Moderação de conteúdo, discurso de ódio e desinformação

Contribuições do IRIS à tomada de subsídios da Advocacia-Geral da União

INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE

Moderação de conteúdo, discurso de ódio e desinformação

Contribuições do IRIS à tomada de subsídios da Advocacia-Geral da União

AUTORIA
Fernanda dos Santos Rodrigues Silva

REVISÃO
Ana Bárbara Gomes Pereira
Luiza Correa de Magalhães Dutra

PROJETO GRÁFICO, CAPA, DIAGRAMAÇÃO E FINALIZAÇÃO
Felipe Duarte
Imagem de capa: Freepik

PRODUÇÃO EDITORIAL
IRIS - Instituto de Referência em Internet e Sociedade

COMO REFERENCIAR EM ABNT
SILVA, Fernanda dos Santos Rodrigues. **Moderação de conteúdo, discurso de ódio e desinformação**: Contribuições do IRIS à tomada de subsídios da Advocacia-Geral da União. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2025. Disponível em: https://bit.ly/3Eft3MS. Acesso em: dd mmm aaaa.

INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE

DIREÇÃO

Ana Bárbara Gomes

Paloma Rocillo

MEMBROS

Felipe Duarte | Coordenador de Comunicação

Fernanda Rodrigues | Coordenadora de Pesquisa e Pesquisadora

Glenda Dantas | Pesquisadora

Júlia Caldeira | Pesquisadora

Júlia Tereza Koole | Estagiária de pesquisa

Luísa Melo | Estagiária de pesquisa

Luiza Correa de Magalhães Dutra | Pesquisadora

Paulo Renã da Silva Santarém | Pesquisador

Thais Moreira | Analista de comunicação

Wilson Guilherme | Pesquisadore

irisbh.com.br

1. Contextualização

Em pronunciamento feito no dia 7 de janeiro de 2025, a Meta anunciou mudanças profundas na moderação de conteúdo em suas plataformas. Defendendo "mais discurso e menos erros", destacou que os esforços por moderar mais e melhor estariam prejudicando usuários inocentes, com a remoção indevida de conteúdo. Assim, enquanto a moderação de conteúdo automatizada seguirá utilizada para casos de "conteúdo ilegal e violações de alta severidade, como terrorismo, exploração sexual infantil, drogas, fraudes e scams", as demais infrações às políticas de comunidade serão analisadas primeiramente a partir de denúncias. A intenção, segundo nota da empresa, é chegar a uma convicção melhor embasada de que o conteúdo é realmente violador, antes de proceder a sua remoção.

Nesse sentido, diferentes restrições sobre temas como imigração, identidade de gênero e sexualidade foram levantadas, uma vez que, afirma, "não é certo que coisas possam ser ditas na TV ou no plenário do Congresso, mas não em nossas plataformas". Ocorre que, desde então, a atualização das diretrizes demonstra um posicionamento que retrocede no reconhecimento de direitos básicos de grupos minoritários, como é o caso da população LGBTQIAPN+. Para esta, as novas diretrizes passaram a autorizar, por exemplo, "alegações de doença mental ou anormalidade quando baseadas em gênero ou orientação sexual, considerando discursos políticos e religiosos sobre transgenerismo e homossexualidade, bem como o uso comum e não literal de termos como 'esquisito'". Além disso, a nova política também permite "conteúdo que defenda limitações baseadas em gênero para empregos militares, policiais e de ensino", além de conteúdo semelhante "relacionado à orientação sexual, desde que fundamentado em crenças religiosas".

A empresa afirmou, ainda, sobre a adoção de mecanismo similar de Notas da Comunidade aplicado pela plataforma X, para anunciar o fim da parceria com agências de checagem para verificação de notícias. Embora esta alteração em específico valha somente para os EUA neste momento, onde a regra será testada, a ideia é que haja usuários colaboradores que escrevem e avaliam o conteúdo postado.

