) [Read more](#)

6. Public participation, to vote, and to be elected

The [right to public participation, to vote, and to be elected](#) in free and fair elections is a cornerstone of democracy. Protecting the integrity of elections on our services and apps is one of our highest priorities. We work hard to protect elections online before, during and after election periods.

Examples of potential inherent salient human rights risks identified in CSRA

Violating content can adversely impact public participation, voting or running for office. This can stem from activities including, but not limited to:

- Coordinated bad actors interfering with elections
- Threat of offline harm and violence to candidates
- Individual efforts to prevent people from voting, increases in spam, foreign interference or reports of content that violates our policies

Examples of how Meta addressed the potential risks in 2024

Elections were a priority in 2024. We [prepared for elections at scale](#), including [highest-risk elections](#), and helped voters find information, among other actions.

[→](#) [Read more](#)



7. Freedom of association and assembly

The [right to freedom of association and assembly](#) is essential to democracy and interdependent with many other rights guaranteed under international human rights law, including the right to freedom of expression and to take part in the conduct of public affairs. For Meta, this right relates to our core values of giving people a voice and building connection and community.

Examples of potential inherent salient human rights risks identified in CSRA

Content or coordinated inauthentic behavior on Meta platforms may lead some people to feel unable to freely gather on Meta apps or offline.

Examples of how Meta addressed the potential risks in 2024

We deployed our [Crisis Policy Protocol](#) to help resource our efforts to address violating content regarding mass demonstrations, for example in [Bangladesh](#) and [Georgia](#).

We also prepared well in advance of [elections](#) to reduce the risk of violating content that could cause people to feel unsafe to gather during and after elections.

Threads [joined](#) the [fediverse](#), an open, global network of social media servers. This enabled people to expand their communities and reach new audiences.

8. Right to health

The [right to health](#) is the right of everyone to the highest attainable standard of physical and mental health. Meta respects this right by increasing access to health information, enabling people with similar health issues to connect with one another, and empowering people to make informed decisions about their health and well-being.

Examples of potential inherent salient human rights risks identified in CSRA

Policy-violating content that incites or is intended to cause offline harm

Examples of how Meta addressed the potential risks in 2024

We launched the [Thrive program](#), a cross-industry signal sharing program designed to prevent the spread of suicide and self-harm content, with Snap and TikTok.

We held a [Policy Forum](#) on commercial content with regulatory-informed health and safety risks.

We updated our [Community Standards](#) and [Advertising Standards](#) to reference recalled goods.

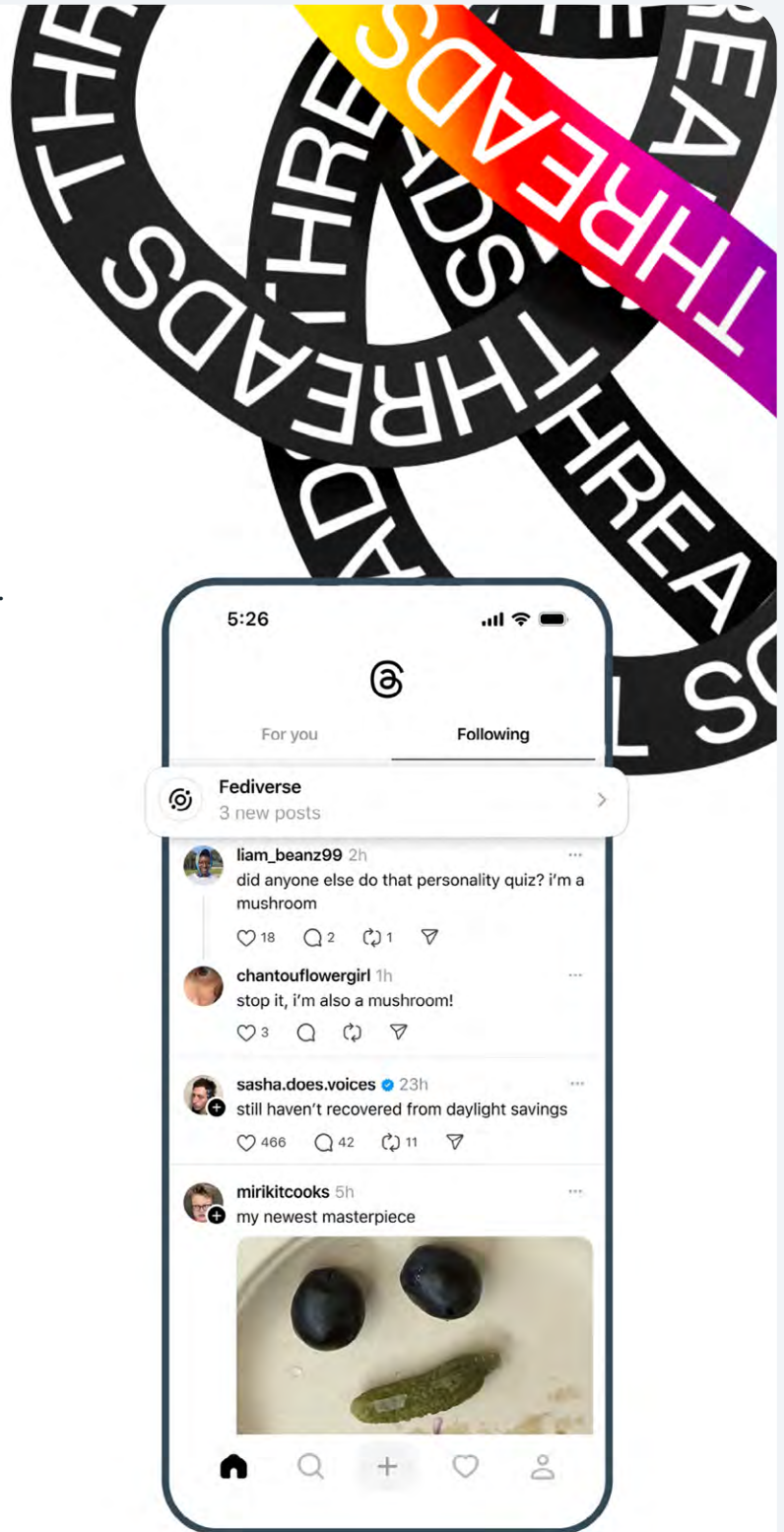


New products and services

As discussed in our [2023 Human Rights Report](#), Meta strives to respect human rights in the design and development of our products and services.


In 2024, Threads [joined](#) the [fediverse](#), an open, global network of social media servers. If a user decides to turn on sharing to the fediverse, people from different platforms (such as Mastodon or Flipboard) can follow and engage with the user's Threads content, even if they don't have a Threads profile. This enables people to exercise their freedom of expression and freedom of association and assembly by reaching new audiences, expanding their communities and joining public discussions on topics they care about. It also helps create a more diverse information ecosystem.

We also provided people who use Threads with education via a dedicated section in our [Help Center](#) and a new fediverse guide in our [Privacy Center](#) on how decentralization and interoperability affect their privacy.





Accelerating AI innovation with respect for human rights



Advancements in artificial intelligence (AI) accelerated in 2024. Generative AI tools and apps became widely used and increasingly transformed the ways we communicate, learn, create and work. At Meta, we recognize that this rapid development and adoption of AI raises important, and often novel, human rights benefits and risks.

Our [long-term vision](#) is to build personal superintelligence, and make it widely available so everyone can benefit.



In 2024, we released our open [Llama 3](#), [Llama 3.1](#), [Llama 3.2](#) and [Llama 3.3](#) large language models (LLMs). We also launched our [Meta AI assistant](#) and integrated it throughout our technologies. [Meta AI Studio](#) debuted as a platform for creating customized AI characters, a [suite of generative AI tools](#) helped advertisers build their businesses, and Meta AI was [integrated with our Ray-Ban Meta glasses](#). We continued to conduct and publish [cutting-edge AI research](#), including our [Movie Gen models](#) that generate videos and enable precise, instruction-based video editing, and our [Video Seal model](#) for durably watermarking AI-generated videos, among other advances.

By the end of 2024, developers had downloaded our Llama open models [more than 650 million times](#), and Meta AI had nearly 600 million monthly active users globally, making our models the world's most widely adopted. This large userbase of both developers and end users underscores our responsibility to build AI with respect for human rights.

Our open approach

We believe that [open source AI is an important part of ensuring everyone benefits from AI advances](#). As we outlined in our [2023 Human Rights Report](#), an open approach has important benefits for human rights. Open source AI models:



Are inherently more resilient to censorship and other restrictions on the right to freedom of expression because they can be downloaded and run offline, reducing the impacts of potential post-release government demands to restrict outputs.



Better enable adaptation and [fine-tuning](#) to reflect local context and nuance, in line with the right to equality, improving accessibility and language inclusion.



Make it easier for developers to build smaller, more efficient models that can lower barriers for traditionally underserved communities, supporting economic, social and cultural rights.



Support critical research into AI safeguards and security by allowing anyone to scrutinize models for potential risks, helping to mitigate potential adverse human rights impacts.

We are already seeing tangible benefits of our open approach. Our [Llama Impact Grants program](#), launched in 2023, continued in 2024. This program, together with our 2024 [Llama Impact Innovation Awards](#), supports and highlights positive social impact use cases of our open models.

For example, developers leveraged Llama to build the [Vax-Llama model](#), a chatbot service designed to provide accurate information on vaccination intended for adoption by healthcare providers around the world. It was also used for the [Llama-Suho project](#), an initiative fine-tuning Llama with Korean-specific data to enhance AI safeguards in the Korean context.



Centering protections

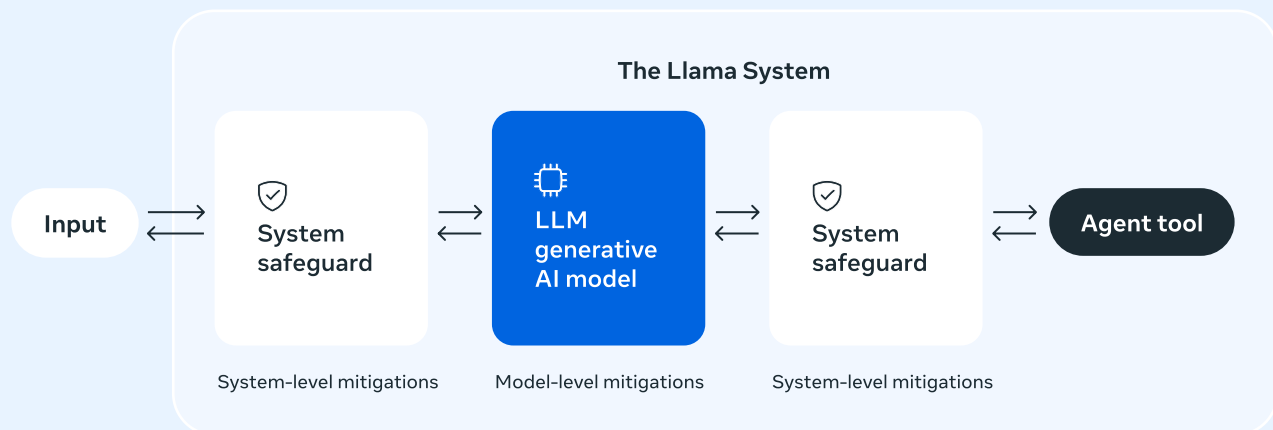
We remain committed to developing and deploying state-of-the-art AI products while considering human rights standards and safeguards against abuse.

Our [Corporate Human Rights Policy](#) explicitly references the applicability of our human rights commitments to AI.

With the release of Llama 3 in April 2024, we began emphasizing a [systems-based approach to AI safeguards](#). A systems-based approach gives developers additional flexibility to apply the appropriate layers of protections for different use cases and audiences. For example, we provide protections for certain types of potentially offensive, but lawful, speech as optional system-level mitigations. This is in addition to continuing to embed baseline protections against the generation of child exploitation content in our foundation models.

