Medidas da Meta para Salvaguardar a Integridade dos Processos Eleitorais do Brasil

Agosto de 2026

Este relatório é publicado de acordo com os requisitos das Resoluções nº 23.610/2019 e nº 23.755/2026, do Tribunal Superior Eleitoral do Brasil. O relatório a seguir descreve as medidas da Meta Platforms, Inc. ("Meta") e do WhatsApp LLC ("WhatsApp") para salvaguardar a integridade das eleições brasileiras no Facebook, Instagram, Threads e WhatsApp antes e durante o período eleitoral, de acordo com o Art. 125-B da Resolução nº 23.610/2019/TSE, com as alterações trazidas pela Portaria nº 463/2026 da Presidência do Tribunal Superior Eleitoral.

Ao longo deste relatório, as referências às políticas e práticas da "Meta" referem-se ao Facebook, Instagram e Threads; a discussão sobre as políticas e práticas do WhatsApp será abordada de forma independente. Nos casos em que as políticas e práticas se estendem a toda a empresa, elas serão atribuídas às "Empresas da Meta" como um todo.

Introdução

Proteger a integridade das eleições do Brasil em nossos aplicativos é uma prioridade para as Empresas da Meta. Embora cada eleição seja única, nosso trabalho se baseia em lições fundamentais aprendidas em mais de 200 eleições em todo o mundo desde 2016. Essas lições nos ajudam a concentrar nossas equipes, tecnologias e investimentos para que tenham o maior impacto possível.

A Meta investiu mais de US$ 30 bilhões em equipes e tecnologia de segurança e proteção na última década, e continua a investir significativamente tanto em sistemas automatizados de IA quanto em operações de análise humana que apoiam a moderação de conteúdo em seus serviços em mais de 70 idiomas, incluindo o português.

Para cada eleição de âmbito nacional, as Empresas da Meta avaliam se nossos mecanismos padrão (que incluem as políticas, ferramentas e processos que estão documentados em uma série de publicações na Sala de Imprensa sobre o assunto) abordam as ameaças específicas daquela eleição. Caso haja riscos atípicos, as Empresas da Meta trabalham com várias equipes para implementar medidas apropriadas a fim de mitigar esses riscos.

No Facebook, Instagram e Threads ("Meta", conforme acima), nossa abordagem em relação às eleições é informada pela experiência e pelos marcos regulatórios aplicáveis em vigor. Na primeira seção deste relatório, descrevemos a abordagem da Meta nos seguintes pilares, levando em consideração a natureza e as viabilidades técnicas de cada aplicativo:

1. Utilização e implementação dos Padrões da Comunidade e moderação de conteúdo, para remover conteúdo nocivo e manter as pessoas seguras
2. A avaliação da Meta de impacto na integridade eleitoral
3. Salvaguardas e esforços de transparência relacionados à publicidade política
4. Abordagem responsável com relação à IA Generativa
5. Proteção de dados pessoais em anúncio eleitoral
6. Canais de denúncia local
7. Cumprimento de ordens judiciais

No WhatsApp, nossa abordagem é igualmente informada pela experiência e por nossas obrigações de conformidade regulatória, mas reflete as diferentes medidas apropriadas para um aplicativo de mensagens privadas. Na segunda seção deste relatório, descrevemos as medidas implementadas no WhatsApp, focadas nas seguintes linhas de esforço:

1. Termos de Serviço, políticas e diretrizes do WhatsApp, incluindo as Diretrizes dos Canais e aplicação dessas diretrizes
2. A avaliação do WhatsApp de impacto na integridade eleitoral
3. Proibição de anúncio político
4. A abordagem do Whatsapp com relação à inteligência artificial
5. Proteção de dados pessoais
6. Canais de denúncia local
7. Cumprimento de ordens judiciais

Em estrita observância aos arts. 8º e 9º da Portaria nº 463/2026/TSE, algumas informações foram classificadas neste Plano como de acesso restrito, conforme indicação específica de "**[INÍCIO DO CONTEÚDO CONFIDENCIAL]**" e "**[FIM DO CONTEÚDO CONFIDENCIAL]**" em cada uma das seções pertinentes.

Conforme indicado especificamente ao longo deste documento, referidas informações abrangem, entre outras, descrições acerca de sistemas de integridade e segurança das plataformas, incluindo mecanismos de detecção, análise e mitigação de riscos, prevenção e disrupção de comportamentos abusivos, contenção de disseminação de conteúdo, bem como proteção contra evasão de controles. O acesso restrito a referidas informações é de rigor, nos termos do art. 8º, §1º, II, da Portaria, pois sua divulgação ampla poderia comprometer a segurança dos sistemas e a eficácia dos mecanismos empregados para proteção da integridade do processo eleitoral.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

A classificação restrita/confidencial encontra amparo na proteção constitucional conferida à livre iniciativa e à livre concorrência (art. 170, caput e inciso IV, da Constituição Federal), bem como na tutela da propriedade industrial e dos demais ativos imateriais utilizados na atividade econômica (art. 5º, XXIX, da Constituição Federal), e ainda na proteção das informações empresariais voltadas à repressão da concorrência desleal e à proteção de segredos de negócio (Lei nº 9.279/1996).

META (FACEBOOK, INSTAGRAM, THREADS)

De acordo com o Artigo 2º da Portaria nº 463/2026/TSE, esta seção especifica o tipo de serviço oferecido por cada aplicativo coberto por este plano de conformidade. Nos termos do parágrafo único do Artigo 2º, o plano é submetido de forma integrada, com seções específicas por serviço onde as funcionalidades, arquiteturas e perfis de risco à integridade eleitoral diferem.

O Facebook, o Instagram e o Threads são aplicativos sociais que permitem a publicação, o compartilhamento, a distribuição, a indexação e o impulsionamento de conteúdo político-eleitoral de terceiros, sendo, portanto, caracterizados como provedores nos termos do Artigo 3º, I, da Portaria. Cada um deles oferece publicidade paga, incluindo o impulsionamento de Anúncios sobre Temas Sociais, Eleições ou Política ("SIEP"). Suas funcionalidades individuais encontram-se descritas abaixo.

SEÇÃO 1: Utilização e Implementação dos Padrões da Comunidade

Esta seção aborda os requisitos estabelecidos nos Artigos 13, 14, 22 e 23 da Portaria nº 463/2026/TSE. Coletivamente, essas disposições exigem que o plano de conformidade descreva: (i) os instrumentos normativos que regem o uso da plataforma e o tratamento de conteúdo político-eleitoral por terceiros, incluindo a arquitetura de moderação de conteúdo do provedor, as medidas de aplicação disponíveis e as estruturas de governança que supervisionam sua aplicação (Art. 22); (ii) os parâmetros, critérios e fluxos operacionais empregados pelo provedor para a identificação, detecção e remoção proativas de conteúdo que dissemine fatos notoriamente falsos ou gravemente descontextualizados capazes de afetar a integridade eleitoral, bem como os mecanismos de monitoramento contínuo e arranjos de verificação de fatos que apoiam esse esforço (Art. 13); (iii) os mecanismos para detectar e interromper o comportamento inautêntico coordenado (Art. 14) e (iv) as medidas de iniciativa própria do provedor para o planejamento de ações corretivas e preventivas, o aprimoramento dos sistemas de recomendação e a transparência dos resultados, incluindo, quando aplicável, os critérios e procedimentos para a remoção de contas falsas, automatizadas ou bots (Art. 23).

Em resposta a esses requisitos, este capítulo descreve como os Padrões da Comunidade e todas as outras políticas aplicáveis às tecnologias da Meta servem como a estrutura fundamental por meio da qual a Meta cumpre seus deveres de moderação proativa, governança de conteúdo e integridade da plataforma, atendendo coletivamente à conformidade dos Artigos 9-D, 28 e 38-A da Resolução nº 23.610/2019/TSE.

Estes instrumentos estabelecem o conteúdo e os comportamentos que são e não são permitidos em nossas plataformas, e são projetados para manter um ambiente seguro para os usuários, preservando ao mesmo tempo o espaço para a expressão. As informações apresentadas abaixo detalham a substância dessas políticas, a arquitetura de moderação construída para aplicá-las — incluindo a interação entre sistemas de detecção automatizados e análise humana — e os mecanismos de governança que garantem sua aplicação consistente e proporcional no contexto das Eleições Gerais Brasileiras de 2026.

1.1 Visão Geral dos Padrões da Comunidade da Meta

Os Padrões da Comunidade da Meta estabelecem regras abrangentes que definem o conteúdo e os comportamentos que são e não são permitidos em nossas plataformas. Eles são projetados para manter um ambiente seguro para os usuários, preservando o espaço para a expressão. Eles também abordam as hipóteses de risco identificadas na Resolução Eleitoral. É importante ressaltar que a estrutura de políticas da Meta não é organizada nas mesmas linhas categóricas da Resolução Eleitoral; em vez disso, cada hipótese de risco pode ser abordada por **múltiplos Padrões da Comunidade sobrepostos** que funcionam em combinação para fornecer cobertura.

Por exemplo, o conteúdo que constitui violência política contra as mulheres pode acionar simultaneamente a análise e a aplicação das políticas da Meta sobre Violência e Incitação, Conduta de Ódio e Bullying e Assédio. Da mesma forma, o conteúdo caracterizado como antidemocrático de acordo com a Resolução Eleitoral pode envolver simultaneamente as políticas da Meta sobre Violência e Incitação, Interferência Eleitoral e Supressão de Eleitores, e Coordenação de Atos Danosos e Incentivo à Prática de Atividades Criminosas. Essa abordagem em camadas garante que o conteúdo seja avaliado em relação a todas as políticas aplicáveis, conforme descrito mais adiante na Seção 1.4.

Os Padrões da Comunidade são organizados por área de política. Cada seção começa com uma "Justificativa da Política" que articula os objetivos da política e a natureza do dano que ela aborda, seguida por descrições detalhadas do conteúdo e dos comportamentos específicos que são proibidos.

Essas políticas são desenvolvidas por meio de consultas com especialistas, acadêmicos, organizações da sociedade civil e partes interessadas em todo o mundo. Elas são revisadas e atualizadas regularmente para lidar com danos emergentes e estão disponíveis publicamente na Central de Transparência da Meta.

A Meta emprega uma combinação de sistemas de detecção automatizados, análise humana, parcerias com verificadores de fatos terceirizados e denúncias de usuários para identificar conteúdos que se enquadram nas categorias de risco estabelecidas pela Resolução. Abaixo, oferecemos uma visão geral dos Padrões da Comunidade mais aplicáveis em relação às categorias de conteúdo eleitoral ilegal.

1.1.1 Coordenação de Atos Danosos e Incentivo à Prática de Atividades Criminosas

Em um esforço para prevenir e interromper danos offline — incluindo danos às instituições democráticas e aos processos eleitorais —, nossas políticas proíbem as pessoas de facilitar, organizar, promover ou permitir a prática de atividades criminosas ou prejudiciais direcionadas a pessoas, empresas, propriedades ou animais. Isso inclui condutas destinadas a subverter a ordem constitucional, perturbar o funcionamento das instituições democráticas ou coordenar interferência ilegal nos processos eleitorais, como ofertas de compra ou venda de votos com dinheiro, presentes, serviços ou outros bens materiais, ou defender, fornecer instruções ou demonstrar intenção explícita de participar ilegalmente de uma votação (por exemplo, votar duas vezes ou forjar sua elegibilidade para votar).
Permitimos que as pessoas debatam e defendam a legalidade das atividades, bem como chamem a atenção para atividades prejudiciais ou criminosas que possam testemunhar ou vivenciar, desde que não defendam ou coordenem atividades danosas.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.1.2 Desinformação

Em um esforço para promover a integridade eleitoral, removemos a desinformação que provavelmente contribuirá diretamente para o risco de interferência na capacidade das pessoas de participar desses processos. Isso inclui o seguinte:

- Desinformação sobre as datas, locais, horários e métodos de votação, registro de eleitores ou participação no censo.

- Desinformação sobre quem pode votar, qualificações para votar, se um voto será contado e quais informações ou materiais devem ser fornecidos para votar.
- Desinformação sobre se um candidato está concorrendo ou não.

1.1.3 Conduta de Ódio

Definimos conduta de ódio como ataques diretos contra pessoas – em vez de conceitos ou instituições – com base no que chamamos de características protegidas (CPs): raça, etnia, origem nacional, deficiência, afiliação religiosa, casta, orientação sexual, sexo, identidade de gênero e doenças graves. Além disso, consideramos a idade uma característica protegida quando referenciada juntamente com outra característica protegida. Também protegemos refugiados, migrantes, imigrantes e solicitantes de asilo dos ataques mais graves, embora permitamos comentários e críticas às políticas de imigração. Da mesma forma, fornecemos algumas proteções para características não protegidas, como ocupação, quando são referenciadas juntamente com uma característica protegida. Às vezes, com base em nuances locais, consideramos certas palavras ou frases como substitutos frequentemente usados para características protegidas.

Removemos discursos desumanizadores, alegações de imoralidade ou criminalidade grave e insultos. Também removemos estereótipos prejudiciais, que definimos como comparações desumanizadoras que têm sido historicamente usadas para atacar, intimidar ou excluir grupos específicos, e que muitas vezes estão ligadas à violência offline. Por fim, removemos insultos graves, expressões de desprezo ou nojo, palavrões e apelos à exclusão ou segregação quando direcionados a pessoas com base em características protegidas.

O conteúdo identificado como violador desta política — seja por meio de denúncias de usuários, encaminhamentos de parceiros de confiança ou outros mecanismos de detecção disponíveis — é removido de nossas plataformas.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.1.4 Bullying e Assédio

A Meta define bullying e assédio como conteúdo ou comportamento que assume muitas formas, desde fazer ameaças e divulgar informações de identificação pessoal até enviar mensagens ameaçadoras e fazer contatos maliciosos indesejados. Não toleramos esse tipo de comportamento porque ele impede que as pessoas se sintam seguras e respeitadas no Facebook, Instagram e Threads. Nosso Padrão da Comunidade sobre Bullying e Assédio aborda especificamente o risco de nossas

plataformas serem usadas para promover conteúdo que degrade ou envergonhe os usuários, ou para fazer contato repetido e indesejado com um usuário — incluindo cyberbullying, perseguição (*stalking*), ameaças de danos, assédio em massa e assédio sexual.

Distinguimos entre figuras públicas e indivíduos privados na aplicação desta política. Para figuras públicas, removemos ataques que são graves, bem como certos ataques em que a figura pública é marcada diretamente na publicação ou comentário. Para indivíduos privados, oferecemos proteções mais fortes e removemos uma gama mais ampla de conteúdo, incluindo conteúdo que os alveja por meio de Páginas, Grupos e Eventos indesejados, bem como conteúdo que conclame ou declare a intenção de praticar bullying. Alguns conteúdos sob esta política exigem contexto adicional — como a confirmação do alvo ou de um representante autorizado — antes que uma medida de aplicação seja tomada.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.1.5 Violência e Incitação

Nossa política de Violência e Incitação proíbe conteúdo que constitua uma ameaça crível à segurança pública ou pessoal, incluindo incitação à violência. Isso abrange:

- Conteúdo que constitua uma ameaça grave, direta e imediata de violência contra indivíduos ou instituições, incluindo o judiciário e as instituições democráticas
- Apelos à violência eleitoral
- Apelos à violência com o objetivo de restringir o exercício dos poderes constitucionais ou o Estado de Direito

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.1.6 Conteúdo Eleitoral Gerado por IA e Manipulado Digitalmente

Estamos abordando ativamente os riscos representados pelo conteúdo gerado por IA no contexto eleitoral por meio de requisitos de rotulagem e divulgação, que serão abordados nas seções 1.2, 1.6 e 4 deste relatório.

1.1.7 Instrumentos Normativos que Regem o Conteúdo Político-Eleitoral no Brasil

Nossa governança de conteúdo político-eleitoral opera por meio de uma estrutura em camadas de Termos de Serviço, Padrões/Diretrizes da Comunidade, Políticas de Publicidade e controles em nível de produto, cada um dos quais se aplica às adaptações específicas de produto do Facebook, Instagram e Threads.

Para as eleições brasileiras, além da aplicação das políticas da Meta, implementamos medidas de integridade direcionadas. Como parte de nosso trabalho para proteger as eleições presidenciais de 2026 no Brasil, aplicamos restrições de conteúdo específicas para o Brasil, de acordo com a regulamentação local e ordens judiciais válidas, conforme documentado em nossos relatórios de transparência sobre restrições de conteúdo e em todo este plano de conformidade.

1.1.8 Prevenção de Mensagens em Massa

Infraestrutura de Detecção Multiplataforma

Empregamos limites de taxa para limitar proativamente a velocidade com que as contas podem realizar várias ações no Facebook e no Instagram (como criar publicações ou enviar mensagens) para evitar o uso de bots de qualquer tipo em ambas as plataformas. Esses limites são definidos por referência a métricas identificadas antes e após a implementação, usando conteúdo e comportamentos que violam nossos Padrões da Comunidade.

Nosso programa anti-raspagem (*anti-scraping*) e infraestrutura de limitação de taxa estão sujeitos a revisão e refinamento contínuos. A inteligência de ameaças, os testes realizados pela equipe de segurança cibernética (*red team*) e os programas de recompensa por falhas encontradas (*bug bounty*) contribuem para a melhoria contínua desses controles, garantindo que eles permaneçam eficazes contra ameaças novas e em evolução.

Nossos sistemas de integridade usam uma abordagem de detecção em várias camadas combinando sinais comportamentais, modelos de *machine learning* baseados em conteúdo, sinais de ator e análise fora da plataforma para identificar comportamento automatizado ou de envio em massa.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.1.9 Processamento Automatizado e Priorização

O processamento automatizado de dados é fundamental para o nosso processo de análise e automatiza as decisões para certas áreas onde o comportamento da conta ou o conteúdo da mensagem denunciada tem grande probabilidade de estar em violação. A automação também ajuda a priorizar e agilizar as análises, encaminhando contas, grupos ou comunidades potencialmente violadores para revisores humanos com o conhecimento adequado do assunto e do idioma, para que as equipes possam se concentrar primeiro nos casos mais importantes.

Equipes de Análise Humana
Quando uma conta, grupo ou comunidade requer análise adicional, nossos sistemas automatizados a priorizam e a enviam para uma equipe de análise humana para tomar a decisão final. As equipes de análise humana estão localizadas em todo o mundo, recebem treinamento aprofundado e muitas vezes se especializam em determinadas áreas de políticas e regiões, e são capazes de analisar as informações relevantes da conta e as mensagens denunciadas.

Encaminhamentos de Parceiros de Confiança e Autoridades Policiais
Contas ou grupos potencialmente violadores também são colocados na fila para análise humana quando identificados por parceiros de confiança, sujeitos a diretrizes operacionais de aplicação que podem resultar na suspensão de grupos violadores e no banimento das contas dos administradores responsáveis pelos grupos, bem como das de outros usuários que facilitem o abuso.

Colaboramos ativamente com parceiros de confiança, como autoridades reguladoras, agências de aplicação da lei e órgãos do setor. Essas parcerias fornecem inteligência mais confiável e acionável, que informa as medidas de aplicação, melhorias de produtos e esforços mais amplos para interromper atividades maliciosas.

1.1.10 Política da Meta sobre Contas Falsas e Aplicação em Todas as Plataformas

A Meta exerce sua capacidade de identificar e remover perfis falsos e contas de bots automatizados em suas plataformas. Essa aplicação é regida por **duas estruturas de políticas distintas, mas complementares** sob nossos Padrões da Comunidade: **Representação de Identidade Autêntica (“AIR”)** — que aborda a dissimulação em nível individual, incluindo personificação e falsificação de identidade; e **Comportamento Inautêntico (“IB”)** — que aborda uma variedade de formas complexas de dissimulação, realizadas por uma rede de ativos inautênticos controlados pelo(s) mesmo(s) indivíduo(s), com o objetivo de enganar a Meta ou nossa comunidade ou de burlar a aplicação dos Padrões da Comunidade, incluindo o Comportamento Inautêntico Coordenado (CIB). Abaixo, fornecemos uma visão geral de ambas as políticas.

A política de **Representação de Identidade Autêntica** não permite o uso de nossos serviços e restringirá ou desativará contas do Facebook, Instagram e Threads ou outras entidades do Facebook (como Páginas, grupos) que:

- **Personifiquem outra pessoa ou entidade ao:**
  - Usar sua(s) imagem(ns), nome ou semelhança com o objetivo de enganar outras pessoas
  - Falar em nome de outra pessoa ou entidade para a qual o usuário não está autorizado a fazê-lo (por exemplo, criando uma Página ou perfil)
- **Pratiquem falsificação de identidade** para enganar ou ludibriar outras pessoas, burlar a aplicação das regras ou violar nossos Padrões da Comunidade. Consideramos vários fatores ao avaliar a falsificação de identidade enganosa, tais como:
  - Alterações repetidas ou significativas nos detalhes de identidade, como nome ou idade
  - Informações de perfil enganosas, como detalhes da biografia e localização do perfil
  - Uso de imagens de bancos de imagens

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Diferentemente da política AIR acima, a política de **Comportamento Inautêntico** aborda formas complexas de dissimulação em nível de rede, realizadas por grupos de ativos inautênticos controlados pelo(s) mesmo(s) indivíduo(s), com o objetivo de enganar a Meta ou nossa comunidade, ou de burlar a aplicação das regras. Abaixo, fornecemos uma visão geral dos critérios e procedimentos de aplicação adotados nesse sentido.

Remoção de Contas Falsas
Nossa abordagem para combater contas falsas está fundamentada em nossa política de identidade e autenticidade sob nossos Padrões da Comunidade.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Não permitimos:

- Contas que se passam por outras pessoas para enganar ou ludibriar terceiros, burlar medidas de fiscalização ou violar nossos Padrões da Comunidade.
- Contas Abusivas Automatizadas: Contas automatizadas ou semiautomatizadas criadas em massa e usadas para realizar atividades abusivas na plataforma.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.1.11 Direitos do Usuário e Recursos

Em consonância com nosso compromisso com a transparência, informamos os usuários cujo conteúdo ou contas tenham sido objeto de medidas de aplicação de regras e oferecemos a eles a possibilidade de recorrer de tais medidas. Tratamos as reclamações de maneira célere, diligente, transparente e objetiva, revertendo as medidas sem demora injustificada quando as reclamações forem consideradas procedentes.

1.1.12 Alterações relevantes dos Termos de Serviço/Políticas de Conteúdo

Nosso processo de notificação aos usuários sobre alterações relevantes em nossos termos de serviço e políticas de conteúdo é regido por uma estrutura de controle interno e por uma avaliação jurídica caso a caso. Esse processo aplica-se uniformemente a todas as alterações relevantes, inclusive àquelas impulsionadas por exigências de conformidade regulatória, como as leis eleitorais, em vez de por meio de um fluxo de trabalho específico e separado para questões eleitorais. Abaixo está um resumo de como esse processo opera em nossas plataformas.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.2 Redução da Disseminação de Informações Falsas

Em referência direta às obrigações estabelecidas no Art. 9-D da Resolução nº 23.610/2019/TSE, a Meta adota várias medidas para reduzir a disseminação de informações falsas. Isso inclui adotar sistemas de rotulagem de conteúdo e informações contextuais, medidas de redução de distribuição e restrições a anúncios que desencorajam a participação eleitoral. Adotamos uma abordagem abrangente para o combate a desinformação que distingue entre diferentes categorias e tratamentos:

- Conteúdo que interfere diretamente nas eleições (por exemplo, informações falsas sobre processos de votação) é removido sumariamente, conforme descrito acima.
- A Meta não permite anúncios que desencorajem as pessoas a votar, questionem a legitimidade ou integridade do processo eleitoral, ou contenham alegações de que condições ou circunstâncias específicas de votação levarão à anulação de um voto.
- Desinformação geral: trabalhamos com uma rede global, fora dos Estados Unidos, de mais de 90 parceiros independentes de verificação de fatos terceirizados em mais de 60 idiomas. Quando os verificadores de fatos classificam o conteúdo, aplicamos rótulos informativos e reduzimos sua distribuição no Facebook, Instagram e Threads para que menos pessoas o vejam.

Nossa tecnologia pode detectar publicações que provavelmente são desinformação com base em vários sinais, incluindo como as pessoas estão reagindo e com que rapidez o conteúdo está se espalhando. Os verificadores de fatos também podem identificar conteúdo para revisão com base em suas próprias apurações. Durante grandes eventos com repercussão midiática, quando a velocidade é especialmente importante, como eleições, usamos detecção de palavras-chave para reunir conteúdo relacionado em um único lugar, facilitando para os verificadores de fatos encontrá-lo.

1.2.1 Parcerias de Verificação de Fatos Terceirizadas

Estamos comprometidos em combater a desinformação no Facebook, Instagram e Threads. Em muitos países e regiões, trabalhamos com organizações independentes de verificação de fatos terceirizadas que são certificadas pela International Fact-Checking Network (IFCN), apartidária, ou, na Europa, pela European Fact-Checking Standards Network (EFCSN).

Escopo Material
Os parceiros de verificação de fatos revisam conteúdo nas plataformas da Meta, incluindo publicações no Feed, Reels, vídeo e anúncios, abrangendo uma variedade de categorias de alegações, como tópicos políticos, de saúde, ambientais e de segurança pública. Os parceiros identificam e selecionam alegações para revisão de forma independente; a Meta não direciona decisões editoriais sobre o que é classificado ou como.

Os verificadores de fatos analisam um conteúdo e avaliam sua exatidão. Este processo ocorre independentemente da Meta e pode incluir consultas a fontes, verificação de dados públicos, autenticação de imagens e vídeos, entre outros. As classificações que

os verificadores de fatos podem usar são **Falso**, **Alterado**, **Parcialmente Falso**, **Falta Contexto**, **Sátira** e **Verdadeiro**.

Quando o conteúdo tiver sido classificado pelos verificadores de fatos, adicionamos um aviso a ele para que as pessoas possam ler contexto adicional. Também notificamos as pessoas antes de tentarem compartilhar este conteúdo ou se o compartilharam no passado.

Uma vez que um verificador de fatos classifica um conteúdo como Falso, Alterado ou Parcialmente Falso, ou detectamos que é quase idêntico, ele pode receber distribuição reduzida no Facebook, Instagram e Threads. Reduzimos drasticamente a distribuição de publicações Falsas e Alteradas, e reduzimos a distribuição de Parcialmente Falso em menor grau. Para Falta Contexto, focamos em apresentar informações dos verificadores de fatos.

Páginas, grupos, perfis e sites que compartilham repetidamente conteúdo classificado como Falso ou Alterado terão algumas restrições impostas por um determinado período de tempo. Isso inclui removê-los das recomendações que mostramos às pessoas, reduzir sua distribuição, **remover sua capacidade de monetizar e anunciar**, e remover sua capacidade de se registrar como uma Página de notícias.

Em conformidade com o Artigo 7º, §4º, da Portaria nº 463/2026/TSE, a Meta monitorará este trabalho de perto e trabalhará para relatar métricas pertinentes às parcerias de verificação de fatos e ao comportamento do conteúdo rotulado nos serviços da Meta.

Escopo Territorial
Atualmente, temos parceria com mais de 90 organizações de verificação de fatos que revisam e classificam conteúdo em mais de 60 idiomas globalmente. A Meta mantém parcerias com organizações independentes de verificação de fatos terceirizadas no Brasil que são certificadas pela International Fact-Checking Network (IFCN). Essas parcerias são projetadas para cobrir a amplitude do conteúdo potencialmente enganoso que circula nas plataformas da Meta no país. A Meta possui **6 parceiros de verificação de fatos terceirizados (3PFC) no Brasil**:

- **Agência Lupa**
- **AFP** (AFP Checamos)
- **Aos Fatos**
- **Estadão Verifica**
- **Reuters Fact Check**
- **UOL Confere**

Escopo Linguístico
Os parceiros são organizações sediadas no Brasil com expertise no ambiente informacional brasileiro. A cobertura aborda conteúdo que circula domesticamente, incluindo conteúdo em língua portuguesa transfronteiriço quando relevante para o público brasileiro. Os parceiros verificadores de fatos no Brasil são certificados pela IFCN, que é apartidária.

Proporcionalidade à Escala de Operação
Durante períodos de alto risco, como ciclos eleitorais, emergências de saúde pública ou eventos noticiosos significativos, a Meta trabalha com parceiros para garantir que tenham os recursos necessários para combater a desinformação viral.

Independência, Transparência e Responsabilização
Os parceiros de verificação de fatos operam com total independência editorial. A Meta não influencia o resultado de qualquer revisão individual. Os parceiros aplicam suas próprias metodologias publicadas, consistentes com os princípios da IFCN de transparência, apartidarismo e responsabilização. A Meta publica uma lista de seus parceiros de verificação de fatos, e cada organização parceira disponibiliza publicamente sua metodologia de revisão. Criadores de conteúdo cujo conteúdo é classificado podem acessar um processo de recurso diretamente com o parceiro de verificação de fatos relevante.

1.3 Prevenção de Interferência e Comportamento Inautêntico

A Meta emprega uma série de medidas para prevenir interferência eleitoral e comportamento inautêntico. Isso inclui medidas contra o Comportamento Inautêntico Coordenado (CIB), remoção de contas falsas e aplicação proativa.

1.3.1 Comportamento Inautêntico Coordenado

Conforme indicado na seção 1.1.10 acima, o Comportamento Inautêntico refere-se a uma variedade de formas complexas de dissimulação, realizadas por uma rede de ativos inautênticos controlados pelo(s) mesmo(s) indivíduo(s), com o objetivo de enganar a Meta ou nossa comunidade ou de burlar a aplicação dos Padrões da Comunidade.

Abordamos o problema de operações de influência encobertas por meio de nossa política de **Comportamento Inautêntico Coordenado (CIB)**, definido como formas particularmente sofisticadas de Comportamento Inautêntico onde identidades falsas são centrais para a operação e os operadores usam métodos para burlar a detecção ou parecer autênticos. Esta seção detalhará ainda mais esta política.

A política de CIB é baseada em comportamento, o que é diferente de grande parte do trabalho de moderação de conteúdo que analisa peças individuais de conteúdo para determinar se violam as políticas da plataforma. Não removemos redes de CIB por causa do conteúdo que compartilham; removemos porque exibem comportamento enganoso ao usar contas falsas para ocultar suas identidades e fingir ser algo que não são.

Abordagem Baseada em Comportamento: Distinguindo Coordenação Legítima de Redes Inautênticas

Nossa política de CIB foca no **comportamento em vez do conteúdo**, independentemente de quem está por trás da atividade, o que publicam ou se são estrangeiros ou domésticos. A distinção crítica entre ação coordenada legítima e redes inautênticas baseia-se em dois critérios centrais: (i) **coordenação entre contas** controladas pelos mesmos atores e (ii) **uso de contas falsas como parte central das operações**

Esta estrutura baseada em comportamento significa que a atividade coordenada legítima — como campanhas políticas, grupos de defesa de causas ou movimentos ativistas — não é alvo de aplicação de CIB, desde que esses atores não dependam de redes de contas falsas para enganar o público. Essas medidas de aplicação e padrões se aplicam independentemente de conteúdo ou ideologia e são projetados para criar um espaço onde as pessoas possam confiar nas pessoas e comunidades com as quais interagem. Redes de CIB podem ser usadas para amplificar conteúdo dentro das categorias de risco identificadas pela Resolução, incluindo desinformação direcionada ao processo eleitoral ou conteúdo antidemocrático.

Métodos de Detecção nas Plataformas

Empregamos várias técnicas complementares para detectar e aplicar medidas contra CIB em nossas plataformas:

- Investigações especializadas e rastreamento a longo prazo de atores que possam representar uma ameaça; de longo prazo
- Sistemas de detecção automatizados para identificar padrões de atividade inautêntica em escala;
- Revisão de indicações (*leads*) de terceiros, incluindo indicações vindas de governos e parceiros da sociedade civil;
- Programas de aplicação contínua que dependem tanto de sistemas automatizados quanto de investigações manuais para detectar e remover tentativas de redes previamente interrompidas de retornar à plataforma

1.3.2 Medidas de Aplicação

Removemos contas falsas, operadores autênticos, páginas,grupos ou outros ativos da Meta diretamente envolvidos em operações de CIB. Em casos onde esses atores

também criam, gerenciam, cooptam, direcionam ou controlam páginas, grupos ou comunidades que representam entidades autênticas não envolvidas no comportamento violador, podemos tomar medidas para remover os indivíduos violadores enquanto permitindo que as pessoas e comunidades não envolvidas permaneçam em nossos serviços.

1.3.3 Mecanismos para Detecção de Padrões de Coordenação

Empregamos uma abordagem com multicamadas para detectar padrões de coordenação entre contas em nossas plataformas. Nossa política de Comportamento Inautêntico aborda uma variedade de formas complexas de dissimulação realizadas por uma rede de ativos inautênticos controlados pelo(s) mesmo(s) indivíduo(s). Quando agentes maliciosos utilizam identidades falsas para praticar formas sofisticadas de Comportamento Inautêntico, configura-se o que definimos como Comportamento Inautêntico Coordenado (CIB). Essas medidas de aplicação e padrões incidem independentemente do conteúdo ou da ideologia.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.3.4 Descrição Funcional dos Procedimentos de Integração, Ponderação e Validação de Sinais de Risco

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Revisão de Especialistas Multifuncional
Além da revisão humana em escala, mantemos equipes especializadas que conduzem análises manuais e individualizadas, fora dos fluxos em escala, de conteúdo e atores em busca de possíveis violações de políticas no contexto da investigação de questões complexas e de alta prioridade, incluindo questões levantadas por meio de escalonamentos internos e externos.

Gatilhos de Revisão Humana e Escalonamento de Especialistas Canais de Escalonamento Interno
Mantemos canais dedicados onde as equipes internas de revisão de conteúdo podem escalonar possíveis pistas de CIB (Comportamento Inautêntico Coordenado) para uma equipe de funcionários experientes com treinamento especializado que analisa as informações disponíveis e faz recomendações para a aplicação de medidas.

1.3.5 Medidas Preventivas para Dificultar a Formação de Redes Inautênticas

Abordagem

Nossa política de Comportamento Inautêntico aborda uma variedade de formas complexas de enganosidade realizadas por redes de ativos inautênticos controlados pelo mesmo indivíduo ou indivíduos, com o objetivo de enganar a Meta ou nossa comunidade ou burlar a aplicação de medidas. Quando agentes de ameaça usam identidades falsas para se envolver em formas sofisticadas de Comportamento Inautêntico, definimos isso como Comportamento Inautêntico Coordenado (CIB).

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.3.7 Transparência e Relatórios Relevantes para as Autoridades Eleitorais Brasileiras

Divulgação Pública

A Meta publica regularmente suas constatações sobre operações de Comportamento Inautêntico Coordenado em seus Relatórios de Ameaças Adversárias. Esses relatórios incluem informações sobre redes interrompidas em todo o mundo, incluindo sua origem atribuída, públicos-alvo, táticas empregadas e escala de atividade, e contribuem para a compreensão pública mais ampla da evolução das ameaças. Esses relatórios não têm a intenção de cobrir todo o universo de aplicações de medidas sob a política de Comportamento Inautêntico, mas ajudam a informar a compreensão de nossa comunidade sobre a natureza em evolução das ameaças que enfrentamos neste espaço.

Comunicação com Autoridades Relevantes

A Meta responde a solicitações legais válidas de dados de usuários de autoridades brasileiras, incluindo o TSE e as autoridades policiais, de acordo com a lei aplicável, os Termos de Serviço da Meta e sua Política de Privacidade. Os processos da Meta para responder a solicitações governamentais de dados são descritos em seus Relatórios de Transparência e em suas Diretrizes para Autoridades Policiais.

A Meta reconhece o importante papel do TSE na proteção da integridade das eleições brasileiras e permanece disponível para participar de forma construtiva, inclusive por meio de mecanismos de transparência pública, participação em diálogos institucionais e respostas tempestivas a solicitações legais de dados, para apoiar o objetivo compartilhado de salvaguardar os processos democráticos.

Aplicação Proativa e Automatizada
Buscamos continuamente ajustes e melhorias em nossas políticas e sistemas, inclusive para melhorar a aplicação das políticas de Comportamento Inautêntico e garantir que os usuários não sejam penalizados erroneamente. Isso inclui atualizações contempladas nos protocolos de Comportamento Inautêntico sobre Construção de Público Inautêntico, Interferência Estrangeira de Empresas (FIB – *Foreign Interference from Businesses*) e Campanhas de Influência e Manipulação (IMC – *Influence and Manipulation Campaign*).

1.4 Arquitetura de Moderação de Conteúdo

1.4.1 Estrutura Geral: Remover, Reduzir e Informar

Empregamos uma abordagem em três partes para a aplicação de medidas no Facebook, Instagram e Threads: Remover, Reduzir e Informar. O conteúdo que vai contra nossos Padrões da Comunidade e outras políticas aplicáveis é removido de todas as nossas tecnologias; a distribuição de conteúdo problemático que não viola as diretrizes é reduzida (ou rebaixada); e/ou os usuários são informados e recebem contexto adicional para tomar decisões sobre o conteúdo que consomem.

Nossos Padrões da Comunidade se aplicam igualmente a todos, em todos os lugares e a todos os tipos de conteúdo. Eles formam a estrutura fundamental do nosso ecossistema de integridade para que possamos ajudar a manter os usuários seguros e manter um ambiente confiável, equitativo e seguro. Revisamos e atualizamos regularmente nossas políticas, que podem ser acessadas por meio da Central de Ajuda e da Central de Transparência.

1.4.2 Medidas de Resposta em Nível de Conteúdo

Remoção
O conteúdo que viola nossos Padrões da Comunidade é removido de todos os nossos serviços. Mantemos uma política de advertências (*strikes*) baseada em conteúdo, incluindo contagem de advertências e restrições de conta, penalidades crescentes com base na gravidade ou frequência das violações e desativação de violadores persistentes, violadores reincidentes e violadores flagrantes.

Distribuição Reduzida (Despriorização)
A distribuição de conteúdo problemático que não viola as diretrizes de nossos Padrões da Comunidade, ou conteúdo onde a aplicação está sendo recalibrada, é reduzida. Isso significa que o conteúdo é rebaixado no Feed e em outras superfícies de recomendação, limitando seu alcance enquanto preserva a expressão do usuário.

Rotulagem e Informação

Telas de aviso, intersticiais e desfoque de conteúdo são abordagens de produto acomodadas sob o pilar "Informar". Elas são aplicadas a possíveis conteúdos que não violam os Padrões da Comunidade, mas ainda assim podem ser perturbadores ou sensíveis, a fim de proteger a liberdade de expressão subjacente, permitindo que os indivíduos escolham se desejam visualizar o conteúdo.

Conforme abordado na seção 1.2, o Programa de Verificação de fatos independente segue em operação no Brasil. Sob este programa, aplicamos avisos a publicações verificadas e enviamos notificações às pessoas que as postaram, permitindo que as pessoas vejam o que os verificadores de fatos concluíram e decidam por si mesmas o que ler, confiar ou compartilhar.

Desmonetização

Os usuários devem cumprir nossas Políticas de Monetização para Parceiros para acessar ferramentas de monetização de criadores, como a Monetização de Conteúdo do Facebook ou Estrelas no Facebook e Instagram. Esta política proíbe a monetização de certas figuras políticas e cívicas, incluindo: atuais funcionários do governo eleitos e nomeados, atuais candidatos políticos, partidos políticos, comitês políticos registrados e agências e departamentos governamentais. Quando identificamos usuários nessas categorias, a Meta desmonetiza suas contas enquanto forem membros dessas categorias, o que significa que não podem obter receita com ferramentas de monetização de criadores.

Restrições de anúncios

O conteúdo pago (Anúncios) está sujeito aos Padrões da Comunidade da Meta, bem como aos Padrões de Publicidade da Meta, que são mais restritivos do que os Padrões da Comunidade que se aplicam ao conteúdo orgânico. Todos os anúncios são revisados antes de irem ao ar e permanecem sujeitos a nova revisão após a publicação. Quando se descobre que os anúncios violam nossas políticas ou a lei aplicável, eles são removidos. Violações repetidas de publicidade também podem resultar na perda de privilégios de publicidade.

Suspensão e Desativação de Conta

As contas podem ser suspensas ou desativadas permanentemente, dependendo da gravidade das violações individuais ou das medidas de aplicação cumulativas. Nosso sistema de advertências opera em limites crescentes: contas que acumulam um número definido de advertências padrão dentro de um período de um ano são desativadas; um limite menor se aplica a violações graves; e certas violações flagrantes (por exemplo, exploração sexual infantil, terrorismo) resultam na desativação imediata da conta após uma única violação. Usuários com advertências ativas também podem enfrentar restrições temporárias de produtos (por exemplo,

incapacidade de publicar em Grupos, criar anúncios ou usar Vídeo ao Vivo) proporcionais ao seu histórico de violações.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.4.3 Abordagem para Conteúdo Localmente Ilegal

Consulte a seção 1.7 abaixo para informações sobre os esforços buscando restringir conteúdo localmente ilegal.

1.4.4 Salvaguardas Contra Intervenção Indevida em Conteúdo Lícito

Salvaguardas Proativas Contra Aplicação Excessiva

Para minimizar a intervenção indevida em conteúdo lícito, nossos sistemas incorporam as seguintes salvaguardas:

- **Garantia de qualidade contínua**: mantemos programas de medição para avaliar a precisão das decisões de aplicação automatizadas e humanas e impulsionar a melhoria contínua na precisão da aplicação.
- **Recalibração para medidas menos restritivas**: conforme descrito na Seção 1.4.2, estamos recalibrando a aplicação para tipos de conteúdo mais adequados à aplicação reativa — mudando da remoção proativa para a distribuição reduzida — reduzindo o risco de remover incorretamente expressões lícitas.

Declaração de Motivos e Transparência

Quando a Meta toma uma medida de aplicação contra o conteúdo ou a conta de um usuário no Facebook ou Instagram, enviamos ao usuário afetado uma notificação junto com uma Declaração de Motivos explicando a ação. A notificação identifica a política específica violada (Padrões da Comunidade ou outros termos relevantes), a ação tomada e, quando aplicável, como o usuário pode solicitar a revisão da decisão.

A notificação é entregue no aplicativo por meio de uma experiência dedicada "Como tomamos essa decisão" no Facebook e no Instagram; o usuário não precisa navegar para um aplicativo ou navegador separado para ver a explicação. Os usuários também podem visualizar seu histórico completo de aplicação e quaisquer restrições de conta em vigor por meio do Status da Conta no Facebook e no Instagram. Isso se aplica igualmente ao Facebook, Instagram e Threads.

Recursos e Vias de Contestação do Usuário

Quando uma decisão de aplicação é passível de recurso, o usuário recebe uma opção dentro do próprio aplicativo, diretamente na Declaração de Motivos, para solicitar uma

reanálise. O botão de recurso integra a notificação da medida aplicada, juntamente com quaisquer outras opções de autocorreção disponíveis. Os usuários também podem iniciar ou acompanhar um recurso por meio do Status da Conta no Facebook e no Instagram, ou pelo fluxo correspondente na Central de Ajuda. Os recursos são analisados e, nos casos em que a decisão inicial estava incorreta, o conteúdo é restaurado e quaisquer advertências ou penalidades associadas são revertidas.

Comitê de Supervisão
O Comitê de Supervisão é um grupo independente de especialistas que protege a liberdade de expressão, tomando decisões independentes e baseadas em princípios em relação aos desafios mais difíceis de moderação de conteúdo no Facebook e no Instagram. Quando os usuários esgotam nosso processo de recurso interno, eles têm o direito de submeter o caso ao Comitê. A Meta também pode encaminhar casos significativos e difíceis diretamente. Quando o Comitê chega a uma decisão, ela é vinculante para a Meta (a menos que a implementação viole as leis locais). O Comitê também emite recomendações sobre políticas e operações relevantes da Meta, informadas por padrões internacionais de direitos humanos.

1.5 Aplicação das Políticas da Meta por meio da Moderação de Conteúdo para Remover Conteúdo Nocivo e Manter as Pessoas Seguras

Empregamos um sistema de moderação de conteúdo em várias camadas que combina tecnologia automatizada com revisão humana para monitorar continuamente o conteúdo em nossas plataformas. Nosso objetivo principal é remover o mais rápido possível o conteúdo violador que possa prejudicar os usuários, sendo que nossos sistemas são dinâmicos, respondendo a ameaças emergentes e problemas específicos.

A Meta aplica suas políticas principalmente por meio de IA e sistemas automatizados, complementados por revisão humana para casos que exigem julgamento contextual.

Os sistemas de IA da Meta detectam, restringem e removem conteúdos e contas violadoras em escala. Esses sistemas automatizados processam milhões de conteúdos diariamente, com a Meta relatando uma redução de aproximadamente 50% nos erros de aplicação após melhorias em seus modelos de IA em 2025. Nos casos em que os sistemas automatizados não conseguem tomar uma determinação com confiança suficiente, o conteúdo é encaminhado para revisão humana, com priorização baseada na gravidade, viralidade e probabilidade de uma violação.

A Meta publica dados de aplicação em seus Relatórios de Transparência semestrais, que fornecem métricas sobre o conteúdo objeto de medidas, taxas de detecção proativa e resultados de recursos em cada área de política.

Aplicação consistente independentemente do ponto de entrada
Aplicação consistente independentemente do ponto de entrada. Quando um conteúdo é identificado como potencialmente violador, ele é avaliado à luz de todos os Padrões da Comunidade aplicáveis, não apenas aquele que motivou a detecção inicial. Isso significa que, mesmo quando uma hipótese de risco abrange múltiplos tipos de Política, o conteúdo fica sujeito à aplicação do padrão aplicável mais protetivo.

Descrição Funcional
Nossa aplicação de conteúdo opera por meio de duas vias complementares: aplicação proativa e aplicação reativa. A aplicação proativa refere-se à ação tomada antes que o conteúdo seja denunciado ou, em alguns casos, antes que seja amplamente visualizado. A aplicação reativa ocorre em resposta a denúncias de usuários ou outros sinais externos.

O conteúdo que pode violar nossos Padrões da Comunidade é inicialmente avaliado por sistemas automatizados, que poderão não tomar ação alguma ou remover automaticamente, aplicar restrição de idade ou empregar outras ações apropriadas, como intersticiais de 'Marcar como Sensível' ou, ainda, se esses sistemas não puderem determinar com confiança suficiente se o conteúdo viola nossas políticas, enviar o conteúdo para revisores humanos treinados para avaliação adicional.

1.5.1 Sistemas Automatizados

Usamos modelos de aprendizado de máquina, conhecidos como classificadores, para identificar conteúdo que potencialmente viola as políticas. Esses classificadores avaliam todo o conteúdo publicado no Facebook, Instagram e Threads, e atribuem uma pontuação ao conteúdo para informar qual medida de aplicação, se houver, deve ser tomada.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Revisão Humana
Nossas equipes de revisão humana em todo o mundo incluem funcionários e revisores de conteúdo de diversas origens, refletindo nossa comunidade diversificada, e incluem especialistas em áreas de políticas como segurança infantil, conduta de ódio e contraterrorismo. O conteúdo sinalizado por ferramentas automatizadas é encaminhado para moderadores humanos com base na pontuação de confiança do

classificador e na gravidade da possível violação. Os revisores humanos passam por treinamento extensivo, calibração e avaliação contínua para garantir que tomem continuamente a ação correta sobre o conteúdo. A garantia de qualidade é gerenciada por meio de:

- Programas de treinamento padronizados e exames de certificação obrigatórios
- Monitoramento contínuo de desempenho e atualizações regulares de treinamento
- Auditorias regulares de garantia de qualidade das decisões tomadas por todos os moderadores de conteúdo, incluindo moderadores terceirizados, para garantir o alinhamento com nossos Padrões da Comunidade e diretrizes operacionais

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Melhoria Contínua e Monitoramento
Nossos classificadores internos são continuamente calibrados e refinados, com novos treinamentos realizados regularmente. Diferentes classificadores têm diferentes tipos de monitoramento que a equipe responsável configura. Além de alertas automatizados, mantemos vários painéis de monitoramento que a equipe responsável de um determinado classificador estabelece e mantém.

Ferramentas internas são aproveitadas para monitoramento e alertas. Quando o erro do classificador excede um certo limite, uma tarefa é criada automaticamente e alerta a equipe de plantão com alta prioridade, que então examina e depura os problemas.

Controles de Qualidade Contínuos

- Verificações por amostragem de decisões automatizadas (taxa de amostragem de 30% durante a implementação inicial).
- Análise de padrões de erro para melhorar continuamente nossos sistemas de detecção.
- Treinamento dedicado para revisores humanos em cada atualização de política.
- Análise de causa raiz quando erros são identificados.

Metodologia de Indicadores de Qualidade de Detecção
Avaliamos nossa detecção e aplicação de conteúdo violador por meio de vários processos de garantia de qualidade e melhorias contínuas do sistema. As métricas usadas para avaliar o desempenho variam de acordo com diferentes fatores, como ferramenta e contexto. Os seguintes indicadores são os mais comumente usados para avaliar a detecção e aplicação da remoção de conteúdo em nossas plataformas.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Processos Adicionais de Avaliação (Facebook, Instagram e Threads)
Nossos processos para avaliar a precisão dos sistemas automatizados de moderação de conteúdo incluem:

- Monitoramento e revisão de métricas, incluindo precisão, revocação (recall), prevalência e/ou proxies de tais métricas, dependendo do que estiver disponível, geralmente no nível da política.
- Testes offline e online, incluindo testes A/B em que diferentes usuários são submetidos a diferentes tecnologias e o desempenho é comparado, bem como testar o desempenho de uma determinada tecnologia em relação ao de revisores humanos.

Também investimos em sistemas para detectar e prevenir erros de moderação de conteúdo antes que ocorram (como nosso sistema de "prevenção de erros"), e em processos de denúncia e recurso de usuários para corrigir erros que venham a surgir, incluindo o Comitê de Supervisão.

1.5.3 Metodologia de Ações Corretivas/Preventivas

Para **sistemas de revisão humana**, nossa metodologia corretiva inclui: (i) treinamento dedicado para cada atualização de política; (ii) monitoramento e resposta a métricas de qualidade que avaliam a precisão dos rótulos atribuídos por revisores humanos; e (iii) quando apropriado, a condução de uma análise de causa raiz de erros específicos.
A garantia de qualidade para o treinamento de moderadores é gerenciada por meio de programas de treinamento padronizados, exames de certificação obrigatórios e monitoramento contínuo de desempenho. Realizamos atualizações regulares de treinamento, bem como treinamentos adicionais conforme necessário quando padrões anormais de aplicação são observados. Sessões contínuas de calibração e ciclos de feedback ajudam a manter o alinhamento e a resolver ambiguidades.

Para **sistemas automatizados de moderação de conteúdo**, avaliamos e melhoramos continuamente por meio de:

- Monitoramento e revisão de métricas, incluindo precisão, recall e prevalência no nível da política.
- Testes offline e online, incluindo testes A/B comparando diferentes tecnologias e testando o desempenho em relação a revisores humanos.

- Um sistema especializado de reavaliação envolvendo um ciclo de feedback entre revisores humanos para garantir a compreensão cultural e contextual. Denúncias e contestações de usuários fornecem um mecanismo para corrigir erros de aplicação, e as decisões de rotulagem de conteúdo por revisores humanos são usadas para treinar e retreinar classificadores para melhorar a precisão, levando em consideração a evolução do comportamento do usuário e as tendências de conteúdo.

Se identificarmos que as métricas monitoradas se desviam significativamente de seu nível normal ou esperado, investigamos para determinar a causa e tomamos medidas para corrigir o desvio (por exemplo, retreinando controles automatizados de conteúdo).

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Propagação de Medidas de Aplicação
Nosso sistema de Propagação é um mecanismo central de aplicação que assegura que medidas tomadas sobre uma entidade sejam aplicadas de forma consistente a entidades relacionadas. Ele gerencia a propagação de medidas do conteúdo-fonte para conteúdos vinculados, por exemplo, garantindo que, quando uma publicação principal sofre uma medida de aplicação, os anexos associados também sejam abrangidos. Esse sistema integra-se à infraestrutura mais ampla de aplicação para gerenciar tanto as decisões de aplicação direta quanto os fluxos de restauração quando medidas são objeto de recurso, operando de forma transversal no Facebook, Instagram e Threads.

1.6 Esforços para remover conteúdo localmente ilegal

1.6.1 Abordagem Proativa para Remover Conteúdo Ilegal

Quando nossas políticas não cobrem substancialmente o conteúdo eleitoral ilegal, empregamos uma abordagem em várias camadas combinando tecnologia de detecção automatizada com revisão humana para identificar e agir proativamente sobre o conteúdo que é localmente ilegal nos termos da legislação eleitoral brasileira, incluindo conteúdo que não viola nossos Padrões da Comunidade globais.

Nosso sistema de detecção proativa usa pipelines baseados em palavras-chave e em declarações especificamente projetadas para identificar o que pode se enquadrar como conteúdo eleitoral localmente ilegal. Semelhante à abordagem descrita na Seção 1.4, usamos um sistema em várias camadas que combina essas tecnologias

para mitigar proativamente os riscos de conteúdo eleitoral ilegal ser publicado em nossas plataformas e, quando necessário, para detectar e aplicar medidas contra esse conteúdo em escala.

Nossa tomada de decisão de aplicação opera por meio de uma estrutura de revisão estruturada e em várias camadas. As revisões manuais de violações de nossas políticas são realizadas seguindo orientações detalhadas sobre conteúdo localmente ilegal, quando aplicável. Para denúncias de conteúdo ilegal, as revisões são conduzidas por equipes de operações especializadas. O processo de revisão de conteúdo ilegal envolve várias camadas. **A seguir está o processo geral de revisão.**

- Quando recebemos uma denúncia, primeiro a revisamos em relação aos Padrões da Comunidade. Se determinarmos que o conteúdo vai contra nossas políticas, nós o removemos. Se o conteúdo não for contra nossas políticas, em linha com nossos compromissos como membro da Global Network Initiative e de nossa Política Corporativa de Direitos Humanos, conduzimos uma análise jurídica cuidadosa para confirmar se a denúncia é válida, bem como a devida diligência em direitos humanos.
- Nos casos em que acreditamos que as denúncias não são legalmente válidas, são excessivamente amplas ou são inconsistentes com os padrões internacionais de direitos humanos, podemos solicitar esclarecimentos ou não tomar nenhuma ação.
- Quando agimos contra conteúdo orgânico com base na lei local em vez de nossos Padrões da Comunidade, restringimos o acesso ao conteúdo apenas na jurisdição onde se alega ser ilegal e não impomos quaisquer outras penalidades ou restrições de funcionalidades. Também notificamos o usuário afetado.

Indisponibilidade Imediata para Tipos de Conteúdo Enumerados
A grande maioria das condutas enumeradas no Art. 9-E da Portaria já constitui uma violação dos Padrões da Comunidade globais da Meta e, portanto, é removida do Facebook, Instagram e Threads em todo o mundo após a detecção, em escala, por meio dos sistemas proativos descritos nas seções anteriores.

Conforme abordado na seção 1.2 Visão Geral dos Padrões da Comunidade da Meta, embora a estrutura de políticas da Meta não seja organizada nas mesmas linhas categóricas da Resolução Eleitoral, cada hipótese de risco pode ser abordada por múltiplos Padrões da Comunidade sobrepostos que funcionam em combinação para fornecer cobertura.

- A conduta antidemocrática é amplamente coberta pelas políticas da Meta sobre Violência e Incitação, Coordenação de Atos Danosos e Incentivo à Prática de Atividades Criminosas e Segurança Cibernética.

- Da mesma forma, o conteúdo caracterizado como incitação a crimes contra o Estado Democrático pode envolver simultânea ou individualmente as políticas da Meta sobre Violência e Incitação, Interferência Eleitoral e Supressão de Eleitores ou Coordenação de Atos Danosos e Incentivo à Prática de Atividades Criminosas de forma mais ampla.
- Fatos falsos/descontextualizados que afetam a integridade da votação e o conteúdo que desacredita o sistema eletrônico de votação são cobertos pela nossa categoria de Interferência na Votação e no Censo, dentro das políticas de Coordenação de Atos Danosos e Incentivo à Prática de Atividades Criminosas e de Desinformação.
- A violência política contra as mulheres pode envolver simultaneamente as políticas da Meta sobre Violência e Incitação, Conduta de Ódio e Bullying e Assédio
- O conteúdo sintético de IA é regido pelas políticas da Meta Aplicáveis a Conteúdo Gerado por IA detalhadas na seção 4, enquanto a seção 7.2 mais abaixo também descreve nossa abordagem para conteúdo anteriormente objeto de ordens judiciais

Onde uma categoria enumerada não é totalmente coextensiva com as políticas globais da Meta, a Meta complementa a aplicação baseada em políticas com a restrição automática de certos tipos de conteúdo orgânico em escala — juntamente com denúncias de usuários e canais de denúncia priorizados disponíveis para a Justiça Eleitoral, entidades políticas e usuários.

1.6.2 Compromisso de Monitoramento

A Meta revisará o desempenho de nossas restrições específicas para eleições durante todo o período de campanha, inclusive durante os períodos de silêncio.

1.7 Sistemas de Recomendação

Nossos sistemas de recomendação são projetados para ajudar os usuários a descobrir conteúdo que possam achar útil, interessante, relevante e valioso. Publicamos Diretrizes de Recomendação nas Centrais de Ajuda para ajudar as pessoas a entender os tipos de conteúdo que recomendamos e por que alguns conteúdos não são incluídos nas recomendações.

Nossos sistemas de recomendação são alimentados por vários sistemas de IA que trabalham separadamente e em conjunto para identificar conteúdo e prever o interesse do usuário. Eles são projetados para evitar a recomendação, recirculação ou amplificação de conteúdo que potencialmente viole as políticas ou que seja de outra forma problemático. Os sistemas:

- Produzem um inventário do conteúdo disponível e o filtram para remover o conteúdo que potencialmente viola políticas e padrões;
- Filtram adicionalmente para excluir conteúdo não elegível para recomendação;
- Rebaixam a classificação do conteúdo por vários motivos e não recomendam contas relacionadas a infratores reincidentes de políticas.

Avaliamos continuamente se nossos sistemas de recomendação contribuem para riscos sistêmicos. Por exemplo, avaliamos que os sistemas de recomendação são projetados para promover interações positivas e melhorar a experiência do usuário, fornecendo informações mais relevantes, como por meio de recursos como o *Feed Deep Dive,* que traz à tona conteúdo existente alinhado aos interesses do usuário sem introduzir novo conteúdo.

1.7.1 Medidas de Transparência para Resultados

Publicamos **Diretrizes de Recomendação** nas Centrais de Ajuda do Facebook e do Instagram para ajudar os usuários a entender os tipos de conteúdo que recomendamos e fornecer contexto sobre por que alguns tipos de conteúdo não são incluídos nas recomendações.

Publicamos um **Relatório de Aplicação dos Padrões da Comunidade (CSER)**, que acompanha nosso desempenho na aplicação de políticas com referência a métricas como:

- Prevalência de conteúdo violador (porcentagem de visualizações de conteúdo que são de conteúdo violador)
- Taxa de moderação proativa (porcentagem de conteúdo violador detectado e alvo de ação antes das denúncias dos usuários)
- Volume de conteúdo alvo de ação
- Volume de conteúdo restaurado após contestação

O CSER fornece dados históricos, permitindo que as partes interessadas entendam como nosso desempenho mudou ao longo do tempo.

Critérios de Suficiência para Medidas Adotadas

O Padrão "Razoável, Proporcional e Eficaz"

Ao avaliar se devemos implantar uma medida de mitigação ou aprimoramento, aplicamos três critérios:

- **Razoável**: A medida pode ser implantada com dependência limitada em partes externas ou entidades não pertencentes à Meta, e é apropriada, justa e projetada para lidar com riscos de integridade.

- **Proporcional:** A medida é adequada, pertinente, apropriada e necessária para enfrentar os riscos sistêmicos especificados, e não é excessiva em relação a uma finalidade declarada e especificada e à exposição a risco residual.
- **Eficaz:** A medida reduz comprovadamente a exposição a risco quando avaliada por meio das pontuações de Eficácia de Concepção (*Design Effectiveness*) e Eficácia de Mitigação (*Mitigation Effectiveness*).

Determinação do Risco Residual
Após aplicarmos nossa metodologia de avaliação de controles, determinamos se os riscos sistêmicos estão suficientemente mitigados. Nossa conclusão é a de que os riscos sistêmicos estão suficientemente mitigados pelos controles razoáveis, proporcionais e eficazes existentes quando a pontuação de Risco Residual, obtida pela multiplicação do Risco Inerente pela Eficácia do Conjunto de Controles (*Control Suite Effectiveness*), demonstra níveis aceitáveis de risco. Quando esse padrão é atendido, determinamos que não são necessárias medidas adicionais de mitigação.

Metas de Desempenho e Monitoramento
Definimos metas de desempenho tanto para os sistemas humanos quanto para os sistemas automatizados de moderação de conteúdo, com foco na redução tanto da aplicação insuficiente (violações não detectadas) quanto da aplicação excessiva (remoção de conteúdo não violador). Essas metas equilibram os direitos dos usuários à liberdade de expressão e de informação com a segurança. As metas estão sujeitas a alterações, em função da evolução dos ambientes regulatórios e da natureza dos vetores de ameaça e dos danos, o que exige o refinamento contínuo das estratégias de avaliação e resposta a riscos.

1.7.2 Sinais do Sistema de Recomendação para Conteúdo Político-Eleitoral

Definição de Conteúdo Político/Cívico
Usamos os termos “cívico” e “político” de forma intercambiável para nos referirmos ao mesmo conteúdo.

Facebook
Usamos sistemas de IA para personalizar o conteúdo que as pessoas veem e, como parte desse trabalho, usamos sinais diretos e indiretos em nossa abordagem de classificação de conteúdo. Implantamos um modelo classificador que detecta se o conteúdo se refere a temas cívicos/políticos. Esse classificador opera sobre o próprio conteúdo (texto, imagens e contexto da publicação) para produzir uma pontuação de predição que indica a probabilidade de determinado conteúdo tratar de temas políticos ou cívicos.

Configurações de Controle do Usuário (atualizadas em janeiro de 2025)
Desde 2021, fizemos alterações para reduzir a quantidade de conteúdo cívico que as pessoas veem. Em janeiro de 2025, começamos a reverter gradualmente essa abordagem, adotando um modelo mais personalizado, para que as pessoas que desejam ver mais conteúdo político em seus feeds possam fazê-lo. A configuração de controle de Conteúdo Político no Facebook permite que as pessoas escolham quanto conteúdo político veem.

Como o rebaixamento funciona na prática
Na configuração "Padrão", o conteúdo cívico pode aparecer nas Superfícies de Recomendação dos usuários, mas será "rebaixado", ou seja, aparecerá com menos destaque do que outras formas de conteúdo não sujeitas a controle. O conteúdo cívico não aparecerá entre as principais publicações recomendadas no Feed do usuário, mas poderá aparecer à medida que o usuário rola a tela.

Instagram
O Instagram aplica o mesmo classificador de conteúdo cívico e a mesma abordagem do Facebook. O objetivo é que o conteúdo político (conforme identificado pelo classificador) não seja recomendado nas superfícies de recomendação do Instagram (incluindo Feed, Reels e Explorar) para usuários na configuração "Ver Menos", enquanto usuários na configuração "Padrão" veem esse conteúdo em posição rebaixada nas Superfícies de Recomendação.

Aplicam-se as mesmas configurações de controle do usuário: "Ver Mais", "Padrão" e "Ver Menos". As alterações introduzidas no controle de conteúdo político do Instagram não afetam o processo de classificação de conteúdo como cívico — afetam apenas o posicionamento desse conteúdo nas Superfícies de Recomendação, de acordo com as preferências manifestadas pelos usuários.

Threads
Aplicamos a mesma abordagem ao Threads. Desde janeiro de 2025, implantamos no Threads uma abordagem mais personalizada, para que as pessoas que desejam ver mais conteúdo político em seus feeds possam fazê-lo.

O mesmo classificador de conteúdo cívico se aplica ao Threads. As configurações de controle do usuário ("Ver Mais", "Padrão", "Ver Menos") funcionam de forma idêntica ao Instagram — afetando o posicionamento nas Superfícies de Recomendação sem alterar o processo de classificação subjacente.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.7.4 Alterações Planejadas para o Período Eleitoral

O classificador de conteúdo cívico e as configurações de controle de conteúdo político não mudarão para o período eleitoral; a abordagem personalizada em vigor desde janeiro de 2025 continua. Duas alterações no comportamento de recomendação se aplicam:

- **Não recomendação de perfis e canais informados pela Justiça Eleitoral.** Em conformidade com o Artigo 28, §1º-A da Resolução nº 23.610/2019/TSE, a partir do início do período de campanha, perfis e canais informados pelos candidatos ao TSE, juntamente com seu conteúdo orgânico não pago, são sistematicamente excluídos de serem recomendados ou exibidos. A lista oficial de candidatos fica disponível em 16 de agosto de 2026 e é inserida nos sistemas da Meta a partir do banco de dados do TSE.
- **Remoção de conteúdo localmente ilegal da elegibilidade de recomendação.** O conteúdo determinado como localmente ilegal, após revisão humana, é restrito no Brasil e registrado ("*banked*") para que o conteúdo correspondente também seja restrito. O conteúdo restrito não é elegível para recomendação no Brasil.

Ambas as alterações operam durante o período eleitoral, incluindo o período de segundo turno, se houver. A implementação ocorrerá após 16 de agosto de 2026, quando a lista oficial de candidatos for disponibilizada pelo Tribunal Superior Eleitoral.

Compromisso de Monitoramento

Visto que a eficácia dessas restrições específicas para as eleições depende de dados gerados durante o período eleitoral, a Meta assume um compromisso de monitoramento nos termos do Art. 7º, §4º, da Portaria nº 463/2026/TSE. A Meta revisará a operação e a eficácia dessas restrições ao longo do período de campanha, inclusive durante os períodos de vedação, e os valores efetivamente mensurados serão informados no relatório consolidado a ser apresentado ao TSE até 19 de dezembro de 2026.

SEÇÃO 2: Avaliação de Impacto da Meta na Integridade Eleitoral

As informações apresentadas nesta seção visam demonstrar o cumprimento, pela Meta, de seu dever de realizar uma avaliação de impacto sobre a integridade do processo eleitoral, abordando especificamente os requisitos do Artigo 20 da Portaria nº 463/2026/TSE, no que tange às obrigações estabelecidas no Artigo 9-D, inciso V, da Resolução nº 23.610/2019/TSE — dispositivos que exigem, respectivamente, que os provedores de aplicações de internet elaborem, em ano eleitoral, uma avaliação do impacto de seus serviços na integridade do processo eleitoral.

Nos itens a seguir, a Meta descreve a metodologia e a periodicidade de sua avaliação de impacto, os resultados consolidados em uma matriz de riscos e a metodologia de mensuração das medidas de mitigação adotadas, em conformidade com os incisos I a III e o parágrafo único do Artigo 20 da referida Portaria.

A Meta utiliza Avaliações de Risco Eleitoral para mitigar riscos sistêmicos que se materializam no contexto de processos eleitorais. Com base em uma priorização anual da atividade de eleições globais, existem três tipos de avaliações conduzidas como parte das Avaliações de Risco Eleitoral, a saber: (i) a avaliação de suporte em escala; (ii) as avaliações de linha de base iniciadas antes de uma eleição para planejar estratégias de mitigação adaptadas para essa eleição específica; e (iii) avaliações aprofundadas para certas eleições que podem apresentar riscos maiores para determinar a necessidade de desenvolver capacidades adicionais.

A Meta expande progressivamente seus esforços de preparação aprofundando-se um nível a cada etapa: (i) primeiro a Meta avalia se suas medidas 'Sempre Ativas' (*Always On*) são suficientes; (ii) se não, a Meta conduzirá uma avaliação de risco de linha de base; e (iii) quando há riscos atípicos, a Meta conduz uma avaliação de risco aprofundada adicional.

Essas avaliações formam uma estrutura que permite à Meta antecipar possíveis lacunas e riscos em seus serviços associados a eleições globalmente, garantir recursos apropriados para mitigar esses riscos e coordenar equipes trabalhando em unidades de negócios e áreas de danos para lidar com esses riscos.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

A avaliação de risco da Meta para as Eleições Gerais do Brasil de 2026 avaliou cinco Áreas Críticas: discurso hostil, desinformação, ameaças adversárias, riscos cívicos e discurso violento. Em todas as Áreas Críticas elevadas, a avaliação de risco identifica duas tendências transversais que informam nossa estratégia de mitigação: primeiro, adaptação adversária, onde atores exploram cada vez mais características específicas da plataforma, incluindo migração de conteúdo entre plataformas e linguagem codificada ou imagens difusas para burlar classificadores e contornar a aplicação de regras; e segundo, o efeito cumulativo de regimes regulatórios simultâneos, criando tanto maior exposição regulatória quanto obrigações reforçadas para ação proativa.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Para gerenciar janelas críticas, a Meta ativará um Centro de Operações de Produtos de Integridade (IPOC) no período que antecede as Eleições Gerais de 2026 no Brasil. Este hub centralizado reúne especialistas multifuncionais de equipes como inteligência de ameaças, ciência de dados, produto, engenharia, operações, políticas e consultores jurídicos, facilitando avaliações rápidas e decisões de aplicação em tempo real. Enquanto ativo, o centro está equipado para identificar, avaliar e mitigar materiais ilegais ou que violem políticas por meio de revisões manuais aceleradas, controles de distribuição programáticos e um canal de comunicação constante com o Tribunal Superior Eleitoral (TSE).

Além dos procedimentos operacionais padrão, a Meta mantém uma estrutura de contingência adaptada para períodos eleitorais para lidar com situações em que os vetores de risco superam as métricas de aplicação de linha de base. Este protocolo institui fluxos de trabalho pré-autorizados, implantação de recursos escalonáveis e parâmetros de tomada de decisão delegados, permitindo execução imediata sob pressão operacional sem a necessidade de buscar aprovações adicionais.

A estrutura de contingência tem como alvo vários cenários específicos: aumentos acentuados em conteúdo problemático que sobrecarregam a filtragem automatizada; o surgimento de métodos adversariais originais não detectados por algoritmos padrão; esforços de influência multiplataforma organizados e exigindo rápida identificação e contenção; e a rápida disseminação de conteúdo excepcionalmente grave em dias de eleição exigindo mitigação imediata.

SEÇÃO 3: Salvaguardas e Esforços de Transparência Relacionados à Publicidade Política

Esta seção apresenta as informações exigidas pelos Artigos 16, 18 e 19 da Portaria nº 463/2026/TSE, abordando três categorias inter-relacionadas de obrigações relativas a conteúdo político-eleitoral pago nas plataformas da Meta.

O Artigo 18 da Portaria especifica os elementos que o plano de conformidade deve conter em relação à transparência e integridade no impulsionamento de conteúdo político-eleitoral — incluindo a manutenção de uma biblioteca de anúncios, identificação do anunciante, a divulgação obrigatória do uso de IA no fluxo de contratação de anúncios — em cumprimento aos deveres estabelecidos nos Artigos 27-A; 28, §§ 7º-A e 7º-B e 29, §§ 6º e 7º, da Resolução nº 23.610/2019/TSE.

O Artigo 16 da Portaria especifica os elementos relativos à gestão do período de silêncio eleitoral, incluindo o desligamento automático de anúncios pagos, em cumprimento aos deveres estabelecidos no Artigo 29, § 11, da mesma Resolução.

Nos termos do Artigo 19 da Portaria nº 463/2026/TSE, ao apresentar este plano de conformidade, a Meta solicita expressamente seu credenciamento perante a Justiça Eleitoral, em cumprimento ao Artigo 29, § 9º, da Resolução nº 23.610/2019/TSE, como provedora que oferece impulsionamento pago de conteúdo sobre questões sociais, eleitorais e políticas (SIEP) no Brasil. Para fins informativos, a Meta declara que não oferece serviços de mecanismos de busca (Artigo 18, inciso V, da Portaria nº 463/2026/TSE).

As subseções abaixo descrevem os mecanismos, salvaguardas e procedimentos operacionais que a Meta mantém para lidar com cada uma dessas obrigações.

3.1 Definindo Anúncios sobre Temas Sociais, Eleições ou Política

A Meta aplica uma definição ampla que abrange todos os anúncios sobre temas sociais, eleições ou política. Categorizamos um anúncio como sendo sobre temas sociais, eleições ou política no Brasil quando ele:

- É feito por, em nome de, ou sobre um candidato a cargo público, figuras políticas, um partido político, um comitê de ação política, ou defende o resultado de uma eleição para cargo público; ou
- É sobre eleições, referendos ou iniciativas de votação, incluindo campanhas de incentivo ao voto ou eleitorais; ou
- É sobre quaisquer temas sociais no local onde o anúncio está sendo veiculado; ou
- É regulamentado como publicidade política.

Permitimos publicidade política no Facebook, Instagram e Threads, desde que os anunciantes cumpram todas as leis aplicáveis, nossos Padrões de Publicidade, Padrões da Comunidade e o processo de autorização exigido pela Meta. Nossa política de Anúncios sobre Temas Sociais, Eleições ou Política (SIEP) promove transparência, responsabilidade e autenticidade, pois exige maior transparência de autoridades eleitas e nomeadas, candidatos a cargos públicos e anunciantes de conteúdo que inclua temas sociais, conteúdo de anúncios eleitorais ou políticos.

3.1.1 Rótulo do Anúncio

No Facebook, Instagram e Threads, todos os anúncios exibem um rótulo "Anúncio" ou "Patrocinado" que é um componente intrínseco e não opcional do bloco de anúncios. Este rótulo é renderizado automaticamente por meio de caminhos de código específicos da plataforma para cada superfície e posicionamento. A mesma lógica de rotulagem se aplica independentemente do dispositivo (iOS, Android, desktop, FBLite) ou tipo de posicionamento.

Os anúncios no Facebook, Instagram e Threads incluem este rótulo "Anúncio" para que os usuários saibam que estão visualizando conteúdo pago, distinguindo-o de conteúdo orgânico publicado por indivíduos, empresas ou criadores.

Além do rótulo "Anúncio", os anúncios sobre temas sociais, eleições ou política no Facebook, Instagram e Threads carregam um aviso "Pago por..." ou "Propaganda Eleitoral" exibido na parte superior do bloco de anúncios. O aviso é preenchido a partir do registro de autorização do anunciante e é um campo obrigatório — um anunciante não pode publicar um anúncio de tema social, eleitoral ou político no Brasil sem primeiro concluir o processo de autorização e anexar um aviso ao anúncio.

O aviso nomeia a entidade ou indivíduo que paga pelo anúncio, que para a publicidade eleitoral brasileira é o candidato, partido, federação ou coligação que contratou a veiculação (veja "Escopo e Definição de Partes Contratantes" e "Processo de Verificação de Anunciante" abaixo). Juntamente com o rótulo "Anúncio", ele distingue a publicidade eleitoral paga do conteúdo político orgânico e identifica a parte contratante para o espectador no momento de entrega do anúncio.

3.1.2 Preservação da Identificação em Conteúdo Compartilhado/ Encaminhado

O rótulo "Anúncio" está intrinsecamente ligado à entrega paga do bloco de anúncios — ele se aplica apenas a conteúdo cuja veiculação foi paga. Quando um usuário compartilha, reposta ou cita um anúncio em sua própria conta, o conteúdo compartilhado resultante se torna uma publicação orgânica (conteúdo gerado pelo usuário) e não é, por si só, um anúncio pago.

O aviso "Pago por" ou "Propaganda Eleitoral" se comporta de maneira diferente do rótulo "Anúncio" quando o conteúdo sai de seu contexto de entrega paga. O rótulo "Anúncio" está vinculado à veiculação paga e, portanto, não é transferido para a republicação orgânica de um usuário. Por outro lado, o aviso de temas sociais, eleitoral ou político é carregado no próprio criativo do anúncio e permanece visível quando o anúncio é compartilhado organicamente.

O efeito prático é que a identidade do pagador acompanha o conteúdo: um espectador que encontra um anúncio sobre temas sociais, eleitoral ou político compartilhado em uma superfície orgânica ainda vê quem pagou por ele, embora o compartilhamento em si não tenha sido uma veiculação paga.

3.2 Critérios para Análise, Aprovação, Recusa, Suspensão e Remoção de Anúncios Políticos no Facebook, Instagram e Threads

3.2.1. Critérios para Análise de Anúncios Políticos

Anúncios Proibidos no Contexto Eleitoral

Rejeitamos anúncios que desencorajam as pessoas de votar ou que questionam a legitimidade das próximas eleições. Anúncios relacionados à votação no contexto das eleições (incluindo eleições primárias, gerais, especiais e de segundo turno) estão sujeitos a proibições adicionais e podem ser rejeitados se violarem nossas políticas. Certos conteúdos relacionados a eleições podem ser proibidos por lei local ou removidos em regiões específicas antes da votação.

Processo de Revisão de Anúncios

Nosso processo de revisão de anúncios é um componente crítico para manter a transparência e a conformidade. Temos várias camadas de análise e detecção, tanto antes quanto depois que um anúncio vai ao ar.

Revisão pré-publicação

Antes que os anúncios sejam veiculados no Facebook, Instagram e Threads, nós os revisamos em relação aos nossos Padrões de Publicidade. O processo de revisão começa automaticamente após um anúncio ter sido criado ou editado. A revisão pode incluir componentes específicos de um anúncio, como imagens, vídeo, texto e informações de direcionamento, bem como a página de destino associada ao anúncio ou outros destinos.

Detecção automatizada: usamos modelos de aprendizado de máquina chamados "classificadores" para revisar todos os anúncios veiculados em nossas plataformas em busca de violações de nossas políticas de anúncios. Os classificadores são elaborados por engenheiros e cientistas de dados para detectar tipos ou categorias específicas de danos, treinados em conteúdo conhecido por ser violador, bem como em conteúdo conhecido por ser benigno.

Revisão pós-publicação: todos os anúncios podem ser revisados novamente a qualquer momento. Anúncios aprovados anteriormente podem ser selecionados para revisão adicional por motivos que incluem amostragem aleatória para garantir a precisão, níveis inesperadamente altos de engajamento (especialmente quando as

pessoas denunciam, bloqueiam ou ocultam anúncios) ou quando um anúncio pode ter escapado ao nosso processo de revisão inicial.

Revisão humana: em alguns casos, a revisão manual complementa nossos sistemas automatizados. Quando revisores humanos avaliam uma contestação, eles podem acolher o recurso e reverter a reprovação, ou manter a decisão inicial.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

3.2.2. Critérios para Aprovação: Autorização e Requisitos de Rótulo

Processo de Autorização para o Brasil
Qualquer anunciante que veicule anúncios sobre temas sociais, eleições ou política localizado ou direcionado a pessoas em países designados — incluindo o Brasil — deve concluir o processo de autorização exigido pela Meta, que inclui verificação de identidade, antes de poderem veicular tal anúncio.

Nosso processo de autorização exige que os anunciantes:

- **Verifiquem sua identidade** por meio de um processo de autenticação em várias etapas confirmando quem são e seu país de residência.
- **Incluam avisos "Pago por" ou "Propaganda Eleitoral"** em todos os anúncios políticos, eleitorais e sobre temas sociais, mostrando quem pagou pelo anúncio.
- **Operem dentro de restrições geográficas** — os anunciantes só podem veicular anúncios SIEP no país onde estão autorizados.

Se o anúncio promove uma candidatura
Um anúncio que promove uma candidatura — isto é, um anúncio feito por, em nome de, ou sobre um candidato a cargo público, ou que defende o resultado de uma eleição — aciona o requisito completo de autorização SIEP. Os anunciantes devem verificar sua identidade por meio do processo de autorização e incluir um aviso verificado "Pago por" ou "Propaganda Eleitoral" nesses anúncios para mostrar a entidade ou pessoa responsável por veicular o anúncio.

Todos os anúncios políticos e sobre temas sociais são armazenados na Biblioteca de Anúncios — um banco de dados público pesquisável de todos os anúncios políticos, eleitorais e sobre temas sociais dos últimos 7 anos — para que o público possa ver o que os candidatos estão dizendo, que públicos estão segmentando e quem pagou por eles.

3.2.3. Critérios Específicos em Relação aos Tipos de Conteúdo Enumerados

Todos os anúncios nas plataformas da Meta estão sujeitos aos nossos Padrões da Comunidade como requisito básico; conteúdo que viola os Padrões da Comunidade em sua forma orgânica é igualmente proibido em forma paga e sujeito às mesmas medidas de aplicação, incluindo remoção imediata. Assim, tudo o que foi capturado na Seção 1.6 - Indisponibilidade Imediata para Tipos de Conteúdo Enumerados (Art. 9-E) - permanece aplicável para o conteúdo pago sob as mesmas categorias.

Especificamente em relação ao Artigo 18, IV, da Portaria, vale ressaltar que as políticas da Meta cobrem, entre os critérios existentes para revisão proativa de conteúdo, conteúdo que possa representar Violência política contra mulheres, Violência e Incitação, Conduta de Ódio e Bullying e Assédio, todos os quais podem abordar preocupações ligadas à propaganda eleitoral negativa.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

A análise para aprovação, rejeição, suspensão ou remoção de conteúdo político-eleitoral impulsionado é realizada de acordo com os critérios já listados e com os requisitos e limitações expressamente direcionados aos provedores de aplicação pela legislação eleitoral e pelas disposições da Resolução nº 23.610/2019/TSE.

3.3 Mecanismo Obrigatório de Declaração de Uso de IA no Fluxo de Impulsionamento

Nossos Padrões de Publicidade de anúncios SIEP exigem que os anunciantes divulguem sempre que um anúncio contenha conteúdo fotorrealista ou com som realista que foi digitalmente criado ou alterado para:

- Representar uma pessoa real dizendo ou fazendo algo que ela não disse ou fez
- Representar uma pessoa com aparência realista que não existe ou um evento com aparência realista que não ocorreu, ou alterar filmagens de um evento real
- Representar um evento realista que supostamente ocorreu, mas que não é uma imagem, vídeo ou áudio verdadeiro do evento

Dentro do fluxo de criação de anúncios (incluindo o fluxo de impulsionamento), o anunciante seleciona "Temas sociais, eleições ou política" na seção de Categoria Especial de Anúncios e então, na seção de criativos do anúncio, seleciona "Incluir mídia

criada ou editada com IA" (rotulada como "Mídia neste anúncio é digitalmente criada ou alterada"). Esta seleção constitui a autodeclaração obrigatória.

Adicionalmente, anunciantes de SIEP são impedidos de usar as ferramentas de IA generativa próprias da Meta (geração de plano de fundo, expansão de imagem, geração de imagem e geração de texto) — tanto no Gerenciador de Anúncios quanto via impulsionamento. Quando um anunciante impulsiona uma publicação orgânica que foi criada ou aprimorada com as ferramentas voltadas ao consumidor de IA generativa da Meta, o anúncio resultante mantém aquele conteúdo criado por IA e os requisitos de divulgação se aplicam.

Mais detalhes sobre as políticas da Meta sobre rotulagem de IA podem ser encontrados na Seção 4.1.1.

Como a Declaração É Associada ao Registro do Anúncio

Rotulagem no Anúncio: Quando um anunciante divulga o uso de mídia criada ou editada por IA através do mecanismo de autodeclaração, um rótulo "Informações de IA" aparece no anúncio próximo ao aviso de "Pago por". Este é um rótulo visível, de "0-cliques", diretamente no próprio anúncio.

Registro na Biblioteca de Anúncios: A seção "Detalhes do anúncio" da Biblioteca de Anúncios inclui uma divulgação afirmando "Mídia neste anúncio criada ou editada com IA". Anúncios SIEP permanecem na Biblioteca de Anúncios por sete anos, preservando esta divulgação como parte do registro do anúncio.

Registro de Detecção Automatizada: Quando a detecção automatizada identifica uso de IA (no lugar da autodeclaração do anunciante), esta informação é registrada e disponibilizada aos usuários através do menu de três pontos no anúncio em "Sobre este Anúncio". Este registro é distinto do rótulo de autodeclaração — a autodeclaração aparece diretamente no anúncio, enquanto os resultados da detecção automatizada aparecem nas informações suplementares do anúncio.

3.4 Biblioteca de Anúncios: Visão Geral e Cobertura de Plataformas

A **Biblioteca de Anúncios da Meta** é um banco de dados abrangente e pesquisável para a transparência de anúncios. As pessoas podem usar a Biblioteca de Anúncios para obter mais informações sobre os anúncios que veem nas tecnologias da Meta.

As pessoas podem pesquisar todos os anúncios ativos em execução nos produtos da Meta. Para anúncios sobre temas sociais, eleições ou política, fornecemos informações adicionais, incluindo gastos, alcance e entidades financiadoras. Esses anúncios são visíveis estejam ativos ou inativos e são armazenados na Biblioteca de Anúncios por sete anos.

Para anúncios sobre temas sociais, eleições ou política no Brasil que entregam uma impressão, oferecemos informações adicionais. Esses anúncios são exibidos na Biblioteca de Anúncios enquanto ativos e arquivados por sete anos após a entrega de sua última impressão.

Para cada anúncio sobre temas sociais, eleitorais e políticos na Biblioteca de Anúncios para o Brasil, as seguintes informações são disponibilizadas:

- Uma **cópia digital do anúncio** (conteúdo/criativo do anúncio)
- **Identificador único** para o anúncio
- **Status ativo/inativo**
- **Datas de início e término** (intervalo de datas em que o anúncio foi veiculado)
- **Plataforma(s)** em que o anúncio foi entregue
- Faixa de gastos
- O **aviso de "pago por" ou "Propaganda Eleitoral"**, incluindo o nome da entidade ou pessoa responsável pela veiculação do anúncio
- **Opções de segmentação incluindo idade, gênero e localização**
- **Alcance total e alcance discriminado por idade, gênero e país**

Informações de Segmentação (Anúncios sobre Temas Sociais, Eleitorais e Políticos)
Desde julho de 2022, nossa Biblioteca de Anúncios publicamente disponível inclui um **resumo das informações de segmentação** para anúncios sobre temas sociais, eleitorais ou políticos. Este resumo de segmentação inclui:

- O número total de anúncios sobre temas sociais, eleitorais e políticos que uma Página veiculou usando cada tipo de segmentação (como localização, dados demográficos e interesses)
- A porcentagem de gastos com anúncios sobre temas sociais, eleitorais e políticos usados para segmentar essas opções
- Se uma Página usou Públicos Personalizados e/ou Públicos Semelhantes

3.5 API da Biblioteca de Anúncios: Funcionalidades e Coleta Sistemática

Capacidades Principais da API

A **API da Biblioteca de Anúncios** é uma interface de programação de aplicações que permite uma análise mais aprofundada de anúncios sobre temas sociais, eleições ou política. Usuários autorizados também podem analisar conteúdo de marca ativo e público em execução no Facebook e Instagram via API.

Os usuários podem realizar **pesquisas personalizadas por palavras-chave** de anúncios armazenados na Biblioteca de Anúncios. Os resultados da API incluem os criativos e dados de desempenho de anúncios que estão arquivados na Biblioteca de Anúncios. Fizemos melhorias em nossa API para que as pessoas possam acessar facilmente anúncios de um determinado país e analisar anunciantes específicos.

Escopo de Acesso à API
A API da Biblioteca de Anúncios permite pesquisas personalizáveis para Anúncios sobre **temas sociais, eleições ou política** que foram entregues no Brasil

A API da Biblioteca de Anúncios permite que reguladores, jornalistas, entidades fiscalizadoras e outras pessoas analisem anúncios políticos e seus respectivos anunciantes. A API é uma ferramenta de transparência pública que permite acesso programático a dados de publicidade, apoiando análises de risco automatizadas e sistemáticas.

3.6 Visão Geral da Abordagem de Pausa de Anúncios da Meta

Em conformidade com o Artigo 16 da Portaria nº 463/2026/TSE, bem como o Artigo 29, § 11, da Resolução nº 23.610/2019/TSE, a Meta está preparada para implementar as restrições do período de silêncio eleitoral, impedindo que candidatos e partidos veiculem anúncios do tipo SIEP no Brasil com base nos períodos designados pela Regulamentação Eleitoral.

Esta medida será aplicada em ambos os turnos de votação, entre 2 e 5 de outubro de 2026 e 23 e 26 de outubro de 2026.

Durante esses períodos, além de desativar todos os Anúncios SIEP no Brasil, a Meta monitorará anúncios em veiculação no Brasil, através de detecção automatizada e revisão humana, buscando anúncios que possam ter sido criados em violação à nossa política de publicidade SIEP descrita acima e que possam tentar contornar os períodos de silêncio eleitoral.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Compromisso de Monitoramento. Em conformidade com o Artigo 7º, §4º, da Portaria nº 463/2026/TSE, a Meta monitorará as medidas de aplicação dessas políticas com base nos dados gerados durante o período eleitoral e poderá reportar as métricas de anúncios do tipo SIEP não declarados que foram detectados, nos quais os usuários não aplicaram o aviso requerido no momento da criação. Os números efetivamente mensurados serão reportados no relatório consolidado a ser apresentado ao TSE até 19 de dezembro de 2026 (Art. 28).

3.7 Anúncios gratuitos para a Justiça Eleitoral

O Artigo 9º-D, §3º, da Resolução nº 23.610/2019/TSE, estabelece que a Justiça Eleitoral poderá, por meio de ordem judicial, determinar o impulsionamento de conteúdo que contrarie anúncios que representam informação considerada notoriamente inverídica ou gravemente descontextualizada, que possam impactar a integridade do processo eleitoral.

A disposição não impõe às plataformas o dever de produzir contra narrativas nem de publicar o conteúdo por conta própria. Neste cenário, a Justiça Eleitoral deverá criar a campanha publicitária, com o conteúdo informativo corretivo que esclareça um anúncio previamente identificado como notoriamente inverídico ou gravemente descontextualizado pela ordem judicial, e este conteúdo precisará ser publicado em uma conta oficial da Justiça Eleitoral.

Este fluxo é consistente com a arquitetura das plataformas da Meta, que são projetadas para que terceiros criem e publiquem seu próprio conteúdo. A Meta não pode ela mesma criar conteúdo em nome de terceiros ou publicar conteúdo através de perfis oficiais da Justiça Eleitoral, pois apenas o titular da conta possui as credenciais e o controle editorial para fazê-lo.

SEÇÃO 4: Abordagem Responsável para IA Generativa

Esta seção apresenta as informações exigidas pelos Artigos 15 e 16 da Portaria nº 463/2026/TSE, abordando duas categorias complementares de obrigações relacionadas a sistemas de inteligência artificial generativa.

O Artigo 15 da Portaria especifica os elementos que o plano de conformidade deve conter em relação à governança de sistemas de IA generativa e tecnologias conversacionais equivalentes oferecidas pelo provedor, em cumprimento aos deveres

estabelecidos nos Artigos 9º-B, § 5º, e 28, § 1º-C, incisos I a IV, da Resolução nº 23.610/2019/TSE.

O Artigo 16 da Portaria aborda a gestão do período de silêncio eleitoral no que se aplica a conteúdo gerado por IA, incluindo a publicação, republicação e impulsionamento de novo conteúdo sintético que use imagem, voz ou semelhança de candidato ou figura pública durante as 72 horas anteriores e 24 horas posteriores à eleição, bem como a suspensão de chatbots e avatares não identificados como tal — em cumprimento aos deveres estabelecidos nos Artigos 9º-B, §§ 3º-A e 4º, da mesma Resolução.

Além do acima exposto, o Artigo 9º-C da Resolução nº 23.610/2019/TSE estabelece uma proibição dirigida aos responsáveis pela propaganda eleitoral, e seu §2º atribui consequências aos candidatos. Os deveres de cuidado correspondentes incumbidos aos provedores em relação à mesma matéria são aqueles estabelecidos nos Artigos 9º-D e 9º-E, abordados ao longo deste relatório.

As subseções abaixo descrevem as salvaguardas, controles e medidas operacionais que a Meta mantém para atender a cada uma dessas obrigações.

4.1 Políticas Aplicáveis a Conteúdo Gerado por IA

Os Padrões da Comunidade e Padrões de Publicidade da Meta se aplicam a todo conteúdo carregado em nossas plataformas independentemente de como foi produzido, incluindo conteúdo gerado por IA. Conteúdo gerado por IA que viole essas políticas está sujeito às mesmas medidas de aplicação que qualquer outro conteúdo violador.

Conteúdo gerado por IA também é elegível para informações contextuais através do programa de rotulagem da Meta, que identifica mídia gerada ou alterada por IA para os usuários. A Meta rotula imagens fotorrealistas criadas usando Meta AI, bem como conteúdo gerado por IA de provedores terceiros. Em 2025, a Meta expandiu sua rotulagem "Informações de IA" para formatos e superfícies adicionais. O Comitê de Supervisão da Meta orientou a empresa a identificar e rotular áudio e vídeo manipulados em escala e em idiomas locais, incluindo português.

Ao publicar conteúdo orgânico com vídeo fotorrealista ou áudio com som realista, os usuários são obrigados a divulgar que esse foi digitalmente criado ou alterado. O não cumprimento pode resultar em penalidades. Anunciantes que veiculam anúncios políticos também devem divulgar conteúdo gerado ou alterado por IA (ver Seção 3.3).

4.1.1 Mecanismos para Identificação e Rotulagem de Conteúdo Sintético

Mecanismos de Autodeclaração - Conteúdo Orgânico (Facebook, Instagram, Threads)
Fornecemos oportunidades de autodeclaração para os usuários informarem quando seu conteúdo foi criado ou modificado usando IA. Usuários que postam conteúdo no Facebook, Instagram e Threads podem voluntariamente divulgar que seu conteúdo foi gerado ou editado com ferramentas de IA. Esta autodeclaração aciona a aplicação de um rótulo no conteúdo.

Conteúdo Pago (Facebook, Instagram, Threads)
Para anúncios relacionados a temas sociais, eleições ou política, os anunciantes são obrigados a divulgar se usam uma imagem ou vídeo fotorrealista, ou áudio com som realista, que foi digitalmente criado ou alterado (inclusive com IA) em certos casos. A Meta também fornece a opção para os anunciantes autodeclararem o uso de IA para gerar ou editar substancialmente o conteúdo em seus anúncios. Quando um anunciante autodeclara o uso de IA, um rótulo é aplicado ao anúncio informando os usuários sobre o envolvimento de IA.

Mecanismos de Detecção e Rotulagem

Conteúdo Próprio (1P) — Ferramentas de IA da Meta (Facebook, Instagram, Threads): Todo conteúdo criado com as próprias ferramentas de IA da Meta (como o gerador de imagens Meta AI) incorpora automaticamente mecanismos de proveniência:

- Marca d'água invisível: Um identificador único oculto embutido diretamente nos pixels ou dados do conteúdo (imagens ou vídeos).
- Marcadores visíveis: Marcadores visíveis colocados no conteúdo gerado para que os usuários possam ver que IA foi envolvida.
- Metadados: Tanto metadados C2PA (Coalition for Content Provenance and Authenticity) quanto metadados IPTC (International Press Telecommunications Council) são incorporados nos arquivos de imagem.

Quando esses sinais técnicos são detectados, o sistema aplica automaticamente um rótulo indicando que o conteúdo é de IA, sem que seja necessária qualquer ação do usuário ou do anunciante

Conteúdo de Terceiros (3P) — Ferramentas de IA Externas (Facebook, Instagram, Threads): Usamos padrões do setor para identificar conteúdo gerado por IA produzido por ferramentas de terceiros:

- A detecção depende da identificação de indicadores padrão do setor de imagens de IA, particularmente credenciais de conteúdo C2PA e metadados IPTC, que outras empresas incluem em seus conteúdos gerados por IA.
- Quando esses sinais são detectados em conteúdo carregado, o sistema aplica um rótulo identificando o conteúdo como de IA. Este é um esforço contínuo e em evolução enquanto a indústria trabalha coletivamente para desenvolver padrões comuns para proveniência de conteúdo.

Conteúdo Pago — Detecção Automatizada para Anúncios (Facebook, Instagram):A Meta exige que os anunciantes autodeclarem quando mídia de anúncios para temas sociais, eleições ou política foi criada ou editada usando ferramentas de IA de terceiros em certos casos, e rotula visivelmente o conteúdo declarado. A Meta não permite que anunciantes gerem mídia de anúncios para temas sociais, eleições ou política usando ferramentas de IA generativa próprias. A Meta também detecta e rotula conteúdo gerado por IA produzido por ferramentas de terceiros baseando-se em credenciais de conteúdo C2PA padrão do setor. Esta detecção usa tecnologia padrão do setor para identificar conteúdo gerado por IA, e nenhuma ação do anunciante é necessária. Se a Meta detectar que mídia de anúncio foi criada ou editada usando ferramentas de IA de terceiros, o anúncio é rotulado de acordo.

Formato e Visibilidade do Rótulo no Produto

Formato de visibilidade (Facebook, Instagram, Threads):

- O rótulo (ex: "Informações de IA") é exibido para conteúdo gerado por IA e autodeclarado com “0-cliques” — ou seja, é imediatamente visível no conteúdo sem exigir qualquer interação do usuário para vê-lo. O rótulo e informações sobre conteúdo editado por IA podem ser encontrados no menu de 3 pontos.
- O rótulo aparece na própria publicação (em publicações do Feed, Reels e Stories).
- Os usuários podem tocar ou clicar no rótulo para ver informações adicionais sobre o motivo do rótulo ter sido aplicado e o que ele significa.
- Conteúdo criado diretamente com ferramentas Meta AI é adicionalmente rotulado com "Meta AI".

Formato de visibilidade (Anúncios / Conteúdo pago):

- Para anúncios relacionados a temas sociais, eleições ou política, o rótulo é automaticamente aplicado quando detectamos imagens fotorrealistas de humanos geradas por IA (1P), ou quando o anunciante autodeclara (3P).
- O rótulo é visível para usuários visualizando o anúncio em seu feed, bem como na Biblioteca de Anúncios.

Programa de Rotulagem de IA

Desde maio de 2024, a Meta tem aplicado rótulos a conteúdo gerado ou editado por IA. Este programa abrange:

- Rotulagem de imagens fotorrealistas criadas usando Meta AI, bem como conteúdo gerado por IA de provedores terceiros
- Detecção de indicadores padrão do setor de imagens de IA (como metadados C2PA) para aplicar rótulos de "Informação de IA" automaticamente
- Uma exigência para os usuários divulgarem ao publicar conteúdo orgânico com vídeo fotorrealista ou áudio com som realista que foi digitalmente criado ou alterado
- Expansão da rotulagem "Informação de IA" para formatos de conteúdo e superfícies de produto adicionais em 2025

Colaboração da Indústria em Transparência de IA
A Meta participa de esforços de todo o setor para desenvolver padrões para identificação de conteúdo gerado por IA. A Meta é membro da Partnership on AI e estava entre 20 empresas que assinaram o Acordo Tecnológico para Combater o Uso Enganoso de IA em Eleições na Conferência de Segurança de Munique em fevereiro de 2024. A Meta também contribui para o desenvolvimento de padrões técnicos como C2PA (Coalition for Content Provenance and Authenticity) para proveniência e autenticidade de conteúdo.

4.2 Preparação de IA para Eleições

A abordagem da Meta para segurança eleitoral incorpora capacidades de detecção e resposta baseadas em IA, desenvolvidas e aperfeiçoadas através de experiência em mais de 200 eleições globalmente. Nosso investimento contínuo em infraestrutura de segurança e proteção apoia o desenvolvimento contínuo de sistemas de detecção de IA, operações de revisão de conteúdo e programas de integridade eleitoral. Esses recursos servem às eleições no Brasil e em outras jurisdições.

Estamos ativamente abordando os riscos apresentados por conteúdo gerado por IA no contexto eleitoral através de requisitos de rotulagem e divulgação, tais como:

- Rotulamos conteúdo no Facebook, Instagram e Threads com "Informações de IA" quando detectamos indicadores padrão do setor (como metadados C2PA) de que foi criado ou significativamente modificado usando IA. Para conteúdo orgânico que detectamos que foi apenas modificado ou editado por ferramentas de IA, o rótulo aparece no menu da publicação; para conteúdo gerado por IA, o rótulo é exibido de forma mais proeminente. Também fornecemos recursos para criadores autodeclararem quando compartilham

conteúdo gerado por IA, e introduzimos rótulos voluntários de perfil "criador de IA" para contas que regularmente publicam mídia gerada por IA.
- Se determinarmos que conteúdo de imagem, vídeo ou áudio digitalmente criado ou alterado cria um risco particularmente elevado de induzir o público a erro, de forma relevante, sobre assunto de interesse público, podemos adicionar um rótulo mais proeminente para que as pessoas tenham mais informação e contexto.
- Anunciantes que veiculam anúncios relacionados a temas sociais, eleições ou política são obrigados a divulgar se usam uma imagem ou vídeo fotorrealista, ou áudio com som realista, que foi criado ou alterado digitalmente, inclusive com IA.

Autodeclaração de IA
Em conformidade com o Artigo 9º-B, §5º, da Resolução nº 23.610/2019/TSE, os anunciantes têm a opção de autodeclarar o uso de mídia gerada por IA em seus anúncios no Facebook, Instagram e Threads. Esta opção aparece para todos os anunciantes. Se um anunciante autodeclarar o uso de mídia gerada por IA em um anúncio, a Meta adicionará um rótulo "Informação de IA" na face do anúncio.

Da mesma forma, qualquer usuário publicando conteúdo orgânico no Facebook, Instagram e Threads também terá a opção de autodeclarar o uso de mídia gerada por IA em seu conteúdo orgânico. Esta opção aparece para todos os usuários.

Além das opções de Autodeclaração de IA fornecidas acima, conteúdo eleitoral digitalmente alterado também está sujeito a requisitos de rotulagem de divulgação de IA sob as políticas da Meta, conforme descrito acima [17].

4.3 Detecção Baseada em IA

A Meta implementa sistemas automatizados para detectar e abordar ameaças relacionadas a eleições em escala:
- **Detecção de Campanhas Coordenadas:** Sistemas de aprendizado de máquina identificam padrões de comportamento inautêntico coordenado e operações de influência
- **Capacidades Multilíngues:** Sistemas de detecção operam em dezenas de idiomas, incluindo português, para identificar conteúdo violador independentemente do idioma

4.4 Padrões de Conteúdo Específicos para Eleições e Salvaguardas em Nível de Modelo

Nossos modelos de IA generativa — incluindo Meta AI, que também está disponível no Facebook, Instagram, Threads e WhatsApp — incorporam padrões de conteúdo específicos para eleições que orientam respostas do modelo sobre tópicos políticos, incluindo informações sobre votação. Esses padrões definem resultados de modelo aceitáveis e inaceitáveis no domínio eleitoral. Testes de segurança (*Red-teaming)* são realizados contra esses padrões para ajudar a garantir que respostas a *prompts* solicitando informações sobre votação ou aconselhamento especializado sejam respondidas de maneira que não viole padrões de conteúdo.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

4.4.1 Salvaguardas Técnicas para Prevenir a Geração de Conteúdo Sexual Não Consensual Envolvendo Candidatos

A Meta implementou um conjunto abrangente e em várias camadas de salvaguardas técnicas projetadas para prevenir a geração de conteúdo sexual não consensual envolvendo pessoas reais ou identificáveis — incluindo candidatos políticos — na Meta AI, incluindo integrações no Facebook, Instagram, Threads e WhatsApp. Essas salvaguardas operam durante todo o ciclo de vida do modelo e em vários estágios de interação do usuário.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

4.4.2. Políticas de Moderação de Conteúdo sobre Violência de Gênero

Nossas políticas proíbem todas as categorias restritas de material de IA generativa, incluindo material de violência de alto impacto e material de instrução de violência, e essas proibições se aplicam ao conteúdo gerado por IA e às experiências habilitadas pela Meta AI no Facebook, Instagram, Threads e WhatsApp. De forma consistente com essas proibições, implementamos controles em camadas em todo o ciclo de vida do modelo — incluindo governança do modelo e salvaguardas de treinamento, avaliações de risco de pré-lançamento e avaliações em escala, *red teaming* e mitigações de ajuste fino — para reduzir a probabilidade de saídas que violem as políticas.

***Red Teaming* para Riscos Eleitorais e de Violência de Gênero**
Nossos protocolos de *red teaming* incluem especificamente testes de vulnerabilidades relacionadas à violência de gênero no contexto da IA generativa. Esses testes adversariais estruturados e baseados em cenários incorporam especialistas em conteúdo multilíngue com experiência em questões de integridade em diferentes mercados, permitindo que os protocolos levem em conta variações linguísticas, expressões locais e nuances culturalmente específicas que podem influenciar como o conteúdo prejudicial é solicitado ou transmitido.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

4.5 Abordagem da Meta para o Período de Silêncio da IA

O Artigo 16 da Portaria especifica os elementos relativos ao gerenciamento dos períodos de silêncio eleitoral, incluindo a proibição de novos conteúdos gerados por IA (publicação e republicação) que usem a imagem, voz ou expressão de um candidato ou figura pública durante o período entre 72 horas antes e 24 horas após o término da votação, estabelecido pelo Artigo 9º-B, §3º-A, da Resolução nº 23.610/2019/TSE.

Padrões de Conteúdo Específicos para Eleições para a Meta AI
Nossos modelos têm padrões de conteúdo específicos para eleições para orientar as respostas do modelo sobre tópicos, incluindo informações de votação. O *red-teaming* é realizado em relação a esses padrões para ajudar a garantir que as respostas aos prompts que solicitam informações de votação ou aconselhamento especializado sejam respondidas de uma maneira que se alinhe com nossas políticas de conteúdo.

Mecanismos de Restrição Automática para a Meta AI
Implementamos várias camadas de mecanismos de restrição automática para a Meta AI no contexto das eleições:

- Engenharia de prompt, listas de bloqueio e filtros: Eles são implementados para bloquear estrategicamente palavras-chave em entradas ou saídas que vão contra nossos padrões de conteúdo eleitoral.
- Respostas prontas (*Canned responses*): Para lidar com riscos especificamente relacionados às eleições, nossos modelos fornecem respostas prontas de

acordo com os padrões de conteúdo, em vez de gerar respostas novas sobre tópicos eleitorais sensíveis.

- Avaliação direcionada: Conduzimos avaliações direcionadas focadas em consultas relacionadas a eleições para ajudar a garantir que as respostas do modelo sigam nossas políticas e padrões.
- Principais métricas e *guardrails* (barreiras de proteção): Definimos as principais métricas e *guardrails*, como segurança de resposta e precisão factual, e implementamos recursos de detecção para monitorar a adesão aos *guardrails*. Se os *guardrails* forem violados, as equipes investigam e tomam as medidas apropriadas.

Conteúdo de Resposta de IA Inaceitável

Nossos padrões especificam que exemplos de respostas inaceitáveis incluem, mas não se limitam a, conteúdo que contenha informações sobre o status de registro de eleitor de um indivíduo específico e aconselhamento definitivo de saúde, jurídico ou financeiro.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Aplicabilidade Específica da Plataforma

Facebook e Instagram

As políticas de publicidade de Anúncios sobre Temas Sociais, Eleições ou Política (SIEP), os mecanismos de restrição automática, os períodos de restrição de anúncios e as restrições específicas para eleições da Meta AI se aplicam integralmente ao Facebook e ao Instagram.

Threads

O Threads está coberto pela nossa abordagem de rotulagem de conteúdo de IA — rotulamos imagens que os usuários publicam no Facebook, Instagram e Threads quando conseguimos detectar indicadores padrão do setor de que foram geradas por IA. Os recursos de Meta AI no Threads também são regidos pelos mesmos padrões de conteúdo e salvaguardas específicos para Eleições.

Compromisso de Monitoramento

Em conformidade com o Artigo 7º, §4º, da Portaria nº 463/2026/TSE, a Meta monitorará o uso de conteúdo gerado por IA durante o período eleitoral e poderá reportar as métricas relacionadas aos conteúdos rotulados. As métricas efetivamente mensuradas serão reportadas no relatório consolidado a ser apresentado ao TSE até 19 de dezembro de 2026 (art. 28).

SEÇÃO 5: Proteção de Dados Pessoais para Publicidade Eleitoral

Esta seção apresenta as informações exigidas pelo Artigo 21 da Portaria nº 463/2026/TSE, que especifica os elementos que o plano de conformidade deve conter com relação aos deveres de proteção de dados pessoais especificamente relacionados às obrigações estabelecidas nos Artigos 33-A e 33-B da Resolução nº 23.610/2019/TSE.

O Artigo 33-A exige que os provedores de aplicativos de internet informem expressamente aos usuários sobre a possibilidade de seus dados pessoais serem tratados para fins de disseminação de propaganda eleitoral, dentro do escopo e dos limites técnicos de cada serviço. O Artigo 33-B estabelece os deveres incumbidos aos provedores, partidos políticos, federações, coligações e candidatos no tratamento de dados pessoais para fins de propaganda eleitoral.

As subseções abaixo descrevem os mecanismos e procedimentos que a Meta mantém para cumprir cada uma dessas obrigações, observando que os aspectos relacionados à interpretação, fiscalização e aplicação da Lei Geral de Proteção de Dados Pessoais (“LGPD") permanecem sujeitos às competências legalmente atribuídas à Agência Nacional de Proteção de Dados (“ANPD").

5.1 Mecanismos para Informar os Titulares de Dados Sobre o Tratamento para Publicidade Eleitoral

Empregamos múltiplos mecanismos de transparência complementares no Facebook, Instagram e Threads para informar os titulares de dados sobre o tratamento de seus dados pessoais para publicidade eleitoral.

Alterações nas Opções de Segmentação (Facebook, Instagram e Threads)
Conforme descrito na Seção 5.2, os anunciantes que veiculam anúncios sobre temas sociais, eleições ou política só podem veicular anúncios no país no qual estão autorizados. Além disso, as opções relacionadas a tópicos que podem ser percebidos como sensíveis foram removidas em toda a nossa plataforma de publicidade (veja a Seção 5.2 para detalhes). Essas alterações simplificam as informações de transparência divulgadas aos titulares de dados sobre como foram alcançados.

Ferramenta "Por que estou vendo este anúncio?" (Facebook, Instagram e Threads)
Fornecemos uma ferramenta de transparência integrada ao produto — "Por que estou vendo este anúncio?" — que oferece aos usuários uma visão geral das razões pelas

quais os anúncios foram exibidos para eles. Essa ferramenta apresenta informações sobre os fatores que consideramos ao exibir anúncios aos usuários, incluindo tópicos, atividade do usuário e seleções de audiência do anunciante.

Política de Privacidade e Política de Cookies (Todas as Plataformas)
Fornecemos transparência aos usuários sobre nosso uso de dados para publicidade por meio tanto da nossa Política de Privacidade quanto da nossa Política de Cookies. Em nossa Política de Privacidade, listamos as informações que coletamos de várias fontes e descrevemos como podem ser usadas para fins publicitários. Em nossa Política de Cookies, explicamos o uso de cookies e tecnologias similares que apoiam a entrega de publicidade.

Exercício dos Direitos de Privacidade da LGPD
Os Direitos de Privacidade são informados na Política de Privacidade da Meta. A Política de Privacidade da Meta fornece aos usuários acesso direto a ferramentas dedicadas de autoatendimento que os permitem exercer seus direitos de privacidade. Além disso, a Política de Privacidade da Meta oferece canais de contato disponíveis para usuários e não usuários enviarem perguntas relacionadas à privacidade e entrarem em contato com o Encarregado de Proteção de Dados da Meta, de acordo com suas preocupações específicas de privacidade. Veja:

- Canal de contato de Privacidade: https://help.meta.com/support/privacy
- Encarregado de Proteção de Dados: https://help.meta.com/support/privacy > Usuários e não usuários devem selecionar o produto (ex.: Facebook, Instagram, etc.) e então selecionar a pergunta > "Como posso receber ajuda do escritório de proteção de dados para uma consulta?"

A Meta também publicou um Portal de Direitos de Privacidade, que fornece um ponto de entrada centralizado para usuários e não usuários enviarem facilmente consultas de privacidade e exercerem seus direitos. Usuários e não usuários podem acessar o Portal através dos pontos de entrada existentes na Política de Privacidade da Meta e na Central de Ajuda.

O Portal permite que indivíduos enviem solicitações relacionadas ao exercício de seus direitos de proteção de dados de maneira simplificada e em autoatendimento. Este mecanismo foi projetado para garantir que os indivíduos possam exercer seus direitos diretamente, sem a necessidade de ação intermediária.

5.2 Mecanismos que Garantem a Limitação de Finalidade

A Meta garante a limitação de finalidade, necessidade, adequação e tratamento não discriminatório para publicidade eleitoral através dos seguintes mecanismos:

(a) Limitação de Finalidade: Os dados são usados de acordo com as finalidades descritas na Política de Privacidade da Meta. Os usuários podem gerenciar a personalização de seus anúncios através das configurações da plataforma (Preferências de Anúncios, Tecnologias Fora da Meta).

(b) Necessidade e Adequação: Os anunciantes que veiculam anúncios sobre temas sociais, eleições ou política só podem veicular anúncios no país no qual estão autorizados. Separadamente, em toda a nossa plataforma de publicidade, removemos opções de segmentação detalhada que se relacionam a tópicos que podem ser percebidos como sensíveis — como opções que fazem referência a causas, organizações ou figuras públicas relacionadas a saúde, raça ou etnia, afiliação política, religião ou orientação sexual. É importante observar que essas opções de segmentação por interesse não eram baseadas nas características físicas ou atributos pessoais das pessoas, mas sim em fatores como as interações das pessoas com conteúdo em nossa plataforma. Elas foram removidas após feedback de especialistas e para reduzir o potencial de abuso.

(c) Não Discriminação: Os Padrões de Publicidade da Meta proíbem anúncios que discriminem pessoas com base em atributos pessoais. Nossos Padrões de Publicidade também proíbem anúncios que contenham conteúdo que afirme ou implique atributos pessoais, incluindo afirmações diretas ou indiretas sobre raça, etnia, religião, crença, idade, orientação ou práticas sexuais, identidade de gênero, deficiência, condições de saúde física ou mental (incluindo condições médicas), situação financeira vulnerável, situação eleitoral, filiação sindical, antecedentes criminais ou nome de uma pessoa.

5.3 Procedimentos de Segurança

Programa Abrangente de Segurança da Informação
Mantemos um programa de segurança em toda a empresa que consiste em um sistema integrado de políticas, padrões e diretrizes, organizado por domínios, objetivos e controles. Essas políticas regem o acesso e o uso de dados de usuários e se aplicam a todos os funcionários da Meta no Facebook, Instagram, Threads e WhatsApp.

Nossas políticas, padrões e controles levam em consideração padrões relevantes e aplicáveis da indústria (como NIST CSF, PCI-DSS e ISO 27001) e são projetados e dimensionados para capturar a abordagem da Meta à segurança da informação com base na natureza específica de nossos sistemas de informação e necessidades de

negócios. Isso garante uma estrutura que atende a medidas técnicas, organizacionais e operacionais para garantir a proteção de dados pessoais de acordo com as melhores práticas internacionais.

Controles Técnicos de Segurança
Mantemos um Programa Abrangente de Segurança da Informação projetado para garantir a confidencialidade, integridade e disponibilidade dos dados pessoais tratados em nossos sistemas e redes. Esses controles incluem:

- **Controles de Acesso:** Implementamos controles de acesso centralizados e automatizados baseados em função que limitam a capacidade de modificar dados pessoais apenas a sistemas e usuários autorizados. O acesso é concedido com base na função de trabalho e necessidade de negócio.
- **Criptografia:** Os dados são criptografados em trânsito usando protocolos padrão do setor (ex.: TLS) como parte de nossa arquitetura de segurança padrão, protegendo dados pessoais conforme se movem por nossos sistemas e redes.
- **Detecção e Monitoramento:** Implantamos mecanismos de detecção, trilhas de auditoria e monitoramento contínuo para identificar e responder a tentativas de acesso não autorizado ou atividades anômalas. Esses controles permitem a identificação tempestiva de ameaças potenciais à segurança dos dados pessoais.
- **Práticas de Desenvolvimento Seguro:** Mantemos estruturas de desenvolvimento seguro integradas ao nosso ciclo de vida de desenvolvimento de produtos, incluindo revisões e testes de segurança, para garantir que os recursos que tratam dados dos usuários sejam projetados e construídos com princípios de segurança desde a concepção.
- **Segurança Física:** Os dados são armazenados em centros de dados da Meta com medidas de segurança física padronizadas, incluindo controles de acesso por cartão, vigilância por CCTV, pessoal de segurança no local, controles ambientais, scanners biométricos para acesso a áreas seguras e destruição segura de mídias de armazenamento descomissionadas.
- **Segmentação de Dados:** Os dados pessoais são segmentados por sensibilidade e finalidade através de medidas técnicas e de políticas que definem o manuseio adequado dos dados ao longo de seu ciclo de vida, garantindo que os dados pessoais estejam sujeitos a proteções apropriadas proporcionais ao seu perfil de risco.

Programa de Gerenciamento de Resposta a Incidentes
Nosso programa de gerenciamento de resposta a incidentes é uma estrutura abrangente projetada para identificar, mitigar, avaliar e remediar incidentes de privacidade. Utilizamos uma variedade de ferramentas e processos manuais e

automatizados para detectar potenciais incidentes de privacidade, incluindo monitoramento de sistemas, relatórios internos, relatórios de terceiros e um programa de recompensa por bugs. Uma vez detectado um incidente, ele passa por um processo de triagem para avaliar a gravidade e o impacto, e, em situações apropriadas, é seguido por um processo de investigação e remediação. Atualizamos e melhoramos continuamente nossos processos de resposta a incidentes para abordar riscos em evolução. O programa inclui análises pós-incidente e ações de acompanhamento, conforme apropriado, para prevenir futuros incidentes e melhorar a estratégia geral de resposta.

Notificação às Autoridades de Supervisão
Quando ocorre um incidente de dados qualificado, notificamos as autoridades de proteção de dados relevantes de acordo com os requisitos legais aplicáveis.

Notificação aos Titulares de Dados Afetados
Nossa abordagem para notificar os titulares de dados após um incidente é guiada pelos requisitos legais aplicáveis e por uma avaliação detalhada de risco para determinar o impacto potencial nos direitos e liberdades dos usuários.

Abordagem de avaliação de risco
Avaliamos fatores incluindo a natureza e sensibilidade dos dados envolvidos, evidências de ações adicionais tomadas sobre ou com os dados, se categorias especiais ou sensíveis de dados foram acessadas, e o escopo e status de contenção do incidente. Nos termos da lei aplicável, consideramos esses fatores e os riscos de danos relevantes ao determinar a notificação formal aos titulares de dados.

Medidas proativas de proteção do usuário
Mesmo quando a notificação formal aos titulares de dados não é legalmente exigida, tomamos medidas proativas para proteger os usuários afetados.

5.4 Mecanismo de Consentimento

O Art. 33-B, §1º exige consentimento específico e expresso para o tratamento de dados pessoais sensíveis para publicidade eleitoral. Os sistemas de publicidade da Meta são projetados para excluir categorias de dados sensíveis de serem usadas como sinais para segmentação e entrega de anúncios. Essas categorias não estão disponíveis como entradas de segmentação nas ferramentas de publicidade da Meta.

A Meta não usa dados explicitamente categorizados como sensíveis sob a LGPD (Art. 5º, II) para segmentar ou entregar anúncios eleitorais. Os anunciantes podem usar seus próprios Públicos Personalizados (*first-party*), e os anunciantes são responsáveis

por garantir uma base legal para seus dados próprios (*first-party*) antes de fornecê-los à Meta. Os Padrões de Publicidade da Meta adicionalmente proíbem anúncios que discriminem pessoas com base em atributos pessoais.

Consequentemente, como os sistemas da Meta são projetados para prevenir o tratamento de dados pessoais sensíveis para fins de publicidade eleitoral, nenhum mecanismo de consentimento separado sob o Art. 33-B, §1º foi implementado para essa categoria de tratamento.

SEÇÃO 6: Canais de Denúncia Locais

Esta seção apresenta as informações exigidas pelos Artigos 12 e 17 da Portaria nº 463/2026/TSE, descrevendo a estrutura e a capacidade operacional dedicada ao recebimento, triagem e processamento de solicitações de reguladores, usuários, candidatos e partidos políticos, para o tratamento de conteúdos e comportamentos sujeitos à Resolução nº 23.610/2019/TSE, Artigos 9-E, I-VII, 28, §§ 4 e 4-B, incluindo os respectivos fluxos de triagem e tempos de resposta estimados.

A Meta opera e mantém canais que permitem a usuários e não usuários denunciar conteúdo que viola políticas e conteúdo localmente ilegal.

Conteúdo que Viola Políticas
Os usuários podem utilizar as opções de denúncia no aplicativo, uma ferramenta chamada Experiência de Feedback e Denúncia (FRX). Essa funcionalidade é acessada clicando no menu de três pontos existente nos conteúdos disponíveis nas respectivas plataformas, selecionando "Denunciar" e seguindo as instruções guiadas que permitem aos usuários categorizar a natureza de sua preocupação.

No Brasil, além dos canais de denúncia que estão disponíveis globalmente aos usuários, as Empresas da Meta mantêm canais dedicados ao público brasileiro, com o objetivo de cumprir a lei brasileira, inclusive a Resolução nº 23.610/2019/TSE, Artigos 9-D, II-VII e 9-E§ 2 . Estes canais incluem:

- Canais dedicados que permitem ao TSE denunciar conteúdo e contas suspeitas de envolvimento em abuso eleitoral (os quais existem desde o ciclo eleitoral de 2020 e foram mantidos em todos os ciclos eleitorais subsequentes, inclusive neste de 2026)
- Um mecanismo de notificação disponível através da Central de Ajuda que permite a indivíduos (usuários ou não-usuários) e entidades o envio de notificações às Empresas da Meta sobre conteúdo que alegam ser ilegal no Brasil, inclusive aqueles potencialmente violadores da regulação eleitoral

- Separadamente, a Meta oferecerá um caminho dedicado para candidatos, partidos políticos, federações e coligações denunciarem conteúdo eleitoral localmente ilegal.
- E finalmente, um canal exclusivo com a Justiça Eleitoral, para receber e processar ordens judiciais eleitorais (abordado em mais detalhes na Seção 7).

Os canais possibilitam denúncias relativas a conduta antidemocrática, desinformação eleitoral, ameaças de violência contra os poderes constitucionais e seus membros, incitação à abolição violenta do Estado Democrático de Direito, violência política de gênero, conduta de ódio e conteúdo político digitalmente alterado potencialmente violador da lei local.

6.1. Canal de Denúncia do Tribunal Superior Eleitoral

Mantemos um canal por meio do qual o Tribunal Superior Eleitoral notifica à Meta conteúdo que, nos termos da legislação brasileira, seria irregular e que foi encontrado no Facebook, no Instagram ou no Threads. As notificações são enviadas através do próprio portal do Tribunal Superior Eleitoral, no contexto do Sistema de Alertas de Desinformação Eleitoral (SIADE) e são processadas pela Meta seguindo o mesmo procedimento aplicado para outros tipos de solicitações regulatórias. O canal operou pela primeira vez durante as eleições gerais de 2022, novamente nas eleições municipais de 2024 e foi mantido para as eleições gerais de 2026.

6.2 Canal de Denúncia de Usuários – Conteúdo localmente ilegal

Para denunciar conteúdo como ilegal, denunciantes podem acessar um formulário de contato dedicado a este envio. Os denunciantes podem acessar este formulário sem precisar criar ou conectar-se a uma conta junto a algum serviço da Meta.

Quando recebemos uma denúncia neste canal, primeiro a analisamos em relação aos Padrões da Comunidade. Se determinarmos que o conteúdo viola nossas políticas, nós o removemos. Se o conteúdo não violar nossas políticas, conduzimos uma análise legal cuidadosa para confirmar se a denúncia é válida, bem como uma *due diligence* de direitos humanos, em linha com nossos compromissos como membro da Global Network Initiative e nossa Política Corporativa de Direitos Humanos.

6.3 Canal de Denúncia para Candidatos e Partidos

A Meta disponibilizará um canal dedicado através do qual candidatos, partidos políticos, federações e coligações podem denunciar conteúdo no Facebook, Instagram e Threads que supostamente constitui conteúdo político-eleitoral ilegal. Este canal utilizará as informações fornecidas por candidatos, partidos políticos, federações e coligações ao Tribunal Superior Eleitoral, de modo que a Meta abrirá o canal após a publicação da lista oficial de candidatos, a partir do dia 16 de agosto.

Empregamos os melhores esforços para garantir que as denúncias recebidas através deste canal sejam analisadas dentro de 24 horas após o recebimento e sejam priorizadas de maneira consistente com a gravidade e o potencial dano aos temas relacionados às eleições. As denúncias são avaliadas em relação aos Padrões da Comunidade da Meta e, quando relevante, à lei eleitoral brasileira, conforme descrito no item anterior.

As decisões tomadas em resposta às denúncias recebidas por meio deste canal estão sujeitas ao processo de contestação descrito acima.

Suporte Dedicado para Candidatos Políticos - Meta Support Pros (Canal Dedicado para Candidatos)

Oferecemos uma equipe de suporte dedicada — *Meta Support Pros* — para auxiliar candidatos e partidos políticos com questões técnicas e de publicidade. Os candidatos podem acessar esse suporte fazendo *login* em uma conta com acesso de administrador à sua página política em www.facebook.com/gpa/help. Esta equipe fornece suporte para dúvidas relacionadas às ferramentas de publicidade, questões relacionadas a anúncios, ao conteúdo orgânico, denúncias, e muito mais.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Engajamento da Equipe Local

Equipes locais de relações governamentais e de parcerias mantém contato e estabelecem canais de comunicação com partidos políticos, realizando sessões de treinamento nas quais informam sobre os recursos e canais de suporte da Meta. Isso inclui engajamento direto com os partidos sobre o uso de nossas plataformas, melhores práticas de segurança de conta e como utilizar as ferramentas de denúncia.

Materiais de treinamento estão publicamente disponíveis no Hub de Eleições http://facebook.com/gpa/brasil2026, e todas as proteções se aplicam às contas declaradas pelos candidatos ao TSE.

Identificação de Candidatos e Partidos
Contatamos proativamente todos os partidos políticos nacionais registrados no TSE através de canais de comunicação divulgados no site do TSE. Como parte dos nossos esforços para disseminar informações, todos os partidos são avisados sobre nossos recursos educativos disponíveis, canais de suporte e oportunidades de engajamento. Além disso, nossas equipes locais de relações governamentais e parcerias mantém canais de comunicação abertos e constantes com partidos e candidatos ao longo do período eleitoral.

6.4 Fluxo de Triagem

Identificação e Roteamento
Quando conteúdo potencialmente violador é denunciado, a automação nos ajuda a encaminhar rapidamente o conteúdo para os revisores humanos, classificando e priorizando o conteúdo para que nossas equipes de revisão possam se concentrar no conteúdo de maior impacto primeiro.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

6.5 Canais de Notificação e Comunicação

Confirmação de Recebimento de Denúncia
Mantemos um processo de confirmação de recebimento de denúncias, para informar aos usuários e não usuários de que suas denúncias foram recebidas. Para denúncias no aplicativo, os usuários recebem confirmações automáticas no aplicativo em sua Caixa de Entrada de Suporte, e há uma exibição do status de suas solicitações. Para denúncias fora do aplicativo usando o formulário de contato, denunciantes recebem a confirmação por mensagem de e-mail enviada ao endereço informado no formulário.

Notificações de Decisões de Aplicação das Regras ou da Lei Local
Mantemos um processo para notificar usuários e denunciantes quando tomamos uma medida de aplicação das nossas regras ou da lei local. A notificação e a declaração de motivos informa o destinatário sobre a decisão, o seu fundamento e as eventuais opções de contestação disponíveis para ele.

Processo de Contestação
Mantemos um processo para que criadores de conteúdo e/ou proprietários de contas recorram de decisões de moderação sobre conteúdo que criaram ou publicaram e que

foi removido. Os criadores de conteúdo e proprietários de contas podem recorrer da decisão original, sendo notificados do resultado. Se houver contestação, o conteúdo é revisado para reavaliar se há violação das nossas políticas e/ou a lei local.

Prazo de Resposta
Embora nos esforcemos para processar as denúncias o mais rapidamente possível, os tempos de resposta podem variar dependendo do volume de envios e da complexidade das questões levantadas. Estamos comprometidos em revisar todas as denúncias em tempo hábil e de acordo com nossas obrigações legais.

Transparência
Os volumes agregados de conteúdo restrito em resposta a solicitações legais no Brasil são publicados no relatório de Restrições de Conteúdo da Meta, que reporta as restrições por país. Os números efetivamente mensurados serão reportados no relatório final consolidado, a ser apresentado ao TSE até 19 de dezembro de 2026 (Art. 28).

SEÇÃO 7: Cumprimento de Ordens Judiciais

Esta seção aborda os requisitos estabelecidos no Artigo 12 da Portaria nº 463/2026/TSE, que especifica os elementos que o plano de conformidade deve conter no que diz respeito aos deveres de cumprimento de ordens e determinações emitidas pela Justiça Eleitoral.

Esses deveres derivam das obrigações substantivas estabelecidas nos Artigos 9-D, §5, 32, 36, 38, §§ 4 e 6, 39 e 40 da Resolução nº 23.610/2019/TSE, que regem, respectivamente: dever de cumprir ordens judiciais para remoção de conteúdo, suspensão de perfil, divulgação de dados ou outras medidas emitidas por autoridades judiciais (Artigo 9-D, §5); a responsabilidade do provedor por não cessar a propaganda irregular após notificação judicial (Artigo 32); a suspensão do acesso a conteúdo em não conformidade com ordem da Justiça Eleitoral (Artigo 36); os padrões e requisitos processuais para ordens judiciais de remoção de conteúdo, incluindo prazos mínimos de cumprimento e requisitos de identificação, como a URL (Artigo 38, §§ 4 e 6); a coleta e o dever de guarda de dados de usuários (Artigo 39); e o cumprimento das ordens judiciais que determinarem o fornecimento de dados e registros eletrônicos para fins probatórios em processos eleitorais (Artigo 40).

As medidas descritas abaixo estabelecem a estrutura, os procedimentos e a capacidade operacional que a Meta mantém para garantir o cumprimento tempestivo e efetivo dessas obrigações durante todo o período eleitoral.

7.1 Estrutura de Assessoria Jurídica Externa

A Meta será representada nas demandas propostas junto à Justiça Eleitoral por advogados integrantes de um escritório de advocacia com sede no Brasil, que atuará como o principal ponto de contato local para receber, fazer a triagem e processar ordens e determinações judiciais.

A cada ciclo eleitoral, o escritório monta uma estrutura exclusiva para lidar com a demanda do contencioso judicial eleitoral. Em 2026, essa estrutura é composta por 112 advogados e 59 funcionários de apoio jurídico, organizados em dois turnos operacionais para garantir a capacidade de atendimento contínuo durante o dia.

Essa estrutura é responsável por: (i) receber ordens e determinações judiciais emitidas pela Justiça Eleitoral por meio do canal exclusivo aberto em cumprimento ao Artigo 10 da Resolução nº 23.610/2019/TSE; (ii) encaminhá-las às equipes internas da Meta para análise técnica e execução; (iii) receber as respostas das equipes internas; e (iv) protocolar as petições correspondentes perante o sistema das cortes eleitorais e gerenciar todo o ciclo de vida processual perante os tribunais em todo o país.

A capacidade desta estrutura é proporcional ao volume e à complexidade das demandas historicamente recebidas durante os períodos eleitorais, e é compartilhada por todas as Empresas da Meta como um todo.

7.2 Estrutura e Capacidade Operacional Interna

Para as eleições gerais brasileiras de outubro de 2026, a Meta opera uma estrutura multifuncional que reúne equipes responsáveis pela elaboração e aplicação das políticas de conteúdo, de integridade, de anúncios, de combate à desinformação, de autenticidade e de suporte jurídico, entre outras. A estrutura inclui pessoas com experiência no mercado brasileiro e no idioma português, e é reforçada com a presença no Brasil durante todo o período.

Não há critérios fixos ou padronizados para determinar o número de profissionais necessários para uma determinada eleição. A capacidade é dimensionada antes de cada período eleitoral a partir de uma previsão de volume de demandas e ajustada de acordo com a avaliação de risco para aquela eleição, com recursos adicionais alocados onde surgem riscos específicos e adicionais. A capacidade é entregue por meio de uma estrutura de revisão compartilhada que também atende a outras jurisdições e tipos de casos; as questões eleitorais são priorizadas dentro dessa estrutura

compartilhada durante o período, e a cobertura é estendida em torno de cada turno de votação.

Recebimento, triagem e execução de ordens judiciais
A Meta mantém canais online dedicados para o atendimento das ordens judiciais. As ordens eleitorais para remoção e produção de dados são enviadas por meio de canais de entrada dedicados para o Brasil, onde a solicitação e a ordem judicial subjacente são fornecidas juntas em um único envio estruturado, identificando o conteúdo objeto da ordem, o prazo estabelecido pelo tribunal para o cumprimento e o serviço em questão. De acordo com a legislação brasileira aplicável, o objeto específico deve ser identificado na própria ordem.

Cada solicitação é analisada em relação às nossas políticas e à legislação local. Quando o conteúdo viola alguma de nossas políticas, nós o removemos. Quando não viola – e uma ordem judicial assim exige –, restringimos o acesso ao conteúdo no Brasil. Solicitações referentes a categorias de discurso mais sensíveis, incluindo contas de interesse público significativo e de imprensa, são encaminhadas para revisão adicional de especialistas e de políticas antes que qualquer ação seja tomada.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

7.3 Procedimento de Escalonamento Interno

Em casos de ordens judiciais que exijam dados ou informações adicionais para possibilitar o cumprimento, o procedimento interno adotado consiste em quatro etapas escalonadas:

1. Verificação imediata da existência e suficiência de identificadores válidos na ordem ou na documentação anexa recebida (ou seja, a respectiva URL do conteúdo ou conta em questão, de acordo com o Artigo 17, III, da Resolução nº 23.608/2019/TSE).
2. Resposta tempestiva ao Tribunal, por meio de petição protocolada nos autos, indicando de forma objetiva e específica quais dados ou elementos são necessários para possibilitar o cumprimento;
3. Uma nova verificação assim que os identificadores suplementares forem recebidos; e
4. Uma vez confirmada a suficiência das informações, encaminhamento da ordem às equipes internas para cumprimento.

A Meta entende que a obrigação prevista no Artigo 9-E, § 1º, da Resolução nº 23.610/2019/TSE deve ser interpretada em harmonia com o arcabouço jurídico mais amplo que rege as ordens de remoção de conteúdo. Em particular, tal obrigação deve ser interpretada de forma consistente com o Artigo 38, § 4º, da mesma Resolução e com o Artigo 19 da Lei nº 12.965/2014 (Marco Civil da Internet) — dispositivos que exigem que as ordens de remoção de conteúdo online identifiquem especificamente o conteúdo em questão por meio de um identificador válido, sob pena de nulidade —, bem como com o Artigo 9-D, § 5º, que reconhece a possibilidade de o provedor indicar, de forma objetiva e dentro do prazo de cumprimento, quaisquer dados complementares necessários à execução integral da ordem.

Assim interpretado, o dever de impedir a recirculação de conteúdo previamente declarado ilícito pressupõe a identificação prévia e precisa — mediante identificadores de conteúdo válidos — do material a ser tornado indisponível, uma vez que são esses identificadores que tornam tecnicamente viável, para os provedores, localizar o conteúdo específico e dar cumprimento efetivo e tempestivo às determinações judiciais de remoção.

A Meta cumpre ordens judiciais válidas e solicitações legais para restringir conteúdo e produzir dados de usuários, recebidas de autoridades brasileiras, de acordo com as leis locais. Mantemos relatórios detalhados de restrição de conteúdo documentando nosso cumprimento de ordens de remoção emitidas por tribunais e autoridades regulatórias. Esses relatórios cobrem o conteúdo restrito no Facebook, Instagram e outros serviços da Meta, de acordo com os requisitos legais locais.

WHATSAPP

O WhatsApp LLC é uma entidade legal separada e distinta da Meta Platforms, Inc. De acordo com o Artigo 2º da Portaria nº 463/2026/TSE, o WhatsApp é um serviço de mensagens privadas que fornece mensagens criptografadas de ponta a ponta por padrão, além de chamadas de voz e de vídeo.

O WhatsApp é a plataforma de mensagens mais popular do mundo, usada para se comunicar de forma privada com amigos, familiares e empresas. Em sua essência, o WhatsApp permite:

- **Mensagens instantâneas** — texto, mensagens de voz, fotos, vídeos e documentos compartilhados em tempo real
- **Chamadas de voz e vídeo** — chamadas gratuitas por Wi-Fi ou dados móveis
- **Conversas em grupo** — conversas com vários participantes para coordenação e comunidade

- **Atualizações de status** — conteúdo efêmero de mensagens compartilhado com contatos mútuos
- **Canais** — um recurso opcional de transmissão unidirecional no WhatsApp, separado das mensagens privadas, projetado para ajudar as pessoas a acompanhar informações de pessoas, organizações e figuras públicas que são importantes para elas
- **Criptografia de ponta a ponta** — todas as mensagens e chamadas são protegidas pelo protocolo de criptografia Signal, garantindo que apenas o remetente e o destinatário pretendido possam lê-las] atualizações de Canais não são criptografadas.

As Três Interfaces
O WhatsApp opera por meio de três interfaces distintas, cada uma projetada para um tipo diferente de usuário e escala de comunicação:

WhatsApp Messenger (Aplicativo para o Consumidor)
Um aplicativo de mensagens padrão e gratuito para uso pessoal. Ele permite que as pessoas conversem, liguem e compartilhem mídia com seus contatos. Esta é a versão com a qual a maioria dos usuários interage diariamente.

Aplicativo WhatsApp Business
Um aplicativo gratuito e independente projetado para proprietários de pequenas empresas. Ele tem a mesma aparência do aplicativo para o consumidor, mas adiciona recursos profissionais, como:

- Um perfil comercial com a marca (endereço, horário, descrição, site)
- Catálogos de produtos e serviços que funcionam como vitrines online
- Mensagens automatizadas de saudação e ausência
- Respostas rápidas e rótulos de mensagens para organizar conversas
- Uma configuração de número de telefone único e usuário único

O Aplicativo Business é ideal para proprietários individuais e pequenas equipes que precisam de uma presença profissional, mas sem complexidade técnica.

Plataforma WhatsApp Business (API)
Um conjunto de APIs e ferramentas de desenvolvedor criadas para médias e grandes empresas que precisam trazer clientes em escala. **É importante observar que partidos políticos, políticos, candidatos políticos e campanhas políticas não estão autorizados a acessar a Plataforma WhatsApp Business (API).** Os principais recursos incluem:

- Vários números de telefone e nomes de exibição
- Mensagens multimídia e de marketing personalizadas

- Integração com sistemas de back-end (CRM, suporte ao cliente, plataformas de marketing)
- Agentes ou bots interagindo com os clientes de forma programática
- Rastreamento e análise' de mensagens enviadas e recebidas

Ao contrário dos aplicativos, a Plataforma Business não é baixada — ela é implementada por desenvolvedores ou por meio de um Parceiro de Mensagens de Negócios terceirizado certificado.

Todas as três interfaces compartilham a mesma base de privacidade: cada mensagem — seja pessoal ou comercial — é protegida pelo mesmo protocolo de criptografia líder do setor.

O WhatsApp não opera um feed de conteúdo, não classifica ou recomenda mensagens e não oferece pesquisa no aplicativo ou descoberta de usuários ou grupos não conectados. As pessoas usam o WhatsApp com mais frequência para se comunicar com outras pessoas que já conhecem, e você precisa do número de telefone delas para enviar mensagens. O WhatsApp não oferece impulsionamento pago de conteúdo político-eleitoral.

Seção 1: Termos de Serviço, Políticas e Diretrizes do WhatsApp

Esta seção aborda os requisitos estabelecidos nos Artigos 13, 14, 22 e 23 da Portaria nº 463/2026/TSE. Coletivamente, essas disposições exigem que o plano de conformidade descreva:

1. Os instrumentos normativos que regem o uso da plataforma e o tratamento de conteúdo político-eleitoral por terceiros, incluindo a arquitetura de moderação de conteúdo do provedor, as medidas de aplicação disponíveis e as estruturas de governança que supervisionam sua aplicação (Art. 22);
2. Os parâmetros, critérios e fluxos operacionais empregados pelo provedor para a identificação, detecção e remoção proativas de conteúdo que dissemine fatos notoriamente falsos ou gravemente descontextualizados capazes de afetar a integridade eleitoral, bem como os mecanismos de monitoramento contínuo e acordos para verificação de fatos que apoiam esse esforço (Art. 13);
3. Os mecanismos para detectar e interromper o comportamento inautêntico coordenado (Art. 14); e
4. As medidas de iniciativa própria do provedor para o planejamento de ações corretivas e preventivas, o aprimoramento dos sistemas de

recomendação e a transparência dos resultados, incluindo, quando aplicável, os critérios e procedimentos para a remoção de contas falsas, automatizadas ou de bots (Art. 23).

O uso do WhatsApp é regido por seus Termos de Serviço, Termos de Serviço Suplementares dos Canais, Termos de Serviço do WhatsApp Business e políticas comerciais, bem como por suas Diretrizes de Mensagens e Canais. Eles são desenvolvidos pelas equipes de Políticas de Produto e Jurídica do WhatsApp e servem como a estrutura de política fundamental que estabelece o conteúdo e os comportamentos que são e não são permitidos no WhatsApp.

1.1 Limitações de Uso Político e Eleitoral

Além das restrições gerais e uso de automação, que serão abordados na Seção 1.3, atores políticos e campanhas políticas estão especificamente proibidos de usar a Plataforma WhatsApp Business ou usar o Recurso de Listas Transmissões Comerciais no Aplicativo WhatsApp Business. Essas medidas evitam o tipo de envio de mensagens políticas em massa proibido pelo Artigo 34 da Resolução nº 23.610/2019/TSE.

A Política de Mensagens do WhatsApp Business proíbe o uso da Plataforma WhatsApp Business por Partidos Políticos, Políticos, Candidatos Políticos e Campanhas Políticas, bem como entidades que ofereçam serviços relacionados à política, candidatos políticos, estratégia de campanha política, serviços de campanha política, empresas privadas que fornecem soluções de votação e sistemas para eleições e provedores de serviços governamentais exclusivos. Eles proíbem ainda o uso da Plataforma por Agências de Segurança Pública, Serviços Militares, Agências de Segurança Nacional e Inteligência.

Para o aplicativo Business, os Termos da Meta para Transmissões Comerciais no Aplicativo WhatsApp Business estabelecem que as Transmissões Comerciais não podem ser usadas por certas categorias de entidades, incluindo Entidades Governamentais e Políticas. De acordo com a Seção 6.g (Rescisão) e 6.l (Uso por Entidades Governamentais e Políticas), se a Meta determinar, a seu exclusivo critério, que o usuário é uma Entidade Governamental e Política, a Meta poderá encerrar imediatamente seu acesso às Transmissões Comerciais. Entidades Governamentais e Políticas incluem, sem limitação, entidades governamentais; entidades políticas (incluindo partidos políticos, políticos, candidatos políticos e campanhas políticas); organizações políticas sem fins lucrativos; entidades que oferecem serviços relacionados à política e campanhas políticas; e empresas privadas que fornecem soluções de votação e sistemas para eleições.

1.2 Lidando com o Abuso Enquanto Mantém a Criptografia de Ponta a Ponta: Limitando a Viralidade e Reduzindo a Desinformação

O WhatsApp se dedica a ajudar as pessoas a falar livremente e reconhece que as instituições democráticas protegem esse direito. O WhatsApp mantém uma equipe dedicada para prevenir abusos, focada especificamente em manter a natureza privada do WhatsApp, prevenir o uso indevido e coordenado do WhatsApp, além de capacitar os usuários a combater a desinformação.

Criptografia de Ponta a Ponta Para Suas Conversas Pessoais
O WhatsApp fornece criptografia de ponta a ponta por padrão para mensagens pessoais, com o objetivo de proteger as conversas das pessoas contra *hackers*, criminosos e outras ameaças cibernéticas. Nenhum terceiro — incluindo o WhatsApp — pode ler ou ouvir conversas ou chamadas pessoais. A verificação em duas etapas oferece segurança adicional à conta. Especialistas em eleições defendem que fornecer segurança forte é fundamental para proteger as mensagens das pessoas. Isso inclui discursos políticos e discussões sobre candidatos e suas campanhas.

Segurança Desde o Início

O WhatsApp criou recursos de segurança para capacitar os usuários. Esses incluem a capacidade de bloquear e reportar contato indesejado, parar de ser adicionado a novos grupos sem permissão, sair de grupos sem notificação e silenciar chamadas de números desconhecidos.

Proteções estruturais adicionais no WhatsApp incluem a impossibilidade de pesquisar usuários individuais no aplicativo. O WhatsApp também desenvolveu os cartões de contexto "Experiência da Primeira Mensagem" e "Experiência de Adição a Grupo" que alertam os usuários quando são contatados por números desconhecidos ou adicionados a grupos por não contatos, fornecendo opções fáceis para bloquear ou denunciar.

A funcionalidade de Verificação de Privacidade é fácil de usar e ajuda os usuários a fazerem escolhas para reforçar a segurança de suas contas, por meio de um guia passo-a-passo sobre as configurações de privacidade.

Prevenção de Abuso em Grupos
Os próprios limites de tamanho do grupo (máximo de 1.024 usuários) restringem o alcance de atores abusivos. Desenvolvemos uma configuração de privacidade para permitir que os usuários decidam quem pode adicioná-los aos grupos. As escolhas

incluem: todos, apenas seus contatos ou apenas contatos selecionados a critério do usuário. Essa configuração impede que as pessoas sejam adicionadas a grupos indesejados que podem ser criados para enviar conteúdo em escala. Também criamos ferramentas para ajudar os administradores de grupos a gerenciar seus grupos, incluindo a capacidade dos administradores de bloquear a foto e o nome do perfil do grupo, aprovar quem pode entrar em um grupo e restringir as configurações para que apenas o administrador possa enviar uma mensagem em um grupo. Atuamos em cima das contas, incluindo rescindindo elas quando determinamos que estão tentando criar grupos em escala para enviar mensagens a usuários.

Bloquear e Denunciar

Ao contrário do SMS tradicional, o WhatsApp fornece uma maneira simples para os usuários bloquearem contas. Os usuários podem facilmente fazer denúncias ao WhatsApp se encontrarem mensagens, grupos, contas ou atualizações de Canais problemáticos. Banimos a grande maioria das contas abusivas por meio de detecção automatizada, inclusive antes que qualquer mensagem seja enviada. No entanto, as denúncias nos ajudam a identificar se uma conta ou Canal está envolvido em violação de comportamento, permitindo uma investigação mais aprofundada para evitar danos.

Limites de Encaminhamento e Rotulagem

O WhatsApp limita o encaminhamento de mensagens e atualizações de Canais a cinco conversas de uma vez. Isso torna o WhatsApp uma das poucas empresas de tecnologia a restringir intencionalmente o compartilhamento. Quando introduzido, isso reduziu as mensagens encaminhadas em mais de 25%.

Estabelecemos limites adicionais a mensagens e atualizações de Canais para mensagens que foram encaminhados muitas vezes. Esses encaminhamentos são rotulados com o rótulo de altamente encaminhadas. Mensagens pessoais ou atualizações que mostram o rótulo de altamente encaminhadas indicam que elas provavelmente não se originaram de um contato próximo e merecem um escrutínio maior. Elas só podem ser encaminhadas para uma outra conversa por vez, uma limitação que reduziu esses tipos de mensagens em mais de 70%. Além disso, as atualizações encaminhadas de um Canal sempre vinculam de volta ao Canal original para fornecer aos destinatários contexto adicional quanto à fonte da atualização.

Pesquisar na web

O WhatsApp fornece uma maneira simples de verificar mensagens pessoais que foram encaminhadas muitas vezes. Isso ajuda nossos usuários a encontrar resultados de notícias ou outras fontes de informação sobre o conteúdo que receberam. Esse recurso funciona permitindo que os usuários toquem em uma lupa que, por sua vez,

direcionará o usuário à pesquisa pelo teor da mensagem em um motor de busca terceiro, a partir do navegador de internet do seu dispositivo.

Suporte para Verificação de fatos no WhatsApp
O WhatsApp se preocupa profundamente com a segurança de nossos usuários e continuamos a nos concentrar na prevenção da desinformação e na conexão dos usuários com informações precisas e confiáveis. Fizemos parceria com a International Fact-Checking Network para disponibilizar verificadores de fatos certificados no WhatsApp, incluindo nos Canais. Por meio dessa parceria, organizações de verificação de fatos pelo mundo, incluindo o Brasil, usam o WhatsApp para ajudar a conectar os usuários a informações confiáveis. Capacitamos as organizações de verificação de fatos e a garantir que elas tenham os recursos necessários para combater a desinformação e garantir que os eleitores tenham informações confiáveis sobre como se registrar para votar e onde votar.

1.3 Prevenção de Abuso, Interferência e Comportamento Inautêntico

Os Termos de Serviço do WhatsApp proíbem especificamente o uso de seus serviços para coordenar danos, publicar falsidades ou deturpações, fazer-se passar por alguém ou enviar comunicações ilegais ou não permitidas, como mensagens em massa, mensagens automáticas ou discagem automática. Os usuários que violarem esses termos podem ter suas contas desativadas ou suspensas.

O Whatsapp utiliza sinais comportamentais disponíveis (como padrões de mensagens em massa e indicadores de atividade automatizada) e denúncias de usuários para identificar contas violadoras em potencial.

O WhatsApp possui a melhor tecnologia de detecção de spam da categoria. Nossa tecnologia detecta contas envolvidas em comportamento anormal para que não possam ser usadas para espalhar *spam* ou desinformação. Isso significa que nossos sistemas automatizados são frequentemente capazes de interromper o abuso mesmo antes de ser denunciado.

O WhatsApp tem mecanismos robustos de detecção automatizada projetados para identificar comportamentos suspeitos desde o registro da conta, na criação de grupos e ao enviar mensagens. Os seguintes mecanismos de salvaguardas de segurança

contra burla são implementados:

- Detecção de criação de contas coordenadas: Os sistemas identificam várias contas que compartilham características comuns, indicando controle pelo mesmo indivíduo.
- Aplicação das regras *(enforcement)* baseada em feedback negativo: A aplicação se concentra em contas com altas taxas de bloqueio e denúncia de outros usuários.
- Aplicação das regras *(enforcement)* graduada: Um sistema graduado leva ao banimento de contas por violações graves ou repetidas.

1.3.1 Medidas Preventivas para Dificultar a Formação de Redes Inautênticas

O WhatsApp foi projetado para que o abuso em escala seja difícil de alcançar, impedindo a capacidade de um ator mal-intencionado de obter e operar um grande número de contas de forma barata e automática no momento do registro da conta.

O WhatsApp bane contas no momento do registro ou antes dele de forma contínua, e essa camada preventiva é responsável por uma parcela substancial de todas as contas abusivas removidas do serviço. As medidas abaixo descrevem como o registro é estruturado para resistir à automação, as categorias de sinal usadas para identificar possíveis abusos e as ações tomadas quando o abuso é detectado — incluindo as salvaguardas que protegem os usuários legítimos de aplicação errônea.

Prevenção — atrito embutido no registro

- **Titularidade do número de telefone**. Toda conta deve verificar o controle de um número de telefone real por meio de um código de uso único entregue por SMS ou chamada de voz. Não há caminho de registro inicial que pule essa etapa.
- **O registro é bloqueado**. Verificações de integridade são executadas na própria solicitação de registro, de modo que tentativas suspeitas de script são interrompidas na porta da frente, em vez de depois que uma conta passe a existir.
- **Apenas clientes oficiais**. Tentativas de registro de aplicativos WhatsApp de terceiros modificados ou não oficiais são detectadas e bloqueadas.

Sinais usados para identificar abusos no registro

- **Reputação do número** — se o número, ou números muito semelhantes a ele, foi usado recentemente de forma abusiva.
- **Reputação da rede** — o endereço IP e as informações da operadora associada, e se essa rede foi vinculada a comportamento suspeito.

- **Integridade do cliente** — detecção de de aplicativos não oficiais ou modificados.
- **Padrões de coordenação** — anomalias nos atributos da solicitação de registro que indicam atividade em massa, orientada por máquina, em vez de humanos individuais.

Medidas adotadas quando se identifica abuso

- **Bloqueio ou banimento preventivo**. Contas suspeitas são barradas durante ou antes da conclusão do cadastro — sendo banidas antes mesmo de enviarem qualquer mensagem.
- **Aumento da fricção para o número ou vetor de acesso**. Nos casos em que há indícios fortes, mas não conclusivos, o cadastro é recusado e o usuário é forçado a tentar novamente mais tarde; isso gera custos e atrasos para o agente malicioso, sem penalizar permanentemente um usuário legítimo.
- **Recursos**. As medidas de controle não são definitivas. Os usuários podem recorrer da decisão e pode ser solicitada a apresentação de informações adicionais para comprovar a titularidade legítima e recuperar o acesso.

1.3.2. Medidas de Aplicação e Mecanismos Para Detectar Padrões de Coordenação

O WhatsApp respondeu ao aumento dos níveis de crimes cibernéticos e atividades maliciosas em toda a Internet melhorando as defesas de seus produtos, investindo pesadamente no aumento da detecção e aplicação de regras, continuando a cooperação com reguladores, autoridades policiais e parceiros do setor, e fornecendo funcionalidades expandidas aos usuários no aplicativo e por meio de outros mecanismos.

A tomada de decisão de aplicação de regras pelo WhatsApp resulta de uma estrutura de revisão formatada e em várias camadas. O WhatsApp usa técnicas de processamento automatizado e equipes de revisão humana para detectar e revisar informações, incluindo informações de perfil de conta, grupo e comunidade, bem como mensagens denunciadas, em busca de possíveis violações.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Procedimentos de Tomada de Decisão

Análises manuais de violações de nossas Diretrizes de Mensagens, Política de Mensagens de Negócios e Diretrizes de Canais são realizadas por revisores humanos e/ou combinadas com detecção automatizada.

Nossas equipes de Operações de Integridade do WhatsApp identificam e tratam o conteúdo violador por meio de classificadores de aprendizado de máquina (detecção proativa) e por meio de medidas reativas projetadas para agir contra conteúdo violador de nossas políticas, denunciado por usuários e terceiros.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

Mecanismo de Contestação no Aplicativo
Mantemos um processo para que os titulares de contas possam contestar as decisões tomadas pelo WhatsApp na aplicação das suas regras. Quando um conteúdo ou uma conta é restringida, notificamos o titular da conta por meio de um Aviso de Aplicação de Regras, explicando o motivo da ação e fornecendo orientações sobre como evitar futuras violações. A partir desse aviso, os titulares de contas têm a opção de solicitar uma revisão da decisão de aplicação de regras (uma contestação). Essa funcionalidade de contestação fica disponível por seis meses (180 dias) a partir da data da ação de aplicação de regras.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.4 Diretrizes de Canais do WhatsApp

Os Canais do WhatsApp são um recurso opcional de transmissão unidirecional dentro do WhatsApp, separado das mensagens privadas. O WhatsApp mantém um conjunto de Diretrizes de Canais do WhatsApp aplicáveis globalmente e os Termos de Serviço aplicáveis que definem o que é e o que não é permitido no serviço. Empregamos uma combinação de design de produto, processamento automatizado de dados e análise humana, bem como uma série de medidas de aplicação, para implementar essas políticas, incluindo a suspensão de entregas futuras de atualizações do canal, a remoção de conteúdo violador dos campos de perfil do canal, o bloqueio geográfico e a suspensão de um canal.

As Diretrizes de Canais do WhatsApp definem o conteúdo permitido. Entre outras coisas, essas Diretrizes proíbem:

- Atividade ou conteúdo que propositalmente engane, se passe por outro, ou deturpe intencionalmente, ou que, de outra forma, aplique golpes, perpetre fraudes ou explore usuários
- Conteúdo que possa causar danos graves às pessoas, incluindo conteúdo que constitua uma ameaça crível à segurança pública, incite a violência ou organize ou coordene atividades violentas ou criminosas
- Conteúdo que apoie organizações ou indivíduos extremistas violentos ou criminosos, facilite a exploração de pessoas, ou retrate, ameace ou promova violência, abuso ou exploração sexual
- Imagens que sejam excessivamente violentas ou explícitas, ou que contenham conteúdo sexualmente explícito ou pornografia
- Conteúdo que constitua uma ameaça grave, direta e imediata de violência contra indivíduos ou instituições, incluindo o judiciário e as instituições democráticas
- Incitação à violência eleitoral

Nos Canais do WhatsApp, as proteções estruturais adicionais incluem:

- Canais são projetados como um recurso de transmissão unidirecional, o que inerentemente impede interações diretas ou mensagens diretas entre os administradores do canal e os seguidores
- Os Canais não permitem mensagens em grupo
- O(s) administrador(es) do canal não são identificáveis para outros usuários. O(s) administrador(es) não podem identificar seguidores que ainda não sejam contatos. Os seguidores não podem identificar quaisquer outros seguidores do mesmo canal.
- Os Canais possuem um regime de verificação para apresentar aos usuários um meio de identificar canais autênticos.
- O WhatsApp limita os novos seguidores a 30 dias de histórico do canal e impõe limites de encaminhamento, reduzindo a viralidade e a utilidade do produto na promoção de conteúdo.
- As notificações do canal são silenciadas por padrão para todos os usuários, reduzindo a eficácia dos administradores de canal que tentam enviar atualizações de "spam" para seus seguidores
- Os Canais limitam a pesquisa na plataforma ao título e à descrição do canal, e implementam o bloqueio de termos de pesquisa para palavras-chave associadas a conteúdo que viole as políticas.

Descoberta de Canais: os usuários não veem conteúdo de Canais até selecionarem ou seguirem um Canal. Não há recomendação de conteúdo específico para os usuários. Recomendações de Canais são apresentadas aos usuários na aba Atualizações (separada da caixa de entrada de mensagens). O WhatsApp utiliza as seguintes informações limitadas para oferecer recomendações relevantes:

- Informações de criação do canal (nome, foto, descrição)
- Atualizações do canal e sua frequência
- Seguidores, visualizadores e sinais de engajamento (reações)
- Informações gerais de localização (país/cidade)
- Informações de dispositivo e conexão (idioma)
- Informações de uso sobre a atividade do usuário nos Canais (tipos de canais com os quais interagiu, pesquisas, recursos usados e tempo gasto)

1.4.1 Abordagem Proativa para a Remoção de Conteúdo Ilegal nos Canais

Os Canais do WhatsApp são um recurso de transmissão onde o conteúdo não é criptografado e é acessível publicamente. O WhatsApp implementou medidas específicas para identificar e agir proativamente contra o conteúdo violador nos Canais, incluindo através do uso de sistemas de detecção automatizada.

O WhatsApp implementou medidas líderes do setor para proteger os Canais e seus usuários de conteúdo nocivo, ilegal e que viole as políticas. Identificamos esse conteúdo e tomamos medidas de aplicação das regras contra a conta e/ou Canal responsável, usando uma combinação de detecção automatizada proativa, análise humana e denúncias de autoridades válidas e/ou usuários.

Nossos sistemas de aplicação de regras para Canais incluem detecção baseada em sinais e classificadores que permitem a identificação e suspensão de canais que distribuam conteúdo violador. Medimos a precisão de nossos sistemas de suspensão de Canais para garantir a exatidão nas medidas de aplicação.

Os Canais do WhatsApp analisam cada atualização do canal, incluindo imagens, texto, vídeo e áudio, para detectar e aplicar medidas contra conteúdo que viole as políticas. As ferramentas de detecção automatizada do WhatsApp incluem classificadores de texto e mídia de aprendizado de máquina projetados para identificar conteúdo potencialmente violador. Quando o conteúdo é identificado por nossos sistemas de detecção automatizada como potencialmente violador, ele pode sofrer aplicação de medidas automaticamente quando há alta confiança, ou ser colocado na fila para análise humana para confirmação de quaisquer violações das Diretrizes de Canais do WhatsApp. O WhatsApp também utiliza bancos de mídia de conteúdo violador conhecido. Quando se descobre que a mídia corresponde a material contido nesses bancos, medidas de aplicação automatizadas são acionadas sem a necessidade de revisão humana.

1.4.2 Tecnologias de Detecção nos Canais

Sistemas Automatizados:

O WhatsApp executa sistemas automatizados que analisam vários elementos do Canal, incluindo metadados e conteúdo, para determinar se ele viola alguma das políticas ou diretrizes do WhatsApp.

Os Canais do WhatsApp possuem medidas extensivas em vigor para, identificar e agir contra conteúdo gerado pelo usuário que possa violar nossas políticas, e tomam medidas para mitigar a disseminação de tal conteúdo. Mantemos uma política rigorosa segundo a qual qualquer Canal contendo conteúdo sexualmente explícito, violento ou ilegal está sujeito a sanções coercitivas, que variam desde o fechamento temporário do Canal até que o administrador exclua o conteúdo violador, até a suspensão completa do Canal. Nosso monitoramento automatizado para os Canais do WhatsApp inclui:

- Classificadores: Classificadores de aprendizado de máquina no nível do conteúdo e no nível da conta para detectar material que viole as políticas
- Sistemas de detecção de imagens: Hashing perceptual e criptográfico para identificar imagens violadoras conhecidas

Também contamos com denúncias de usuários, para detectar e prevenir abusos, e trabalhamos continuamente para aprimorar nossas capacidades de detecção.

Revisão Humana: A revisão humana pode ser aplicada ao conteúdo sinalizado por meio de detecção automatizada ou denúncias de usuários nos Canais. Os canais denunciados são avaliados por revisores treinados em relação às nossas políticas, e medidas de aplicação são tomadas conforme apropriado, incluindo restrições temporárias ou suspensão permanente.

Monitoramento Contínuo: monitoramos a precisão de nossas decisões de suspensão de Canais por meio do rastreamento contínuo de métricas, incluindo medições de precisão média de 7 dias que avaliam a proporção de Canais do WhatsApp suspensos que foram identificados corretamente como violadores de nossas políticas. Esse monitoramento contínuo de desempenho nos permite manter e melhorar a precisão de nossos sistemas de aplicação automatizados.

Precisão da detecção: monitoramos com que frequência nossos sistemas identificam corretamente conteúdos como infratores. Nossos controles de qualidade incluem:

- Verificações por amostragem de decisões automatizadas.
- Análise de padrões de erro para melhorar continuamente nossos sistemas de detecção.
- Treinamento dedicado para revisores humanos em cada atualização de política.
- Análise de causa raiz quando erros são identificados.

Metodologia de Indicadores de Qualidade de Detecção
Avaliamos nossa detecção e aplicação de conteúdo violador por meio de vários processos de garantia de qualidade e melhorias contínuas do sistema. As métricas usadas para avaliar o desempenho variam de acordo com diferentes fatores, como ferramenta e contexto. Os seguintes indicadores são mais comumente usados para avaliar a detecção e aplicação da remoção de conteúdo em nossas plataformas.

Principais Indicadores de Qualidade de Detecção

- A precisão mede a proporção de verdadeiros positivos (conteúdo violador identificado corretamente) entre todas as identificações positivas feitas por nossos sistemas. Isso nos ajuda a determinar com que frequência o conteúdo sinalizado como violador está de fato violando nossas políticas
- A prevalência é a proporção estimada de conteúdo violador presente na plataforma relevante em um determinado momento.
- A taxa de reversão é a porcentagem de medidas de aplicação que são revertidas após uma análise mais aprofundada, o que destaca possíveis erros nas decisões iniciais.

1.4.3 Aplicação de Regras (*Enforcement*) nos Canais do WhatsApp

Conforme informado acima, analisamos atualizações de Canais quanto à conformidade com nossas Diretrizes de Canais e tomamos medidas corretivas contra canais violadores, incluindo suspensão. Canais que compartilham conteúdo que viola as políticas, conteúdo limítrofe associado a danos de alta gravidade ou conteúdo de baixa qualidade e impróprio são ocultados da descoberta por novos usuários, a fim de mitigar ainda mais os riscos.

Quando violações de políticas são identificadas, as medidas de aplicação incluem: remoção das informações do perfil do canal; suspensão do envio de atualizações do canal aos usuários (fechamento do canal); remoção de conteúdo infrator dos campos do perfil do canal; revogação do link de convite do canal; suspensão do canal (globalmente ou em uma jurisdição específica); ou banimento de administradores do canal por violações de alta gravidade ou reincidentes. O WhatsApp também atende a solicitações válidas de autoridades policiais.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

1.5 Outras informações relevantes

Transparência e Relatórios Relevantes para as Autoridades Eleitorais Brasileiras
O WhatsApp responde a solicitações legais válidas de dados de usuários de autoridades brasileiras — incluindo o TSE e as autoridades policiais — de acordo com a lei aplicável, os Termos de Serviço do WhatsApp e sua Política de Privacidade. Os processos do WhatsApp para responder a solicitações governamentais de dados estão descritos em suas Informações para Autoridades Policiais.

O WhatsApp reconhece o importante papel do TSE na proteção da integridade das eleições brasileiras e continua disponível para se envolver de forma construtiva — inclusive por meio de mecanismos de transparência pública, participação em diálogos institucionais e respostas tempestivas a solicitações legais de dados — para apoiar o objetivo compartilhado de salvaguardar os processos democráticos.

Preservação de provas
O WhatsApp preserva os registros de contas de acordo com a lei aplicável e seus termos de serviço, incluindo sua Política de Privacidade. Para contas violadoras detectadas como resultado de medidas de detecção e aplicação em escala (incluindo contas identificadas por meio de denúncias de usuários), o WhatsApp exclui os dados associados à conta banida após 30 dias.

Alterações Relevantes nos Termos de Serviço/Políticas de Conteúdo
O WhatsApp pode fazer alterações em seus termos de serviço e/ou políticas de conteúdo de tempos em tempos devido a mudanças na aplicação, atualizações de políticas ou requisitos de conformidade regulatória. Quando o faz, o WhatsApp notifica os usuários sobre alterações relevantes e atualiza a "Data de Vigência" na parte superior dos termos do WhatsApp. A notificação de alterações relevantes pode assumir a forma de notificações no aplicativo, páginas de destino e/ou conteúdo da central de ajuda.

O uso contínuo de nossos Serviços por um usuário é considerado aceitação de nossos Termos e de quaisquer alterações subsequentes. Os usuários que não concordarem com os Termos alterados podem parar de usar o serviço excluindo sua conta.

Seção 2: Avaliação de Impacto do WhatsApp na Integridade Eleitoral

2.1 Avaliação de Risco

2.1.1 Metodologia e frequência da avaliação de impacto na integridade eleitoral

A abordagem geral de avaliação de risco da Meta abrange o WhatsApp, aplicando a mesma estrutura de avaliação em níveis e estruturada em toda a plataforma. Isso significa que a metodologia busca mitigar riscos sistêmicos que possam surgir no contexto de processos eleitorais. Principais elementos desta abordagem: Priorização, Cadência e Avaliação Eleitoral Personalizada.

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

[INÍCIO DO CONTEÚDO CONFIDENCIAL]

[FIM DO CONTEÚDO CONFIDENCIAL]

2.1.3 Metodologia para medir resultados de mitigação

O WhatsApp mede a efetividade de suas mitigações de integridade eleitoral através de uma estrutura organizada que combina monitoramento contínuo com avaliação pós-eleitoral.

Os principais indicadores monitorados incluem o volume de contas banidas, taxas de Mensagens Altamente Encaminhadas, o volume de denúncias, tempo médio de resposta a escalonamentos, volume de mensagens em massa bloqueadas e medidas temporárias tomadas para lidar com risco emergente (TRRM).

A linha de base é estabelecida usando métricas de períodos não eleitorais e dados históricos de ciclos eleitorais brasileiros anteriores. Esta linha de base dupla permite que a equipe distinga picos do período eleitoral de padrões de estado estável e compare o desempenho atual com eleições anteriores.

As metas esperadas incluem manter o nível atual de mitigação de risco em mensagens privadas e cumprir os SLAs de tempo de resposta para escalonamentos de parceiros confiáveis e autoridades.

Os critérios de cálculo dependem de comparação período a período entre períodos eleitorais e não eleitorais, mecanismos de detecção de picos e análise de eficácia de TRRM conduzida comparando métricas antes e depois da ativação da medida. Cada Medida de Redução Temporária de Risco possui critérios de entrada e saída claramente definidos que permitem medição rigorosa de impacto.

O processo de avaliação contínua é apoiado por painéis de monitoramento interno, medição de detecção e aplicação de regras em tempo real durante IPOCs (Centros de Operações de Produtos de Integridade) e revisões pós-eleitorais da efetividade de medidas temporárias.

2.1.4 Capacidade de intervir em conteúdo de gravidade extraordinária

A Meta (incluindo o WhatsApp) implementa e aprimora continuamente estratégias de mitigação dinâmicas para abordar vulnerabilidades identificadas em todos os domínios de risco. Com base em ampla experiência operacional, esses mecanismos, estruturas e fluxos de trabalho passam por atualizações contínuas para combater efetivamente desafios novos e em evolução. Para gerenciar janelas críticas, a Meta ativará um Centro de Operações de Produtos de Integridade (IPOC) na semana anterior às Eleições Gerais de 2026 no Brasil. Este centro centralizado une especialistas multifuncionais de inteligência de ameaças, ciência de dados, produto, engenharia, operações, políticas e assessoria jurídica, facilitando avaliações rápidas e decisões de aplicação de regras em tempo real sem os atrasos dos caminhos tradicionais de escalação entre múltiplas equipes. Enquanto ativo, o centro está equipado para identificar, avaliar e mitigar materiais ilegais ou que violem políticas por meio de revisões manuais aceleradas, controles programáticos de distribuição e um canal de comunicação direta com o Tribunal Superior Eleitoral (TSE).

Além dos procedimentos operacionais padrão, a Meta mantém uma estrutura de contingência adaptada para períodos eleitorais para abordar situações em que os vetores de risco ultrapassam as métricas de aplicação de regras de linha de base. Este protocolo institui fluxos de trabalho pré-autorizados, implantação escalável de recursos e parâmetros delegados de tomada de decisão, permitindo execução imediata sob pressão operacional sem buscar aprovações adicionais.

A estrutura de contingência visa vários cenários específicos: aumentos acentuados em conteúdo problemático que sobrecarregam a filtragem automatizada; o surgimento de métodos adversariais originais não detectados por algoritmos padrão; esforços de influência organizados e multiplataforma que exigem identificação e contenção rápidas; e a disseminação rápida de conteúdo excepcionalmente grave em dias de eleição que requer mitigação imediata.

2.1.5 Prevenção de Envio de Mensagens em Massa

Mecanismos para identificar padrões de atividade anômalos ou automatizados que indiquem o uso de ferramentas externas de envio em massa:
O WhatsApp utiliza mecanismos de detecção automatizada e classificadores de aprendizado de máquina projetados para identificar comportamentos suspeitos desde o momento do cadastro da conta, passando pela criação de grupos até o envio de

mensagens. Informações da conta — incluindo dados básicos do perfil da conta, do grupo ou da comunidade — também são analisadas.

Os sistemas analisam sinais comportamentais e podem banir automaticamente contas que realizam envios de mensagens em massa, utilizam ferramentas automatizadas ou apresentam sinais de comprometimento da conta para contatar um grande número de usuários. O WhatsApp pode restringir a capacidade de contas suspeitas iniciarem conversas com usuários.

As proteções estruturais que limitam a capacidade de envio em massa incluem:

- Verificação do número de telefone via SMS ou chamada no momento do cadastro
- Verificação em duas etapas (proteção opcional por PIN)
- O WhatsApp não permite que usuários pesquisem e encontrem outros usuários individuais a partir de uma lista de contas; é necessário ter o número de telefone ou compartilhar a participação em um grupo para enviar mensagens
- Limites para o encaminhamento de mensagens e para o tamanho dos grupos (máximo de 1.024 usuários)
- Cartões de contexto de "Experiência com a Primeira Mensagem", que fornecem informações sobre remetentes desconhecidos antes de o usuário responder, oferecendo opções para bloquear ou denunciar
- Cartões de "Experiência de Adição a Grupos", exibidos quando um usuário é adicionado a um grupo por alguém que não está em sua lista de contatos

Os sistemas do WhatsApp conseguem identificar e banir a maioria das contas infratoras envolvidas em fraudes e golpes antes mesmo de receberem qualquer denúncia de usuários contra elas. Nossas ações de aplicação de regras contra contas envolvidas em golpes utilizam detecção automatizada por meio de aprendizado de máquina para analisar sinais de metadados (comportamento da conta, padrões de rede, anomalias no cadastro) e denúncias de usuários. Nossos sistemas empregam técnicas avançadas de aprendizado de máquina para identificar padrões associados a comportamentos fraudulentos, como *phishing* e envio de mensagens em massa.

Seção 3: Proibição de Publicidade Política

Anúncios podem ser veiculados no Status e nos Canais do WhatsApp, separadamente de chats pessoais e chamadas. Todavia, anúncios com Temas Sociais, Eleições e Política (SIEP) não são atualmente para serem colocados nas superfícies de publicidade do WhatsApp. Mais informações sobre anúncios relacionados a Temas Sociais, Eleições e Política podem ser encontradas na Seção 3 do relatório da Meta (Facebook, Instagram e Threads).

Seção 4: Abordagem do WhatsApp para Inteligência Artificial

O WhatsApp detectará metadados seguindo padrões do setor em atualizações de Canais do WhatsApp e rotulará conteúdo gerado ou editado por IA que apareça em atualizações de Canais do WhatsApp, além de oferecer autodeclaração aos usuários no Brasil.

Conteúdo Orgânico (Canais do WhatsApp)
Forneceremos oportunidades de autodeclaração para que os usuários declarem quando seu conteúdo foi criado ou modificado usando IA. Usuários que fazem upload de conteúdo para os Canais do WhatsApp podem divulgar voluntariamente que seu conteúdo foi gerado ou editado com ferramentas de IA. Esta autodeclaração aciona a aplicação de um rótulo no conteúdo.

Conteúdo de Terceiros (3P) — Ferramentas de IA Externas (Canais do WhatsApp)
Usaremos padrões do setor para identificar conteúdo gerado por IA produzido por ferramentas de terceiros:

- A detecção depende da identificação de indicadores de imagem de IA padrão do setor, particularmente credenciais de conteúdo C2PA e metadados IPTC, que outras empresas incluem em suas saídas geradas por IA.
- Quando esses sinais são detectados em conteúdo carregado, o sistema aplica um rótulo identificando o conteúdo como IA. Este é um esforço contínuo e em evolução, à medida que a indústria trabalha coletivamente para desenvolver padrões comuns para proveniência de conteúdo.

Seção 5: Proteção de Dados Pessoais

Esta seção apresenta as informações exigidas pelo Art. 21 da Portaria nº 463/2026/TSE, que especifica os elementos que o plano de conformidade deve conter no que tange aos deveres de proteção de dados pessoais relacionados especificamente às obrigações estabelecidas nos arts. 33-A e 33-B da Resolução nº 23.610/2019/TSE.

5.1 Mecanismos para Informar os Titulares de Dados Sobre Tratamento para Publicidade

Ferramenta "Por que você está vendo este anúncio?"

No WhatsApp, anúncios são veiculados somente na Aba Atualizações (Status e Canais). Os usuários podem tocar ou clicar em "Por que você está vendo este anúncio" para visualizar alguns dos fatores que consideramos ao mostrar um determinado anúncio — por exemplo, quando um anunciante deseja mostrar seu anúncio para pessoas em um país que corresponde à localização do usuário. Os usuários também podem gerenciar suas preferências de anúncios diretamente dentro do aplicativo. Conforme descrito na Seção 3 e 5.2 abaixo, anúncios sobre eleições e política não estão atualmente disponíveis nas superfícies de publicidade do WhatsApp.

Exercício dos Direitos de Privacidade da LGPD
Os Direitos de Privacidade e as formas de exercê-los estão informados na Política de Privacidade do WhatsApp. Além disso, a Política de Privacidade do WhatsApp oferece canais de contato disponíveis para usuários e não usuários enviarem perguntas relacionadas à privacidade e contatarem o Encarregado de Proteção de Dados do WhatsApp. Consulte:

- Canal de Contato de Privacidade: https://www.whatsapp.com/contact/forms/915483389072145
- Encarregado de Proteção de Dados: http://whatsapp.com/contact/forms/2958559137738747/

Os canais de Consultas de Privacidade e Encarregado de Proteção de Dados são gerenciados pela equipe de Resposta aos Direitos dos Usuários, com apoio da equipe de DPO.

5.2 Mecanismos que Garantem a Limitação de Finalidade

O WhatsApp veicula publicidade no Brasil através do sistema de publicidade da Meta. No entanto, anúncios em Categorias Especiais de Anúncios — incluindo anúncios SIEP — não estão atualmente disponíveis no WhatsApp.

Os princípios de limitação de finalidade, necessidade, adequação e não discriminação para publicidade eleitoral são, portanto, estruturalmente satisfeitos: nenhum perfilamento ou tratamento de dados pessoais para a entrega de anúncios eleitorais ocorre no WhatsApp no Brasil, pois tais anúncios são inelegíveis para as superfícies de publicidade da plataforma.

5.3 Procedimentos de Segurança

Programa Abrangente de Segurança da Informação

Mantemos um programa de segurança em toda a empresa que consiste em um sistema integrado de políticas, padrões e diretrizes, organizados por domínios, objetivos e controles.

Essas políticas regem o acesso e o uso de dados de usuários e se aplicam a todo o pessoal da Meta no Facebook, Instagram, Threads e WhatsApp.

Nossas políticas, padrões e controles levam em consideração padrões relevantes e aplicáveis da indústria (como NIST CSF, PCI-DSS e ISO 27001) e são projetados e dimensionados para capturar a abordagem da Meta para segurança da informação com base na natureza específica de nossos sistemas de informação e necessidades de negócios. Isso garante uma estrutura que atende a medidas técnicas, organizacionais e operacionais para garantir a proteção de dados pessoais de acordo com as melhores práticas internacionais.

Controles Técnicos de Segurança
Mantemos um Programa Abrangente de Segurança da Informação projetado para garantir a confidencialidade, integridade e disponibilidade de dados pessoais processados em nossos sistemas e redes. Esses controles incluem:

- Controles de Acesso: Implementamos controles de acesso centralizados e automatizados baseados em funções que limitam a capacidade de modificar dados pessoais apenas a sistemas e usuários autorizados. O acesso é concedido com base na função de trabalho e necessidade de negócios.
- Criptografia: Os dados são criptografados em trânsito usando protocolos padrão do setor (por exemplo, TLS) como parte de nossa arquitetura de segurança padrão, protegendo dados pessoais enquanto se movem através de nossos sistemas e redes.
- Detecção e Monitoramento: Implantamos mecanismos de detecção, trilhas de auditoria e monitoramento contínuo para identificar e responder a tentativas de acesso não autorizado ou atividade anômala. Esses controles permitem a identificação tempestiva de potenciais ameaças à segurança dos dados pessoais.
- Práticas de Desenvolvimento Seguro: Mantemos estruturas de desenvolvimento seguro integradas ao nosso ciclo de vida de desenvolvimento de produtos, incluindo revisões de segurança e testes, para garantir que recursos que processam dados de usuários sejam projetados e construídos com princípios de segurança por design.
- Segurança Física: Os dados são armazenados em data centers da Meta com medidas padronizadas de segurança física, incluindo controles de acesso por cartão, vigilância por CCTV, pessoal de segurança no local,

controles ambientais, scanners biométricos para acesso a áreas seguras e destruição segura de mídias de armazenamento descomissionadas.

- Segmentação de Dados: Os dados pessoais são segmentados por sensibilidade e finalidade através de medidas técnicas e de política que definem o manuseio adequado de dados ao longo de seu ciclo de vida, garantindo que os dados pessoais estejam sujeitos a proteções apropriadas proporcionais ao seu perfil de risco.

Programa de Gestão de Resposta a Incidentes

Nosso programa de gestão de resposta a incidentes é uma estrutura abrangente projetada para identificar, mitigar, avaliar e remediar incidentes de privacidade. Utilizamos uma variedade de ferramentas e processos manuais e automatizados para detectar potenciais incidentes de privacidade, incluindo monitoramento de sistemas, relatórios internos, relatórios de terceiros e um programa de recompensa por bugs. Uma vez que um incidente é detectado, ele passa por um processo de triagem para avaliar a gravidade e o impacto e, em situações apropriadas, é seguido por um processo de investigação e remediação. Atualizamos e melhoramos continuamente nossos processos de resposta a incidentes para abordar riscos em evolução. O programa inclui análises retrospectivas (*post-mortems*) e ações de acompanhamento, conforme apropriado, para prevenir futuros incidentes e melhorar a estratégia geral de resposta.

Notificação às Autoridades de Supervisão

Quando ocorre um incidente de dados qualificado, notificamos as autoridades de proteção de dados relevantes de acordo com os requisitos legais aplicáveis.

Notificação aos Titulares de Dados Afetados

Nossa abordagem para notificar os titulares de dados após um incidente é guiada pelos requisitos legais aplicáveis e uma avaliação de risco detalhada para determinar o impacto potencial aos direitos e liberdades dos usuários.

Abordagem de avaliação de risco

Avaliamos fatores incluindo a natureza e sensibilidade dos dados envolvidos, evidência de ações adicionais tomadas sobre ou com os dados, se categorias especiais ou sensíveis de dados foram acessadas, e o escopo e status de contenção do incidente. Nos termos da lei aplicável, consideramos esses fatores e os riscos de danos relevantes ao determinar notificar formalmente os titulares de dados.

Medidas proativas de proteção do usuário

Mesmo quando a notificação formal aos titulares de dados não é legalmente exigida, tomamos medidas proativas para proteger os usuários afetados.

5.4 Mecanismo de Consentimento

Conforme descrito nas Seções 3 e 5.2, anúncios SIEP não estão atualmente disponíveis nas superfícies de publicidade do WhatsApp.

Assim, não há tratamento de dados pessoais sensíveis (conforme definido no Art. 5º, II, da LGPD) no serviço para fins de anúncio eleitoral; assim, a exigência de consentimento específico e expresso prevista no Art. 33-B, § 1º, é estruturalmente inaplicável.

Seção 6: Canais de Denúncia Locais

Esta seção apresenta as informações exigidas pelos artigos 12 e 17 da Portaria nº 463/2026/TSE, descrevendo a estrutura e a capacidade operacional dedicadas ao recebimento, à triagem e ao tratamento de solicitações provenientes de órgãos reguladores, usuários, candidatos e partidos políticos, referentes a conteúdos e comportamentos sujeitos à Resolução nº 23.610/2019/TSE(arts. 9º-E, I-VII; 28, §§ 4º e 4º-B; e 34, II), incluindo os respectivos fluxos de triagem e os prazos estimados de resposta.

6.1 Canal de Denúncia do Tribunal Superior Eleitoral

O WhatsApp mantém um canal de denúncia com o Tribunal Superior Eleitoral para receber e analisar denúncias sobre contas suspeitas de realizar envio de mensagens em massa. O TSE informará ao WhatsApp os números de telefone das contas de WhatsApp, com base em denúncias de usuários, que são suspeitas de envolvimento em disseminação em massa de conteúdo eleitoral.

6.2 Canal de Denúncia de Usuários – Conteúdo localmente ilegal

Um formulário dedicado hospedado no site do WhatsApp estará disponível para usuários e não usuários para denunciar conteúdo nos Canais do WhatsApp. Este formulário será separado do canal que atende Facebook, Instagram e Threads.

Ao enviar, a parte solicitante recebe uma confirmação automática de recebimento contendo um número de referência do caso, que pode ser usado em correspondências subsequentes sobre o assunto. Respostas substantivas são retornadas à parte que enviou através do mesmo canal.

As denúncias serão inicialmente revisadas por nossas equipes de revisão humana em escala em relação às nossas Políticas de Canais do WhatsApp. Após a conclusão da revisão de política, o conteúdo denunciado é então revisado para conteúdo localmente ilegal, onde é encaminhado a uma equipe de revisão especializada para avaliação em relação à lei local.

Mantemos um processo de confirmação de recebimento de denúncia, via e-mail, para informar usuários e não usuários que suas denúncias foram recebidas e sobre o resultado final de nossa investigação.

6.3 Canal de Denúncia para Candidatos e Partidos

Um caminho dedicado no canal de denúncias do WhatsApp estará disponível para candidatos, partidos, federações e coligações para denunciar conteúdo nos Canais do WhatsApp. Este canal aproveitará as informações fornecidas por candidatos, partidos políticos, federações e coligações ao Tribunal Superior Eleitoral, de modo que a Meta poderá abrir o canal após a publicação da lista oficial de candidatos depois de 16 de agosto.

Ao enviar, a parte solicitante recebe uma confirmação automática de recebimento contendo um número de referência do caso, que pode ser usado em correspondências subsequentes sobre a ordem. Respostas substantivas são retornadas à parte que enviou através do mesmo canal.

As denúncias serão inicialmente revisadas por nossas equipes de revisão humana em escala em relação às nossas Políticas de Canais do WhatsApp. Após a conclusão da revisão de política, o conteúdo denunciado é então revisado para conteúdo localmente ilegal, onde é encaminhado a uma equipe de revisão especializada para avaliação em relação à lei local.

As denúncias enviadas através do canal de candidatos políticos são revisadas de forma contínua. Embora os tempos de resposta possam variar dependendo do volume de envios e da complexidade das questões levantadas, nos esforçamos para revisar todas as denúncias dentro de 24 horas. Estamos comprometidos em revisar todas as denúncias de maneira tempestiva e de acordo com nossas obrigações.

Mantemos um processo de confirmação de recebimento de denúncia, via e-mail, para informar usuários e não usuários que suas denúncias foram recebidas e sobre o resultado final de nossa investigação.

Transparência: O volume agregado de denúncias recebidas por meio desses canais será informado pelo WhatsApp no relatório consolidado a ser apresentado ao TSE até 19 de dezembro de 2026 (Art. 28).

Seção 7: Cumprimento de Ordens Judiciais

Esta seção aborda os requisitos estabelecidos no Artigo 12 da Portaria nº 463/2026/TSE, que especifica os elementos que o plano de conformidade deve conter com relação aos deveres de cumprir ordens e determinações emitidas pela Justiça Eleitoral.

Esses deveres derivam das obrigações substantivas estabelecidas nos Artigos 9-D, §5, 32, 36, 38, §§4º e 6º, 39 e 40 da Resolução nº 23.610/2019/TSE, conforme descrito na Seção 7 do relatório Meta.

7.1 Estrutura de Assessoria Jurídica Externa

O WhatsApp será representado nas demandas propostas junto à Justiça Eleitoral por advogados integrantes de um escritório de advocacia com sede no Brasil, que atuará como o principal ponto de contato local para receber, fazer a triagem e processar ordens e determinações judiciais.

A cada ciclo eleitoral, o escritório monta uma estrutura exclusiva para lidar com a demanda do contencioso judicial eleitoral. Em 2026, essa estrutura é composta por 112 advogados e 59 funcionários de apoio jurídico, organizados em dois turnos operacionais para garantir a capacidade de atendimento contínuo durante o dia.

Essa estrutura é responsável por: (i) receber ordens e determinações judiciais emitidas pela Justiça Eleitoral por meio do canal exclusivo aberto em cumprimento ao Artigo 10 da Resolução nº 23.610/2019/TSE; (ii) encaminhá-las às equipes internas da Meta para análise técnica e execução; (iii) receber as respostas das equipes internas; e (iv) protocolar as petições correspondentes perante o sistema das cortes eleitorais e gerenciar todo o ciclo de vida processual perante os tribunais em todo o país.

A capacidade desta estrutura é proporcional ao volume e à complexidade das demandas historicamente recebidas durante os períodos eleitorais, e é compartilhada por todas as Empresas da Meta como um todo.

7.2 Estrutura e Capacidade Operacional Interna

Para as eleições gerais brasileiras de outubro de 2026, o WhatsApp opera uma estrutura multifuncional dedicada para o período eleitoral, reunindo as equipes responsáveis pela aplicação de todas as políticas descritas acima.

Não existem critérios padronizados para determinar o número de revisores necessários para uma determinada eleição. A capacidade é dimensionada antes de cada período eleitoral a partir de uma previsão de volume e ajustada de acordo com a avaliação de risco para aquela eleição, com recursos adicionais alocados onde riscos específicos emergem. A capacidade é entregue através de uma estrutura de revisão compartilhada que também atende outras jurisdições e tipos de casos; assuntos eleitorais são priorizados dentro dessa estrutura compartilhada pela duração do período, e a cobertura é estendida em torno de cada turno de votação.

Recebimento, triagem e execução de ordens judiciais
O WhatsApp mantém canais online dedicados para solicitações judiciais que possam exigir a remoção de contas e a produção de dados. Advogados externos enviam essas solicitações por meio de canais de recebimento dedicados para o Brasil, onde a solicitação e a ordem judicial subjacente são fornecidas juntas em um único envio estruturado, identificando as entidades do WhatsApp em questão, o prazo estabelecido pelo tribunal e o serviço afetado. Em conformidade com a legislação brasileira aplicável, o alvo específico deve ser identificado na própria ordem.

Classificação no Recebimento: As ordens são classificadas no momento do envio pela natureza da ordem (restrição, restauração ou outra) e pela superfície afetada (contas, grupos, canais ou identificadores de conteúdo). Quando uma ordem faz referência a identificadores de conteúdo, o envio do documento judicial de origem é obrigatório, para que os identificadores possam ser validados em relação à ordem emissora antes que qualquer ação seja tomada.

Direcionamento e Escalonamento: Cada envio gera um registro de processo direcionado à equipe responsável pela execução.

Cada solicitação é analisada em relação aos nossos Termos de Serviço e à legislação local. Quando uma conta ou grupo viola nossas políticas ou uma ordem válida exige o cumprimento, nós a cumprimos. Solicitações referentes a contas ou grupos de interesse público significativo e à imprensa são escalonadas para análise adicional de especialistas e de políticas antes que qualquer ação seja tomada. As solicitações sobre os Canais do WhatsApp são analisadas com base nas diretrizes de Canais e na legislação local. Quando o conteúdo viola nossas políticas de forma independente, nós o removemos. Quando isso não ocorre, e uma ordem válida assim o exige, tomaremos medidas adicionais para cumpri-la sob nossa obrigação legal

7.3 Procedimento de Escalonamento Interno

Em casos de ordens judiciais que exijam dados ou informações adicionais para viabilizar o cumprimento, o procedimento interno adotado consiste em quatro etapas escalonadas:

1. Verificação imediata da existência e suficiência de identificadores válidos na ordem ou na documentação anexa recebida.
2. Resposta tempestiva ao Tribunal, por meio de petição protocolada nos autos, indicando objetiva e especificamente quais dados ou elementos são necessários para viabilizar o cumprimento;
3. Uma nova verificação assim que os identificadores suplementares forem recebidos; e
4. Uma vez confirmada a suficiência das informações, encaminhamento da ordem às equipes internas para cumprimento.

O WhatsApp cumpre ordens judiciais e solicitações legais válidas, de acordo com as leis locais.