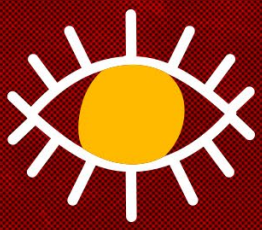


DE OLHO NOS DADOS!

POLÍTICAS E MEDIDAS DE SEGURANÇA NAS PLATAFORMAS DIGITAIS

PARTE 2

OBSERVATÓRIO DA DESINFORMAÇÃO

SLEEPING GIANTS BRASIL

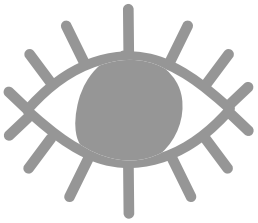
FICHA TÉCNICA

RELATÓRIO:
De olho nos dados! Políticas de Transparência, Bibliotecas de Anúncios e Acesso a dados nas Plataformas Digitais
(Maio, 2024)

DESENVOLVIMENTO:
Observatório da Desinformação - Diretoria de Pesquisa
Sleeping Giants Brasil
Av. Guido Caloi, 1000 – Bl. 5 – 4º. Andar - Jd. São Luis
São Paulo – SP
CEP 05.802-140

contato@sleepinggiantsbrasil.com
www.sleepinggiantsbrasil.com

PROJETO GRÁFICO
Sleeping Giants Brasil



DE OLHO NOS DADOS!

POLÍTICAS E MEDIDAS DE SEGURANÇA NAS PLATAFORMAS DIGITAIS

OBSERVATÓRIO DA DES INFOR MAÇÃO

SLEEPING GIANTS BRASIL

RESUMO EXECUTIVO DOS RESULTADOS

SLEEPING GIANTS BRASIL

1. Políticas Gerais de Segurança - as plataformas analisadas possuem políticas de segurança nos tópicos destrinchados. Destacamos que, no campo de política dedicadas a contas de autores de ataques ou atos violentos, o Youtube (Google) menciona esses atores dentro da sua Política de Organizações Extremistas, mas não há especificações de consequências para **as contas de usuário** para autores de ataques, como exclusão e/ou suspensão de suas contas.
2. No campo de definição de Grupos Protegidos, destaca-se que, diferente das outras Plataformas, a Meta considera que **refugiados, migrantes, imigrantes e pessoas que buscam asilo** têm proteção de ataques mais graves. Entendemos que o TikTok, a Google e o X devem também considerar esses grupos dentro de seus grupos protegidos, considerando a vulnerabilidade adicional em que se encontram.
3. Indicamos que o X inclua **desumanização** dentro da sua definição de discurso violento, algo feito pelas outras plataformas. Também indicamos que a Meta e o X incluam **a promoção de teorias da conspiração contra grupos protegidos nas suas políticas de segurança, dentro de Política de Desinformação** (caso de TikTok) **ou de Discurso de Ódio** (caso de Tiktok e Google) Tomamos como exemplo a política da Google, que caracteriza afirmando que pessoas ou grupos são maus, corruptos ou maliciosos com base no status de grupo protegido.
4. No mesmo sentido, o TikTok e o Google contempla, dentro de suas políticas, a oposição demarcada contra a **negação ou minimização da dimensão de eventos históricos bem documentados que prejudicaram grupos protegidos**. O X e a Meta também proíbem a negação de eventos violentos, **mas não a minimização**, que também é uma forma de relativização e negacionismo. Indicamos que isso seja contemplado.
5. A Google - Youtube **não possui política de comportamento inautêntico coordenado. Indicamos fortemente que isso passe a ser contemplado em suas políticas.**
6. Nem o X, nem a Google noticiaram um conjunto de medidas específicas em relação às eleições municipais brasileiras, enquanto a Google anunciou medidas para as eleições na União Europeia deste ano. Nota-se que a Meta é a única das plataformas analisadas que seguirá permitindo anúncios políticos: é de interesse monitorar a adequação da revisão de conteúdo da Meta ao atualmente exigido pelo TSE.

7. A Google e o TikTok não possuem políticas de protocolo de crise. Também recomendamos a definição e divulgação de seus protocolos.

8. Dentro das Políticas de Desinformação, um dos tópicos indicados pelo TikTok é a desinformação **climática**, que não é mencionada nominalmente nas políticas de desinformação das demais plataformas. Entende-se que as demais plataformas podem incorporar essa questão como agenda prioritária, dado o contexto de emergência climática no qual vivemos.

9. Meta e Youtube-Google não possuem nenhuma política específica para as contas políticas ou de mídia afiliadas ao Estado. Ainda, o X tem a política de mídia, mas tem sido noticiado uma não-adequação entre a existência da política e sinalização das contas na plataforma. Indicamos que seja desenvolvida uma política para contas políticas, nos diferentes níveis de governo.

10. Apenas o TikTok possui **portal de fácil acesso para sinalizadores confiáveis segundo o Digital Services Act da UE.**

11. O TikTok tem um Modo Restrito com proteções adicionais para crianças menores de 13 anos **disponível nos Estados Unidos** e um disponível nos Estados Unidos, e um Portal de Segurança Online na Europa[1], com informações e recursos de transparência, etc.

1 https://www.tiktok.com/euonlinesafety/en/

INTRODUÇÃO

A internet não é terra de ninguém. Para além das legislações existentes que regulam crimes cibernéticos, é cada vez mais urgente a existência de políticas de segurança e mecanismos nas plataformas digitais para a garantia da segurança dos usuários, contemplando o combate à desinformação e ao discurso de ódio. Com as eleições municipais que se aproximam no Brasil, isso se torna especialmente relevante. A sociedade civil tem se mobilizado junto às instituições públicas na cobrança pela implementação nas plataformas digitais de políticas e práticas de segurança. Essa pesquisa trata disso.

O debate sobre regulação das Big Techs reflete a necessidade da colaboração entre o campo cívico, as instituições públicas e as empresas que detêm o monopólio dos dados e do espaço digital. Nesse sentido, é preciso considerar quais mecanismos têm sido implementados pelas plataformas para combater a circulação de desinformação e o discurso de ódio, garantindo a defesa dos direitos e da dignidade humana, e a segurança das disputas eleitorais quando é vigente o uso das plataformas de forma perniciosa. Interessa-nos especialmente a consideração de quais ferramentas estão disponíveis internacionalmente e não estão presentes no cenário brasileiro.

As eleições de 2024 se darão nas maiores democracias do mundo, como nos Estados Unidos, Índia, México, Indonésia e na União Europeia, tal como em países em situação de conflito e declínio democrático, como a Etiópia, o Egito e a Tunísia. **As eleições municipais brasileiras acontecerão entre os dias 6 e 27 de outubro de 2024** e especialistas já têm alertado sobre os riscos da desinformação para o ciclo eleitoral que se aproxima. Desse modo, é preciso que sejam mapeadas as políticas e mecanismos de segurança e de combate à desinformação e ao discurso de ódio para garantir a saúde dos processos democráticos.

O presente relatório “De olho nos dados! Políticas de Segurança e Combate ao Discurso de Ódio e Desinformação” é a segunda parte do projeto DE OLHO NOS DADOS, precedido pelo relatório “De olho nos dados! Políticas de transparência, bibliotecas de anúncio e acesso a dados nas plataformas digitais”. Tratamos aqui do levantamento de políticas de segurança, mecanismos de segurança - tais como protocolos de crise -, parcerias com checadores e demais ferramentas implementadas atualmente nas plataformas X (antigo Twitter), a Meta, a Google (focalizado no YouTube) e o TikTok. Fazemos, assim, a continuação da sistematização do panorama comparativo internacional da análise de quatro relevantes plataformas, com foco na comparação entre os mecanismos e políticas existentes na União Europeia e no Brasil. Objetivamos, assim, mapear as disparidades existentes de forma a pleitear o melhoramento da situação brasileira.

METODOLOGIA

Em 2024, mais de 2 bilhões de pessoas irão às urnas em mais de 65 eleições em todo o mundo. No Brasil, as eleições municipais acontecem em outubro e as questões referentes à desinformação e o discurso de ódio nas redes sociais já é a principal pauta do debate público sobre esse processo. Como destacaram ativistas do movimento Year of Democracy[2], há grandes chances do Brasil ter novamente nestas eleições campanhas de desinformação e discursos de ódio sendo disseminadas, sobretudo, nas redes sociais. Os efeitos dos ataques ocorridos em 8 de janeiro de 2023 à Praça dos Três Poderes em Brasília não foram o bastante para as Big Techs desenvolverem mecanismos de segurança que possam frear essa onda de violência.

Para a presente pesquisa, nosso primeiro ponto é a definição de conceitos de **desinformação, discurso de ódio e sobre as políticas e mecanismos de segurança** e contra os mesmos. Em seguida, trazemos uma breve análise do tema a partir de relatórios de organizações nacionais e internacionais que trabalham com a temática. A partir disso, mergulhamos no levantamento de dados sobre as plataformas escolhidas (**X, Meta, TikTok, YouTube - Google**).

O segundo ponto da pesquisa consiste na aplicação da análise de redes. A análise de redes é uma metodologia que visa compreender a estrutura, os padrões e as dinâmicas das interações entre entidades (como indivíduos, organizações, países, etc.) representadas por meio de conexões. Os passos metodológicos a partir da análise de redes seguirá:

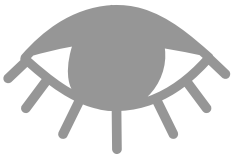
1. **Coleta de Dados** sobre as políticas e mecanismos de segurança nas plataformas: X (Twitter); Meta; Google - Youtube; TikTok;
2. **Representação de Rede**, com transformação dos dados qualitativos em tabelas para análise comparativa da situação entre as plataformas;
3. **Análise descritiva**, interpretações e implicações dos dados encontrados.

1 https://yearofdemocracy.org/case-study/brazil-municipal-elections-have-big-tech-companies-learnt-anything-from-the-january-8th-attacks/

A partir da sistematização dos dados, a análise descritiva comparativa aqui disposta segue os seguintes eixos de análise:

1. Políticas de Segurança Gerais
2. Políticas de Mídia Sintética e Manipulada
3. Políticas de Integridade Cívica e/ou Eleitoral
4. Políticas contra Desinformação
5. Contas políticas e/ou afiliadas ao Estado
6. Mecanismos de Detecção e correção de conteúdo nocivo
7. Trusted Flaggers / Sinalizadores Confiáveis
8. Medidas/Ferramentas de segurança específicas por plataforma

JUSTIFICATIVA



O Sleeping Giants Brasil tem o papel, enquanto organização da sociedade civil e movimento de consumidores que fazem uso das plataformas digitais das mais variadas formas, atuando em uma política de ***desmonetização* de conteúdos com desinformação e discurso de ódio nas redes sociais**, como uma das alternativas concretas e efetivas de combate a esse problema social, além de conscientizar e responsabilizar as Big Techs por políticas que assegurem a confiabilidade da informação no território brasileiro. Tanto a desinformação quanto o discurso de ódio representam sérias ameaças para a sociedade, pois podem minar a coesão social, prejudicar a democracia, fomentar o preconceito e causar danos às comunidades. Combatê-los requer uma abordagem multifacetada que envolve a educação pública, o fortalecimento da alfabetização midiática, o desenvolvimento de políticas e regulamentações eficazes, o engajamento da sociedade civil e o apoio de plataformas digitais na moderação de conteúdo e na promoção de um ambiente online seguro e inclusivo. Os **mecanismos de revisão e denúncia de conteúdo** e as **políticas de segurança** são justamente a incorporação desse combate dentro das plataformas digitais.

DIGITAL SERVICES ACT
O MODELO DA UNIÃO EUROPEIA

O modelo de regulamentação mais rigoroso aplicado sobre as Big Techs atualmente é o *Digital Services Act* (DSA, ou Regulamento de Serviços Digitais). De acordo com o relatório "Regula Big Techs: Recomendações do Sleeping Giants Brasil para uma proposta de regulação que mire modelos de negócio e não no conteúdo" (SGBR, 2023), o DSA é uma emenda à diretiva de comércio eletrônico (Diretiva 2000/31/EC), que inaugurou obrigações básicas para todos os provedores de Internet, especialmente de responsabilidade, transparência e processo de moderação. A legislação criou uma regulação assimétrica com obrigações ainda mais específicas para plataformas de diferentes tamanhos, as VLOPs - *very large online platforms* (redes sociais/plataformas de grande dimensão) e as VLSEs - *very large search engines* (ferramentas de busca de grande dimensão). Diante do quadro comparativo entre o DSA e o PL 2630, elaborada pelo SGBR no referido relatório, temos que:

No caso de Regras Aplicáveis Aos Provedores De Redes Sociais, o DSA estabelece:

- Sistema interno de gestão de reclamações dos destinatários e entidades notificantes.
- Direito de solucionar conflito em órgão extrajudicial.
- Sinalizadores de confiança e priorização de notificações enviadas por esses atores. Proteção da segurança do menor, inclusive com vedação à publicidade.
- Rastreabilidade de usuários que coloquem produtos ou serviços disponíveis em marketplaces.

Nas Regras Aplicáveis A Redes Sociais E Ferramentas De Busca:

- Elaboração de relatórios de avaliação de riscos.
- Medidas de atenuação de riscos.
- Mecanismo de respostas e protocolos de crise.
- Transparência de impulsionamento eleitoral, com obrigações de fornecimento de informações específicas de fácil acesso e integráveis com máquinas.
- Informações sobre recursos humanos empenhados na atividade de moderação, com indicação de qualificações e conhecimentos linguísticos

Dentre as **Indicações do SGBR**, destacamos em primeiro lugar a responsabilidade civil objetiva para impulsionamento e

publicidade. O Artigo 19 do Marco Civil da Internet estabelece a imunidade dos provedores por conteúdos de terceiro, contudo, há a exceção taxativa, em seu Artigo 21, impondo responsabilidade subsidiária aos provedores caso não ajam diligentemente após tomarem conhecimento de conteúdos relativos à pornografia de vingança e pedofilia. O SGBR sugere a criação de uma nova hipótese de exceção ao regime de responsabilidade, impedindo que os provedores gozem de imunidade por conteúdos ilícitos que permitam impulsionar ou patrocinar. Nota-se que **o conteúdo apenas tem seu impulsionamento/patrocínio realizado após autorização do provedor o que, em alguns casos, pode superar 48 horas**. Assim, por desempenhar um papel ativo na amplificação do conteúdo, não consideramos razoável que os provedores gozem, nessas circunstâncias, da mesma imunidade conferida aos conteúdos orgânicos de terceiros.

Ainda, recomenda-se a obrigação de produção de relatórios e medidas de mitigação de riscos, além de mecanismos e protocolos de crise. Ao criar obrigação semelhante, a legislação europeia estabeleceu em seu Art. 34 que os provedores mencionados devem inserir necessariamente, em seu processo de avaliação, os sistemas algorítmicos aptos a produzir e ampliar tais riscos. Os relatórios devem incluir aqueles riscos específicos que seus serviços impõem à difusão de conteúdos ilegais, ao exercício de direitos fundamentais, aos processos eleitorais, à segurança e à saúde públicas. Também indica-se a implementação de sinalizadores de confiança, plataforma de gestão de notificações/ reclamações e dever de fundamentação. A sociedade é parte fundamental para a mitigação dos riscos e é a ela que devem ser conferidas ferramentas necessárias e eficientes para o aprimoramento das precárias atividades de moderação das plataformas. Recomenda-se que seja incorporado na legislação o dever de cuidado de que provedores criem um sistema de gestão de reclamações e notificações, que permitam ao usuário acompanhar seus pedidos de revisão de atividades de moderação com impactos sobre seus direitos fundamentais. Ainda, defende-se que qualquer usuário, indivíduo ou entidade da sociedade civil tenha disponibilizado canais necessários à realização de denúncias sobre conteúdos potencialmente inadequados e ilícitos nos serviços, bem como que possam acompanhar o desenvolvimento de sua denúncia através do sistema de gestão de reclamações.

Também recomenda-se estabelecer a obrigação legal de que as denúncias e pedidos de revisão sejam respondidos de maneira fundamentada, além da criação de prazos para o

oferecimento da resposta, o que admitimos que seja estabelecido no Código de Condutas. Recomenda-se que a condição de sinalizadores de confiança seja concedida às entidades que cumpram determinados requisitos de especialização e independência, sendo tal estatuto conferido a elas por meio de um órgão independente.

Nota-se que, segundo Giovanni De Gregorio (2022)[3], o **DSA foi concebido para abordar o processo de moderação de conteúdo. O caso da moderação de conteúdos é um exemplo paradigmático de governança de plataformas, em que as plataformas definem os padrões e regras que regem o fluxo de conteúdo online nos seus espaços digitais. A organização, incluindo a remoção de conteúdo, é aplicada diretamente por elas a partir de uma combinação de tecnologias algorítmicas e moderadores humanos**. Para o autor, a limitação do poder da plataforma advém de um novo sistema de auditoria independente e de aplicação pública que combina a cooperação a nível nacional e da UE. Cada Estado-Membro é obrigado a nomear um Coordenador de Serviços Digitais, uma autoridade independente que será responsável pela supervisão dos serviços intermediários e pela aplicação de multas. **Outro papel importante do DSA é o aumento da transparência no domínio da publicidade direcionada. A DSA reconhece que os sistemas de publicidade utilizados por plataformas de grande dimensão apresentam riscos específicos relacionados - mas não limitados - à propagação da desinformação, com impactos potencialmente de longo alcance em áreas tão diversas como a saúde pública, a segurança pública, o discurso civil, a participação política e a igualdade.**

O DSA também trata de circunstâncias extraordinárias que afetam a segurança pública e a saúde pública. Nestes casos, a Comissão tem o poder de confiar em protocolos de crise para coordenar uma resposta rápida, colectiva e transfronteiriça, especialmente quando as plataformas podem ser utilizadas indevidamente para a rápida propagação de conteúdos ilegais ou desinformação ou quando surge a necessidade de uma disseminação rápida de informações fiáveis (Art 37.º). Nestes casos, as plataformas de grande dimensão são obrigadas a adotar estes protocolos, mesmo que sejam aplicados apenas temporariamente e não levem as plataformas a uma obrigação geral e contínua de monitorização dos conteúdos em linha.

3 *How does the DSA contribute to platform governance and tackle disinformation? | Heinrich-Böll-Stiftung | Tel Aviv - Israel (boell.org)*

SEGURANÇA NAS PLATAFORMAS DIGITAIS

A segurança da informação nas redes sociais tem sido tema em um debate imprescindível sobre a integridade da informação, seus efeitos, causas e controvérsias no cenário público mundial. Para além de uma retórica individualista que coloca usuários como auto-gestores da sua segurança nas redes, é importante pressionar para que cada vez mais as Big Techs também tomem como papel fundamental para a manutenção de um estado democrático de direito, fornecer políticas de segurança que garantam a integridade mencionada.

Como mencionamos no primeiro relatório do projeto "DE OLHO NOS DADOS", políticas de privacidade, acesso e segurança são peças essenciais para proporcionar a integridade das informações que circulam nas redes sociais, bem como seus efeitos no cotidiano dos usuários. Segundo o informe do **Centro de Informação das Nações Unidas para o Brasil (UNIC Rio)**[4], a integridade da informação refere-se à precisão, consistência e confiabilidade da informação, sendo ameaçada pela desinformação, informação falsa e discurso de ódio. Neste sentido, é essencial que haja uma política de combate à desinformação, além do esforço coletivo entre as Big Techs para assegurar aos usuários políticas de transparência e segurança.

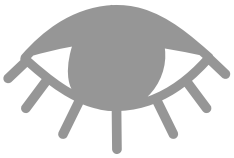
As políticas de segurança nas redes sociais podem incluir medidas como verificações de fatos, restrições ao compartilhamento de conteúdo enganoso ou prejudicial, remoção de contas fraudulentas ou maliciosas, e transparência na divulgação de propaganda política paga. Ao implementar essas políticas, as plataformas digitais podem ajudar a proteger a integridade do processo eleitoral, promovendo um ambiente online mais seguro e confiável para o debate público.

No contexto das redes sociais, as políticas de segurança desempenham um papel crucial na mitigação da disseminação de desinformação e discurso de ódio, especialmente em períodos eleitorais, como nas futuras eleições municipais de 2024 no Brasil. Quando as informações são manipuladas ou distorcidas nas plataformas digitais, isso pode influenciar negativamente o processo democrático, afetando a tomada de decisões dos eleitores e minando a confiança nas instituições democráticas.

Se por um lado a maioria das propostas de segurança nas redes sociais é entendida a partir de como os indivíduos devem se proteger, nossa ideia central é investigar e exigir que políticas de segurança eficientes sejam asseguradas pelas Big Techs aos seus usuários.

4 https://brasil.un.org/pt-br/249995-como-proteger-integridade-da-informa%C3%A7%C3%A3o-nas-plataformas-digitais-onu-publica-orienta%C3%A7%C3%B5es-do

1. POLÍTICAS GERAIS DE SEGURANÇA



Na Tabela 1, há o panorama geral das políticas de segurança das plataformas, contemplando a existência ou não das políticas e qual escopo elas contemplam. Os tópicos abrangidos são política de discurso de ódio, política de entidades ou organizações violentas e odiosas, política contra glorificação de violência, política contra abuso, política contra assédio, política contra conduta odiosa, política para contas de autores de ataques violentos, política para mídia sensível. Também abrangemos aqui como cada plataforma coloca a definição de **grupo protegido**.

Nota-se que, no campo de política dedicadas a contas de autores de ataques ou atos violentos, o Youtube (Google) menciona esses atores dentro da sua Política de organizações extremistas, mas não há especificações de consequências para as contas de usuário para autores. **Já no campo de Grupos Protegidos, destaca-se que, diferente das outras Plataformas, a Meta considera refugiados, migrantes, imigrantes e pessoas que buscam asilo têm proteção de ataques mais graves**. Além disso, há algumas proteções para aspectos como ocupação, quando são mencionados juntamente com uma característica protegida. Outro ponto interessante é a linguagem empregada para a caracterização do **discurso de ódio** quando temos que o X **não** elenca desumanização dentro dos parâmetros de discurso violento, diferente do TikTok, Meta e Youtube.

TABELA 1

POLÍTICAS DE SEGURANÇA				
Política/Plataforma	**X**	**TikTok**	**Meta**	**Google - Youtube**
Há política de discurso de ódio?	sim	sim	sim	sim
O que viola?	Ameaças violentas; Desejo de danos (como morte, doença, tragédias); Incitação de violência; Glorificação da violência;	Não permitimos nenhum discurso de ódio, comportamento de ódio ou promoção de ideologias de ódio (como a supremacia racial, a misoginia, a LGBTQ+fobia, o antissemitismo ou a islamofobia). Inclui discursos supremacistas, declarações conspiratórias contra grupos protegidos, desumanização, negação ou minimização a dimensão de eventos históricos bem documentados que prejudicaram grupos protegidos; Uso de Nomes e Pronomes Anteriores;	Enquandrado como "conteúdo questionável". Definido como ataque direto a pessoas, e não a conceitos e instituições, baseado no que chamamos de características protegidas; Definimos ataques como discursos violentos ou desumanizantes, estereótipos prejudiciais, declarações de inferioridade, expressões de desprezo, repulsa ou rejeição, xingamentos e incitações à exclusão ou segregação. Também proibimos o uso de estereótipos prejudiciais, que definimos como comparações desumanizantes historicamente usadas para atacar, intimidar ou excluir grupos específicos, e que muitas vezes estão ligadas à violência no meio físico. É proibido Discurso violento ou apoio de forma escrita ou visual, Generalizações afirmando inferioridade (por escrito ou visuais) sob as seguintes formas de Deficiências físicas, Deficiências mentais, Deficiências morais; Segregação e exclusão; etc.	Incentivar a violência contra pessoas ou grupos com base no status de grupo protegido. Incitações implícitas à violência como ameaças reais. Desumanizar indivíduos, chamando essas pessoas de sub-humanos ou fazendo comparações; exaltar ou promover a violência contra indivíduos ou grupos com base no status de grupo protegido; Usar insultos raciais, religiosos ou de qualquer outro tipo e estereótipos para incitar ou promover o ódio com base no status de grupo protegido de alguém. Afirmar que pessoas ou grupos são fisicamente ou mentalmente inferiores, deficientes ou doentes com base no status de grupo protegido. Promover a supremacia de ódio ao alegar a superioridade de um grupo sobre outros com status de grupo protegido para justificar violência, discriminação, segregação ou exclusão. Promover teorias da conspiração afirmando que pessoas ou grupos são maus, corruptos ou maliciosos com base no status de grupo protegido. Negar ou minimizar um conflito violento bem documentado ou a vitimização de pessoas causada por um evento desse tipo.Atacar os interesses românticos, emocionais e/ou sexuais de uma pessoa ou grupo.

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há política de entidades ou organizações violentas e odiosas?	sim	sim	sim	sim
O que viola?	É proibido se afiliar a entidades violentas e odiosas ou promover as atividades delas. É proibido Envolver-se ou promover atos violentos Recrutar, fornecer ou distribuir serviços (como mídia/ propaganda) para objetivos declarados	Contempla indivíduos ou organizações incluem extremistas violentos, organizações criminosas violentas, organizações políticas violentas, organizações que propagam ódio e indivíduos que praticam violência em série ou em massa. É proibido se afiliar a entidades violentas e odiosas ou promover as atividades delas; envolver-se ou promover atos violentos e recrutar, fornecer ou distribuir serviços (como mídia/ propaganda) para objetivos declarados"	Glorificação, apoio e representação de várias organizações e indivíduos perigosos. Esses conceitos se aplicam às próprias organizações, suas atividades e seus membros e não proíbem a defesa pacífica de resultados políticos específicos. Apoio. Representação (Declarar que você é membro de uma entidade ou criar entidade que se pretenda representar organização ou individuo). Contempla terrorismo, ódio organizado, atividade criminosa em larga escala, ato criminoso com várias vítimas, tentativas de ato criminoso com várias vítimas, assassinatos em série e eventos violentos que vão contra nossas política, e agentes violentos não estatais e entidades que incitam a violência Organizações e indivíduos designados pela Meta como Agentes violentos não estatais ou Entidades que incitam a violência não estão autorizados a ter presença no Facebook nem ter presença mantida por outras pessoas em seu nome.	Conteúdo produzido por organizações extremistas, criminosas ou terroristas violentas; que apoie ou enalteça figuras terroristas, extremistas ou criminosas com o objetivo de incentivar outras pessoas a realizar atos de violência; que enalteça ou justifique atos violentos realizados por organizações extremistas, criminosas ou terroristas violentas;destinado a recrutar novos membros para organizações extremistas, criminosas ou terroristas violentas; que mostre reféns ou que tenha a intenção de aliciar, ameaçar ou intimidar pessoas em nome de uma organização extremista, criminosa ou terrorista violenta; que mostre os emblemas, logotipos ou símbolos de organizações extremistas, criminosas ou terroristas violentas como forma de elogiá-las ou promovê-las/ que exalte ou promova tragédias violentas, como tiroteios em escolas O YouTube se baseia em vários fatores, incluindo classificações governamentais e de instituições internacionais para determinar o que constitui uma organização criminosa ou terrorista. Por exemplo, encerramos qualquer canal se tivermos razões para acreditar que o proprietário da conta é membro de uma organização classificada dessa forma, como uma Organização Terrorista Estrangeira (EUA) ou uma que seja identificada pelas Nações Unidas.

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há política contra glorificação de violência?	sim	sim	sim	sim
O que viola?	Contemplada na Política de Discurso Violento: Você não pode exaltar, elogiar nem celebrar atos de violência em que tenha ocorrido danos, o que inclui, entre outros, expressar gratidão por alguém ter sido vítima de ferimentos físicos ou elogiar entidades violentas e autores de ataques violentos. Inclui-se aí exaltar crueldade ou abuso contra animais.	Contemplada na política contra Comportamento Violento e Criminoso, Promover ou incitar a violência, como incentivar um ataque ou outras pessoas a atacar, enaltecer um ato violento ou recomendar que as pessoas levem armas a um local para intimidar outras pessoas	Contemplada na Política sobre Violência e incitação; é proibido também compartilhar imagens explícitas para prazer sádico ou para glorificar a violência	Contemplado nas Políticas de conteúdo violento ou gráfico. Não são permitidos no YouTube conteúdos violentos ou sangrentos destinados a chocar ou repugnar os espectadores, ou conteúdos que incentivem outras pessoas a cometer atos violentos.
Há política contra abuso?	sim	sim	sim	sim
O que viola?	Assédio direcionado, Negação de eventos violentos, Incitamento ao Assédio, Conteúdo sexual indesejado ou objetificação gráfica, conduta sexual não desejada e objetificação gráfica que sexualize um indivíduo sem seu consentimento, Insultos, Uso de Nomes e Pronomes Anteriores	Abuso refere-se a abuso sexual e físico contra jovens e contra adultos. Isso inclui material de abuso sexual infantil (CSAM, sigla em inglês), aliciamento, extorsão sexual, prostituição, pedofilia e danos físicos ou psicológicos a jovens. Isso inclui atos sexuais não consensuais, abuso sexual baseado em imagens, extorsão sexual, abuso físico e assédio sexual.	Contempla abuso de imagem íntima e sextorção; e política contra exploração sexual infantil, abuso e nudez. Não permitimos conteúdo ou atividade que explore crianças sexualmente ou as coloque em perigo. Quando tomamos conhecimento de um caso aparente de exploração infantil, fazemos uma denúncia ao National Center for Missing and Exploited Children, em conformidade com a legislação aplicável.	Contemplada nas políticas de assédio e bullying, na política de nudez e conteúdo sexual, e na política de segurança infantil online, enquadradas em "conteúdo sensível".

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há política contra assédio?	sim	sim	sim	sim
O que viola?	Assédio direcionado, Negação de eventos violentos, Incitamento ao Assédio, Conteúdo sexual indesejado ou objetificação gráfica, conduta sexual não desejada e objetificação gráfica que sexualize um indivíduo sem seu consentimento, Insultos, Uso de Nomes e Pronomes Anteriores	Contempla assédio e bullying. Ataque de doxxing envolve a publicação online de informações pessoais sobre alguém com intenção maliciosa. Degradar um indivíduo que passou por sofrimento físico ou devido a sua aparência pessoal, inteligência ou circunstâncias pessoais (como higiene, saúde ou histórico médico); Mostrar alguém sendo maltratado fisicamente por outra pessoa ou grupo; Degradar ou revitimizar pessoas que vivenciaram uma tragédia, Comprometer a segurança física de um indivíduo, Incitar outras pessoas a assediar alguém ou promover assédio coordenado, Ameaçar ou incitar outras pessoas a praticar doxxing ou chantagear alguém ou a compartilhar ou violar informações da conta	Removemos conteúdo publicado com o objetivo de degradar ou constranger, como alegações sobre a atividade sexual de alguém. Contato não desejado; Apelos à automutilação ou suicídio; Ataques a vítimas de agressão ou exploração sexual, assédio sexual ou violência doméstica.; Declarações mostrando intenção de participar de atividade sexual ou solicitando a participação em atividade sexual; Comentários com teor altamente sexual; Photoshop ou desenhos sexualizados depreciativo; Ataques com uso de termos depreciativos relacionados a atividade sexual; Alegações negando a ocorrência de uma tragédia violenta; Alegações de que indivíduos estão mentindo sobre terem sido vítimas de uma tragédia violenta ou ataque terrorista; Ameaças de divulgação de informação privada; Chamadas à ação ou declarações de intenção de participar em atos de bullying e/ou assédio; Conteúdo que degrade ou manifeste repugnância para com indivíduos que são expostos menstruando, urinando, vomitando ou defecando. Há proteções adicionais para todos os menores de idade, pessoas físicas adultas e figuras públicas de escopo limitado (por exemplo, pessoas cuja fama primária esteja restrita ao seu ativismo, jornalismo, ou que ficaram famosas involuntariamente) e Proteções adicionais para menores de idade, pessoas físicas adultas e figuras públicas involuntárias menores de idade.	Conteúdo que contenha insultos ou insultos prolongados com base nos atributos intrínsecos de alguém. Esses atributos incluem seu status de grupo protegido, atributos físicos ou seu status como sobrevivente de agressão sexual, distribuição de imagens íntimas não consensuais, abuso doméstico, abuso infantil e muito mais. Conteúdo carregado com a intenção de envergonhar, enganar ou insultar um menor. Isso significa ter a intenção de fazer com que um menor sinta emoções desagradáveis, como angústia, vergonha ou inutilidade; pretendendo enganá-los para que se comportem de maneiras que possam prejudicar a si mesmos ou a seus bens; ou xingamentos em relação a eles. Um menor é alguém menor de 18 anos. Há outros comportamentos proibidos, como conteúdo que compartilha, ameaça compartilhar ou incentiva outras pessoas a compartilhar informações de identificação pessoal (PII) não públicas, que incentive comportamentos abusivos coordenados, que promova teorias conspiratórias prejudiciais ou que tenha como alvo alguém alegando que faz parte de uma teoria da conspiração prejudicial, que ameace um indivíduo identificável ou sua propriedade, que retrata um encontro encenado que é usado para acusar um indivíduo identificável de má conduta flagrante com um menor, sem a presença de autoridades policiais, que revele ou zombe da morte ou ferimentos graves de um indivíduo identificável, que simule realisticamente menores falecidos ou vítimas de grandes eventos violentos mortais ou bem documentados descrevendo sua morte ou violência experimentada, que retrata criadores simulando atos de violência grave contra outras pessoas, que retrata criadores simulando atos de violência grave contra outras pessoas, que contenha perseguição de um indivíduo identificável, que negue ou minimize o papel de alguém como vítima de um grande evento violento bem documentado e que contenha sexualização indesejada de um indivíduo identificável.

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há política contra conduta odiosa?	sim	sim	sim	sim
O que viola?	É proibido atacar outras pessoas com base em raça, etnia, nacionalidade, casta, orientação sexual, gênero, identidade de gênero, crença religiosa, idade, deficiência ou doença grave. É proibido direcionar a alguém ou a grupos conteúdo que faz referência a formas de violência ou eventos violentos em que uma categoria protegida seja o principal alvo ou a vítima, em que a intenção seja o assédio. Isso inclui Desumanização, Imagens de propagação de ódio, Perfil de propagação de ódio, Mídias que retratam imagens de propagação de ódio não são permitidas em vídeos ao vivo, na bio da conta ou em imagens de capa ou do perfil.	Comportamento Violento e Criminoso - Ameaçar ou expressar o desejo de causar dano físico a uma pessoa ou grupo Promover ou incitar a violência, como incentivar um ataque ou outras pessoas a atacar, enaltecer um ato violento ou recomendar que as pessoas levem armas a um local para intimidar outras pessoas Promover o roubo ou a destruição de propriedades ou do meio ambiente Fornecer instruções sobre como cometer atividades criminosas que possam causar danos a pessoas, animais ou propriedades.	Removemos ameaças de violência contra vários alvos. Ameaças de violência são declarações ou elementos visuais que representam uma intenção, aspiração ou incitação de violência contra um alvo. As ameaças podem ser expressas por vários tipos de declarações, como declarações de intenção, chamadas para ação, defesa, declarações aspiracionais e condicionais.	Política de conteúdo violento ou explícito, sendo conteúdo explícito ou violento, contemplando também abuso animal e conteúdo encenado ou ficcional em que o espectador não tem discernimento da ficcionalidade.
Há política contra conduta odiosa?	sim	sim	sim	sim
O que viola?	Removeremos permanentemente quaisquer contas mantidas por autores de ataques terroristas, extremistas ou de violência em massa, e poderemos também remover posts que disseminem manifestos ou outro conteúdo produzido por esses autores. Também remoção de contas dedicadas ao compartilhamento de conteúdo prejudicial e violento associado aos autores ou ao ataque violento.	Contemplada na Politica de Organizações e indivíduos violentos e odiosos. Não permitimos a presença de organizações ou indivíduos violentos e odiosos em nossa plataforma. Esses atores incluem extremistas violentos, organizações criminosas violentas, organizações políticas violentas, organizações odiosas e indivíduos que causam violência em série ou em massa. Se tomarmos conhecimento de que qualquer um desses atores pode estar em nossa plataforma, realizaremos uma revisão completa - incluindo o comportamento fora da plataforma - que pode resultar em um banimento da conta.	Contemplada na Politica de Organizações e indivíduos perigosos, no campo de indivíduos terroristas; entidade de ódio; autores de ato criminosos com várias vítimas e assassinatos em série; agentes violentos não estatais e entidades que incitam a violência. Parte-se também de organizações listadas pelos EUA como Comandos Especiais de Tráfico de Drogas, Organizações Terroristas Estrangeiras ou Terroristas Globais Especialmente Designados	Contemplado na Política de organizações extremistas ou criminosas violentas, mas não há especificação sobre consequências para autores de ataques violentos, o foco é na reprodução de conteúdo e recrutamento.

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há política para mídia sensível?	sim	sim	sim	sim
O que viola?	É proibido postar mídias que sejam excessivamente sangrentas ou compartilhar conteúdo adulto ou violento em vídeos ao vivo, no perfil ou em imagens de cabeçalhos. Mídias que representem violência e/ou agressão sexual também são proibidas.	Contempla Atividade e serviços sexuais; Nudez e exposição do corpo; Conteúdo sexualmente sugestivo; Conteúdo explícito e impactante; Abuso animal	Imagens violentas ou explícitas. Publicações que contenham descrições de bullying ou assédio, se compartilhadas para aumentar o reconhecimento. Algumas formas de nudez. Publicações relacionadas a suicídio ou tentativas de suicídio.Há telas de aviso para conteúdo potencialmente sensível	Contempla Políticas de Nudez e Política de Conteúdo Sexual; Política de miniaturas; Política de segurança infantil; Política de suicídio, automutilação e transtornos alimentares; Política linguística vulgar

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Em caso da violação de alguma política de segurança...	Medidas: Limitação da visibilidade do post, com exclusão dos resultados de pesquisa, remoção de timelines, restrição da visibilidade, rebaixamento do post; Impedir posts de terem anúncios adjacentes; Solicitação de remoção do post; Identificação de um post; Aviso de exceção devido a interesse público; em contas, colocação de uma conta no modo somente leitura e Verificação da propriedade da conta: podemos exigir que o proprietário da conta verifique a propriedade com um número de celular ou endereço de e-mail. O conteúdo também pode ser ocultado com aviso, ser retido conforme idade ou retenção em país. Suspender contas cujo único propósito seja violar nossa política de Conteúdo Sexual Indesejado e Objetificação Gráfica, ou contas dedicadas a assediar indivíduos. Impedir posts de terem anúncios adjacentes: A partir de abril de 2023, os posts identificados como em violação de nossas regras começarão a receber identificações.	Medidas: Banimento temporário ou permanente de conta; conteúdo tornado inelegível para o feed "Para você" quando ele indiretamente rebaixa grupos protegidos. Priorizar a remoção mais rápida de conteúdo altamente grave e explícito, como material sobre abuso sexual infantil e extremismo violento. Minimizar a visão geral de conteúdos que violam nossas Diretrizes da Comunidade. Garantir precisão, consistência e imparcialidade para os criadores. Sob o novo sistema de moderação de contas, se alguém postar um conteúdo que viole uma das Diretrizes da Comunidade, o conteúdo será removido e sua conta receberá uma advertência. Se uma conta atingir o limite de advertências dentro de uma mesma ferramenta ou política, ela será banida permanentemente. Esses limites podem variar dependendo do potencial de uma violação em causar danos aos membros da comunidade. Dentro de banimento, você tem uma violação grave na sua conta se publicar, promover ou facilitar a exploração de jovens ou material de abuso sexual de crianças; promover ou ameaçar violência; publicar ou promover conteúdo que represente atos sexuais sem consentimento, como estupro ou importunação; publicar conteúdo que facilite o tráfico humano; publicar conteúdo que represente tortura no mundo real.	Medidas: Usamos um sistema de advertências para contabilizar as violações e responsabilizar você pelo conteúdo que publicar. Sua conta também pode ser sofrer restrição ou ser desativada, dependendo da política que seu conteúdo viola, do seu histórico anterior de violações e do número de advertências que você recebeu. Para garantir que o nosso sistema de advertência seja justo e adequado, não contabilizaremos advertências na maioria das violações de conteúdo publicadas há mais de 90 dias ou, no caso de violações mais graves, há mais de quatro anos. Para a maioria das violações, sua primeira advertência resultará em um aviso sem outras restrições. Dentre as restrições, temos de uso de recursos e criação de conteúdo por tempo limitado. Em caso de agitação civil, também podemos restringir contas de figuras públicas por períodos mais longos se eles incitarem ou exaltarem violência. Determinaremos o período de restrição após avaliar a gravidade da violação, o histórico de violações e restrição da conta, além do risco geral para a segurança pública. Também Reduzindo a distribuição de conteúdo problemático.	Medidas: remoção ou restrição do acesso de idade de conteúdo; remoção de canal se dá se acumular três avisos das diretrizes da comunidade em 90 dias, tiver um único caso de abuso grave (como comportamento predatório) ou for determinado como totalmente dedicado à violação de nossas diretrizes (como costuma acontecer com contas de spam). Quando um canal é encerrado, todos os seus vídeos são removidos.

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há delimitação de grupos protegidos?	sim	sim	sim	sim
Se sim, quais grupos são contemplados?	Você não pode atacar diretamente outras pessoas com base em raça, etnia, origem nacional, casta, orientação sexual, gênero, identidade de gênero, afiliação religiosa, idade, deficiência ou doença grave.	Grupos protegidos são indivíduos ou comunidades que compartilham atributos protegidos.Atributos protegidos significam caraterísticas pessoais com as quais você nasceu, que são imutáveis ou causaria danos psicológicos graves se você fosse forçado(a) a alterá-las ou fosse atacado por causa delas. Isso inclui: Casta, Etnia, Nacionalidade, Raça, Religião, Tribo, Status de imigração, Gênero, Identidade de gênero, Sexo, Orientação sexual, Deficiência, Doença grave. Além disso, também oferecemos algumas proteções relacionadas à idade e podemos considerar outros atributos protegidos quando tivermos contexto adicional, como informações específicas fornecidas por uma organização não governamental (ONG) local.	Proteções adicionais para pessoas físicas adultas, todos os menores de idade, pessoas de alto risco e pessoas ou grupos com base nas suas características protegidas: raça, etnia, nacionalidade, deficiência, religião, casta, orientação sexual, sexo, identidade de gênero e doença grave. Nota-se que a idade é característica protegida quando referenciada junto de outra protegida. Também protegemos de ataques mais graves refugiados, migrantes, imigrantes e pessoas que buscam asilo, embora permitamos comentários e críticas às políticas de imigração. Da mesma forma, há algumas proteções para aspectos como ocupação, quando eles são mencionados juntamente com uma característica protegida.	Status de grupo protegido pela política do YouTube: • Idade • Classe social • Deficiência • Etnia • Identidade e expressão de gênero • Nacionalidade • Raça • Situação de imigração • Religião • Sexo/gênero • Orientação sexual • Vítimas de um conflito violento em grande escala e os familiares dessas pessoas • Veteranos de guerra

2. POLÍTICAS DE MÍDIA SINTÉTICA E MANIPULADA; COMPORTAMENTO INAUTÊNTICO

A Tabela 2 analisa a presença de mídia sintética ou manipulada, política de spam e política de comportamento inautêntico coordenado. A mídia sintética nas redes sociais refere-se a conteúdos gerados por inteligência artificial ou outras tecnologias que criam imagens, vídeos ou áudios de maneira artificial, muitas vezes indistinguível de conteúdos reais produzidos por humanos. De acordo com Spiandorin et al (2023)[5], a mídia sintética consiste em conteúdos por algoritmos de inteligência artificial, onde modelos de aprendizagem de máquina reconhecem e analisam padrões de dados e criam conteúdos digitais, como imagens, vídeos, ou áudios a partir de dados pré-existentes. Esses algoritmos são capazes de processar dados disponíveis e gerar representações realistas de pessoas e diálogos que nunca ocorreram na realidade.

5 A era da mídia sintética: o conflito entre inovação e regulação (insper.edu.br)

TABELA 2

POLÍTICAS DE MÍDIA SINTÉTICA E MANIPULADA; COMPORTAMENTO INAUTÊNTICO				
Política/Plataforma	**X**	**TikTok**	**Meta**	**Google - Youtube**
Tem política sobre mídia sintética e manipulada?	sim	sim	sim	sim
O que viola a política?	Conteúdo gerado por IA que mostre cenas realistas; conteúdo gerado por IA que contenha a imagem de uma figura pública caso o conteúdo seja usado para endossos ou viole qualquer outra política; Incluir mídias que sejam significativamente ou enganosamente alteradas, manipuladas ou fabricadas, ou incluir mídias que sejam compartilhadas de modo enganoso ou com falso contexto, e Incluir mídias que tenham potencial para resultar em confusão generalizada sobre questões públicas, ter impacto na segurança pública ou causar danos graves	conteúdo gerado por IA que contêm a semelhança (visual ou sonora) de uma pessoa real ou fictícia não são permitidos, mesmo se informados com a indicação de conteúdo gerado por IA; Pessoas com aparência realista com menos de 18 anos; Conteúdo que parece ser proveniente de uma fonte de autoridade, como um canal de notícias respeitável; Um evento de crise, como um conflito ou desastre natural.	Descrever uma pessoa real dizendo ou fazendo algo que não disse ou fez; retratar uma pessoa, com aparência realista, que não existe ou um evento realista que não aconteceu, ou altere a gravação de um evento real que aconteceu; ou descreva um evento real que supostamente tenha ocorrido, mas que não seja uma imagem, vídeo ou gravação de áudio verdadeira do evento.	Youtube - conteúdos alterado ou sintético que não seja realista; anúncios políticos que não divulguem que o material é alterado; mídia manipulada, falsificações profundas e outras formas de conteúdo adulterado com objetivo de enganar, fraudar ou enganar os usuários.
Quais são as medidas corretivas?	Exclusão de post; Rotulação; Bloqueios de contas.	Banimento de conta e banimento de qualquer conta alternativa que esteja send usada ou novas contas que forem criadas. link	Remoção de conteúdos que violam as políticas da plataforma, com auxilio de parceiros de checagem de fatos.	Remoção de conteúdos que violam as políticas. Há o SynthID, uma ferramenta do Google DeepMind, que incorpora diretamente uma marca d'água digital em imagens e áudio gerados por IA. As políticas de desinformação do YouTube proíbem conteúdo tecnicamente manipulado que engana os usuários e pode representar um sério risco de danos flagrantes, o YouTube exigirá que os criadores divulguem quando criaram conteúdo alterado ou sintético realista e exibirá um rótulo que indica para as pessoas quando o conteúdo que estão assistindo é sintético.

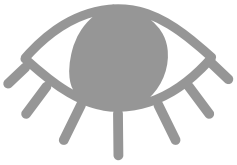
Política/Plataforma	X	TikTok	Meta	Google - Youtube
Tem política sobre spam?	sim	sim	sim	sim
Tem política sobre comportamento inautêntico coordenado? (CIB)	sim	sim	sim	não
O que viola a política?	Contemplado na Política de Manipulação e Spam, dentro de engajamentos inautênticos, que tentam fazer com que as contas ou o conteúdo pareçam mais populares ou ativos do que são; atividade coordenada, que tenta influenciar artificialmente as conversas através do uso de múltiplas contas, contas falsas, automação e/ou script;	Comportamentos que possam acarretar spam ou enganar a comunidade; sso inclui a realização de operações de influência disfarçadas, a manipulação de sinais de engajamento para ampliar o alcance de determinado conteúdo e a operação de contas de spam ou que se passem por outra pessoa.	Atividades motivadas financeiramente, como spam ou táticas de engajamento falsas que dependem de amplificação inautêntica ou evadiamento da aplicação da lei, em vez de um uso principal de contas falsas. uso de vários ativos do Facebook ou Instagram em conjunto para se envolver em comportamento não autêntico (conforme definido acima) em que o uso de contas falsas é essencial para a operação. Interferência Estrangeira ou Governamental (FGI) abrangem esforços de estrangeiros para manipular o debate público em outro país e operações dirigidas por um governo para atingir seus próprios cidadãos. Estes podem ser particularmente preocupantes quando combinam técnicas enganosas com o poder do mundo real de um estado. Nesse caso, aplicam-se remoção de todas as propriedades on-plataforma conectadas à operação em si e às pessoas e organizações por trás dela.	N/A

Para além da tabela comparativa, mas ainda em assunto correlato, destacamos no campo de navegadores que, a partir de 5 de maio de 2024, a Google começará a punir abuso de reputação de site. Isso significa quando conteúdo de baixo valor produzido por terceiros é publicado sem revisão adequada. “Conteúdo de baixo valor” inclui páginas patrocinadas, propagandas ou parcerias com objetivo de se aproveitar da reputação e autoridade de quem hospeda as páginas.

O abuso de reputação do site ocorre quando páginas de terceiros são publicadas com pouca ou nenhuma supervisão ou envolvimento do site primário, e o objetivo é manipular a classificação da Pesquisa aproveitando os indicadores de classificação do site primário. Essas páginas de terceiros incluem páginas patrocinadas, de publicidade, de parceiros ou outras páginas de terceiros que normalmente são independentes da finalidade principal do site host ou são produzidas sem supervisão ou envolvimento do host, e oferecem pouco ou nenhum valor para os usuários.

* **A Google-Youtube não possui uma política contra comportamento inautêntico coordenado.**
* **Nota-se que a Google, a Meta e o TikTok são parte da Parceria em Inteligência Artificial sobre Práticas Responsáveis para Mídia Sintética**

3. POLÍTICA DE INTEGRIDADE CÍVICA E OU ELEITORAL



A política faz parte de um conjunto de esforços de organizações e movimentos presentes na internet para garantir a integridade eleitoral e os processos democráticos sob ameaças das redes de desinformação nesse período. Segundo relatório "Compromisso com a Democracia" do InternetLab, ressalvadas as diferenças entre empresas e seus serviços, parte fundamental da atividade das plataformas de internet consiste na organização de seus usuários e usuárias e do conteúdo por eles(as) produzido. Isto, por sua vez, possui impacto em direitos humanos, como a liberdade de expressão, o livre desenvolvimento da personalidade e a soberania popular, de modo que a primeira pergunta que surge, nesse cenário, é se tais serviços devem ou não possuir políticas de conteúdo específicas abordando questões sobre processos cívicos e democráticos, como eleições e a transmissão pacífica de poder entre representantes eleitos(as).

TABELA 3

POLÍTICA DE INTEGRIDADE CÍVICA/ELEITORAL				
Política/Plataforma	**X**	**TikTok**	**Meta**	**Google - Youtube**
Há política de integridade cívica?	sim	sim	sim	sim
O que ela contempla?	Dispõe de atos cívicos: Informações enganosas sobre como participar; informações falsas ou enganosas sobre as circunstâncias envolvendo um ato cívico para intimidar ou dissuadir pessoas da participação; incitar, promover nem incentivar outras pessoas a ameaças ou coagir outras pessoas a participar ou se abster da participação em um ato cívico; criar contas falsas que representem afiliação de forma indevida ou compartilhar conteúdo que represente de forma falsa a afiliação a um candidato, representante público eleito, partido político, autoridade eleitoral ou entidade governamental.	Contempla desinformação, integridade cívica e eleitoral, mídia editada e conteúdo geraldo por IA, engajamento falso, conteúdo não-original e spam e comportamento fraudulento. Não é permitido propaganda política paga, publicidade política ou arrecadação de fundos por políticos e partidos políticos (para eles ou para outras pessoas). Também se indica que a política de anúncios políticos inclui tanto anúncios pagos tradicionais quanto criadores de conteúdo que recebem remuneração para apoiar ou se opor a um candidato a um cargo público. O conteúdo poderá não ser recomendado para o feed "Para você" se contiver desinformações que possam prejudicar a capacidade de um eleitor tomar uma decisão informada. Por cautela, alegações não verificadas sobre uma eleição e conteúdo temporariamente sob revisão de verificadores de fatos também podem não ser qualificados para o feed "Para você".	Integridade eleitoral - ações no campo de prevenir a interferência; combater a desinformação; incentivar as pessoas a votar e aumentar a transparência	Política de desinformação eleitoral, nota-se: Integridade das eleições: conteúdo com alegações falsas de que fraudes, erros ou problemas técnicos generalizados ocorreram em determinadas eleições passadas para determinar os chefes de governo. Ou conteúdo que afirma que os resultados certificados dessas eleições são falsos. Atualmente, essa política se aplica a: qualquer eleição presidencial dos EUA; eleições federais da Alemanha de 2021; eleições presidenciais do Brasil de 2014, 2018 e 2022. (indicação de que é uma lista não-completa)

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há medidas ou protocolos de proteção para eleições?	sim	sim	sim	sim
Quais medidas ou protocolos?	As violações da política de integridade cívica; as propagandas políticas (anúncios de conteúdo político e de campanha política) devem cumprir os requisitos legais específicos do país, bem como as respectivas leis eleitorais e regras de períodos de silêncio eleitoral. De acordo com as medidas para as eleições de 2022 no Brasil, temos etiquetas de identificação de candidatos(as) para estar a par de quem está concorrendo às eleições; sessão do Explorar dedicada às eleições; proteção proativa adicional para contas de candidatos.	Medidas contra desinformação; colaboração com especialistas externos; Hashtag PSA (utilidade pública) - avisos em diversas páginas de hashtags relacionadas a eleições para relembrar as pessoas a seguir as Diretrizes da Comunidade, verificação de fatos e denúncia de conteúdos que possam violar nossas políticas; ferramenta para relatar conteúdo falso sobre eleições diretamente no aplicativo. Essas denúncias são repassadas a um time brasileiro dedicado a revisar o conteúdo de acordo nas nossas regras relacionadas à desinformação; Guia no aplicativo - para as eleições brasileiras de 2022, lançamento do Guia Eleições, conectando os usuários com informações oficiais do Tribunal Superior Eleitoral.	Medidas contra desinformação; Prevenção de interferência; Aumento de transparência; Incentivo às pessoas a votar; Incentivos financeiros para a proteção das eleições nos EUA; Parceria com checadores de fatos independentes (no Brasil com AFP, Agência Lupa, Aos Fatos e Estadão Verifica).	Medidas contra desinformação; iniciativas de fomento ao combate à desinformação (Jogo Limpo 2.0 em 2023, parceria com International Center for Journalists (ICFJ). iniciativa que tem por objetivo promover métodos e modelos de produção factual para combater os impactos negativos e impedir a disseminação de conteúdo desinformativo. Destinado a jornalistas, checadores de fatos e equipes colaborativas que trabalham contra a desinformação).

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há medidas sendo elaboradas para as eleições da União Europeia de 2024?	não	sim	sim	sim
Quais medidas ou protocolos?	N/A	À medida que a eleição se aproxima, ativação de um Centro de Operações Eleitorais para identificar ameaças potenciais e implementar mitigações em tempo real. Rede de verificação de fatos em expansão com 3 novos parceiros na Bulgária, França e Eslováquia. Comprometimento a adotar uma abordagem responsável para novas tecnologias, como a GenAI, e assinamos o acordo de tecnologia para combater a disseminação de conteúdo de IA enganoso nas eleições.	Ativação de Centro de Operações Eleitorais para identificar ameaças potenciais e implementar mitigações em tempo real. Temos a maior rede de verificação de fatos de qualquer plataforma e atualmente estamos expandindo-a com 3 novos parceiros na Bulgária, França e Eslováquia. Comprometemo-nos a adotar uma abordagem responsável para novas tecnologias, como a GenAI, e assinamos o acordo de tecnologia para combater a disseminação de conteúdo de IA enganoso nas eleições.	Detalhes da votação na Pesquisa Google (quando as pessoas pesquisarem por tópicos como “como votar”, encontrarão detalhes sobre como podem votar — como requisitos de identificação, registro, prazos de votação, votação no exterior e orientação para diferentes meios de votação, como pessoalmente ou via correio); Informações confiáveis no YouTube: Para notícias e informações relacionadas a eleições, nossos sistemas destacam conteúdo de fontes confiáveis, na página inicial do YouTube, nos resultados de pesquisa e no painel “Up Next”; Transparência contínua sobre anúncios eleitorais; investimento no Centro de Engenharia de Segurança da Google para Responsabilidade de Conteúdo em Dublin, dedicado à segurança online na Europa e em todo o mundo; Uso de IA para combater o abuso em escala; fomento financeiro no combate a desinformação via Fundo Europeu de Informação sobre os Meios de Comunicação Social, o Global Fact Check Fund; Prebunking para antecipar a manipulação online:campanha de antecipação antes das eleições para o Parlamento Europeu na França, Alemanha, Itália, Bélgica e Polônia. Os vídeos serão igualmente traduzidos e disponibilizados em todas as línguas da UE; expansão das políticas de conteúdo político para exigir que os anunciantes divulguem quando seus anúncios eleitorais incluem conteúdo sintético que retrata pessoas ou eventos reais ou realistas de forma inautêntica; Rótulos de conteúdo sintético ou alterado no YouTube; fornecimento aos usuários contexto adicional: Programa de Proteção Avançada; parcerias com a PUBLIC, a International Foundation for Electoral Systems (IFES) e a Deutschland sicher im Netz (DSIN) para escalar o treinamento de segurança da conta e fornecer ferramentas de segurança; ação contra operações coordenadas de influência; e recursos úteis em euelections.withgoogle: hub específico da UE com recursos e treinamentos futuros para ajudar as campanhas a se conectarem com os eleitores e gerenciar sua segurança e presença digital.

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há medidas sendo elaboradas para as eleições municipais no Brasil?	não	sim	sim	não
Quais medidas ou protocolos?	N/A	Ao longo de 2024, continuamos a fazer parcerias com especialistas e organizações de checagem de fatos em todo o mundo para oferecer campanhas de educação midiática sobre desinformação, identificação de conteúdo gerado por Inteligência Artificial (IA) e mais. Além disso, seguimos desenvolvendo recursos que fornecem informações adicionais sobre conteúdo e contas no TikTok. Por exemplo, os selos de verificação azul confirmam que contas notáveis são quem dizem ser. Rotulamos o conteúdo que nossa moderação determina como não verificado e expandiremos os recursos de educação midiática para essas classificações este ano. Trabalho com Combate à desinformação; Impedindo operações de influência disfarçadas: Endereçando conteúdo enganoso gerado por IA; Adaptando nossa abordagem de contas políticas e de notícias	Colaboramos com autoridades eleitorais e trabalhamos para combater a desinformação em nossas plataformas. Medidas para os riscos da Inteligência Artificial. Trabalhamos para conter a viralização de mensagens com medidas implementadas diretamente no WhatsApp, como a limitação de encaminhamento por padrão. Não permitimos o uso da Plataforma do WhatsApp Business (API) por candidatos e campanhas políticas. Ativaremos nosso Centro de Operações para Eleições. No Brasil, a Meta possui seis parceiros para essa iniciativa: Agência Lupa, AFP, Aos Fatos, Estadão Verifica, Reuters Fact Check e UOL	N/A

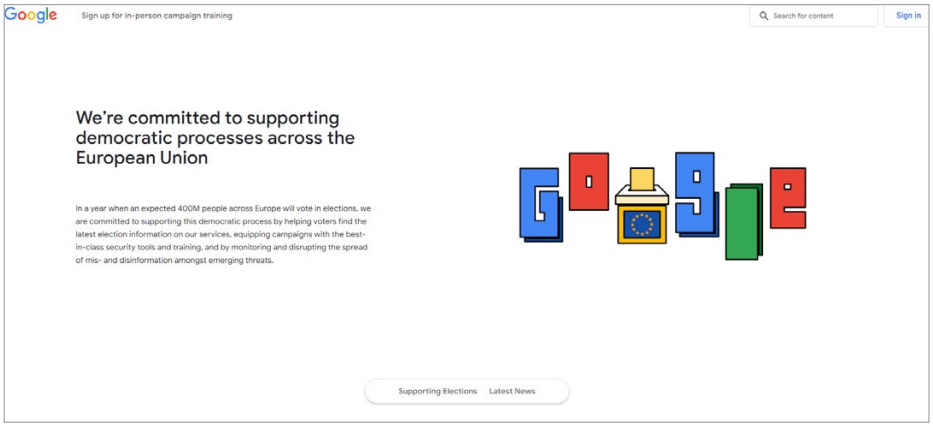
Política/Plataforma	X	TikTok	Meta	Google - Youtube
Serão permitidos anúncios políticos nas eleições no Brasil?	não	não	sim	não

Destaca-se que o X não dispõe de políticas e ferramentas para as eleições municipais brasileiras de 2024, segundo pesquisas no Google e canais oficiais da plataforma. Ainda, enquanto o Tiktok proíbe oficialmente anúncios políticos, a plataforma reconhece que há ocasiões em que os governos poderão necessitar de acesso aos nossos serviços de publicidade, como para apoiar a saúde e a segurança públicas e o acesso à informação, como a publicidade de campanhas de reforço da COVID-19. Continuaremos a permitir que organizações governamentais anunciem em circunstâncias limitadas e será necessário que trabalhem com um representante da TikTok.

Nota-se que há esforços, também, para as eleições presidenciais nos Estados Unidos. O TikTok, por exemplo, fará parcerias com comissões eleitorais e organizações de verificação de fatos para construir Centros Eleitorais que conectem pessoas a informações confiáveis sobre votação. Nossos centros eleitorais locais atingiram mais de 55 milhões de pessoas em todo o mundo no ano passado. Nos próximos dias, lançaremos nosso Centro de Eleições nos EUA, em parceria com a Democracy Works, uma organização sem fins lucrativos. O Centro fornecerá aos mais de 150 milhões de membros da comunidade dos EUA informações confiáveis de votação para todos os 50 estados e Washington, DC. Haverá direcionamento para as pessoas para o Centro de Eleições por meio de solicitações sobre o conteúdo e as pesquisas eleitorais relevantes. Continuaremos a adicionar informações ao longo do ano, incluindo os resultados das eleições.

Dentre os recursos da Google para as eleições na União Europeia, o hub específico da UE em euelections.withgoogle com recursos e treinamentos futuros para ajudar as campanhas a se conectarem com os eleitores e gerenciar sua segurança e presença digital. Antes das eleições para o Parlamento Europeu em 2019, realizamos treinamentos de segurança presenciais e online para mais de 2.500 funcionários de campanha e eleitorais e, em 2024, pretendemos aproveitar esses números.

Figura 1. *Printscreen com informações do comitê de suporte à União Europeia*



O Programa de Proteção Avançada está disponível para autoridades eleitas, candidatos, funcionários de campanha, jornalistas, funcionários eleitorais e outros indivíduos de alto risco. Para as eleições estadunidenses, a plataforma anunciou a expansão da parceria com a Defending Digital Campaigns (defendcampaigns.org) para fornecer às campanhas as ferramentas de segurança necessárias para permanecerem seguras on-line, incluindo ferramentas para configurar rapidamente os recursos de segurança do Google Workspace.

A Google destaca que irá continuar a trabalhar com parceiros como a Democracy Works para exibir informação de qualidade sobre as eleições locais e estaduais no topo da busca quando as pessoas pesquisarem tópicos sobre como e onde votar. No campo dos resultados eleitorais, a plataforma anunciou que fará uma parceria com a The Associated Press para apresentá-los.

Nota-se como as informações são apresentadas sobre as eleições nos EUA e no Brasil:

Figura 2. *Printscreen feito por equipe de pesquisa, com resultados das eleições nos Estados Unidos*

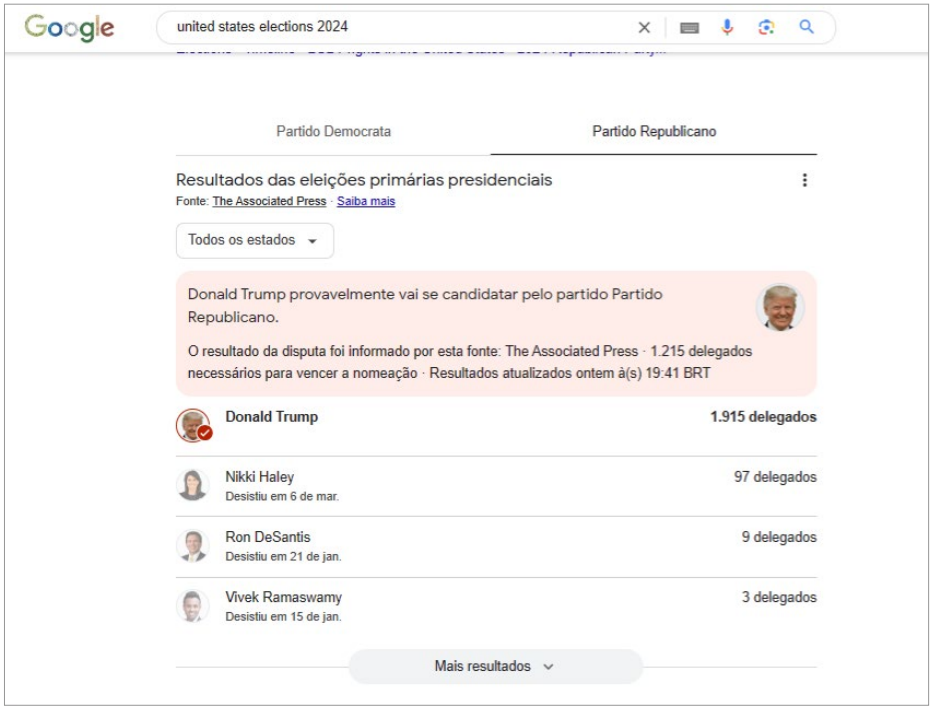
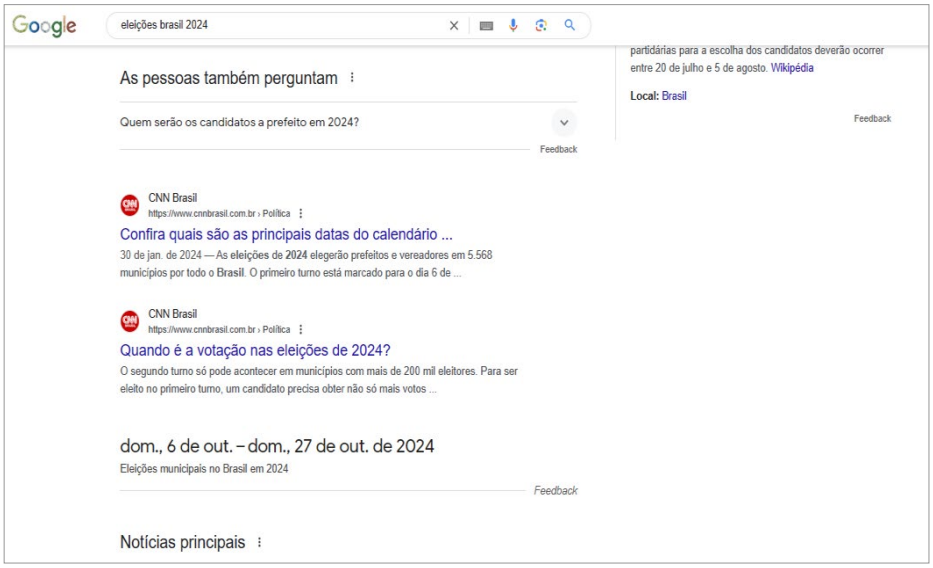


Figura 3. *Printscreen feito por equipe de pesquisa, com resultados das eleições no Brasil.*



4. POLÍTICAS DE DESINFORMAÇÃO

Neste tópico, focamos nas políticas sobre desinformação apresentadas na tabela X. Entende-se como desinformação "a disseminação deliberada de informações falsas, enganosas ou distorcidas com o objetivo de induzir ao erro, manipular opiniões ou influenciar comportamentos.". A desinformação refere-se à disseminação de informações incorretas, enganosas ou falsas, muitas vezes com o objetivo de manipular a opinião pública, influenciar decisões políticas, promover agendas específicas ou causar danos a indivíduos, grupos ou instituições. A desinformação pode se manifestar de várias formas, incluindo notícias falsas, boatos, teorias da conspiração, manipulação de mídia e propaganda enganosa. É importante ressaltar que a desinformação pode ser disseminada intencionalmente por atores mal-intencionados, mas também pode ser inadvertida ou resultar de erros de interpretação.Ainda sobre a noção de desinformação, o Sleeping Giants Brasil também entende que o uso desse mecanismo para conseguir retornos financeiros e políticos é um dos maiores desafios no seu combate, uma vez que Big Techs também tem lucrado com esse mercado. **Nota-se que a existência de políticas contra desinformação é necessário, mas não assegura que esse tipo de conteúdo não circule livremente**: um estudo do projeto Mídia e Democracia da Fundação Getulio Vargas, por exemplo, testou 38 peças publicitárias com ataques ao sistema eleitoral e outras abordagens teoricamente vetadas pela plataforma por estarem em desacordo com as políticas da empresa. Quase todas elas passaram no filtro do Google e foram autorizadas.

TABELA 4

POLÍTICAS DE DESINFORMAÇÃO				
Política/Plataforma	**X**	**TikTok**	**Meta**	**Google - Youtube**
Tem política contra desinformação?	sim	sim	sim	sim
O que a política de desinformação proibe?	Abrange política de Desinformação em Momentos de Crise; política de mídia sintética e manipulada e Política de integridade cívica; Política de desinformação contra a COVID-19	Informações enganosas que representam um risco à segurança pública ou podem causar pânico sobre um evento de crise ou emergência; Desinformações sobre a saúde; Desinformação sobre mudanças climáticas; Teorias de conspiração	Alegações sobre uma eleição ou crise; Alegações direcionadas a determinado grupo étnico, social ou religioso. Desinformação sobre saúde, no campo de vacinas, desinformação sobre saúde pública durante emergências de saúde pública e promover ou defender curas milagrosas prejudiciais à saúde. Alegações sobre produtos ou empregos que podem oferecer risco de grandes perdas financeiras. Desinformação que tenha potencial direto para riscos de danos físicos ou violência. Mídia manipulada.	Supressão de participantes de um censo: informações incorretas sobre o horário, local, meios ou requisitos de qualificação para o censo ou alegações falsas que podem desencorajar a participação. Conteúdo manipulado: mídia editada ou adulterada para enganar os usuários e que oferece risco de danos graves. Conteúdo atribuído erroneamente: vídeos que podem representar risco de danos graves ao afirmar falsamente que imagens de um acontecimento passado fazem parte de um evento atual; Desinformação sobre saúde (Desinformação sobre prevenção, sobre tratamentos e sobre negação).
Quais são as medidas corretivas?	Suspensão temporária ou permanente de conta a depender da incidência de violação; limitação da amplificação do conteúdo; remoção do conteúdo caso consequências offline imediatas e graves; rotulação de conteúdo com conteúdo adicional, redução de visibilidade; desmonetização do post se corrigido pelas Notas Comunitárias; durante acontecimentos importantes podem destacar proativamente mensagens informativas ou atualizações que contrariem narrativas enganosas	Remoção, redução de visibilidade, inelegível para o feed "Para você"	Para conteúdo que siga os Padrões da Comunidade, poderemos colocar um rótulo informativo na frente dele ou rejeitar o conteúdo enviado como publicidade quando: ele for uma imagem ou um vídeo fotorrealista ou um áudio com som realista, que tenha sido criado ou alterado digitalmente e que gera um risco particularmente elevado de enganar, de forma material, o público sobre uma questão de importância pública.	Conteúdo removido notificação no seu e-mail. Se essa for sua primeira violação das nossas diretrizes da comunidade, seu canal vai receber apenas um alerta sem penalidades. Acúmulo de notificações leva a maiores penalidades.

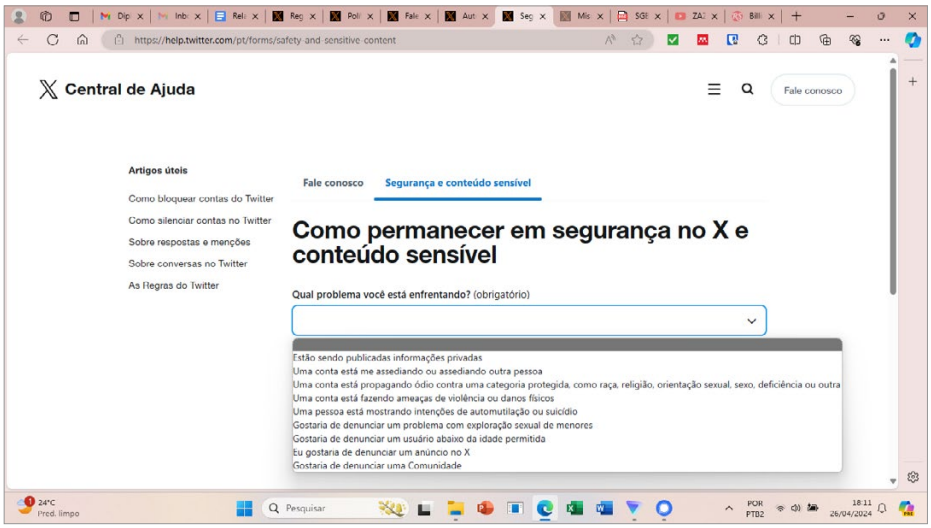
Política/Plataforma	X	TikTok	Meta	Google - Youtube
Canal de denúncia específico?	Alguns usuários podem denunciar posts por conterem desinformações. Isso atualmente está disponível em testes limitados a pessoas na Austrália, no Brasil, na Coreia do Sul, na Espanha, nos Estados Unidos e nas Filipinas.	É possível reportar desinformação, classificada como informação falsa sobre eleição; informação falsas prejudiciais; deepfakes mídia sintética e mídia manipulada	É possível reportar conteúdo com desinformação, classificada como informação falsa nos campos de saúde, política, tema social, criado ou alterado digitalmente ou outra coisa.	Qualquer usuário pode fazer uma denúncia. O Youtube disponibiliza um canal para denúncia. [link]
Tem política específica de desinformação eleitoral?	sim	sim	sim	sim

Política/Plataforma	X	TikTok	Meta	Google - Youtube
O que a política de desinformação eleitoral contempla?	Manipular ou interferir nas eleições ou outros processos cívicos. Isso inclui postar ou compartilhar conteúdo que impede a participação de eleitores ou engana pessoas sobre quando, onde ou como participar de um processo cívico, ou causa violência presencial durante uma eleição. Atos cívicos são os eventos ou procedimentos compulsórios, organizados e conduzidos pelo corpo governante e/ ou eleitoral de um país, estado, região, distrito ou município para tratar de uma questão de preocupação em comum por meio da participação pública. A política aborda Informações enganosas sobre como participar; informações falsas ou enganosas sobre as circunstâncias envolvendo um ato cívico destinado a intimidar ou dissuadir pessoas da participação em uma eleição ou outro ato cívico; incitação, promoção ou incentivo de a ameaças ou coação outras pessoas a participar ou se abster da participação em um ato cívico e criar contas falsas que representem afiliação de forma indevida ou compartilhar conteúdo que represente de forma falsa a afiliação a um candidato, representante público eleito, partido político, autoridade eleitoral ou entidade governamental. Leia mais sobre nossa Política de identidades enganosas e que induzem ao erro.	Desinformação sobre como votar ou concorrer a um cargo; sobre requisitos de elegibilidade dos eleitores para participar de uma eleição e qualificações dos candidatos para concorrer a um cargo; sobre leis, processos e procedimentos que regem a organização e a implementação de eleições e processos cívicos, como referendos, propostas de votação ou censos; sobre os resultados finais das eleições; promoção ou instrução sobre a interferência eleitoral e participação ilegal, incluindo intimidação de eleitores, observadores eleitorais ou de pessoas que trabalham nas eleições; exigir que se interrompa um resultado legítimo de uma eleição fora do sistema legal, como através de um golpe	Na Política de desinformação, para promover a integridade das eleições e dos censos, removemos a desinformação que pode contribuir ou contribua diretamente para o risco de interferência na capacidade das pessoas participarem desses processos. Isso inclui o seguinte: Desinformação sobre datas, lugares, horários e métodos de votação, registro de eleitores ou participação em censos; sobre quem pode votar, quais são os requisitos eleitorais, se um voto é contabilizado e quais informações ou materiais devem ser apresentados para votar; sobre a participação ou não de um candidato em uma eleição; sobre quem pode participar do censo e quais informações ou materiais devem ser apresentados para participar dele; sobre envolvimento governamental no censo, incluindo, se aplicável, que as informações censitárias de uma pessoa serão compartilhadas com outra agência do governo não censitária; afirmações falsas ou não verificadas de que o Serviço de Imigração e Controle Alfandegário dos EUA (ICE, na sigla em inglês) está em um local de votação; afirmações falsas explícitas de que as pessoas serão infectadas pela COVID-19 ou por outra doença transmissível se participarem do processo eleitoral; afirmações falsas sobre as condições atuais de um local de votação nos EUA que impossibilitariam a votação, como é verificado por uma autoridade eleitoral. Temos políticas adicionais com a intenção de cobrir apelos à violência, apologia à participação ilegal e apelos à interferência coordenada nas eleições, que estão representadas em outras seções dos nossos Padrões da Comunidade.	Supressão de eleitores: proíbe conteúdo destinado a enganar os eleitores relativamente ao horário, local, meios ou requisitos de elegibilidade para votar ou que inclua alegações falsas que desincentivam substancialmente o voto. Por exemplo, um vídeo que diga aos eleitores que podem votar através de métodos falsos, como enviar o voto por mensagem de texto para um determinado número. Elegibilidade de candidatos: proíbe conteúdo que expõe afirmações falsas em matéria de requisitos de elegibilidade técnica de atuais candidatos políticos e funcionários governamentais eleitos efetivos. Os requisitos de elegibilidade considerados baseiam-se na lei nacional aplicável e incluem a idade, o estado de cidadania e saúde. Incitamento à interferência em processos democráticos: proíbe conteúdo que incentive outras pessoas a interferirem com os processos democráticos. Isto inclui a obstrução ou a interrupção de procedimentos de votação.

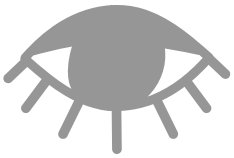
Política/Plataforma	X	TikTok	Meta	Google - Youtube
Tem política de protocolo de crise?	sim	não	sim	não
O que a política contempla?	Situações em que existe uma ameaça generalizada à vida, à segurança física, à saúde ou à subsistência básica. Havendo evidências de que uma reivindicação pode ser enganosa, não amplificaremos nem recomendaremos conteúdo coberto por esta política no Twitter, inclusive na linha do tempo inicial, na Pesquisa e na Exploração. Além disso, priorizaremos a adição de avisos em Tweets altamente visíveis e a Tweets de contas de alto perfil, como contas afiliadas ao Estado e contas oficiais verificadas do governo.	N/A	Durante momentos de crise, a Meta avalia os riscos de danos iminentes dentro e fora de nossa plataforma para que possamos aplicar políticas específicas e tomar medidas que ajudarão a manter as pessoas em segurança. Nosso Protocolo de Política de Crise (PPC) interno nos ajuda a realizar esse trabalho. Criamos esse protocolo com base em uma recomendação do Comitê de Supervisão como forma de aumentar os esforços já existentes nesta área. Embora a Meta analise regularmente os conteúdos que as pessoas publicam para avaliar se violam ou não nossas políticas, os riscos podem ser maiores durante momentos de crises, e diferentes respostas podem ser necessárias. Diante disso, usamos o PPC para fazer uma avaliação. A estrutura interna do PPC nos ajuda a equilibrar a necessidade de agir rapidamente, oferecendo uma resposta de política de conteúdo global consistente e permitindo flexibilidade para nos adaptarmos a condições em rápida mudança. O PPC prevê riscos e é baseado em crises anteriores, o que garante que os principais aprendizados sejam incorporados a ele. O desenvolvimento do protocolo incluiu uma pesquisa original e consultas com mais de 50 especialistas externos globais em segurança nacional, prevenção de conflitos, discurso de ódio, resposta humanitária e direitos humanos.	N/A

É possível observar que boa parte das plataformas, atualmente, têm desenvolvido políticas para o problema da desinformação. As plataformas citadas seguem a mesma noção de desinformação, focando no que pode ser entendido como "informações incorretas" ou "alegações falsas" que podem ser prejudiciais aos usuários ou contextos específicos, como é o caso das eleições. No que diz respeito às medidas corretivas, ambas têm optado pela remoção do conteúdo e suspensão temporária ou permanente do usuário que dissemina informações falsas nas redes.

Figura 4. *Printscreen com aba da Central de Ajuda do X sobre segurança*



5. CONTAS POLÍTICAS OU AFILIADAS AO ESTADO



Algumas plataformas possuem políticas para o que designam como "contas políticas" ou contas afiliadas ao Estado. Geralmente, há a sinalização dessas contas e a aplicação de certas medidas específicas para elas. Temos que o **TikTok possui a política mais detalhada de contas políticas**, não só abrangendo diversas figuras dentro de sua caracterização como "políticas", mas aplicando uma série de restrições para o uso dessas contas. Nota-se que embora muitas contas políticas já tenham adicionado o selo verificado ao seu perfil, isso é geralmente opcional. Nos EUA, porém, o TikTok implementou a verificação obrigatória de contas pertencentes a governos, políticos e partidos políticos durante as eleições intercalares. O **X também possui sinalização de conta de mídia afiliada ao Estado**[6], porém tem sido noticiado que diversas mídias afiliadas ao Estado perderam os seus selos com a compra da plataforma por Musk. Segundo relatório da organização NewsGuard[7], nos 90 dias seguintes à remoção de rótulos de mídia afiliada pelo Estado em X, o engajamento (o número de curtidas e compartilhamentos) em postagens de contas de mídia estatal russa, chinesa e iraniana em inglês disparou 70%, de acordo com uma análise da NewsGuard usando dados da plataforma de monitoramento Meltwater. A Meta, por outro lado, enquanto não oferece essa categorização ou restrições, possui uma plataforma de recursos para governo e organizações não-governamentais, com ferramentas de gerência de conta, treinamentos, recomendações e instruções para campanhas. Já a Google, não possui sinalização em contas políticas ou de mídias afiliadas aos Estados.

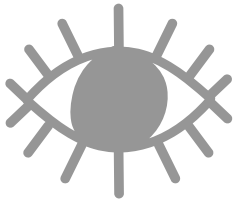
6 https://help.x.com/en/rules-and-policies/state-affiliated-iran
7 https://www.newsguardtech.com/misinformation-monitor/september-2023/

TABELA 5

CONTAS POLÍTICAS E/OU AFILIADAS AO ESTADO				
Política/Plataforma	**X**	**TikTok**	**Meta**	**Google - Youtube**
Há política para contas políticas e/ou afiliadas ao Estado?	sim	sim	não	não
A quem se refere?	Contas governamentais - funcionários do alto escalão e nas entidades que representem a voz de uma nação no exterior, especificamente contas de funcionários-chave do governo, como ministro de relações exteriores, entidades institucionais, porta-vozes oficiais e principais líderes diplomáticos. mídia afiliada ao Estado são meios em que o Estado exerce controle sobre o conteúdo editorial por meio de recursos financeiros, pressões diretas ou indiretas e/ou controle sobre produção e distribuição. Contas pertencentes a entidades de mídia afiliadas ao Estado, seus editores-chefes e/ou equipe de destaque podem receber etiquetas.	Entidades nacionais/federais administradas pelo governo, como agências, ministérios e secretarias; Entidades do governo estadual/provincial e local; Candidatos e representantes eleitos em nível federal/nacional; Funcionários do governo no nível federal/nacional, como ministros e embaixadores; Porta-vozes oficiais ou membros do quadro superior de um candidato no nível nacional/estadual ou um representante eleito/nomeado. Por exemplo: chefes de gabinete, diretores de campanha ou diretores digitais; Porta-vozes oficiais, membros do quadro superior ou líderes executivos de um partido político. Por exemplo: presidentes de partidos ou diretores de finanças; Partidos políticos; Membros de famílias reais com competências oficiais no governo; Associações políticas juvenis (para grandes partidos políticos a critério da política pública regional); Antigos chefes de estado e/ou chefes de governo; Comitês de ação política (PACs, na sigla em inglês) ou equivalentes específicos do país; Candidatos e representantes eleitos nos níveis estadual/provincial e local, conforme determinado por política pública regional baseada em fatores de mercado; Funcionários do governo nos níveis estadual/provincial e local, conforme determinado por política pública regional baseada em fatores de mercado	Não há política específica para as contas governamentais, mas uma plataforma de recursos para governo e organizações não-governamentais, contemplando grupo de defesa de interesses sociis, agência, representantes eleitos, órgãos reguladores de eleições; organizações governamentais; organizações intergovernamentais, organizações sem fins lucrativos e organizações políticas e candidatos a cargos públicos	N/A

Política/Plataforma	X	TikTok	Meta	Google - Youtube
Há limitações para essas contas?	não	sim	não	não
Quais restrições?	N/A	Restrição e/ou proibição nos campos: Programas de incentivo e recursos de monetização para criadores; Publicidade; Arrecadação de fundos para campanhas (eleições); Música; Diferente aplicação de moderação de conteúdo	N/A	N/A
Há sinalização nessas contas?	sim	sim	não	não
Detalhamento	De acordo com O Globo, a AFP notificou em abril de 2023 que o Twitter retirou uma série de etiquetas de contas de mídias governamentais quando da compra da plataforma por Musk. Alguns casos de retirada da etiqueta são: Rádio Pública dos Estados Unidos (NPR), da agência de notícias oficial chinesa Xinhua, da russa RT, da canadense CBC e da espanhola RTVE.	Embora muitas contas políticas já tenham adicionado o selo verificado ao seu perfil, atualmente isso é opcional. A partir de hoje nos EUA, testaremos a verificação obrigatória de contas pertencentes a governos, políticos e partidos políticos durante as eleições intercalares.	N/A	N/A

6. DETECÇÃO E CORREÇÃO DE CONTEÚDO NOCIVO



Segundo Silva e Cesar (2022)[8], a moderação de conteúdo é uma atividade de intervenção das plataformas para filtrar informações postadas por usuários impedindo a sua publicação ou reduzindo o alcance daqueles que violam os termos de uso e políticas da empresa. Para isso, são empregadas diferentes estratégias de moderação de conteúdo. Na tabela a seguir, analisamos o campo de detecção automática de conteúdo e revisão humana, ambas medidas adotadas nas plataformas para a detecção e correção de conteúdo. Nota-se que **todas as plataformas possuem mecanismos de detecção automatizada e revisão por equipes humanas de conteúdo que viole as suas Políticas de Segurança.**

8 https://revista.ibict.br/liinc/article/view/6080/5719

TABELA 6

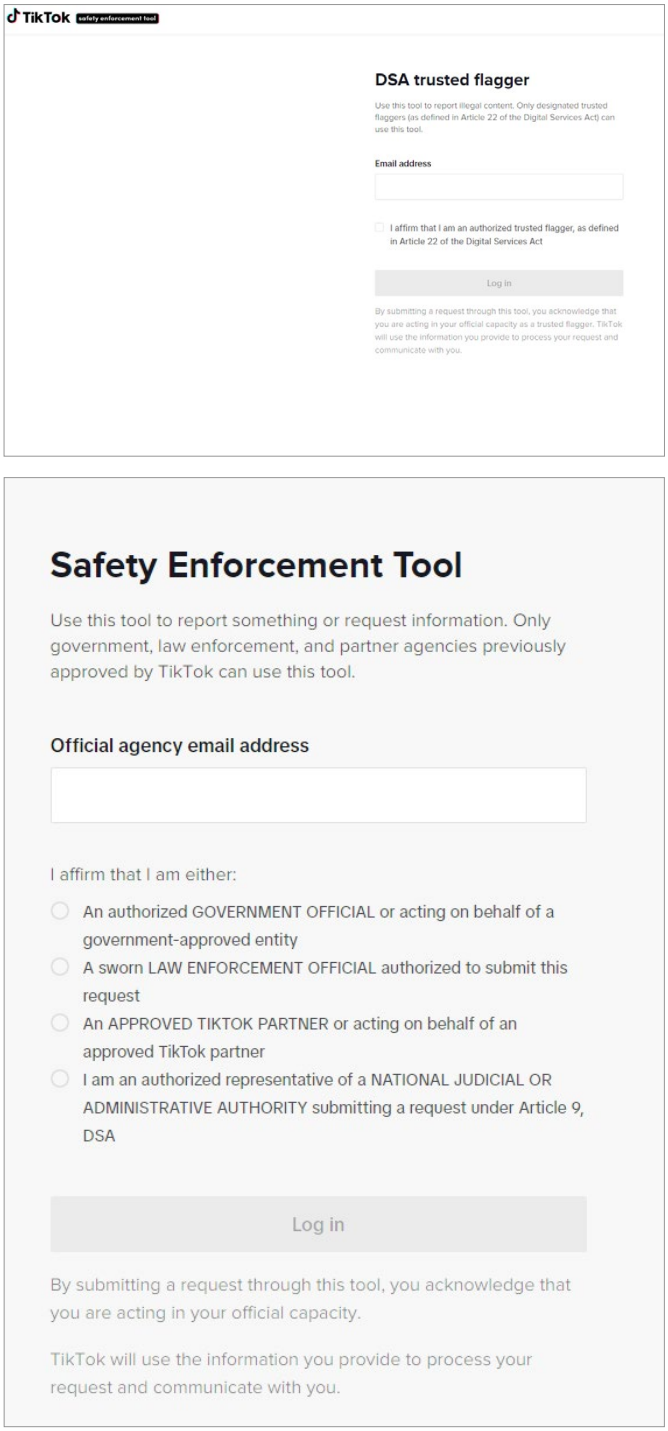
DETECÇÃO E CORREÇÃO DE CONTEÚDO NOCIVO				
Política/Plataforma	**X**	**TikTok**	**Meta**	**Google - Youtube**
Há detecção automática de conteúdo?	sim	sim	sim	sim
Há revisão humana de conteúdo?	sim	sim	sim	sim
Quais os processos de revisão?	O X tem um canal próprio de denúncias feitas diretamente por usuários da plataforma. Além disso, a plataforma segue políticas de privacidade.	Redirecionamento de resultados de pesquisa e hashtags para nossas Diretrizes da Comunidade [sistema automatizado]	1) Detecção Proativa; 2) Automação com IA; 3) Priorização. A meta tem um esquema divido por etapas, onde a Inteligência Artificial prioriza questões como as políticas e diretrizes da rede social, bem como tem uma espécie de "priorização" a partir de fatores como: viralidade, gravidade de dano e probabilidade de violação	1) Políticas de Spam; 2) Sistemas automatizados; 3) Revisão humana.

7. TRUSTED FLAGGERS - SINALIZADORES CONFIÁVEIS

De acordo com a European Commission: “Os sinalizadores confiáveis são entidades especiais sob o DSA. Eles são especialistas em detectar certos tipos de conteúdo ilegal online, como discurso de ódio ou conteúdo terrorista, e notificá-lo às plataformas online. Os avisos enviados por eles devem ser tratados com prioridade, pois se espera que sejam mais precisos do que os avisos enviados por um usuário médio.”[9] É responsabilidade do Coordenador de Serviços Digitais (CSD) do Estado-Membro de estabelecimento da entidade requerente atribui o estatuto de sinalizador de confiança. Os CSDs supervisionam o processo de inscrição, garantindo que as entidades atendam aos critérioss de perícia e competência: os sinalizadores confiáveis devem demonstrar experiência e competência especiais na detecção, identificação e notificação de conteúdo ilegal on-line; independência: os sinalizadores confiáveis devem operar independentemente de quaisquer provedores de plataformas on-line para garantir que suas avaliações sejam imparciais, diligência, precisão e objetividade: sinalizadores confiáveis devem trabalhar de forma diligente, precisa e objetiva, seguindo padrões e procedimentos estabelecidos. Apenas as entidades sediadas na UE podem candidatar-se ao estatuto de sinalizador de confiança. O estatuto de sinalizador de confiança é válido em toda a UE, em relação a qualquer plataforma em linha abrangida pelo artigo 22.º do DSA, independentemente dos locais de estabelecimento. De acordo com a Tabela sobre Sinalizadores Confiáveis, apenas o TikTok possui um portal de fácil acesso para o envio de mensagens e denúncias conforme estabelecido pela política.

9 Sinalizadores confiáveis sob a Lei de Serviços Digitais (DSA) | Moldar o futuro digital da Europa

Figura 6. *Interface do TikTok com canal de denúncia.*



A Google, por outro lado, possui um campo explicativo sobre os “Trusted Flaggers”, enquanto o X indica que aguarda a liberação da lista de Sinalizadores Confiáveis pela União Europeia. Não foi possível encontrar na plataforma da Meta informações sobre a sua política específica para a questão dos sinalizadores.

TABELA 7

TRUSTED FLAGGERS / SINALIZADORES CONFIÁVEIS				
Política/Plataforma	**X**	**TikTok**	**Meta**	**Google - Youtube**
Possui política de sinalizadores confiáveis?	não	sim	não	sim
Link da ferramenta/ informações	Segundo o relatório do DSA de abril de 2024, assim que as informações sobre sinalizadores confiáveis concedidas pelo Artigo 22 do DSA forem publicadas, serão inscritos no programa de sinalizadores confiáveis, o que garante a priorização da revisão humana.	Há ferramenta de aplicação de segurança para denunciar algo ou solicitar informações. Apenas governo, autoridades policiais e agências parceiras previamente aprovadas pelo TikTok podem usar essa ferramenta. [link]	N/A	A plataforma indica que na União Europeia, entidades nacionais chamadas Coordenadores de Serviços Digitais designaram "Trusted Flaggers", que são entidades encarregadas de sinalizar conteúdo supostamente ilegal em nossas plataformas. É provável que os Sinalizadores Confiáveis tenham experiência em um ou mais campos relevantes para a moderação de conteúdo, como privacidade ou segurança infantil. Priorizamos solicitações de Sinalizadores Confiáveis de acordo com a Lei de Serviços Digitais da UE. A Comissão Europeia manterá uma lista de Sinalizadores de Confiança designados numa base de dados acessível ao público. O Programa de Notificações Prioritárias ajuda a fornecer ferramentas avançadas a órgãos governamentais e ONGs. Essas organizações são particularmente eficientes em informar o YouTube sobre conteúdo que viola nossas diretrizes da comunidade. O Programa inclui: um formulário da Web que os órgãos governamentais e as ONGs podem usar para entrar em contato diretamente com o YouTube; visibilidade nas decisões sobre conteúdo denunciado; análises prioritárias de sinalizações para ação mais rápida; discussão e feedback contínuos sobre as áreas de conteúdo do YouTube; treinamentos on-line ocasionais. Os candidatos ideais têm experiência em pelo menos uma categoria de políticas; sinalizam conteúdo frequentemente com uma taxa de precisão alta; estão abertos a discussões e feedback contínuos sobre as áreas de conteúdo do YouTube. Algumas organizações estão sujeitas a uma análise mais detalhada, inclusive as de países/regiões onde há um histórico de violações dos direitos humanos ou de repressão à liberdade de expressão.

8. MEDIDAS ESPECÍFICAS POR PLATAFORMA - TERRITÓRIO

Dentre algumas ferramentas elencadas neste relatório, destacamos as políticas para menores de idade no X, que possui regras específicas para a Austrália, e no TikTok disponível nos Estados Unidos, o qual possui um Modo Restrito com proteções adicionais para crianças menores de 13 anos. Também destacamos o Portal de Segurança Europeia do TikTok.

Figura 8. *Interface do TikTok com Portal de Segurança Europeia.*

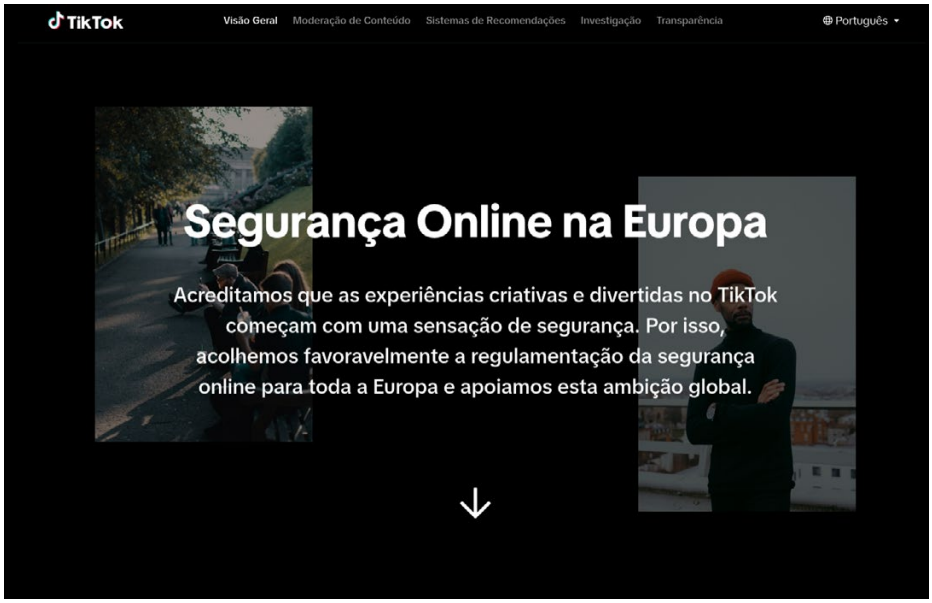


TABELA 8

MEDIDAS/FERRAMENTAS DE SEGURANÇA ESPECÍFICAS POR PLATAFORMA		
X		
Plataforma	**Medida**	**Disponível em quais territórios?**
Notas comunitárias	Notas Comunitárias - forma colaborativa de adicionar contexto útil aos posts e manter as pessoas mais bem-informadas O programa tem como objetivo criar um mundo mais bem-informado, capacitando pessoas no X para adicionar colaborativamente notas úteis a posts que possam ser enganosos. Requisitos para participar: Nenhuma violação de Regras do X desde 1º de janeiro de 2023; Entrou para o X há pelo menos 6 meses; Um número de telefone verificado (o número de celular é de uma operadora de celular confiável e não está associado a outras contas).	Há colaboradores em 69 países, o último a ser incluído foi a Índia. O Brasil faz parte do Programa.
Política para menores	Para crianças australianas com menos de 16 anos que se inscrevem no X, há medidas de segurança. Alguns recursos de segurança e privacidade são: As contas pertencentes a menores conhecidos terão como padrão "Postagens protegidas"; as contas pertencentes a menores de idade conhecidos estarão restritas a receber mensagens diretas de contas que eles seguem por padrão; proibição do acesso de menores de idade a mídias sensíveis; os usuários podem bloquear contas instantaneamente; podem silenciar uma conta se não quiserem ver suas postagens, mas não quiserem deixar de seguir a conta, podem escolher quem poderá responder às suas postagens quando postadas. A posição padrão é que todos possam responder, mas há opções para desativar todas as respostas ou permitir que apenas as contas mencionadas na postagem respondam. Um usuário também pode alterar quem pode responder às suas postagens ou desativar as respostas após a publicação da postagem. O X permite que os usuários denunciem outros usuários que eles acreditam ter menos de 13 anos via um formulário de denúncia dedicado para permitir que qualquer usuário denuncie uma conta que eles suspeitam estar sendo usada por um menor de 13 anos. X tem uma política que proíbe os usuários de promover ou encorajar o suicídio ou a automutilação. Quando alguém pesquisa termos associados a suicídio ou automutilação, o principal resultado da pesquisa é uma notificação que o incentiva a pedir ajuda.	Austrália
Desinformação na França	Informações falsas sobre a votação ou o registro para votar - As Regras do X proíbem a publicação de conteúdo que forneça informações falsas sobre a votação ou o registro para a votar. Se você denunciar esse tipo de conteúdo, analisaremos o post denunciado. Se determinarmos que o post viola nossas políticas, tomaremos providências (desde a exigência da retirada do conteúdo proibido até a suspensão permanente da conta). Você receberá uma notificação de acompanhamento de nossa parte se precisarmos de informações adicionais ou se tomarmos providências com relação ao post denunciado. Informações falsas podem alterar o resultado da votação ou perturbar a ordem pública. A legislação francesa exige que também ofereçamos um meio de você denunciar falsas informações que possam alterar o resultado da votação ou perturbar a ordem pública	França

TIKTOK		
Plataforma	**Medida**	**Disponível em quais territórios?**
Modo para menores de 13 anos	Nos Estados Unidos, há uma experiência separada do TikTok para menores de 13 anos que oferece proteções adicionais, que incluem restrição de recursos interativos, avaliações de adequação de conteúdo da Common Sense Networks e uma Política de Privacidade dedicada. Ao criar uma conta nos Estados Unidos com uma data de nascimento que mostre que você tem menos de 13 anos, você entrará automaticamente nessa experiência.	Estados Unidos
Restrição para usuários de até 16 anos	Estabelecer requisitos de idade mínima para acesso a determinados recursos do produto, incluindo ter 16 anos ou mais para que seu conteúdo seja elegível para o feed “Para você”	Geral
Modo Restrito	O Modo restrito limita a exposição a conteúdo que pode não ser adequado para todos — por exemplo, por incluir temas maduros ou complexos. Alguns recursos estão indisponíveis no Modo restrito, como o feed “Seguindo” e Presentes em LIVEs. Os pais e responsáveis também podem ativar o Modo restrito para seus filhos adolescentes através do Pareamento familiar.	Geral
Portal de Segurança Europeia	Portal com informações e recursos de transparência e segurança da UE. Informações sobre a moderação de conteúdo, sistema de recomendação, pesquisa, transparência e segurança para usuários adolescentes.	União Europeia

AVISOS NO X[10]

No X, pode haver a notificação a uma conta ou a um post para dar mais contexto sobre as ações que nossos sistemas ou equipes podem tomar.

- Avisos em posts - Nossos sistemas e equipes podem adicionar avisos aos posts para fornecer mais contexto ou um aviso antes de você clicar.
- Conteúdo com restrição de idade
- Ocultação de um post em violação por um pequeno aviso

O X não escreve, não avalia nem faz a moderação das notas comunitárias (a não ser que elas violem as regras do X). Acreditamos que permitir que as pessoas façam essas escolhas em conjunto é uma forma eficiente e justa de adicionar informações que ajudem as pessoas a ficarem mais bem informadas.

O programa se pauta pela transparência: todas as colaborações são publicadas diariamente, e nosso algoritmo de avaliação pode ser inspecionado por qualquer pessoa.

Musk anunciou que posts corrigidos pelo Community Notes serão desmonetizados, como um incentivo contra a desinformação.

Figura 9. *Printscreen tweet de Elon Musk*

10 Notificações da conta do X e seus significados - suspensões e mais (twitter.com)

CONSIDERAÇÕES FINAIS

No cenário contemporâneo, onde a interconexão digital permeia todos os aspectos da vida, a segurança nas plataformas digitais emerge como uma preocupação central. Este relatório visa analisar as políticas e medidas de segurança implementadas em diversas plataformas digitais, abordando os desafios enfrentados e as soluções propostas para garantir a proteção dos usuários e a integridade dos dados em um ambiente virtual em constante evolução.

Com base na análise das políticas de segurança das plataformas examinadas, é evidente que cada uma delas aborda questões específicas de maneira diferente, demonstrando variados níveis de comprometimento com a proteção dos usuários e a mitigação de danos potenciais. Destacamos que o YouTube (Google) inclui referências a autores de ataques em sua política de organizações extremistas, mas carece de especificações sobre as consequências para as contas de usuários envolvidas em tais atos.

É notável que a Meta se destaca ao considerar grupos vulneráveis, como refugiados e migrantes, em suas políticas de proteção contra ataques graves, enquanto outras plataformas devem seguir esse exemplo. Recomendamos que o X amplie sua definição de discurso violento para incluir desumanização, à semelhança das outras plataformas.

Além disso, a falta de políticas específicas para negacionismo de eventos históricos, protocolos de crise e contas políticas ou de mídia afiliadas ao Estado em algumas plataformas sugere áreas para melhoria. A disponibilidade de recursos como portais de segurança e modos restritos para proteger grupos específicos, como crianças, destaca-se como uma prática positiva do TikTok, que outras plataformas podem considerar em seus próprios sistemas de segurança, expandindo-as globalmente, sem restrição a um país.

Concluímos, portanto, que há oportunidades para aprimorar as políticas de segurança em todas as plataformas analisadas, buscando garantir um ambiente online mais seguro e inclusivo para todos os usuários.

OBSERVATÓRIO DA DESINFORMAÇÃO

SLEEPING GIANTS BRASIL