Assim, a presente contribuição tem por objetivo sistematizar e complementar os pontos trazidos por ocasião da participação do IRIS em audiência pública promovida pela Advocacia-Geral da União, no dia 22 de janeiro de 2025, acerca das alterações nas políticas de comunidade de plataformas da Meta. São apresentados os achados mais recentes de pesquisas do Instituto sobre moderação de conteúdo, a fim de contribuir para a elaboração de soluções.

2. Os principais problemas da moderação de conteúdo em redes sociais pela perspectiva dos usuários

Segundo pesquisa inédita conduzida pelo IRIS em 2024, as principais reclamações dos usuários de redes sociais sobre o procedimento de moderação de conteúdo se dividem da seguinte forma:

- **54,34%** são reclamações sobre o **procedimento de moderação no caso de remoção de postagens e suspensão/bloqueio de contas.** As principais razões da reclamação nesse tópico são:

  A. fundamentação inadequada de decisões de moderação (52,46%);

  B. contestação de decisão de moderação não respondida (22,54%);

  C. falta de notificação ou aviso sobre decisão de moderação (9,02%);

  D. ausência de ferramentas para contestar a decisão de moderação (7,38%);

  E. design da plataforma inacessível em relação aos mecanismos de revisão de decisão de moderação (4,51%);

  F. outros (4,1%)

- Os outros **45,66%** correspondem a reclamações sobre:

  A. reclamações genéricas sobre a moderação de conteúdo das plataformas (36,1%);

  B. pedidos de moderação de conteúdo de terceiro (30,24%);

  C. problemas com monetização (10,24%);

  D. problemas com a recomendação ou o alcance do conteúdo (10,24%);

  E. restrição de funcionalidades na plataforma em virtude de decisão de moderação (10,24%);

  F. problemas gerados por limitação etária da plataforma (1,95%);

  G. pedidos de moderação de anúncios (1,46%)

O cenário acima demonstra que **a própria forma com que a moderação de conteúdo tem sido conduzida pelas plataformas digitais** é a principal reclamação dos usuários. Decisões que não apontam com exatidão a razão para a remoção da postagem ou conta, a falta de retorno sobre os pedidos de revisão e a falta de notificação sobre uma remoção ou suspensão, são alguns dos principais motivos de insatisfação e desconfiança dos usuários.

No entanto, **afrouxar a moderação automatizada sobre vários conteúdos, incluindo discurso de ódio, não é a solução para esses problemas**. Pelo contrário, autorizar a circulação de mais conteúdo nocivo pode gerar o efeito rebote de **mais ações judiciais solicitando a remoção de contas e postagens**. Enquanto hoje os pedidos para moderação de conteúdo de terceiro representam uma pequena parcela das reclamações analisadas, esse número pode subir significativamente a partir das alterações da política de plataforma.

3. Propostas recomendadas para melhorar a moderação de conteúdo em plataformas digitais

3.1. A aplicação de um direito ao devido processo na moderação de conteúdo

Considerando haver uma percepção tanto por parte da Meta quanto dos usuários de que a moderação de conteúdo em redes sociais tem sido deficiente, o que propomos é a aplicação de um **direito ao devido processo na moderação de conteúdo.**

O que significa um direito ao devido processo na moderação de conteúdo?

O IRIS define o direito ao devido processo aplicado à moderação de conteúdo como um conjunto de mecanismos e procedimentos voltados a legitimar o processo de gerenciamento de conteúdo de terceiro, a ser disponibilizado pelas plataformas digitais, e seu modo de exibição e recomendação.

Como aplicar um direito ao devido processo na moderação de conteúdo?

Para verificarmos a aplicação de um direito ao devido processo na moderação de conteúdo é necessário, pelo menos, a implementação de 5 elementos:

1. **Obrigatoriedade de fundamentação sobre decisão de moderação:** as decisões de moderação de conteúdo devem apontar com exatidão de que forma a postagem ou conta violou os termos da política de comunidade e qual o dispositivo violado especificamente.