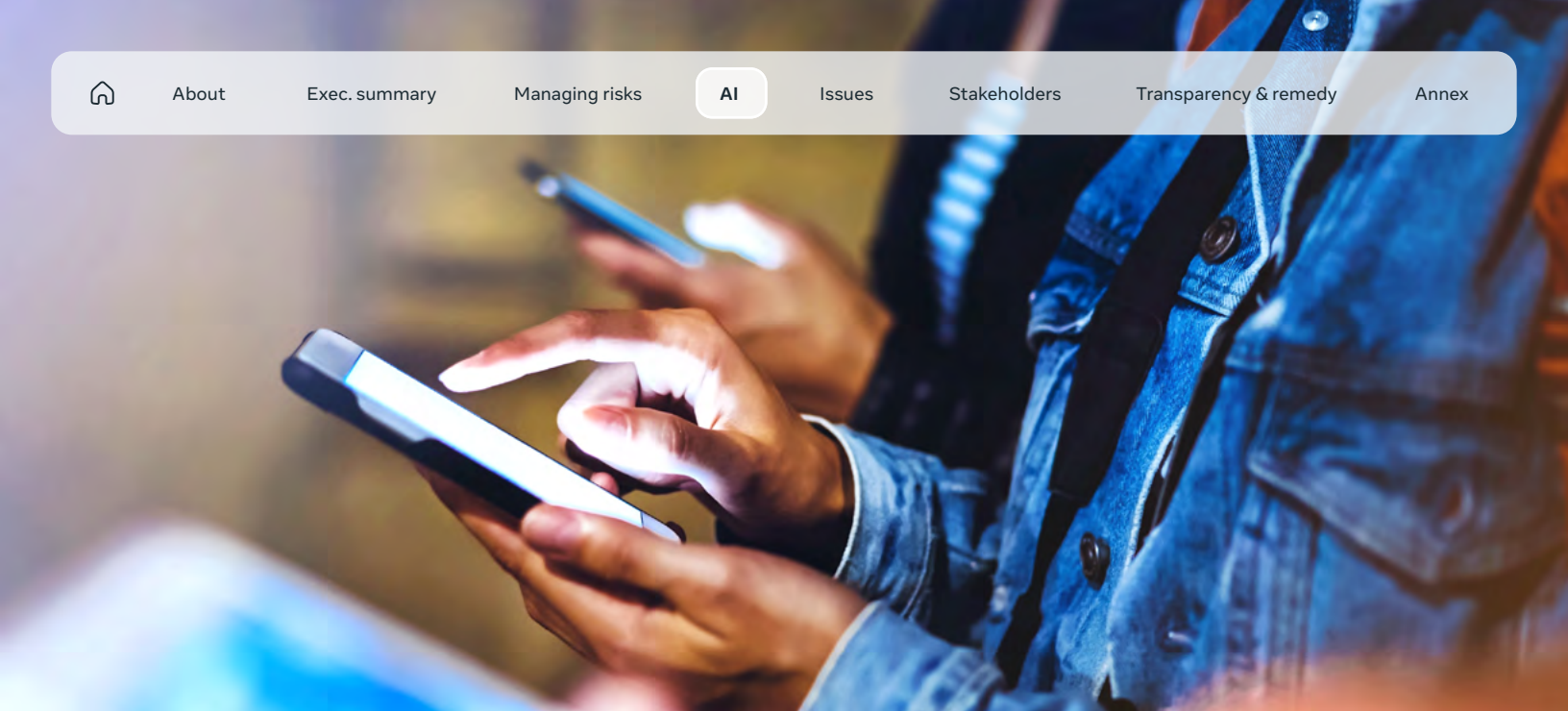
We believe this systems-based approach supports an appropriate balance between freedom of expression and other human rights.

AI safety systems

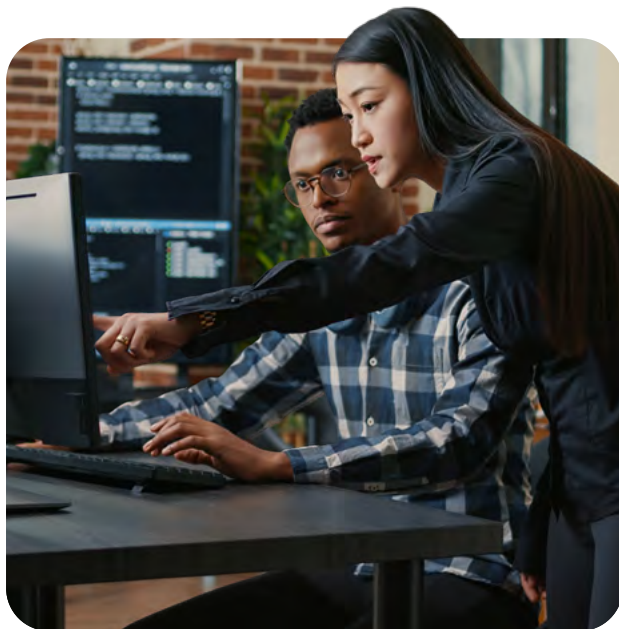


As part of our systems-based approach, we have open sourced three key tools ([Llama Guard](#), [Prompt Guard](#) and [Code Shield](#)) that can be customized and used together or independently by developers to implement protections against abuse.

Our [Developer Use Guide](#) provides detailed guidance on how to responsibly deploy our foundation models and safety systems in a range of contexts. Our [Acceptable Use policy](#) continues to govern deployments of our open models.



In addition to providing these tools in 2024, we also undertook important steps to mitigate risks associated with our own deployments of generative AI. Our practices in 2024 included:



Extensive red-teaming of our models and first-party products prior to release to identify and mitigate potential risks, including those related to potential adverse human rights impacts.



Updating [how we handle manipulated media](#) based on [feedback](#) from the independent Oversight Board, including [adding “AI Info” labels](#) and [context](#) to a wider range of video, audio and image content, and requiring creators to disclose the use of AI.



Refining the internal guidelines and process we use to test for acceptable model output to better reflect real-world use cases and align with international human rights standards.

We also recognize that AI protections require cross-industry and multistakeholder collaboration. In February 2024, together with industry peers, we signed the [AI Elections Accord](#), committing to help prevent deceptive AI-generated content from interfering with global elections. In May 2024, we [joined the Frontier Model Forum](#), an industry-supported body dedicated to advancing the security of frontier AI models.



Addressing false refusals

False refusals occur when a model refuses to produce requested output in response to a benign prompt, often as a result of a model's well-intentioned safety protections. For example, a model may incorrectly refuse to discuss a piece of classic literature that contains an offensive stereotype or slur or to answer a basic high school chemistry question because of safeguards designed to prevent aiding in the creation of chemical, biological, radiological, nuclear and high-yield explosives. While model safety is important and refusals can be necessary to limit the generation of harmful content, false refusals can have negative impacts on freedom of expression, access to information and other rights.

Beginning with Llama 3, we undertook significant work to reduce false refusals by Llama and Meta AI and made substantial progress in the course of 2024.



Internationalizing with care

In 2024, we made Meta AI available in [more than 40 additional countries and multiple new languages](#), including Arabic, Bahasa Indonesia, Filipino, French, German, Hindi, Italian, Portuguese, Spanish, Thai and Vietnamese.

Prior to launching in each country and language, we evaluated potential human rights risks and carried out context-specific [red-teaming](#), a common practice for mitigating unsafe behaviors in LLMs.

Not every country where Meta AI is available enjoys robust protections for freedom of expression in domestic law. As part of our internationalization work in 2024, we developed a human rights-based approach for responding to government requests to restrict or limit Meta AI output, building on our [longstanding policies](#) and in line with our commitments as a member of the [Global Network Initiative](#) and our [Corporate Human Rights Policy](#).

Engaging with stakeholders

Any space where technology is advancing as rapidly as AI poses novel challenges for stakeholder engagement. Throughout 2024, we focused on educating stakeholders and soliciting meaningful feedback.

Among our efforts:



We held AI roundtables in the U.S. to get feedback from multidisciplinary groups on product and model releases, including from experts from the U.S., Brazil, Brussels, Jordan, Mexico and across Africa.



We shared findings from [Community Forums](#) held in the U.S., Brazil, Germany and Spain to explore principles for generative AI chatbots, conducted in partnership with Stanford University's [Deliberative Democracy Lab](#).



We held a series of [Open Loop workshops](#) designed to address the complexities and harness the opportunities of open source AI. These workshops brought together policymakers, industry leaders, academics and civil society representatives from around the world to collaboratively shape effective and responsible AI policies.



Alongside the [13th UN Forum on Business and Human Rights](#) in Geneva, we developed and led an interactive multistakeholder simulation on human rights due diligence for generative AI products, sharing our approach and building an improved mutual understanding of risks and challenges.

As we continue to innovate on AI, we remain committed to globally inclusive and robust stakeholder engagement and consultation.



[Read more](#)



Issues spotlight

2024: The year of elections

2024 was the largest election year in history. More than 70 countries, home to more than half the world's population, held national elections, and roughly 2 billion people were eligible to vote.

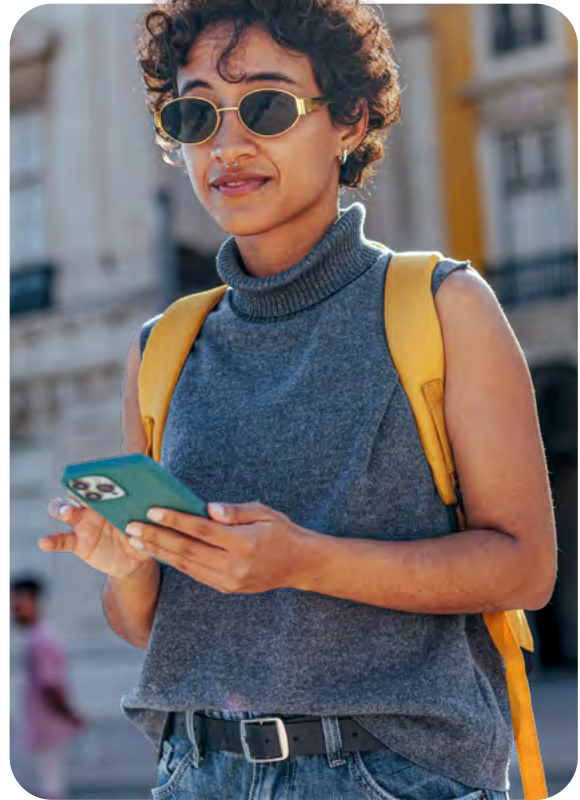
Meta recognizes the importance of enabling the rights to freedom of expression, to vote and to participate in public affairs. Our electoral work was a major focus throughout the year. We [prepared](#) for the scale, spread and cadence of elections, and worked to mitigate related risks to users — including potential risks related to increasing use of AI.

We review our 2024 efforts on the following pages and provide illustrative summaries from elections in the European Union, India, Mexico and the U.S.

Preparing for elections at scale

Meta has iterated our core [approach](#) to elections over the past several years. We deploy it in all countries where our services are used, adapting our strategy to local needs and risks. As part of our preparations for the 2024 elections, a dedicated team was responsible for cross-company efforts, which included experts from our intelligence, data science, product and engineering, research, operations, content, human rights, public policy and legal teams. Throughout the year, we sought to enable people's ability to express themselves and to vote and be elected.

Our approach included efforts to manage AI risks, enforce our [policies on voter interference](#), disrupt adversarial networks, provide transparency of political advertising and connect voters to reliable information. We also evaluated the appropriate language coverage of classifiers and human reviewers for the countries holding elections to support our efforts to take action on policy-violating content. Some highlights of our work included:



Managing AI influence risks

At the start of the year, many people were concerned about the risks generative AI could pose to fair elections, including the risk of widespread deepfakes and AI-enabled disinformation campaigns. We prepared for and closely monitored adversarial threats and [potential election disruption from AI](#). From what we monitored across our services, it appeared these risks did not materialize significantly and any such impact was modest and limited in scope. For example, during the election period in a group of major elections, ratings on AI content related to elections, politics and social topics represented less than 1% of all fact-checked misinformation. Our existing policies and processes appeared sufficient to reduce risks around generative AI content.

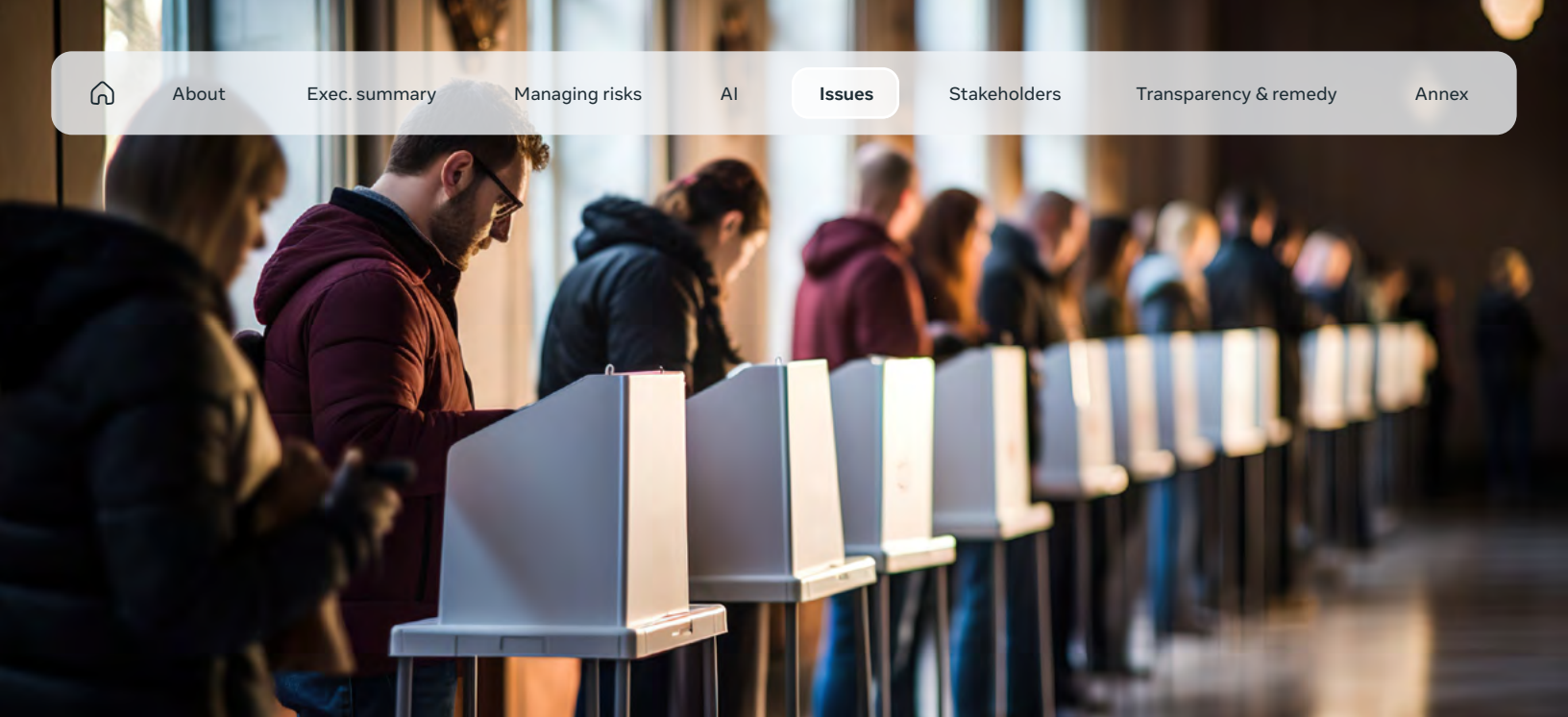
Throughout the year, our efforts focused on stopping influence operations and leveraging global partnerships to help preserve the integrity of elections.



We closely monitored the potential misuse of generative AI by [coordinated campaigns using fake accounts](#). We [found](#) they made only incremental productivity and content-generation gains using generative AI. These incremental gains did not impede our ability to disrupt these influence operations because we focus on behavior when we investigate and take down these campaigns, not on the content they post — whether created with AI or not.



We also [cooperated](#) with others in our industry to combat potential threats from the use of generative AI. For example, in February 2024, we signed the [AI Elections Accord](#) alongside dozens of other industry leaders, pledging to help prevent deceptive AI content from interfering with the 2024 global elections. Examples of country-specific AI initiatives are described on the following pages.



Other election integrity efforts

In addition to mitigating risks of potential AI influence on elections, we also sought to empower voters, prevent foreign interference, enhance candidate safety, build partnerships and help ensure advertiser transparency.

Empowering voters



Access to reliable information and the responsible use of online platforms are especially important during elections. In many countries, we provided people with voter information and election day reminders through in-app notifications on Facebook and Instagram. These features enabled people to access authoritative information from official election authorities about how, where and when to vote on election day. For example, in local elections in Brazil, people engaged with these notifications approximately 9.7 million times across Facebook and Instagram. More than 63 million people on Facebook and 118 million people on Instagram saw the voter registration sticker redirecting them to authoritative information about the election and voting.

Preventing foreign interference



Our security teams investigated and took down coordinated networks of inauthentic accounts, Pages and Groups. In addition, we estimated that every day our automated fake account detection [prevented](#) millions of fake accounts from ever being created. Our teams took down approximately [20 covert](#) influence operations around the world, including in the Middle East, Asia, Europe and the U.S. For example, in [Moldova](#), we removed a network targeting Russian-speaking audiences as part of our investigation into suspected coordinated inauthentic behavior in the region.

Candidate safety



Meta also provided enhanced protection against hacking, impersonation and harassment for the accounts of elected officials, candidates and their staff. We held multiple trainings for candidates on safety, outlining the [guidance](#) available to address harassment on our platforms, and [published](#) educational content so that it was widely available to all election participants.



Outreach and partnerships



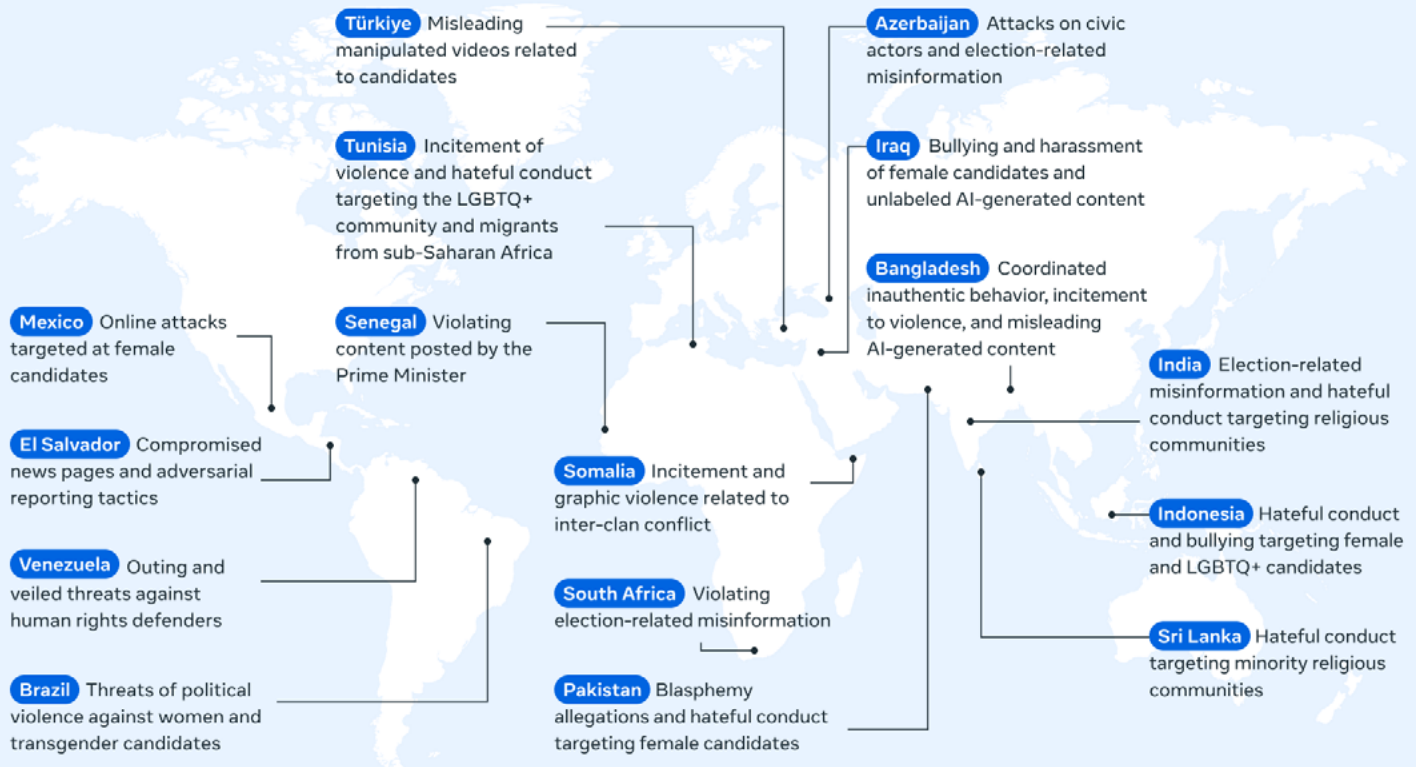
We conducted outreach and established communication channels with government authorities and law enforcement agencies so that they could report content potentially violating our Community Standards or local law. We also partnered with civil society groups, fact-checkers and other technology companies to help us identify and stop emerging threats and the spread of [false information](#).

Ads transparency



We continued to offer industry-leading transparency for ads about social issues, elections and politics. In most markets where we offer such ads, advertisers go through an [authorization process](#) and must include a [“paid for by” disclaimer](#) on their content to run their ads. The disclaimer might include information on the organization or individual responsible for the ad, although requirements may vary by country. The ads are then stored in our publicly available [Ad Library](#). In 2024, we added the requirement that advertisers needed to [disclose when they use AI](#) or other digital techniques to create or alter a social issue, electoral or political ad in certain cases.

Trusted Partners supported the elections integrity efforts in 25 countries in 2024





Preparing for the highest-risk elections

We deemed some elections higher-risk, requiring additional preparation, extra resources and bespoke work. We considered, for example, the type of election, the size of the country relative to our userbase, risks of political violence, targeting of vulnerable groups and our operational capacity. The additional efforts included setting up dedicated monitoring efforts and temporary risk response measures that could be designed and applied across countries and languages.

We ran a number of election operations centers around the world to monitor and react swiftly to issues that arose, including in higher-risk elections. You can find more detail online about our efforts in [Brazil](#), [France](#), [India](#), [Indonesia](#), [Mexico](#), [Pakistan](#), [South Africa](#), the [UK](#), the [U.S.](#) and the [European Parliament](#).

Examples from national elections

The four brief country examples on the following pages help illustrate how we sought to manage election risks in 2024. In each context, we began preparations at least a year in advance.

United States

In preparation for the U.S. election, our [efforts](#) focused on helping connect people with reliable voter information, addressing foreign interference and ensuring advertiser transparency.

Voter information



During the 2024 U.S. general election, top-of-feed reminders on Facebook and Instagram received more than 1 billion impressions. These reminders included information on registering to vote, voting by mail, voting early in person and voting on election day. People clicked on these reminders more than 20 million times to visit official government websites for more information.

Foreign interference



We prepared for online [foreign interference](#) in elections, expanding our ongoing enforcement against Russian state-controlled media entities, and continued to disrupt one of the largest and most persistent [covert influence campaigns](#), called Doppelganger. The vast majority of Doppelganger's attempts to target the U.S. in October and November were proactively stopped before anyone saw their content.

Advertisement restriction period



During the final week of the election campaign, we prohibited new political, electoral and social issue ads — a practice we've maintained since 2020. The [rationale](#) behind this restriction period remained the same as previous years — in the final days of an election, we recognized there might not be enough time to contest new claims made in ads.



Mexico

2024 was the largest election year in Mexico’s history, with about 90,000 candidates running for over 20,000 public offices. Violence during election campaigning was also at an all-time high. At least [37 candidates](#) were killed, and more than [828 non-lethal attacks](#) were recorded. More [women](#) ran for office than during any other election cycle in Mexico’s history, and female candidates suffered high rates of gender-based [violence](#) and killings.

Our [efforts](#) were similar to those in other high-risk settings, and benefited from Meta experts on the ground. We removed higher levels of violating content than usual before and during the election. Violating content included voter inference, vote selling, hateful content, and threats of gender-based harassment and violence against women candidates on Facebook and Instagram.

To help prevent disruption and reduce the risk of offline harm, our efforts centered on candidate safety, providing easy-to-use voter information and media literacy.

Candidate safety



We enrolled more than 3,000 candidates, including all federal-level and governorship candidates, in our [cross-check program](#) to help prevent enforcement mistakes and/or applied [Advanced Protection](#) to their accounts. This included monitoring for potential hacking threats. We developed “[Vote Against Violence](#),” an educational campaign in collaboration with nonprofits and media groups that aimed to deter gender-based violence online. This [campaign](#) reached over 1.2 million people on our platforms and was further amplified across other channels. Authorities sent [takedown requests](#) when they found violence or threats of violence against candidates.

Voter information

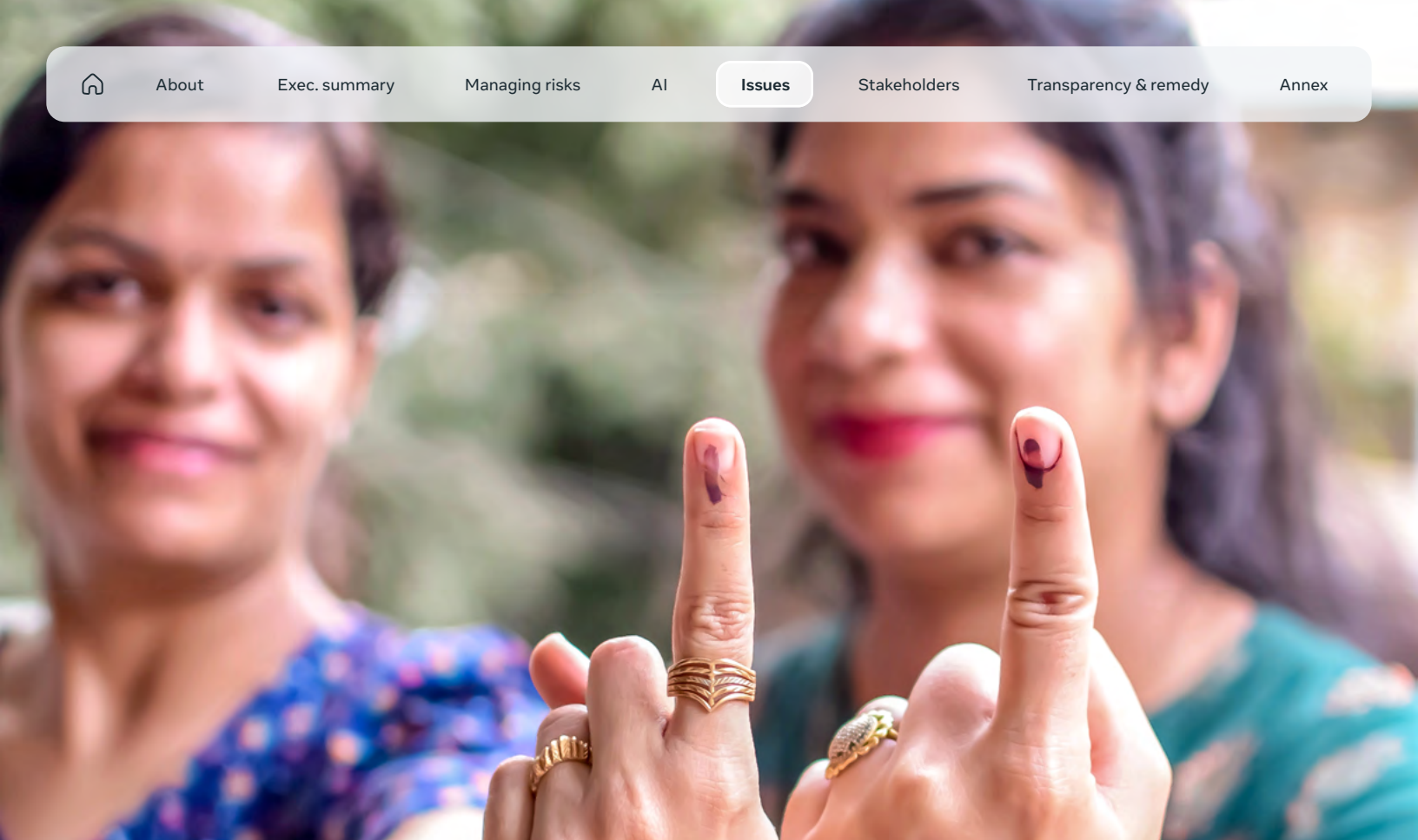


Together with the National Electoral Institute (INE), we launched the “Inés” chatbot on WhatsApp to assist voters. The chatbot answered questions about the electoral process, such as where and how to vote, how to process voter ID cards and the voting procedure for Mexicans living abroad. On voting day, we sent reminders on Facebook and Instagram and launched stickers on both apps to encourage voting.

Media literacy



To help prevent the spread of false information, we launched the “[Soy Digital](#)” (“[We Think Digital](#)”) [campaign](#) in collaboration with INE and Movilizadorio, a civil society organization. The campaign provided accessible learning modules and resources to build digital citizenship and information literacy skills — including how to stay safe online. The campaign reached more than 15 million people. We also trained 300 election district-level leaders who then trained thousands of poll workers in media literacy.



India

Meta began [preparing](#) for the 2024 Indian general elections 18 months in advance. The focus was on ensuring platform integrity and promoting voter education. Our approach was flexible and able to sustain a 60-day election period, during which more than 640 million votes were cast. Our preparations included:

Voter education and awareness



The Voting Alert notification launched from the Election Commission of India's Facebook page and reached 145 million people. The Election Commission of India also deployed the WhatsApp application programming interface (API) to run voting reminder campaigns, reaching around 400 million people.

Ensuring platform integrity



We took actions to prevent the misuse of our platforms. Content reviewers worked across content on Facebook, Instagram and Threads in more than [20 Indian languages](#) and English. We removed fake accounts and honored our commitments under the Voluntary Code of Ethics that we, along with other social media companies, joined in 2019.

Combating misinformation



We launched a dedicated fact-checking helpline on WhatsApp with the cross-industry Misinformation Combat Alliance (MCA) to combat AI-generated misinformation. We launched a [WhatsApp helpline](#) with the MCA, which set up a world-first [Deepfakes Analysis Unit](#) to assess any audio or video content that people suspected could be a deepfake. We also trained hundreds of law enforcement agencies with the MCA to combat deepfakes.



European Parliament elections

Meta's preparations for the European Parliament elections drew on key lessons learned from previous elections around the world, as well as the regulatory framework set out under the Digital Services Act and our commitments in the European Union (EU) Code of Practice on Disinformation.

Our EU-specific elections actions focused on:

Promoting voting information and civic engagement



We provided reliable election information and directed people to information on the electoral process through in-app "Voter Information Units" and "Election Day Information." People engaged with these notifications more than [41 million](#) times on Facebook and more than 58 million times on Instagram.

Tackling influence operations



Our [efforts](#) to stop coordinated inauthentic behavior focused on threats specifically associated with the European Parliament elections. We disabled several [networks that targeted the EU](#), including enforcement multiple times on the Russian-origin network known as Doppelganger.

Combating misinformation



We partnered with the European Fact-Checking Standards Network to help combat AI-generated and digitally altered media, and on a media literacy campaign to raise public awareness of related risks.

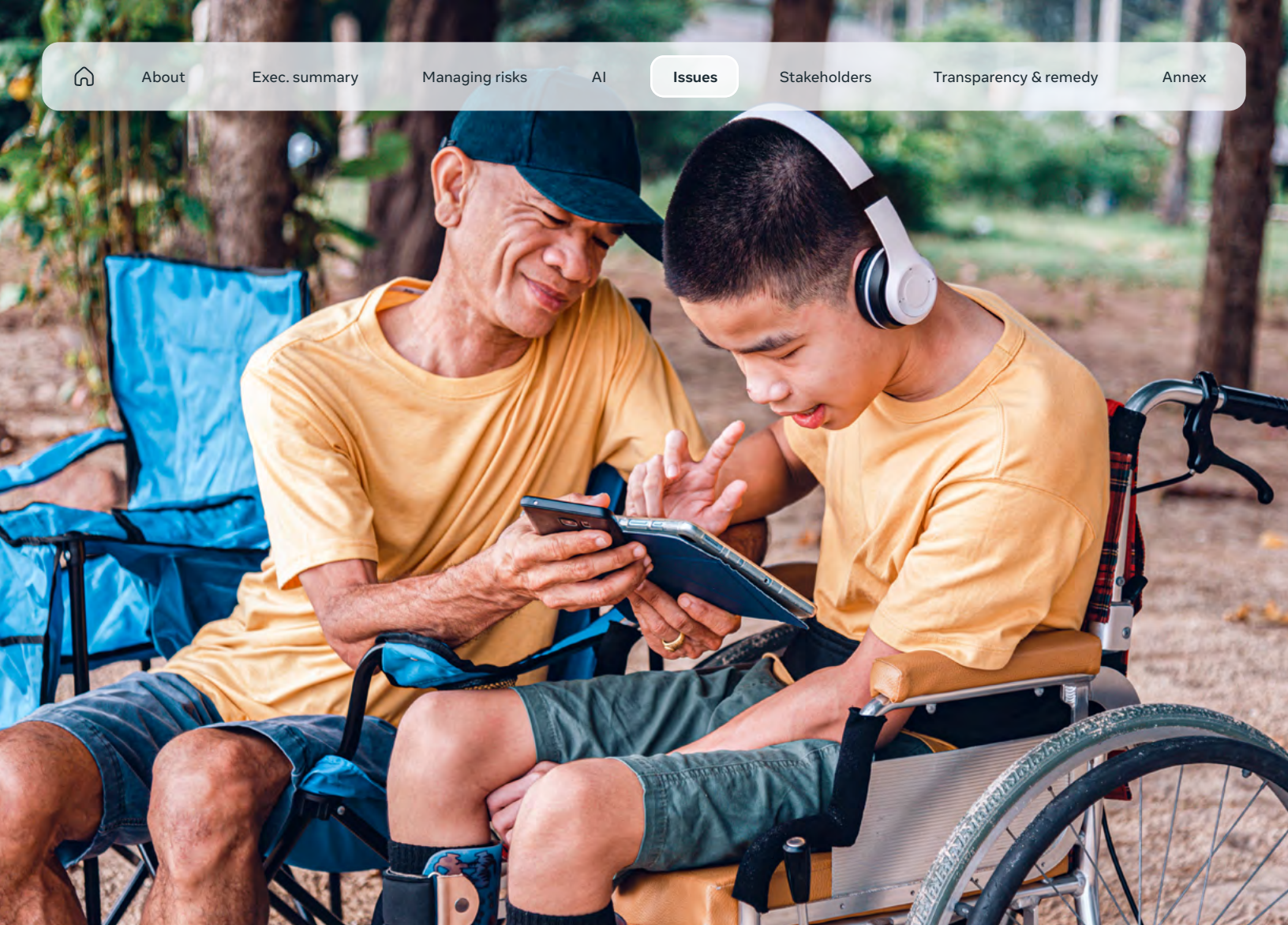
Countering the risks related to the abuse of generative AI technologies



As a result of the policies and measures we put in place to tackle generative AI content, nearly 6,000 ads about social issues, elections or politics and over 5.7 million pieces of content across Facebook and Instagram in the EU were labeled with AI-related disclaimers around the European Parliament elections, providing enhanced transparency.

See more details in our [Transparency Center](#).

[Go to Transparency Center](#)



Child and youth safety

Online child safety is a top priority for Meta. We offer built-in protections, as well as tools for teens and their parents, to help keep teens safe on our apps and services.

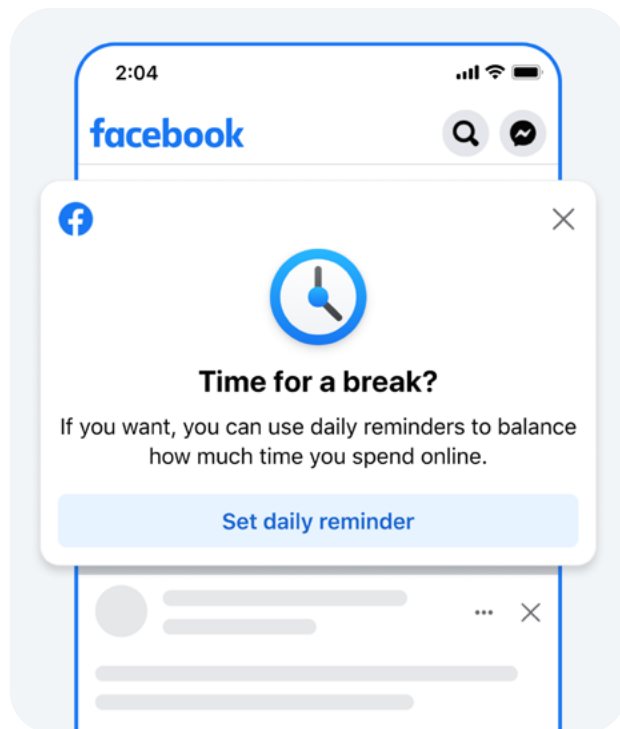
Built-in protections for teens

Keeping teens safe online requires collaboration across multiple stakeholders around the world, including parents, child experts, academics, industry peers, government, civil society and others. We remain committed to helping protect teens while providing space for them to exercise freedom of expression and access to information while being guided by their parents.

We've developed more than [50 tools and resources](#) over many years to support teens and their parents and guardians, and spent over a decade developing policies and technology to address content and behavior that break our rules.



In 2024, we updated our policies and product design to create a unique and more clearly differentiated experience for teens. These updates continue to help teens see [age-appropriate content](#) based on our [Best Interests of the Child Framework](#). We added to existing protections on Instagram by rolling out [Instagram Teen Accounts](#) in the U.S., UK, Canada and Australia, with global rollouts to follow. The redesigned Teen Accounts come with built-in protections to limit who can contact teens and the content teens see, and ways to help manage how much time teens spend on the app. The changes also provide new ways for teens to explore their interests, guided by parents. This new Instagram Teen Account experience is in line with expert guidance and with the principle of the evolving capacities of the child outlined in the [UN Convention on the Rights of the Child](#).



We developed and [launched](#) globally a parental supervision dashboard where parents and guardians who use our supervisory tools can see and manage the accounts belonging to their children — all in one place. This allows parents and guardians to set controls through their own accounts to [see](#) and manage any unwanted contact or inappropriate content, and set limits for screen time.

We also conducted a series of workshops in the U.S. around our [Screen Smart](#) program to help parents navigate conversations with their families about using devices safely and learn more about parental supervision tools from Meta, using boundaries and protections that work best for them.



“Meta’s new Instagram Teen Accounts go a long way towards empowering parents with the ability to guide their teens without taking away the autonomy of older teens. The new settings, along with enhanced safety and privacy tools and advice, are a major step forward.”

— Larry Magid, CEO, ConnectSafely





Fighting sextortion

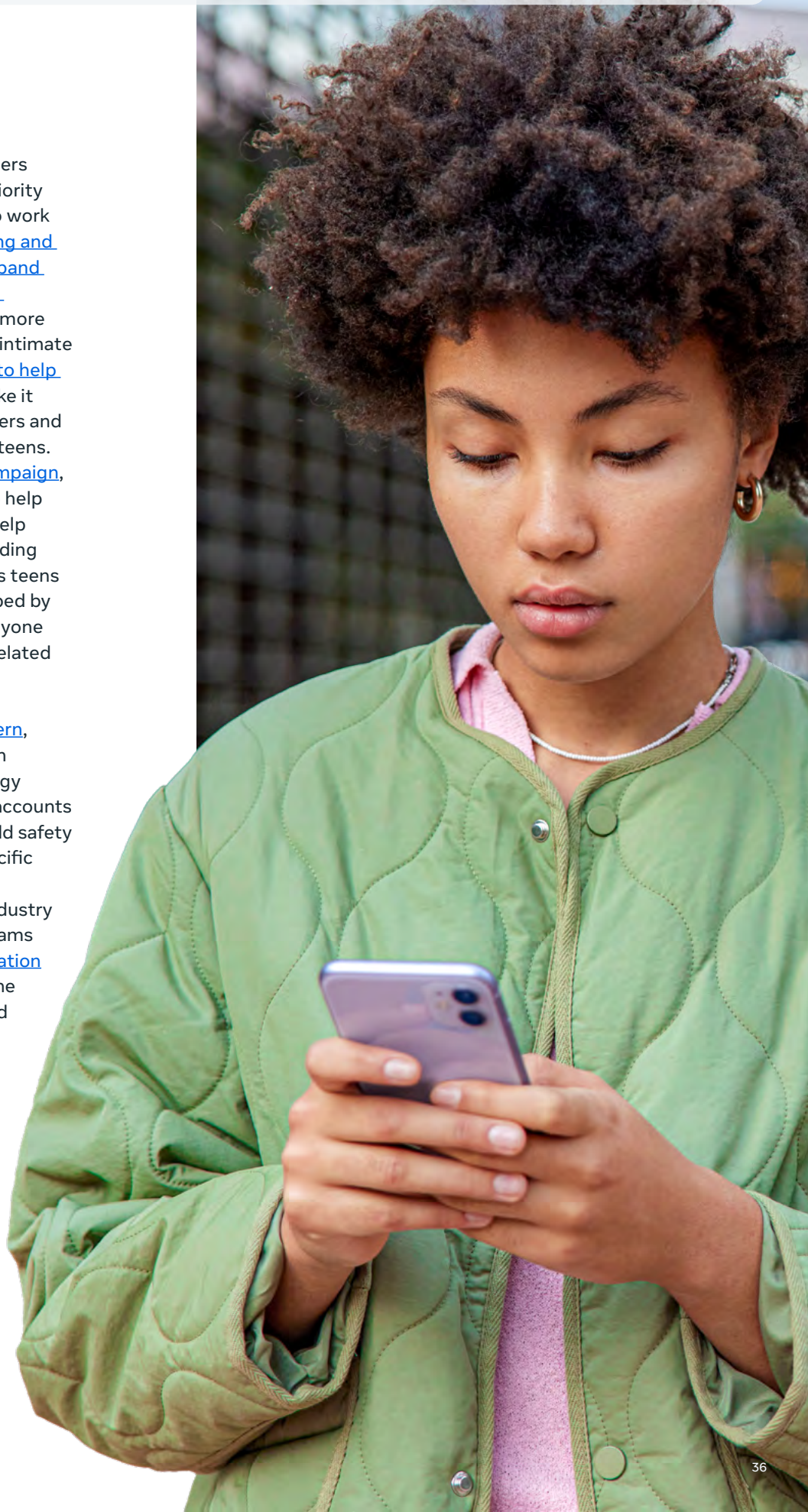
Helping keep children safe from users trying to do harm remains a top priority for Meta. In 2024, we continued to work with the [National Center for Missing and Exploited Children](#) (NCMEC) to [expand the Take It Down program to more countries and languages](#), allowing more teens to take back control of their intimate imagery. We developed [new tools to help protect against sextortion](#) and make it more difficult for potential scammers and criminals to find and interact with teens. We also launched an [education campaign](#), informed by NCMEC and [Thorn](#), to help teens spot sextortion scams and help parents support their teens in avoiding these scams. The campaign directs teens and parents to [expert tips](#), developed by Thorn and adapted by Meta, for anyone seeking support and information related to sextortion.

We are [founding members of Lantern](#), a program run by the Tech Coalition that enables us and other technology companies to share signals about accounts and behaviors that violate their child safety policies. We shared sextortion-specific signals to Lantern to build on this important cooperation between industry players to try to stop sextortion scams across platforms. In 2024, [participation](#) in the program doubled, bringing the total number of companies enrolled with Lantern to 26.

See [here](#) for a full list of our tools, features and resources to help support teens and parents.



[Read more](#)





How we prepare for and respond to crises

We prepare for and respond to many crises, including conflicts, intra-communal violence, civil unrest, mass protests and environmental disasters, as well as terrorist attacks and shootings, around the world. In 2024, we initiated and coordinated crisis responses in Bangladesh, Georgia, Kenya, New Caledonia, Nigeria, South Korea, the UK and Venezuela, among other countries and territories. We continued our efforts for designated crises under the [Crisis Policy Protocol](#) for the conflicts in Ukraine, Sudan and the Middle East.

The Crisis Policy Protocol is a key tool that we use during a period of crisis. The protocol guides our expedited use of levers to mitigate potential harm in the following areas:



Policy, such as issuing additional guidance to reviewers. An example is providing guidance on withholding strikes for certain violations of our violent and graphic content policies in order to avoid overly penalizing or restricting users who are trying to raise awareness of a conflict's impacts.



Product, such as changing the product experience. An example is changing the settings so only friends and family can comment on posts.



People, including moving resources to focus on specific issues.

The Crisis Policy Protocol helps us conduct an offline assessment of situations that may result in on-platform risks. Once designated, we conduct an assessment to identify on-platform risks and determine if any additional levers are required. The specific types of responses deployed are consistent with observed risks. They are informed by past crisis interventions, human rights principles and the law of armed conflict.

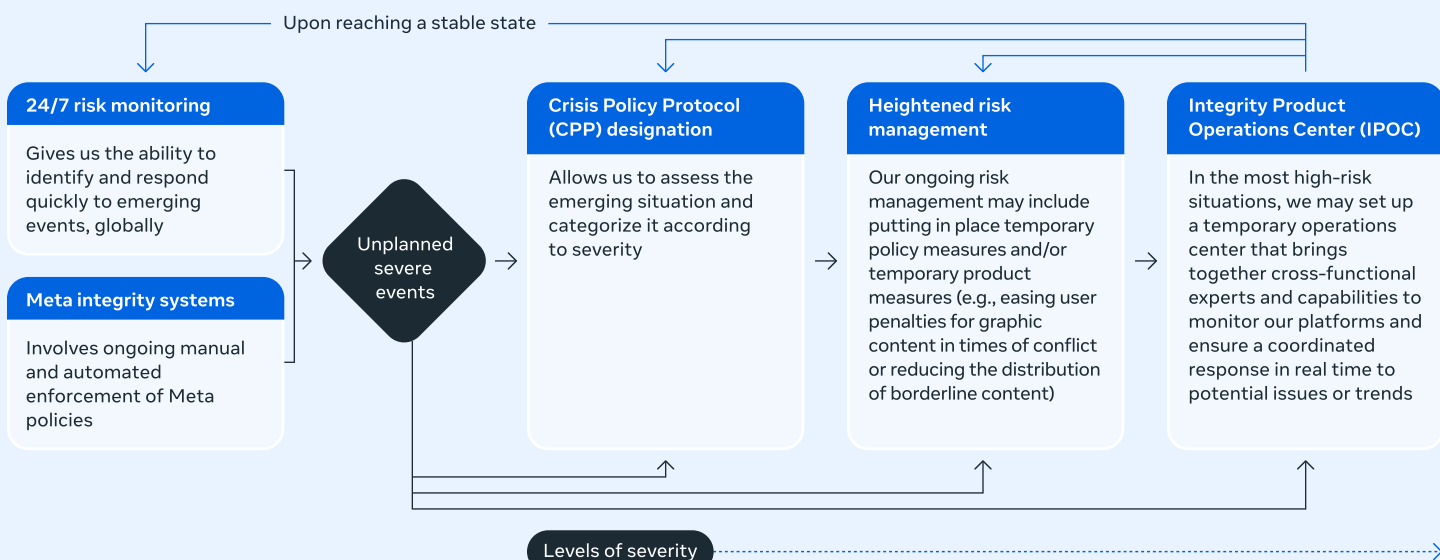
On the next page, we illustrate how we prepare for and respond to crises and conflicts. We also give examples to show the ways we use our Crisis Policy Protocol and the geographical diversity of our efforts.

Preparing for and responding to crises and conflicts⁴

Our Crisis Policy Protocol and work on countries at risk are key tools we use to prevent, detect and mitigate risks. Our product, policy and operations teams assess evolving on-the-ground dynamics to guide effective, proportionate responses.

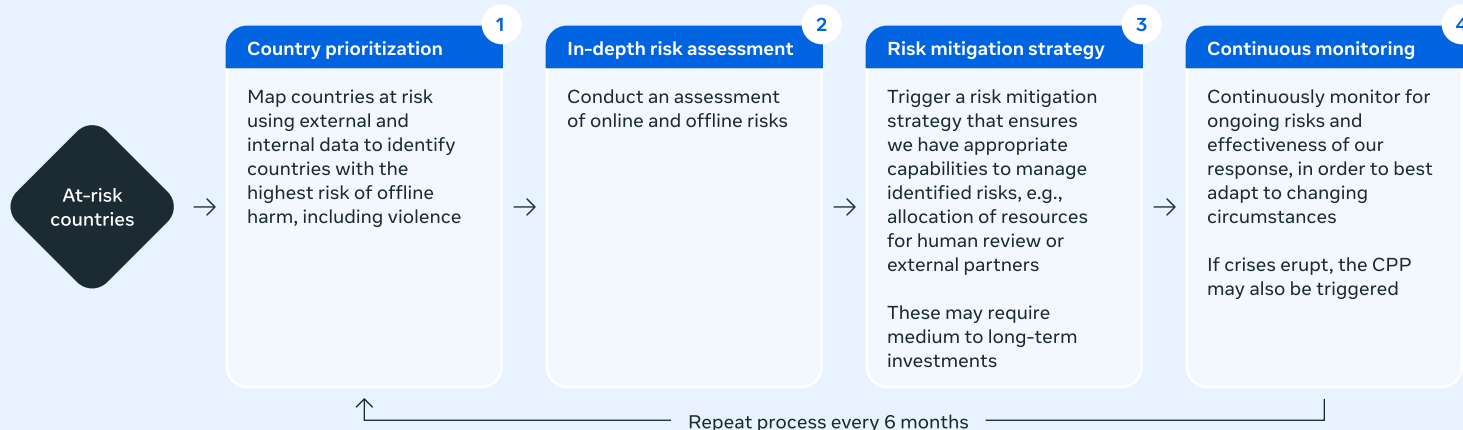
Reactive response

How do we respond quickly to unplanned severe events?



Long-term steps

How do we take long-term steps to mitigate risks of conflict?



⁴ Our crisis work includes many situations, including conflicts, intra-communal violence, civil unrest, mass protests and environmental disasters, as well as terrorist or other criminal attacks, around the world.



Sudan

In 2024, Sudan’s conflict between the Sudanese Armed Forces (SAF) and the Rapid Support Forces (RSF) escalated further, exacerbating the nation’s instability and humanitarian crisis. Volumes of violating content, including violence and incitement, coordinated harm, human exploitation and dangerous organizations and individuals, increased from pre-conflict levels and remained high throughout the year.

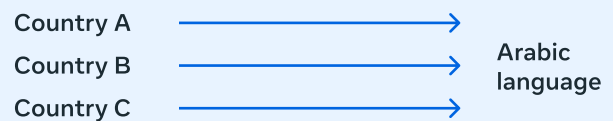
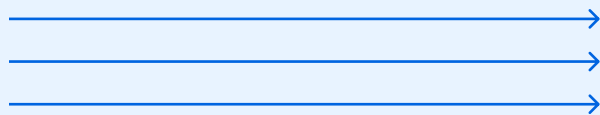
To reduce the prevalence of policy-violating content, we built upon the actions we [took in 2023](#), guided by our Crisis Policy Protocol. As the conflict continued, we deployed temporary measures and developed long-term mitigations to address the risks of ongoing high volumes of violating content.

One of these long-term mitigations was to design, build and launch a system that identifies specific Arabic dialects and prioritizes the content to reviewers who are more likely to understand the linguistic nuances and local context. The previous system identified Arabic as a single language and directed it to moderators with capacity to review content. The new system can identify the particular dialect of Arabic used, and directs the content to the reviewer most likely to understand it. For Sudan, this change resulted in higher volumes of content being reviewed at higher precision, reducing errors in enforcement. This work was informed by and built on outcomes from the [Israel Palestine Human Rights Due Diligence](#).

Routing for linguistic nuance

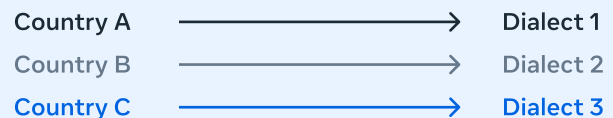
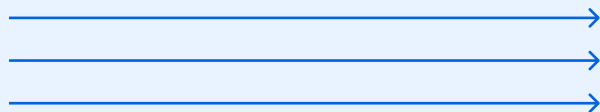
Before

Content to be reviewed



After

Content to be reviewed





During 2024, both sides of the conflict increasingly outed identities of prisoners of war (POWs) online. Revealing their identities increased the risk of real-world harm and undermined the protection of the dignity and safety of POWs under [the Geneva Convention relative to the Treatment of Prisoners of War](#). Guided by an [Oversight Board recommendation](#) made in 2023 and again in 2024 in the [Sudan Rapid Support Forces Video Captive case](#), we also recognized that some POW content could be in the public interest, for example raising awareness about potential human rights abuses or helping locate missing POWs. As a result, Meta provided guidance to content reviewers on POWs in the [Coordinating Harm and Promoting Crime policy](#) so that they were better able to address potentially violating content in the region at scale.

[↗ Read H1 2024 Oversight Board Report](#)

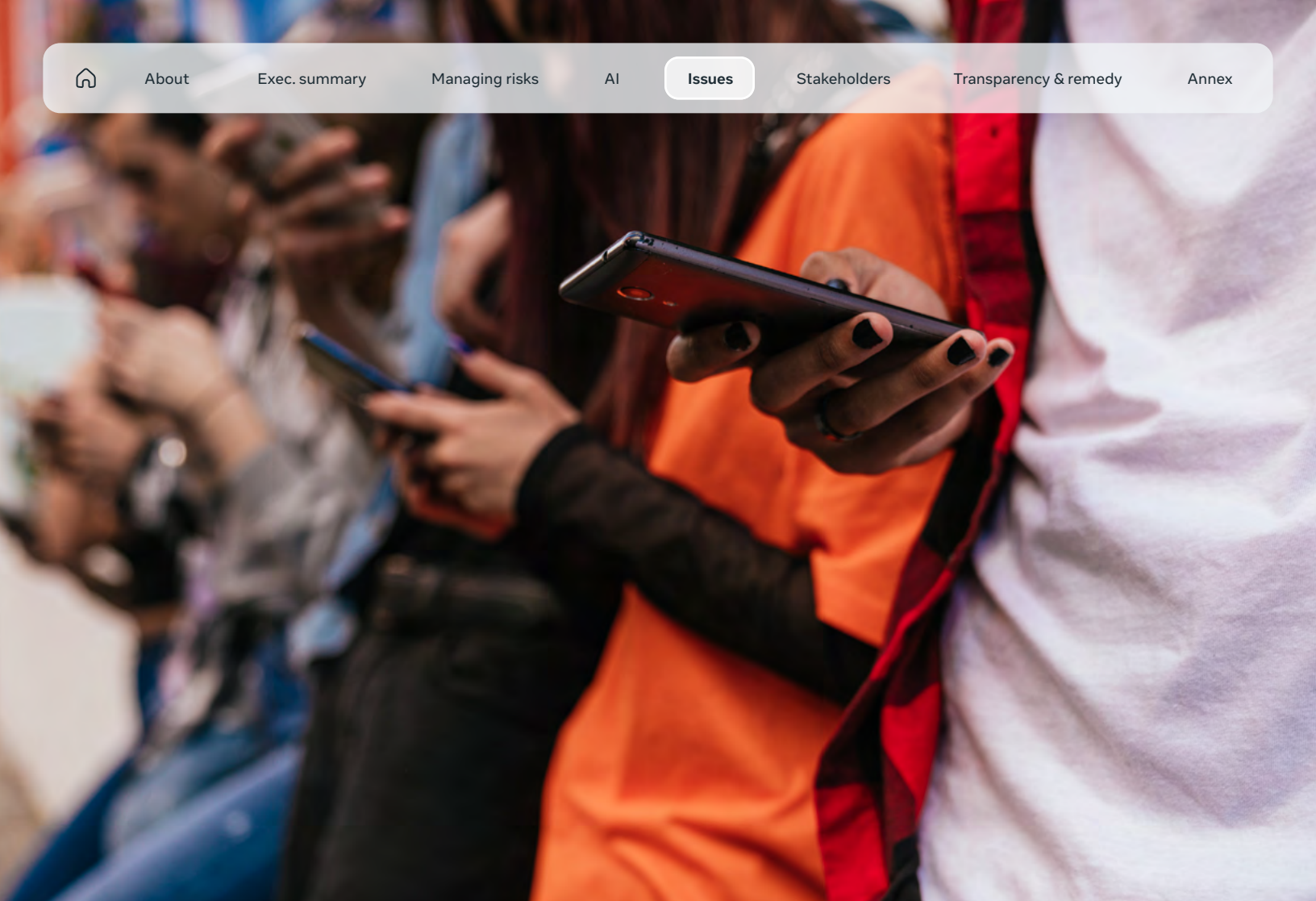
[↗ See Oversight Board case](#)



Trusted Partners played a crucial role by providing critical insights on local developments and potentially violating content related to the conflict. These insights helped enforce the relevant Meta policies, including Hateful Conduct, Bullying and Harassment, Human Exploitation and the designation of pre-reviewed potentially harmful claims under the [Misinformation and Harm policy](#), and ultimately contributed to a safer online environment. Under our Human Exploitation policy, we were able to identify potential risks related to child soldier imagery and recruitment, and remove that content as well as reduce its prevalence.

We conducted training sessions with human rights defenders, journalists, and national and diaspora organizations to help them educate Sudanese users, including migrants and refugees. These sessions focused on content policies, digital security and enhancing their presence on Meta platforms.

Armed conflicts trigger mass displacement of people, so in Sudan we focused on identifying potential human exploitation, including human smuggling and trafficking, sexual exploitation of women and girls, and forced marriage. We removed over 19,100 group posts offering human smuggling services and removed content glorifying forced marriage. Engaging with the diaspora continued to be an important strategy for surfacing content trends and providing clarity for country-level monitoring. This included over 30 unique engagements to help safeguard users on our platforms, including efforts to identify new and emerging words and phrases related to hate speech and human smuggling. These insights enabled us to more effectively detect and respond to content that violated our policies.



Middle East

The conflict in the Middle East remained a priority for Meta. In 2024, we focused on the risks stemming from the ongoing violence in Israel and Gaza as the war expanded to the wider region and as other regional actors became further involved and escalated the conflict. We worked to help ensure that our platforms could be used for freedom of expression, while seeking to prevent the spread of content inciting terrorism, violence and other real-world harms.

Our core approach did not change from 2023. This approach included maintaining the designation of the October 7, 2023, attack by Hamas as a Terrorist Attack under our [Dangerous Organizations and Individuals \(DOI\) policy](#) and addressing violating content under our policies. We discontinued the temporary [product changes](#) introduced in 2023.

In the immediate aftermath of the October 7 terrorist attacks, Meta designated the violence and ensuing conflict at the highest level of our Crisis Policy Protocol and implemented immediate crisis response measures. Actions taken included a dedicated 24/7 cross-functional team and temporary product and policy measures. We looked to the [UN Guiding Principles on Business and Human Rights](#), as foundational to our [Corporate Human Rights Policy](#), and our [diligence work in 2022](#) to inform our approach. Details of our response can be found in our [2023 Human Rights Report](#) and [Newsroom posts](#).



Throughout 2024, we continued to engage with a variety of actors from government, civil society and others across Israel and Arab countries in the Middle East, but also globally, to demonstrate transparency and responsiveness. We also responded to multiple [Oversight Board cases](#).

In addition, we continued implementing the recommendations from the [2022 Human Rights Due Diligence](#) report. We reported [Meta's progress](#) for the period of June 30, 2023, to June 30, 2024, including an increase to our content moderation resources in the Hebrew language and an [update](#) to our DOI policy to allow for more social and political discourse in certain situations. This was in response to feedback that too often our DOI policy captured content such as news reporting, neutral discussion of current events or even condemnation of terrorist and hate groups. Content that praises or supports dangerous organizations or individuals or their violent actions or missions is still prohibited.

Between June 30, 2024, and June 30, 2025, we launched a system that detects and prioritizes content to moderators most likely to understand that particular Arabic dialect. We also revamped our Trusted Partner escalation channel, leading to an improved rapid response to escalations. Details of our progress for this period can be found in our [December 2025 Final Update: Israel Palestine Human Rights Due Diligence](#).

[2023 progress](#)[2024 progress](#)

Bangladesh

Our preparations for the January 2024 elections helped us anticipate and address multiple risks during the reporting period. Our goal was to help safeguard users while supporting their ability to vote and express themselves. These electoral preparations enabled us to respond mid-year to the student protests, violent crackdown and subsequent change of government.

Given the severity of unrest, we deployed our Crisis Policy Protocol. We proactively identified risks, including hate speech, incitement targeting minority religious communities, misinformation and coordinated inauthentic behavior. We put into place mitigations, including the use of our [Temporary High-Risk Locations policy](#), engagement with our Trusted Partners and third-party fact-checking network, and the application of enhanced protections to human rights defenders' accounts.





Our other actions included:



Establishing precise detection signals to identify content spikes related to violations that could be enforced in real time, such as graphic violence and hateful conduct.



Applying tools and techniques including AI detection to identify and enforce on violating content and keyword searches.



Designating additional pre-reviewed potentially harmful claims under the [Misinformation and Harm policy](#).

We did not comply with government takedown requests relating to content about the protests that were inconsistent with international human rights standards. This was in line with our commitments as a member of the [Global Network Initiative](#) and our Corporate Human Rights Policy.

Georgia

We deployed the [Crisis Policy Protocol](#) twice for Georgia in 2024. In March 2024, we first implemented it after a series of mass demonstrations in opposition to the proposed Law on Transparency of Foreign Influence. We reactivated it in December 2024 after the national elections, when further mass demonstrations took place accompanied by escalating violence from police and other security forces.

Deploying the Crisis Policy Protocol allowed our team to enhance risk mitigation efforts, address spikes in violating content and increased risks of physical violence, and help protect human rights defenders. We audited our slur list — words that are historically used to attack certain groups — to identify and manage hateful content across our platforms. We took down fake accounts that were designed to manipulate public opinion or distribute potentially harmful content. We also disrupted a coordinated inauthentic behavior (CIB) network targeting Georgia, as well as other inauthentic accounts.

Throughout the crisis periods, we collaborated with civil society organizations, fact-checkers and Trusted Partners, who helped us understand the developing situation and facilitated information-sharing with wider civil society and the opposition in Georgia. Trusted Partners provided critical signals and insights into violating content targeting opposition groups. We also engaged with Trusted Partners to help them better understand what constitutes violating content according to our Community Standards. In addition, we reached out to civil society partners to identify human rights defenders at risk to help ensure enhanced account protections.



Cybersecurity

Our security policies are important for users' rights to freedom of expression, access to information and privacy, among other rights. We continued to work across the company to identify and defend against adversarial platform threats, including influence operations, cyber espionage, surveillance, and fraud and scams. An important element of our security work is to disrupt adversarial networks that engage in malicious activity.

In 2024, we took down [20 CIB networks](#) in the Middle East, Asia, Europe and the U.S. for violating our [Coordinated Inauthentic Behavior policy](#). These are networks that worked to manipulate public debate for a strategic goal using fake accounts or misleading tactics. We monitor for and remove attempts by networks we previously removed to reconstitute on our platforms, publicly share information through our [threat reports](#), and work to build insights from investigations back into our detection systems and product design to make them more resilient.

We continued to detect and remove CIB networks that targeted and/or posed as specific ethnic or religious groups. In 2024, one of many examples was a network originating in Bangladesh that was removed for coordinated inauthentic behavior. It targeted domestic audiences using fake accounts to post content and manage pages. The network posed as fictitious news entities and used names of existing news organizations to spread anti-Bangladesh-Nationalist-Party content and support the ruling party. The operation was linked to individuals associated with the Awami League party and a nonprofit in Bangladesh.



Another example included a network from China targeting the global Sikh community, using compromised and fake accounts to pose as Sikhs and promote a fictitious activist movement called Operation K, which called for pro-Sikh protests, including in New Zealand and Australia. The operation used AI-generated images and posts, in English and Hindi, about floods in the Punjab region, the Sikh community worldwide, the Khalistan independence movement, the assassination of Hardeep Singh Nijjar and criticism of the Indian government.

[Read more](#)

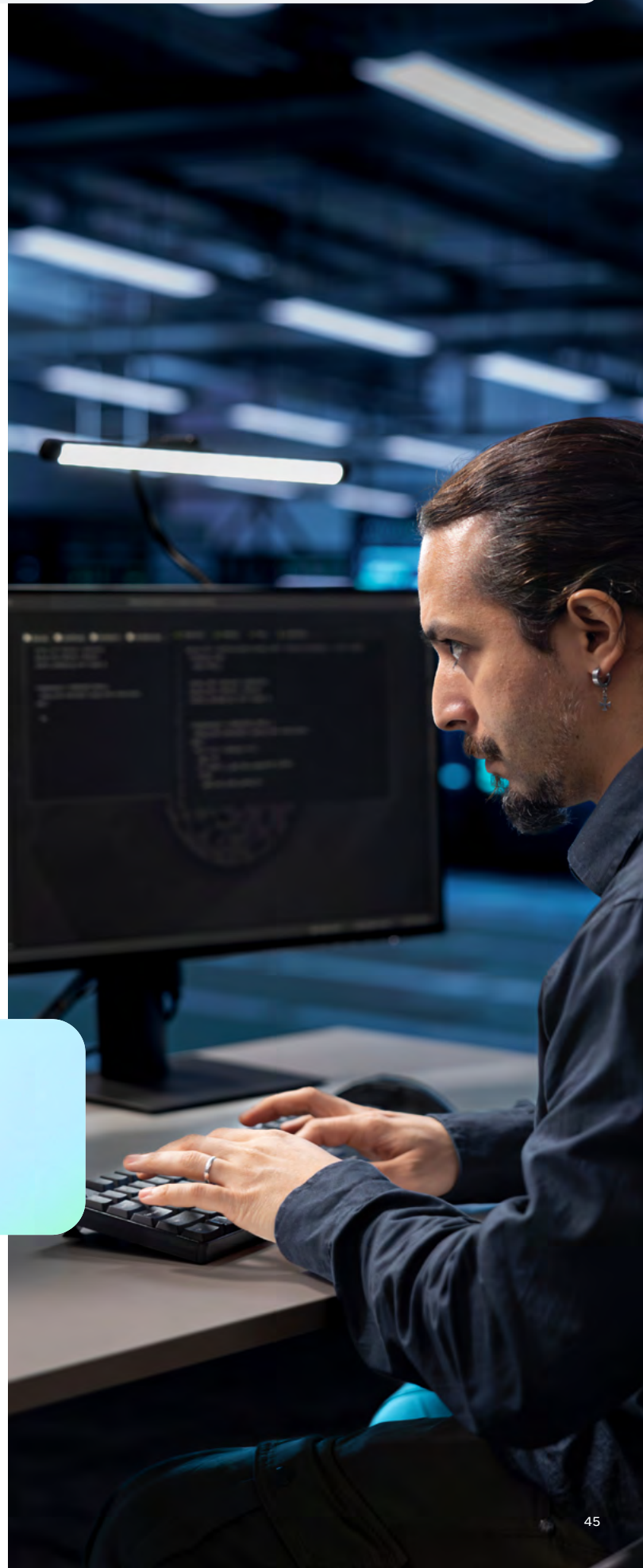


As part of our enforcement efforts against spyware companies, we disrupted and took down activity by Paragon Solutions, a spyware vendor that was targeting a number of users on WhatsApp, including journalists and members of civil society. We reached out to WhatsApp users who may have been affected and provided them with resources on how to protect themselves. We also provided them with information about [The Citizen Lab](#) at the University of Toronto, which provides additional resources for members of civil society. In 2024, we were a founding signatory to the [Pall Mall Memorandum](#), a multinational effort to restrain the abuse of spyware.

In December 2024, a U.S. federal judge [found NSO Group liable](#) for violating state and federal laws, and breaching WhatsApp's terms of service. This was the first time a spyware company was found liable under U.S. law. Meta and WhatsApp filed this case in 2019 against NSO Group, which had accessed WhatsApp servers without authorization in order to install Pegasus spyware on the mobile devices of more than 1,400 users of WhatsApp, including journalists, human rights activists, political dissidents and others.

20

CIB networks taken down





Stakeholder engagement

Proactive, structured [engagement](#) with our global community of users helps shape Meta policies and is central to our human rights risk management.

We engaged with a broad range of stakeholders in 2024, including civil society members, academics, think tanks, human rights experts and regulators. Key policy questions included our approach to responsible artificial intelligence (AI) and election integrity, as well as our designation signals for dangerous organizations and individuals, and violent events.

For example, to assess whether our policy related to [the word “Zionist”](#) was appropriate, we completed consultations with 145 stakeholders from civil society and academia globally. Participants included political scientists, historians, legal scholars, digital and civil rights groups, freedom of expression advocates and human rights experts. We also interacted with stakeholders including non-governmental organizations from our [Trusted Partner program](#) as well as a wide range of diaspora communities representing different viewpoints.



In 2024, we created voice and expression working groups with local civil society organizations for the Sub-Saharan Africa and Middle East and North Africa regions to understand their concerns with legislative proposals in the Kingdom of Saudi Arabia, Jordan, Nigeria and Senegal, among other countries. During these sessions, we explored how to safeguard access to our platforms while navigating content restrictions based on local law and our [Global Network Initiative](#) commitments to uphold freedom of expression and user privacy.

We also piloted a Human Rights Commissions program, which included national human rights institutions from Ethiopia, Ghana, Kenya, Nigeria and South Africa and focused on how Meta addresses potentially harmful content and online content regulation.

In addition, we conducted conflict response workshops in Ethiopia, Palestine, Somalia, Sudan and Tunisia. We trained human rights defenders and journalists from countries holding elections, equipping them with the tools to protect their digital presence.

Through our [Open Loop India program](#) and [Open Loop Sprint](#) work, we collaborated with companies, policymakers and AI experts to produce insights about the role of stakeholder engagement across the AI lifecycle and value chain.

Our approach to stakeholder engagement



Incorporate a broad range of perspectives and expertise: Uncover important insights and engage subject matter experts in all regions for diverse global perspectives as well as local nuances.



Provide transparency: Talk through challenges and improvements with external stakeholders.



Create a feedback loop: Show how our policies evolve over time.



Build trust: Engender legitimacy for our policies and enforcement.



464

stakeholders from **34** countries contributed to **6** Policy Forum workstreams.

121

stakeholders contributed to other Meta development work.

100+

stakeholders engaged in election briefings, with **7** election newsletters produced.

290+

journalists, human rights defenders and activists were trained.

Meta policy development cycle

Constant review

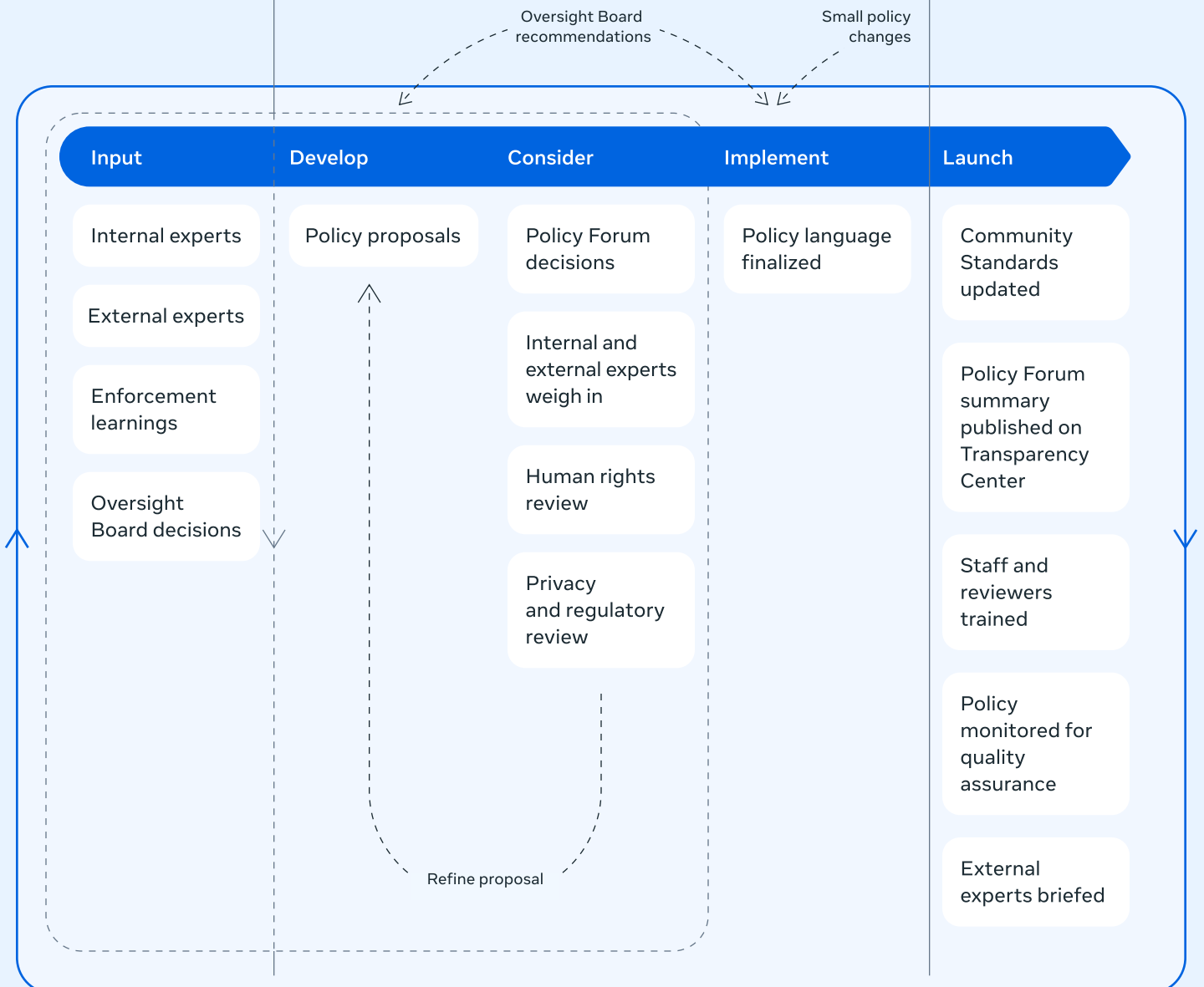
We constantly review our policies based on inputs from a variety of sources.

Development

Proposals pass through a rigorous development process to ensure they are principled, operable and easy to explain.

Launch

Enforcement systems are updated and policies “go live” on our services.





Policy Forums

We seek to craft policies that respect human rights and seek to embrace diverse perspectives, where multiple views and beliefs can be heard and reflected. Meta's [Policy Forum](#) is a regular meeting where subject matter experts discuss potential changes to Community Standards and Advertising Standards. These meetings involve proposing new policies or amending existing ones, following a policy development process that includes extensive engagement with global stakeholders and a review of both external and internal research.

We conducted six Policy Forums in 2024:

1. [“Zionist”](#) as a proxy term for hateful conduct
2. Violating violent events
3. Commercial content with potential health and safety risks
4. Removing sensitive imagery
5. Disordered-eating content
6. Condolences for designated dangerous individuals



Community Forums

Meta's Community Forums are rooted in deliberative governance and are designed to leverage public input on issues where there are competing tradeoffs and no clear answers. Our approach empowers voices outside the company to have a greater say in our decision-making and allows us to see how public opinion may evolve in the future.

In 2024, Meta held a Community Forum, in partnership with [Stanford's Deliberative Democracy Lab](#), focused on principles users want to see underpin the development of AI agents. The forum engaged approximately 1,000 people across India, Nigeria, Saudi Arabia, South Africa and Türkiye. A detailed report is available [here](#).

As a part of the forum, participants were able to hear directly from subject matter experts, deliberate with one another and provide valuable feedback to Meta. The deliberative method enabled participants to grapple with the tensions inherent in providing personalized experiences, weighing the value of personalization with tradeoffs such as data collection and storage.

Our approach to user controls and personalized experiences

The findings have informed our approach to user controls and personalized experiences with AI agents. These included:



Participants supported AI agents remembering their prior conversations to personalize their experience, especially if transparency and user controls were in place.

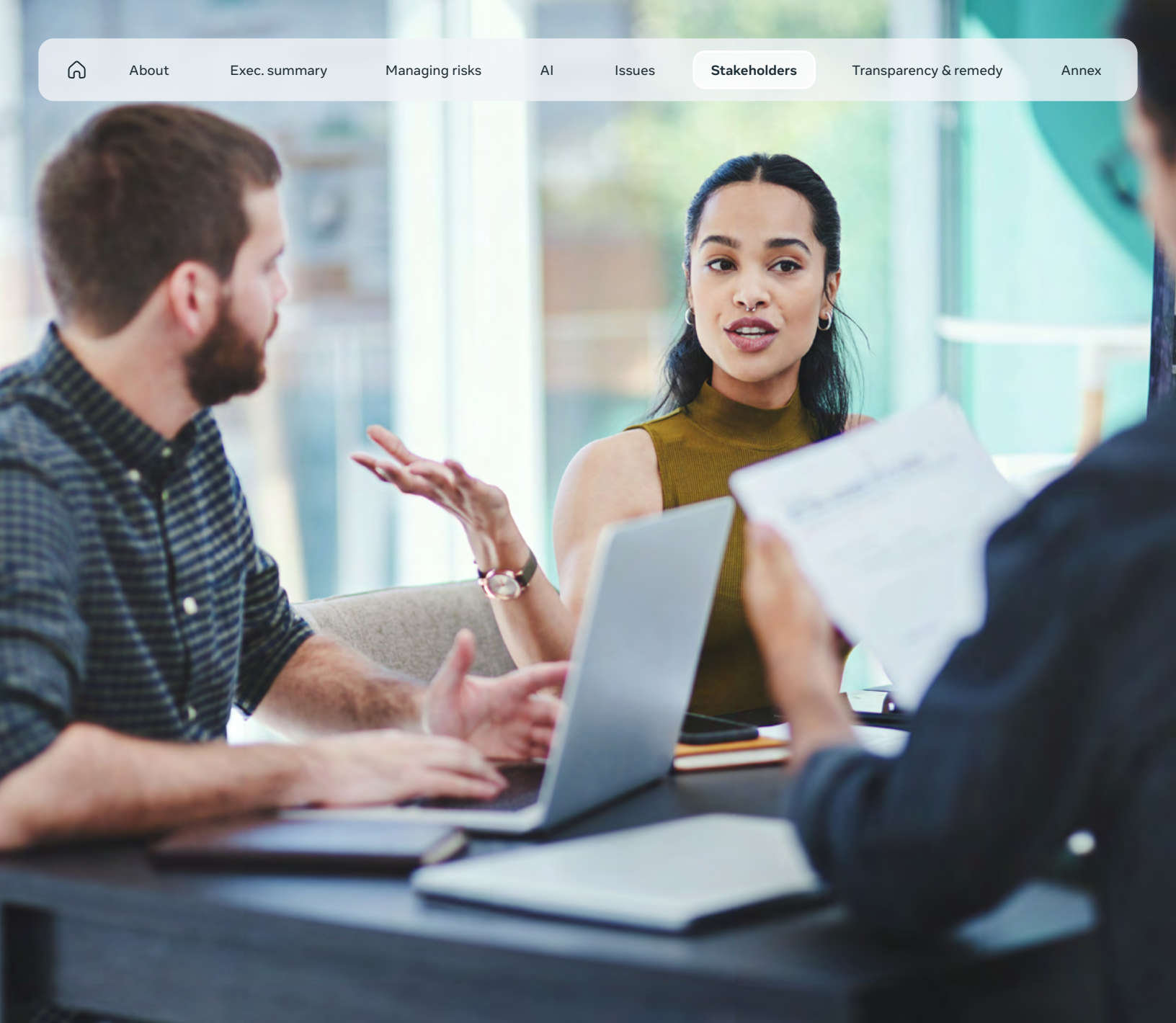


Participants were more supportive of culturally/regionally tailored AI agents compared to standardized AI agents.



Participants were in favor of human-like AI agents that can respond to emotional cues.

In addition, we started a pilot to engage the public on what they think contributes to a culturally relevant AI model, develop preference datasets based on this feedback and open source the data for developers' use. The result would be a readily available collection of datasets to make our Llama large language model more relevant and helpful in different cultural contexts.



Trusted Partners

We continued to engage our [Trusted Partners](#) to identify trends, better understand the impact of online content and behavior on local communities, and explore how we can strengthen our civil society escalation channels.

Trusted Partners are important allies in identifying high-severity violations of our Community Standards and were particularly helpful during 2024, the year of elections. They provided insights and identified harmful content in countries that experienced heightened unrest. These included Bangladesh, Brazil, Côte d'Ivoire, Democratic Republic of Congo, France, Greece, India, Indonesia, Kenya, Kurdistan-Iraq, Mexico, Nigeria, Pakistan, Senegal, South Africa, Syria and Venezuela, among other countries and regions.



In 2024, we removed more than 100,000 pieces of policy-violating content as a result of our Trusted Partner program.

Trusted Partners provided insights into elections-related content trends to inform integrity efforts, help us detect and remove violating content, and identify high-risk users of our platforms for [additional protections](#). Trusted Partners were effective at identifying spikes in hostile speech targeted at marginalized communities and attacks on journalists and human rights defenders, as well as misuse of AI content.

To manage the risk of hateful conduct, we removed designated slurs. We worked with our Trusted Partners to better understand the context in which slurs were used, so we could enforce our policies more accurately.

We consulted more than 40 Trusted Partners across 20 countries to inform policy and product development processes related to removing sensitive imagery, human exploitation, dangerous organizations and individuals designation signals, [“Zionist” as a proxy term](#) for hateful conduct, AI chatbots and more.

In response to the Oversight Board’s [recommendation](#), Meta evaluated the [timeliness and effectiveness](#) of responses to content reported through the Trusted Partner program. Over a two-year period spanning from Q2 2022 to Q4 2024, Meta made substantial improvements in the response time to content reported through the Trusted Partner program.

Investment in training, streamlined enforcement systems and new tooling helped to improve the volume of reports and efficiency of review in 2024.

Global results

Globally, the Trusted Partner program received over **11,800** reported pieces of content in Q2 2022, which increased to over **49,200** reported pieces of content in Q2 2024, reflecting a **fourfold** increase.

Trusted Partner channel’s global growth across 2 years

Q2 2022 - Q4 2024

+4x

Pieces of content reported through Trusted Partners Program

+12pp

Percentage of cases resolved within 5 days of escalation

+15%

Efficiency gain on median turnaround time in days

+15x

Reported content for further policy review

Examples of Trusted Partner impact in Pakistan, Syria and Venezuela can be found on the following pages.

CASE STUDY

Reporting on insights from Syria



In the aftermath of the collapse of the [Assad regime](#) in December 2024, Trusted Partners played a critical role in reporting and analyzing local developments, providing insights on local content trends and escalating high-severity violations.

Reporting by Trusted Partners, with their local expertise, informed Meta's crisis response efforts and enabled us to enforce our policies, as well as to mitigate risk in a more timely and efficient manner. Trusted Partners raised concerns about outing risks and claims of affiliation with the ousted regime targeting ethnic and religious minorities, including the Alawite, Christian and Kurdish groups, and flagged the rise of different extremist factions within the former Syrian Army. These insights supported our efforts to mitigate the risk of [dangerous organizations and individuals](#) on our platforms and of physical attacks based on personal characteristics.



CASE STUDY

Mitigating risks for civic actors in Venezuela



In the run-up to the July 28, 2024, elections in Venezuela, we worked with our Trusted Partners and built new relationships with civil society organizations to prepare for elections-related risks and to increase reporting of potentially violating content.

In the post-election period, protests broke out, followed by the government's repressive response. This included mass detentions and targeted arrests of political opponents. Our Trusted Partners provided critical insights into developments on the ground. They reported harmful content, including veiled threats and outing of protestors and opposition supporters, which placed them at risk of arbitrary detention and physical harm. Additionally, Trusted Partners reported attacks on the accounts of civic actors, such as journalists, opposition members and human rights defenders, among others.

These insights informed proactive detection and helped us enable [Advanced Protection](#) to these accounts and prevent enforcement mistakes through applying cross-check. These measures helped support news reporting and civic engagement in a repressive environment.

CASE STUDY

Trusted Partners tackle blasphemy allegations and hostile speech in Pakistan



In Pakistan, Trusted Partners played an important role in alerting us to potentially harmful content targeted at marginalized communities, including religious and gender minorities.

During the February 2024 election period, Trusted Partners reported elections-related content with a focus on hostile speech [targeted at political candidates](#) and blasphemy allegations that could amount to incitement. In Pakistan, allegations of blasphemy may trigger legal proceedings and physical violence.

As a result of these reports, we were able to remove content related to blasphemy allegations under our [Coordinating Harm and Promoting Crime policy](#).

Trusted Partners also provided signals and insights during other critical moments, such as outbreaks of sectarian violence. Their work enabled us to respond swiftly, remove violating content on our platforms, and strengthen our detection and enforcement.



Engaging stakeholders in Pakistan

Meta undertook a series of engagements in Pakistan with several government and non-governmental stakeholders as part of our human rights due diligence, with a focus on balancing safety with freedom of expression. Highlights included:



A roundtable discussion on online youth safety, cohosted with the Ministry of Human Rights, National Commission on the Rights of Children, National Commission on Human Rights and Digital Rights Foundation. We discussed Teen Accounts and the launch of the [Take It Down portal](#) in Urdu for Pakistani users.



Engagement with a diverse group of human rights defenders to gather insights about internet disruptions and explore possible avenues for collaboration on advocacy. They provided valuable insights into the impact of the “firewall” and whitelisting of virtual private networks (VPNs) by the government.



A roundtable discussion with civil society organizations to provide a deep dive into Meta’s key human rights commitments and the work of our human rights team. This included a discussion on ways to respond to situations without shutting down the internet or throttling social media platforms, including our family of apps.

Every interlocutor, at every event, spoke to the impact of targeting users with adversarial and frivolous accusations of blasphemy. They were reassured when they learned of Meta’s outing risk policy and ongoing work to keep targets safe.





International organizations

In 2024, member states of the [United Nations](#) negotiated and adopted the [Global Digital Compact](#) (GDC) — a comprehensive framework for global governance of digital technologies and AI. We worked with the UN member states, UN agencies and industry coalitions to finalize the text of the GDC. Our work aimed to support freedom of expression while creating a safer, more inclusive and open digital future for everyone.

Meta also engaged across the UN system in other ways throughout the year. Our work included contributing to [UNICEF’s Digital Technologies, Child Rights and Well-Being](#) for due diligence in the tech sector, and with [UNESCO](#) on digital platform governance on disinformation. We also [supported](#) UNESCO with a [translation interface](#) built on Meta No Language Left Behind (NLLB) AI model to support high-quality translation in 200 languages. These included marginalized languages like Asturian, Luganda, Maori, Swahili and Urdu, helping promote linguistic diversity and access to information.

Meta continued to work closely with the [Office of the High Commissioner for Human Rights](#) (OHCHR). We met regularly with OHCHR staff and actively engaged in the [B-Tech Project](#), which provides authoritative guidance and resources for implementing the [UN Guiding Principles on Business and Human Rights](#) in the technology sector and its [Community of Practice](#), a space for confidential dialogue with other tech companies. We actively participated in ongoing discussions around AI and human rights standards. In addition, we participated in the 2024 [UN Forum on Business and Human Rights](#) and presented at panels on “Online Hate Speech” and “Protecting Press Freedom.”



Meta joined policy discussions on the sidelines of the Summit of the Future and the 79th United Nations General Assembly. Topics included the role of AI in global governance, empowerment of digital creators, economic diaspora-led innovation, and the impact of cybercrime and content laws on freedom of expression. We also participated in discussions around the protection of human rights defenders and the use of social media to provide lifesaving information in humanitarian crises.

In addition, we consulted with UN Human Rights Special Procedures — independent human rights experts — including the Special Rapporteurs for Freedom of Opinion and Expression and for Human Rights Defenders, among others.

Throughout the year, Meta collaborated with the G7, G20, UNESCO and the OECD on workstreams related to AI inclusion and governance. We also joined conversations with governments around the importance of information integrity. We continued our participation in the [World Economic Forum's](#) Global Coalition for Digital Safety, which resulted in the publication of [The Intervention Journey: A Roadmap to Effective Digital Safety Measures](#).

In addition, we actively participated and engaged with stakeholders in several multistakeholder forums, including the [Eradicate Hate Summit](#), [Forum on Internet Freedom in Africa](#) (FIFAfrica), [Global Internet Forum to Counter Terrorism](#), [Internet Governance Forum](#) (IGF), [RightsCon](#), [Tech Against Trafficking Conference](#) and the [UN Commission on the Status of Women](#).

Through our membership in the [Global Network Initiative](#) and participation in the [Digital Trust and Safety Partnership](#), Meta attended the “European Rights and Risks: Stakeholder Engagement Forum,” which helped inform our systemic risk assessments under the [Digital Services Act](#).





Transparency & remedy



The [Oversight Board](#), as an independent body, helps us resolve some of the most difficult questions around freedom of expression online: what to take down, what to leave up, and why. It reviews cases referred by Meta or appealed by individuals on Facebook, Instagram or Threads who disagree with our content moderation decisions and provides binding rulings on whether to remove or leave up content. The Oversight Board also provides recommendations to enhance our content moderation practices and offers policy advisory opinions upon request.



Where to find information about the Oversight Board's impact

In 2024, we moved from a quarterly to a bi-annual cadence for [reporting](#) about cases that Meta has referred to the Oversight Board and updates on our progress on implementing its recommendations. Additionally, we launched a [Transparency Center page](#) that tracks the impact of the Oversight Board's recommendations. This is in addition to our [Oversight Board recommendations page](#), where we outline the recommendations related to a case received from the Oversight Board, our commitment level and the implementation status.





Actions related to Oversight Board recommendations in 2024



Oversight Board
recommendations issued

48

(66 in 2023)



Meta assessment and/or
implementation in progress⁵

70

(69 in 2023)



Recommendations
implemented⁵

41

(61 in 2023)

In 2024, the Oversight Board considered cases concerning our content enforcement in light of the international human rights framework, including freedom of expression, the right to health, and the right to equality and non-discrimination, among others. Here are a few illustrative examples of our actions in response to decisions made by the Oversight Board in 2024. Details are in [Meta's Bi-Annual Reports on the Oversight Board](#).

Examples of the Oversight Board's decisions in 2024 included:



The Oversight Board overturned Meta's decisions to remove three Facebook posts showing footage of the March 2024 [Moscow terrorist attack](#), requiring the content to be restored with "Mark as Disturbing" warning screens. The Oversight Board found that while the posts violated Meta's policy on showing the moment of designated attacks on visible victims, removing them was not consistent with the company's human rights responsibilities.



The Oversight Board upheld Meta's decision to remove a video of a [Pakistani politician delivering a speech](#) with text claiming the politician was "crossing all limits of faithfulness" and using the word "kufr" to suggest blasphemy, due to the risk of offline harm.

⁵ Some assessments and/or implementations in progress or recommendations fully implemented include recommendations from prior years (see our [2023 Human Rights Report](#) for further details).



Examples of actions following Oversight Board recommendations included:



Following a series of [recommendations](#) (e.g., [here](#)) on AI, we made changes to the way we handle [AI-generated content](#), including updated labels and policies, such as our [Misinformation policy](#).



Following [recommendations](#) from the Oversight Board on content policy, Meta changed the [Dangerous Organizations and Individuals policy](#) to allow content using the term “[shaheed](#)” in all languages that use this term, except when the content is accompanied by signals of violence or otherwise violates our policies (for example, by glorifying designated dangerous individuals).



In 2024, the Oversight Board also experimented with a shorter timeline for urgent cases. For example, after July’s presidential election in Venezuela when violence broke out, we referred [two pieces](#) of content regarding “Colectivos” for expedited review. “Colectivos” is an umbrella term used to describe irregular armed gangs or paramilitary-style groups closely aligned with the government. These cases were decided on an accelerated timeline of 14 days.

We also partnered with the Oversight Board to engage regulators and civil society organizations including in the Africa, Latin America, and Middle East and Türkiye regions to raise awareness of the Board’s mandate and case selection process.



[About](#)[Exec. summary](#)[Managing risks](#)[AI](#)[Issues](#)[Stakeholders](#)[Transparency & remedy](#)[Annex](#)

Annex



How human rights are governed and managed at Meta

Our human rights experts guide the implementation of our [Corporate Human Rights Policy](#), which is overseen by the President of Global Affairs (now Chief Global Affairs Officer) and Chief Legal Officer.

The human rights experts' tasks include promoting the policy's integration into existing and developing policies, programs and services; undertaking due diligence; and supporting the training of employees on the policy. The policy gives guidance to build rights-respecting products, respond to emerging crises, and work with speed and agility to embed human rights at scale.

Our Corporate Human Rights Policy commits us to periodic reporting to the Board of Directors on key human rights issues. In 2024, the Director of Human Rights briefed the Audit and Risk Oversight Committee of the Board.

In 2024, Meta launched the human rights risk vertical of the company's third-party risk management program. This control demonstrates our commitment to continuously improve our human rights risk management and strive for third-party engagements that are responsible and respectful of human rights.

Training Meta employees on human rights

How we build at Meta is as important as what we build. Our human rights training highlights the potential and actual real-life impacts of our services, policies and business decisions on human rights. It seeks to promote a human rights mindset in our day-to-day work, encouraging respect for human rights to the benefit of all who use our services.

We launched our *Bigger than Meta: Human Rights* training in 2022, which continued throughout 2024. Our privacy training also supports our human rights training objectives by focusing on developing our collective ability to protect individuals — including, in particular, marginalized categories of individuals — against harms emanating out of the processing of people's data.

Links to referenced reports

[2025 Responsible Business Practices Report](#)

[2025 Sustainability Report](#)

[2023 Human Rights Report](#), [2022 Human Rights Report](#), [2021 Human Rights Report](#)

[2024 Anti-Slavery and Human Trafficking Report](#)

[2024 Conflict Minerals Report](#)

[Meta Transparency Reports](#)

[Regulatory and Other Transparency Reports](#)

Previously published Human Rights Impact Assessments: [End-to-End Encryption](#), [Philippines](#), [Myanmar](#), [Indonesia](#), [Cambodia](#), [India](#), [Sri Lanka](#) and [Israel and Palestine](#)