2. **Notificação ao usuário sobre decisão de moderação:** os usuários sempre devem ser notificados acerca de decisões de moderação de conteúdo, contendo informações sobre como recorrer e o prazo para uma resposta pela plataforma.

3. **Definição de prazos procedimentais:** a plataforma deve estabelecer prazos para que o usuário moderado recorra da decisão, bem como para que ela mesma responda aos recursos. Este prazo observar a proporcionalidade e razoabilidade.

4. **Recursos à decisão de moderação:** a plataforma deve estabelecer mecanismos acessíveis e intuitivos para contestação e solicitação de revisão de decisões de moderação pelos usuários.

5. **Design e acessibilidade a ferramentas relacionadas à moderação de conteúdo:** a plataforma deve oferecer ao usuário mecanismos para acompanhamento de solicitações de revisão de decisões de moderação e para realização de denúncias de conteúdo nocivo.

Os itens acima são fundamentais para assegurar a legitimidade do procedimento e para que usuários possam se sentir mais seguros e confiantes em relação ao trabalho da plataforma.

3.2. Aprovação de uma regulação para plataformas digitais e uma regulação para sistemas de inteligência artificial

Atualmente, **o Brasil não possui uma regulação que aborde de forma direta a moderação de conteúdo.** O Marco Civil da Internet, dentre outros pontos, regula somente a forma de responsabilização de provedores de aplicação de internet (o que, frisa-se, não se restringe somente a redes sociais) no caso de conteúdo publicado por terceiro. No entanto, nada é dito sobre como a moderação deve acontecer.

Nesse sentido, **entendemos como fundamental que a regulação de plataformas digitais seja uma das pautas prioritárias do Governo Federal e Congresso Nacional**. O vácuo legislativo abre margem para que as plataformas adotem suas próprias regras de moderação, sem parâmetros mínimos a serem observados. O Projeto de Lei 2.630/2020, popularmente conhecido como PL das Fake News, aguarda andamento na Câmara dos Deputados, após a criação de um Grupo de Trabalho específico para a elaboração de um substitutivo. O seu texto previa, inclusive, um **capítulo para o devido processo**

na moderação de conteúdo, que poderia ser melhorado, mas já representava uma importante inovação. Além disso, a perspectiva de **dever de cuidado** poderia ser essencial para um compromisso maior das plataformas digitais em torno de evitar a circulação de conteúdo nocivo em seus ambientes digitais, por meio da adoção de estratégias sistêmicas que fossem devidamente avaliadas e monitoradas por autoridade competente.

O recente caso da Meta, assim como de tantos outros problemas envolvendo o uso malicioso das redes sociais (como os ataques antidemocráticos de 8 de janeiro de 2022 e os ataques a escolas coordenados pela internet), demonstram que **a necessidade de uma regulação tem sido cada vez mais latente.** Assim, é preciso que medidas urgentes sejam adotadas, a fim de evitar o escalonamento de problemas no futuro.

Ainda, em se tratando a moderação de conteúdo de um procedimento em que é largamente empregado o uso de sistemas algorítmicos, é essencial que se avance em uma **regulação de inteligência artificial que também dê conta de parâmetros mínimos para essas ferramentas.** Em discussão recente em torno do Projeto de Lei 2.338/2023, voltado a ser o marco regulatório da inteligência artificial no Brasil, sistemas de recomendação de conteúdo, como aqueles empregados em redes sociais, chegaram a constar na lista de tecnologias de alto risco, o que significava a adoção de medidas de governança mais rígidas, a fim de evitar potencialidades negativas. No entanto, após intensos debates, a redação foi alterada e essa previsão, excluída.

Ocorre que para promover um uso e desenvolvimento responsável da IA é relevante que **diretrizes mínimas sejam estabelecidas e que disposições mais criteriosas sejam aplicadas para sistemas que podem causar maior impacto comprovado sobre a sociedade e grupos afetados**. Assim, **o avanço do Projeto de Lei 2.338/2023 também se destaca como uma das ações indispensáveis** para promover maior transparência e *accountability* em torno de sistemas

INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE