REDE
DE PESQUISA
EM GOVERNANÇA
DA INTERNET

# Anais do Encontro Anual da Rede de Pesquisa em Governança da Internet

Maio de 2024

# VII ENCONTRO DA REDE DE PESQUISA EM GOVERNANÇA DA INTERNET - REDE 2024

## FICHA TÉCNICA


### ANAIS DA REDE DE PESQUISA EM GOVERNANÇA DA INTERNET-VOL. 7, ISSN 2675-1690

**Editores**

Diego Vicentin
Hemanuel Jhosé Alves Veras
Maria Vitoria Pereira de Jesus


**Autores**

Andressa de Bittencourt Siqueira
Ane Carine Meurer
Diná Santana Santos
Elen Nas
Elizabeth Machado Veloso
Henrique S. Xavier
Isabelle Brito Bezerra Mendes
Jeiel de Santana Barbosa
João Araújo Monteiro Neto
Josiane Ribeiro
Kauã Arruda Wioppiold
Lucas Zanotto
Luis Henrique de Menezes Acioly
Luísa Carli de Lacerda
Matheus Fernandes da Silva
Pedro Odebrecht Khauaja
Sérgio Braga
Tatiana Nascimento Heim
Yago Rabelo Daltiba

### COMITÊ ORGANIZADOR

**Núcleo de Coordenação da Rede de Pesquisa em Governança da Internet**

Alexandre Arns Gonzales
Carolina Batista Israel
Diego Vicentin
Fernanda R. Rosa
Hemanuel Jhosé Alves Veras
Kimberly de Aguiar Anastácio
Maria Vitoria Pereira de Jesus
Nathan Paschoalini Ribeiro Batista
Rodolfo Avelino

**Comitê Científico**

Alexandre Arns Gonzales
Ana Paula Camelo
Carolina Batista Israel
Daniela Araújo
Diego Vicentin
Fernanda R. Rosa
Flávio Rech Wagner
Gills Vilar-Lopes
Jean Ferreira dos Santos
Leonardo Ribeiro da Cruz
Maria Mónica Arroyo
Martha Mourão Kanashiro
Rafael Evangelista
Raquel Gatto
Rodolfo Avelino
Sergio Marcos de Ávila Negri
Rodrigo Firmino

**Moderadores**

Daniela Araújo
Diego Vicentin
Hemanuel Jhosé Alves Veras
Laura Conde Tresca

**Debatedores**

Diná Santana Santos
Elen Nas
Elizabeth Machado Veloso
Henrique S. Xavier
Jeiel de Santana Barbosa
Josiane Ribeiro
Kauã Arruda Wioppiold
Lucas Zanotto
Luísa Carli de Lacerda
Matheus Fernandes da Silva
Pedro Odebrecht Khauaja
Sérgio Braga
Tatiana Nascimento Heim
Yago Rabelo Daltiba


O VII Encontro da Rede de Pesquisa em Governança da Internet e a publicação dos Anais resultantes deste encontro receberam apoio de: Comitê Gestor da Internet no Brasil - CGI.br, Internet Society - Capítulo Brasil e do Programa de Pós-graduação em Ciências Sociais da Universidade Estadual de Campinas (Unicamp), através de recursos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)


Internet Society Capítulo Brasil

CAPES

cgi.br


# DETECÇÃO AUTOMATIZADA DE CONTEÚDO INFRINGENTE NO COMITÊ DE SUPERVISÃO DA META: LIÇÕES E PERSPECTIVAS PARA A MODERAÇÃO DE CONTEÚDO

Andressa de Bittencourt Siqueira

## SUMÁRIO


# DETECÇÃO AUTOMATIZADA DE CONTEÚDO INFRINGENTE NO COMITÊ DE SUPERVISÃO DA META: LIÇÕES E PERSPECTIVAS PARA A MODERAÇÃO DE CONTEÚDO

Andressa de Bittencourt Siqueira[1]

**RESUMO**
Os sistemas automatizados de detecção de conteúdo se tornaram uma ferramenta poderosa para identificar rapidamente o conteúdo infringente, uma vez que os provedores de plataformas digitais têm gradualmente assumido um papel mais ativo na moderação do discurso frente ao grande volume de conteúdos gerados por usuários. Assim, a estrutura de tomada de decisão do Comitê de Supervisão da Meta (doravante Comitê) foi escolhida como objeto deste estudo, com base na seleção de 29 “casos” (do inglês, *case decisions*) e “opiniões consultivas” (do inglês, *policy opinion*), ambos doravante denominados conjuntamente como “decisões”, publicadas pelo Comitê até 31 de março de 2024, envolvendo sistemas automatizados de moderação de conteúdo. Assim, surge o seguinte problema de pesquisa: como a estrutura decisória do Comitê pode contribuir para demonstrar como os sistemas automáticos de detecção de conteúdos infringentes podem ser aprimorados de modo a não causar ingerências indevidas à liberdade de expressão ou a outros direitos fundamentais colidentes? Com base nos resultados, é possível agrupar as recomendações do Board em dois grandes grupos: (i) melhorias quanto à transparência; (ii) melhorias procedimentais.

**PALAVRAS-CHAVE**
Liberdade de expressão; Plataformas digitais; Bloqueio de conteúdos; Sistemas automatizados.

**ABSTRACT**
Automated content detection systems have become a powerful tool for quickly categorizing infringing content, as online platform providers have gradually taken a more active role in moderating speech given the large volume of user-generated content. Therefore,

[1] Doutoranda e Mestre no Programa de Pós-Graduação em Direito da PUCRS. Pesquisadora no Grupo de Estudos e Pesquisas em Direitos Fundamentais (GEDF/CNPq). Bolsista CAPES. Advogada.

the decision-making structure of the Meta's Oversight Board (hereinafter "the Board") was chosen as the subject of this study, based on the selection of 29 case decisions and policy opinions (both hereinafter collectively referred to as "decisions") issued by the Board as of March 31st, 2024, involving automated content moderation systems. Thus, the following research question arises: How can the decision-making structure of the Board contribute to demonstrating how automated systems for the detection of infringing content can be improved so that they do not lead to undue interference with freedom of expression or other conflicting fundamental rights? Based on the results, it is possible to categorize the Board's recommendations into two main groups: (i) improvements in transparency; (ii) procedural improvements.

**KEY WORDS**
Freedom of expression; Digital platforms; Content blocking; Automated systems.

## INTRODUÇÃO

Os sistemas automatizados de detecção de conteúdos tornaram-se uma ferramenta poderosa para classificar rapidamente os conteúdos ilícitos, uma vez que os provedores de plataformas online têm gradualmente assumido um papel mais ativo na moderação do discurso. O termo "infringente" é adotado uma vez que os provedores podem restringir o âmbito dos conteúdos permitidos nas respectivas plataformas que gerenciam, de modo que a expressão engloba conteúdos ilícitos, proibidos pela legislação, e conteúdos não autorizados, proibidos pelas regras internas das plataformas.

Nesse cenário, a estrutura de tomada de decisão do Comitê de Supervisão da Meta (doravante Comitê) foi escolhida como objeto deste estudo, com base na seleção de 29 "casos" (do inglês, *case decisions*) e "opiniões consultivas" (do inglês, *policy opinion*), ambos doravante denominados conjuntamente como "decisões", publicadas pelo Comitê até 31 de março de 2024, envolvendo sistemas automatizados de moderação de conteúdo.

Sendo assim, apresenta-se o seguinte problema de pesquisa: como a estrutura decisória do Comitê pode contribuir para demonstrar como os sistemas automáticos de detecção de conteúdos infringentes podem ser aprimorados de modo a não causar ingerências indevidas à liberdade de expressão ou a outros direitos fundamentais colidentes? Para responder ao problema de pesquisa proposto, foram estabelecidos três objetivos específicos: (i) identificar quais decisões envolvem casos de detecção automatizada de conteúdo, especialmente aquelas que incluem ou reproduzem reco-

mendações para o aprimoramento dos sistemas que realizam esse procedimento; (ii) estruturar as recomendações do Comitê quanto às decisões que envolvem detecção automatizada de conteúdo infringente de forma a permitir a avaliação de avanços e retrocessos na área de moderação de conteúdo por meio do uso de sistemas automatizados; e, por fim, (iii) explicar de que modo a tomada de decisão do Comitê demonstra como a moderação de conteúdo realizada mediante o uso de detecção automatizada para identificação de conteúdo infringente pode ser aprimorada.

A presente pesquisa utiliza em sua abordagem o método indutivo, tendo em vista que, com base no aporte decisório do Comitê sobre o emprego de sistemas automatizados para a identificação de conteúdo infringente, propõe-se uma sistematização dos argumentos adotados para solucionar controvérsias na moderação de conteúdo que envolvam o uso dessas ferramentas (FINCATO; GILLET, 2018). Adota-se, também, o método estruturalista de procedimento, uma vez que se analisa o fenômeno concretamente, para, após esse exercício, atingir um nível de abstração e, em seguida, retornar à análise concreta mediante uma realidade estruturada. A pesquisa também se encaixa no tipo de pesquisa explicativa, uma vez que além de coletar e organizar dados, almeja-se elucidar de que maneira o Comitê fundamenta suas decisões que envolvem o emprego de sistemas automatizados para a identificação de conteúdo infringente (FINCATO; GILLET, 2018).

Ao analisar a amostra de decisões do Comitê, é possível extrair algumas conclusões sobre como as plataformas têm usado a detecção automatizada de conteúdo para identificar conteúdo infringente e como a tecnologia pode ser aprimorada. Em suma, o estudo explora os riscos da adoção da inteligência artificial como uma solução para encontrar e combater conteúdos infringentes.

## O USO DE MÉTODOS DE DETECÇÃO AUTOMATIZADOS PARA COMBATER CONTEÚDO INFRINGENTE NA MODERAÇÃO DE CONTEÚDO ONLINE

### Os desafios da moderação frente ao volume de conteúdo gerado pelo usuário

A vasta quantidade de conteúdo gerado por usuários de plataformas digitais todos os dias é inegável e é por isso que o uso de ferramentas algorítmicas de moderação de conteúdo que “visam identificar, combinar, prever algum conteúdo com base em suas propriedades exatas ou características gerais” (MARTÍNEZ, 2023, p. 212)

está se tornando cada vez mais urgente. Portanto, provedores de plataformas digitais passaram a atuar mais ativamente contra conteúdos proibidos.

Nesse contexto, a moderação de conteúdo está se tornando, de acordo com Tarleton Gillespie, a nova *commodity* das plataformas (GILLESPIE, 2018). Dado o imenso volume de conteúdo gerado pelo usuário, a detecção e remoção automatizada de conteúdo, que remove rapidamente o conteúdo infringente, surge como uma solução – mas não sem reservas. Embora as ferramentas tenham sido desenvolvidas para identificar o conteúdo infringente, elas também podem identificar erroneamente o conteúdo legítimo como infringente, resultando em uma interferência mais intrusiva na liberdade de expressão dos usuários, consagrada no Art. 5, inciso IX da Constituição Federal (doravante CF).

Os provedores de plataformas digitais, inclusive, passaram a ser denominados como "guardiões" (*gatekeepers*) (CELESTE, 2019, p. 79) ou *custodians* (GILLESPIE, 2018, p. 209) e "novos governantes" (*new governors)* (KLONICK, 2017, p. 1662), do ambiente online, isto é, deixaram de ser neutros para se tornarem atuantes. Nesse contexto, o amplo poder dos provedores de plataformas é caracterizado, sobretudo, pelo desequilíbrio frente ao usuário isoladamente considerado (HARTMANN; SARLET, 2019, p. 99) e pelas medidas adotadas para desenvolvimento e manutenção do seu modelo de negócio. Uma vez reconhecido seu protagonismo, os provedores de plataformas digitais constantemente adaptam seus Termos de Uso e Padrões/Diretrizes da Comunidade, bem como desenvolvem procedimentos estruturados para moderar as postagens e perfis dos usuários (REINHARDT, 2020, p. 254), a exemplo do Comitê sob análise.

O Comitê de Supervisão da Meta: o contraste entre o pioneirismo e suas limitações

Popularmente conhecido como "Suprema Corte do Facebook", o Comitê é um exemplo pioneiro e paradigmático até hoje, pois tem origem em uma proposta de autorregulação da própria Meta em meados de 2019 (à época, ainda Facebook Inc.), e é executado por um painel independente de membros que tomam a decisão final sobre (i) como e em que medida a Meta vem aplicando corretamente suas políticas, valores e compromissos de direitos humanos no Facebook, Instagram e Threads, e (ii) se o Comitê mantém ou anula a decisão original da Meta contra conteúdos gerados por usuários nas plataformas gerenciadas pela companhia.

Apesar do caráter pioneiro do Comitê, que é considerado por alguns como “um dos projetos constitucionais mais ambiciosos da era moderna” (DOUEK, 2019, p. 3), é importante reconhecer suas limitações, em particular no que diz respeito à sua atuação em alguns casos representativos e às dificuldades de delinear padrões mínimos de proteção a serem aplicados pela Meta em nível global, especialmente em relação à liberdade de expressão.

Os desafios de desenvolver mecanismos processuais no campo da moderação de conteúdo são compreensíveis, mas a adoção de medidas para garantir a liberdade de expressão online sem violar outros direitos fundamentais é inevitável. Evelyn Douek afirma que há três razões pelas quais “um sistema altamente formalista de moderação de conteúdo” não funciona na prática: (i) tendo em vista que é impossível monitorar o procedimento na prática, (ii) os provedores não têm a intenção de estar vinculadas a precedentes, e (iii) isso isoladamente não cria uma cultura de confiança e legitimidade na moderação de conteúdo realizada por plataformas (DOUEK, 2022, p. 10).

Os provedores de plataformas digitais, por óbvio, não têm intenção em se autolimitar. Enquanto empresas, o seu escopo é manter o âmbito de alcance da livre iniciativa o mais amplo possível (Art. 170, CF). É exatamente por esse motivo, entre outros, que é necessário estabelecer parâmetros regulatórios coesos, que sirvam para equilibrar os tensionamentos existentes entre o âmbito de ação dos provedores de plataformas digitais e a proteção dos direitos fundamentais dos usuários.

Propostas no âmbito legislativo: avanços ou retrocessos?

Em que pese as ressalvas, que se deslocam para momento oportuno, tendo em vista que não serão aqui aprofundadas, a busca para amenização dos tensionamentos acima descritos é justamente o que vem sendo debatido no âmbito do Projeto de Lei (PL) n. 2.630/2020, popularmente conhecido como “PL das Fake News”, tendo em vista que o Marco Civil da Internet (Lei n. 12.965/2014) já não fornece, isoladamente, respostas aos desafios lançados no ambiente digital anos após a sua promulgação. Lançando o olhar sobre o que se passa em outras ordens jurídicas, são relevantes as inovações legislativas na Alemanha (com a *Netzwerkdurchsetzsungsgesetz* – NetzDG), no Reino Unido (com o *Online Safety Act* – OSA ) e, mais recentemente, na União Europeia (com o Digital Services Act – DSA, e o *Artifical Intelligence Act* – AI Act). Tais respostas já endereçadas para a resolução dos problemas envolvendo a mo-

deração de conteúdo são interessantes, especialmente na perspectiva de direito comparado, a fim de evitar regulações fragmentadas e colidentes. Ainda assim, ressalte-se que não é prudente, muito menos salutar, implementar tais inovações no ordenamento jurídico brasileiro sem qualquer filtragem. Deve-se buscar alternativas legislativas que resgatem o protagonismo brasileiro no cenário internacional, a fim de trazer respostas aos crescentes desafios desencadeados no ambiente digital, com os olhos voltados a solucionar problemas identificados no cenário, acima de tudo, brasileiro.

Quanto à moderação de conteúdo, é essencialmente importante levar em conta alguns aspectos. Conforme se verá adiante, o escopo de conscientizar a população sobre saúde, ataques terroristas, considerar as especificidades dos contextos locais e motivação do interlocutor, entre outros, são relevantes aspectos a serem levados em consideração na moderação de conteúdo, para não se classifique automática e erroneamente conteúdos como ofensivos. O cenário ora descrito torna-se sensível, tendo em vista tais aspectos, uma vez que os meios empregados pelas plataformas não são suficientes para a proteção do discurso de minorias sociais. A título de exemplo, usuários LGBTQIA+ atribuem significados diferentes a palavras que podem ser consideradas ofensivas em contexto diverso (i.e. ressignificação de palavras comumente consideradas ofensivas). Portanto, as publicações e os comentários LGBTQIA+ podem ser erroneamente rotulados como ofensivos se as tecnologias de moderação de conteúdo não perceberem adequadamente o contexto em que estão inseridos (OLIVA; ANTONIALLI; GOMES, 2021). Assim, torna-se evidente, desde logo, que a falta de transparência consiste em um dos obstáculos a serem superados para que consequências jurídicas devido a práticas de discriminação algorítmica tenham efeitos práticos (MENDES; MATTIUZZO, 2019, p. 47).

Com o uso de sistemas automatizados, é impossível hoje atingir um nível 0 de detecção e remoção indevida. É possível afirmar, portanto, que os sistemas automatizados são um “mal necessário” (MARTÍNEZ, 2023) para evitar a disseminação de conteúdo infringente. Contudo, é evidente que a liberdade de expressão passa a sofrer maior ingerência, uma vez que conteúdos legítimos são removidos todos os dias. Nesse sentido, as recomendações do Comitê lançam luzes sobre como os sistemas automatizados podem ser aprimorados para evitar derrubadas indevidas de conteúdo.

## MÉTODOS

A pesquisa está focada nas decisões do Comitê sobre o uso de sistemas automatizados para identificação de conteúdo infringente. Este estudo tem como objetivo estruturar as recomendações a partir da análise das políticas até então adotadas pela Meta, de modo que seja possível avaliar a viabilidade de tais medidas e suas possíveis melhorias. Por essa razão, são analisados os casos publicados pelo Comitê até 31 de março de 2024, envolvendo detecção automatizada de conteúdo.

Para desenvolver a pesquisa, foi acessado o site oficial do Comitê (CASE, 2024), onde são publicados os casos. Devido à falta de um mecanismo de busca entre as decisões, cada uma delas teve de ser aberta e analisada.

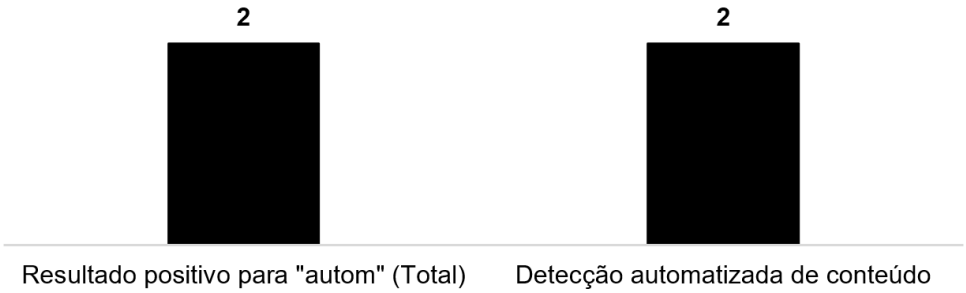
As decisões do Comitê são categorizadas de duas maneiras diferentes: (i) opiniões consultivas sobre políticas, quando a Meta faz perguntas ao Comitê para obter orientação sobre questões específicas a serem incorporadas ao processo de desenvolvimento de políticas da empresa; (ii) casos, quando o Comitê analisa as decisões de conteúdo da Meta para verificar a conformidade com suas próprias regras, valores e compromissos de direitos humanos, e o Comitê tem o direito de manter ou, em caso de não conformidade, anular a decisão da Meta em questões de moderação de conteúdo. Ambos denominados conjuntamente como “decisões”, conforme já referido no início.

Se a análise do caso ou da opinião consultiva confirmar a abordagem dos sistemas automatizados (resultado positivo), esses resultados são combinados em um formato de tabela, que descreve o tipo de decisão (caso ou opinião consultiva) (“TIPO”), a plataforma em que a controvérsia ocorreu (“PLATAFORMA”), as recomendações sobre ferramentas automatizadas feitas pelo Comitê (“RECOMENDAÇÃO”), se a decisão tomada pela Meta foi mantida ou anulada (“MANTIDA/ANULADA”), a política adotada pela Meta nas Diretrizes da Comunidade do Facebook ou Instagram envolvidas na controvérsia (“POLÍTICA”), qual ferramenta automatizada foi usada (“FERRAMENTA AUTOMATIZADA”), a principal questão discutida na decisão ou opinião da política em relação ao uso de ferramentas automatizadas (“QUESTÃO”), o resumo dos fatos que explicam a controvérsia (“RESUMO”), quaisquer citações relevantes das decisões do Comitê que não se encaixem nas outras colunas (“CITAÇÕES RELEVANTES”), bem como quaisquer notas necessárias para melhor compreensão (“NOTAS”) (TABELA, 2024).

Como já mencionado, cada decisão foi aberta e analisada. Embora não haja um mecanismo de busca, o radical “autom” foi pesquisado em cada uma das decisões, a fim de identificar as decisões pertinentes ao estudo.

Até 31 de março de 2024, 4 opiniões consultivas haviam sido publicadas. Duas delas abordam sistemas automatizados de moderação de conteúdo (*PAO-2023-01 Uso do termo “shaheed” em referência a indivíduos designados como perigosos* e *PAO-2022-01 Remoção de desinformação sobre a COVID-19*).

**Gráfico 1: Opiniões consultivas que abordam sistemas automatizados de moderação de conteúdo**




Fonte: Elaboração própria.

Entre os casos, até 31 de março de 2024, foram publicados 87, mas apenas 33 deles envolvem mecanismos automatizados, ou seja, em 33 casos foi encontrado o radical “autom”.

Entre as políticas Meta, objeto dos casos analisados, “Discurso de ódio” tem a maior incidência (7), seguida por “Indivíduos e organizações perigosas” (6) e “Violência e incitação” (5). Deve-se observar que cada decisão pode estar relacionada a até duas políticas, o que explica o número superior a 35 (33 decisões de casos, mais 2 opiniões de políticas). O Gráfico 2 abaixo descreve graficamente o impacto parcial das políticas. As políticas que foram identificadas apenas uma vez não estão incluídas: “Arte/escrita/poesia, Cultura, Comunidades marginalizadas”, “Arte/escrita/poesia, Governos”, “Crianças/direitos das crianças, Segurança”, “Coordenação de danos e divulgação de crimes”, “Eventos culturais, Saúde, Religião”, “Cultura, Discriminação, Religião”, “Eleições, Jornalismo, Desastres naturais”, “Governos, Maus-tratos”, “Saúde”, “Saúde, LGBT, Sexo e igualdade de gênero”, “Saúde, Segurança”, “Jornalismo, Desastres naturais”, “Jornalismo, Eventos de notícias, Política”, “Comunidades marginalizadas, Desinformação”, “Desinformação”, “Maus-tratos, Segurança, Guerra e conflito”, “Even-

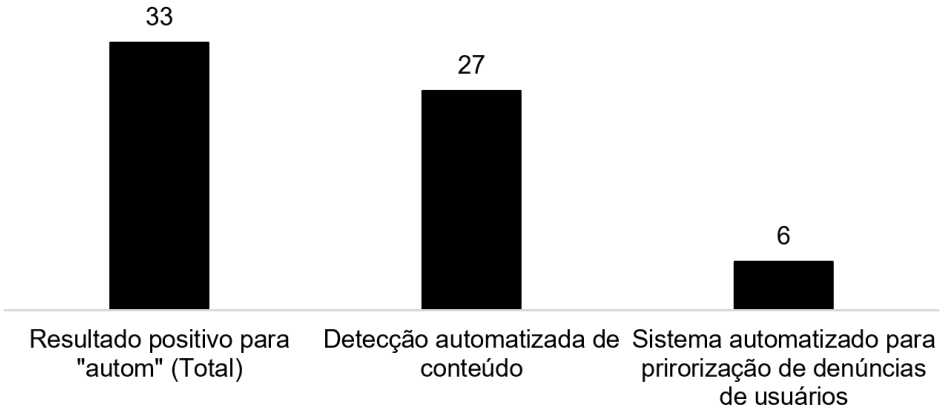
tos de notícias, Política, Guerra e conflito", "Protestos, Segurança", "Protestos, Sexo e igualdade de gênero", "Segurança, Guerra e conflito", "Sexo e igualdade de gênero" e "Violência, Guerra e conflito". Embora algumas delas sejam, de fato, semelhantes, elas não foram combinadas por uma questão de precisão.

**Gráfico 2: Incidência das políticas da Meta nas decisões analisadas**


Fonte: elaboração própria.

Desses 33, 27 envolvem casos de detecção automatizada de conteúdo e 6 envolvem o uso de sistema automatizado para priorização de reclamações de usuários, sendo que algumas incluem ou reproduzem recomendações para aprimoramento de procedimentos envolvendo detecção automatizada de conteúdo infringente.

**Gráfico 3: Casos que abordam sistemas automatizados de moderação de conteúdo**




Fonte: elaboração própria.

Para esta análise, o foco está nas 2 opiniões consultivas e nas 27 decisões de casos que envolvem detecção automatizada de conteúdo.

O sistema automatizado para priorizar as reclamações dos usuários é aqui desconsiderado, pois não está relacionado ao conteúdo da publicação criada pelo usuário, mas apenas à organização de denúncias realizadas por usuários, ou seja, o conteúdo não foi detectado por ferramentas automatizadas, mas sim reportado por usuário e a organização desta denúncia entre as demais é feita por ferramenta automatizada. Por esse motivo, essas 6 decisões não fazem parte da análise daqui em diante.

De fato, há diversos aspectos nas decisões e opiniões do Comitê. Opta-se, aqui, por abordar as recomendações sobre o uso de ferramentas automatizadas para detecção de conteúdo infringente. Há outros aspectos da moderação de conteúdo que merecem uma análise própria, a exemplo do processo de revisão de conteúdo "sob escalonamento" (por exemplo, Caso *2022-007-IG-MR - Gênero Musical Drill do Reino Unido*) e o programa de verificação cruzada (por exemplo, Caso *2023-007-FB-UA, 2023-008-FB-UA, 2023-009-IG-UA (decisão conjunta) - Disputa política antes das eleições turcas*).

## RESULTADOS E DISCUSSÃO

Dos 27 casos envolvendo detecção automatizada de conteúdo, 23 foram anulados pelo Comitê e 4 foram mantidos. O fato de a decisão ter sido anulada não significa que o problema esteja, em todas as ocasiões, na operação da ferramenta automatizada, pois o problema pode residir na análise humana após a detecção do conteúdo infringente.

Deve-se mencionar que a pesquisa, tanto em seus aspectos quantitativos quanto qualitativos, enfrentou desafios em sua primeira fase devido à ausência de um mecanismo de busca no conteúdo dos casos. Em junho, na segunda fase da pesquisa, o site do Comitê foi reformulado e um mecanismo de busca chamado *Recommendation Tracker* (Rastreador de recomendações, em tradução livre) foi implementado (RECOMMENDATION TRACKER, 2024).

Apesar da inovação ser bem-vinda, ainda não há um mecanismo de busca no conteúdo dos casos. Devido a essa falta de um design amigável para pesquisadores das decisões do Comitê – o que ainda pode ser aprimorado –, é possível que certos detalhes do caso não tenham sido divulgados ou adequadamente esclarecidos pelo

Comitê. Em outras palavras, certos fatos podem ter sido omitidos, o que pode afetar os resultados da pesquisa, uma vez que ela se baseou apenas nas informações disponibilizadas pelo Comitê.

No Caso *2021-012-FB-UA Cinto Wampum*, por exemplo, o Comitê declarou que os moderadores humanos devem estar cientes de que o conteúdo está sujeito a uma revisão secundária. No caso em tela, embora uma ferramenta de detecção automatizada tenha identificado o conteúdo como potencialmente violador do Padrão Comunitário de Discurso de Ódio do Facebook (mesmo que não tenha sido denunciado por nenhum usuário e tenha descrito expressamente a abordagem educativa da publicação!), dois revisores humanos concordaram com a remoção, o que foi anulado pelo Comitê. Em outro caso (*2022-009-IG-UA e 2022-010-IG-UA - Identidade de gênero e nudez (decisão conjunta)),* restou evidenciado que a política de nudez adulta e atividade sexual da Meta não é clara, levando a uma análise automática conflitante.

Outros casos mostram que o problema identificado na moderação de conteúdo não pode ser atribuído exclusivamente à ferramenta automatizada. Por exemplo, no Caso *2023-014-IG-UA - Convocação das mulheres para protesto em Cuba*, a revisão humana foi atrasada porque o conteúdo gerado pelo usuário foi restringido no programa de verificação cruzada e não foi enviado para avaliação posterior. Na decisão do Caso *2021-006-IG-UA - Isolamento de Öcalan*, o Comitê recomendou a exclusão de dados de treinamento derivados de erros anteriores, pois isso pode resultar em falhas na avaliação humana após a detecção automatizada de conteúdo. O problema, portanto, não está na detecção automatizada em si, mas na falha em atualizar o banco de dados para moderação humana posterior, especialmente com relação às exceções às políticas de conteúdo proibido na plataforma.

De todo modo, as ferramentas automatizadas não são imunes a erros e, por esse motivo, o Caso *2020-004-IG-UA - Sintomas de câncer de mama e nudez* é paradigmático, uma vez que envolve remoção totalmente automatizada. Nessa decisão, o Comitê recomendou à Meta (i) alertar os usuários se medidas automatizadas forem adotadas para moderar seu conteúdo, (ii) garantir que um recurso (apelação) de decisões automatizadas para um ser humano seja permitido em instâncias específicas, (iii) aprimorar a detecção automatizada de imagens com sobreposição de texto para que as postagens de conscientização, por exemplo, sobre doenças, não sejam sinalizadas erroneamente, (iv) aprimorar seus relatórios de transparência em relação à implemen-

tação de medidas de aplicação automatizadas. O Comitê reafirma tais recomendações ao analisar os casos *2023-049-IG-UA Hospital Al-Shifa* e *2024-011-FB-UA Enchentes na Líbia*.

Assim, ficou demonstrado que a Meta vem adotando as recomendações do Comitê no que diz respeito a informar os usuários em caso de detecção automatizada. No entanto, ainda podem ser feitas melhorias com relação à publicação de relatórios de transparência envolvendo a aplicação automatizada. Por exemplo, no Caso *2022-002-FB-MR Vídeo explícito do Sudão*, a Meta respondeu parcialmente a todas as perguntas relativas ao impacto do sistema automatizado sobre conteúdos na plataforma, demonstrando imprecisão na resposta às perguntas do Comitê.

Na Opinião Consultiva *PAO-2023-01 Uso do termo "shaheed" em referência a indivíduos designados como perigosos*, as mesmas recomendações estão incluídas, concentrando-se na necessidade de expandir os relatórios de transparência para incluir dados sobre o número de decisões de remoção automatizadas e a taxa de anulação após revisão humana. Além disso, os casos *2024-011-FB-UA Enchentes na Líbia*, *2023-044-IG-UA - Remoção de Azov* e *2023-047-FB-UA - Ilustração retratando o golpe no Níger* foram reiteradas a recomendação da Decisão de Caso *2020-004-IG--UA - Sintomas de câncer de mama e nudez* para implementar uma auditoria a fim de analisar continuamente, uma amostra representativa de remoções automatizadas, para que seja possível aprender com erros anteriores.

É importante mencionar as recomendações no Caso 2023-*036-IG-UA, 2023-037-FB-UA Publicações educativas sobre ovulação (decisão conjunta)* reitera a recomendação de "melhorar a detecção automatizada de imagens com sobreposição de texto para garantir que as postagens de conscientização sobre os sintomas do câncer de mama não sejam erroneamente sinalizadas para revisão", feita pelo Comitê no Caso *2020-004-IG-UA - Sintomas de câncer de mama e nudez.* Na mencionada decisão conjunta de 2023, o Comitê menciona expressamente que "a Meta implantou um novo classificador de conteúdo de saúde baseado em imagem e aprimorou um classificador de sobreposição de texto existente para melhorar ainda mais as técnicas do Instagram a fim de identificar conteúdo contextual sobre câncer de mama. Ao longo de 30 dias em 2023, essas melhorias contribuíram para o envio de mais 3.500 itens de conteúdo para análise humana que anteriormente teriam sido removidos automaticamente", o que demonstra uma melhoria significativa nesse aspecto.

Embora o Caso *2020-004-IG-UA - Sintomas de câncer de mama e nudez* demonstre a importância da revisão humana nos processos de identificação automatizada de conteúdo, dado o volume de conteúdo produzido pelos usuários, isso não impediu que a Meta buscasse a identificação e a remoção totalmente automatizadas. Um exemplo disso é a introdução do Media Matching Service (MMS), ou Serviço de Correspondência de Mídias, em tradução livre. Essa ferramenta usa um banco de dados com conteúdos que foram anteriormente considerados infringentes para comparar novos conteúdos que correspondam àqueles do banco de dados, permitindo que uma ação automatizada seja tomada, por exemplo, a remoção do conteúdo. O MMS aparece em 6 casos como a ferramenta automatizada no centro da controvérsia. Por exemplo, a decisão do caso *2023-049-IG-UA Hospital Al-Shifa* mostra um exemplo de uso indevido do MMS, já que "a supervisão humana insuficiente da moderação automatizada durante uma resposta a crises pode levar à remoção errônea de discursos que podem ser de utilidade pública significativa".

Com relação ao uso do MMS, o *Caso 2022-004-FB-UA Ilustração retratando policiais colombianos* é paradigmático. A partir dele, o Comitê afirmou que a remoção errônea de conteúdo poderia exacerbar os efeitos de decisões imprecisas, pois até então era permitido que os revisores incluíssem conteúdos inadvertidamente em um banco, levando à remoção automática de conteúdo idêntico, mesmo que não violasse política alguma. Entre as recomendações, o Comitê sugeriu limitar o tempo entre o momento em que o conteúdo armazenado é identificado para revisão adicional e o momento em que, se considerado não violador, é removido do banco, bem como a necessidade de incluir em seus relatórios de transparência as taxas de erro para conteúdo adicionado erroneamente aos bancos de MMS por conteúdo violador, categorizado por política, seguindo as etapas do Caso *2020-004-IG-UA - Sintomas de câncer de mama e nudez*, em relação aos relatórios de transparência.

Essas recomendações são repetidas nos casos *2022-007-IG-MR - Gênero Musical Drill do Reino Unido* e *2022-011-IG-UA Vídeo após ataque a igreja na Nigéria*. No último caso, é interessante observar que a decisão mostra que há mais de um banco de dados para MMS. Depois que uma equipe regional alertou sobre um contexto de crise na Nigéria, a equipe de políticas da Meta adicionou vídeos do incidente a um banco de dados MMS diferente. Nesse caso, não foi aplicada uma remoção automática, mas um revisor humano permitiu o conteúdo no Instagram adicionando uma tela de

aviso de “conteúdo perturbador”. O Comitê anulou a decisão da Meta pela remoção do conteúdo, corroborando a medida tomada pelo revisor para incluir uma tela de aviso.

Além disso, no Caso *2023-050-FB-UA - Reféns sequestrados de Israel,* surgiu novamente a problemática de remoção de conteúdo indevida que poderia ampliar o impacto de decisões incorretas mediante o uso do MMS, mas não foram feitas novas recomendações ou reiteração das anteriores. O interessante é que, nesse caso, mesmo antes das decisões emitidas pelo Comitê, a Meta se absteve de emitir “strikes” normalmente relacionados a remoções automatizadas com base nos bancos de dados do MMS.

Isso mostra um abrandamento das sanções aplicadas depois que o conteúdo supostamente infringente foi identificado com base no MMS, por exemplo, adicionando uma tela de aviso em vez de removê-lo automaticamente, retendo “strikes” automáticos. No entanto, conforme demonstrado no Caso *2023-004-FB-MR Vídeo de prisioneiros de guerra armênios*, a adição automática de tela de aviso pelo MMS falhou e teve de ser feita por humanos um mês depois, mostrando que ainda podem e devem ser feitas melhorias.

Em outro caso (*2022-005-FB-UA - Menção do Talibã na divulgação de notícias*), o conteúdo que foi removido não atingiu o mínimo em pontuação para ser avaliado prioritariamente por um revisor humano. Isso ocorreu devido a um erro no sistema HIPO (*High Impact False Positive Override*) da Meta, ou Anulação de Falso Positivo de Alto Impacto, em tradução livre, que tem o objetivo de retificar possíveis erros de falsos positivos depois que o conteúdo é processado. Isso demonstra os perigos das ações automatizadas para remoção de conteúdo tomadas antes de uma avaliação secundária por um moderador humano. Portanto, o Comitê recomendou que a Meta revise o sistema HIPO para garantir que ele seja capaz de priorizar possíveis erros na aplicação de exceções, bem como para permitir a revisão do HIPO em todos os idiomas.

Em outro caso (*2021-013-IG-UA - Cerveja de Ayahuasca*), houve uma conexão equivocada ao considerar que uma publicação popular continha conteúdo violador. Apesar de o conteúdo ter sido sinalizado para revisão pelos sistemas automatizados da Meta devido à sua alta audiência (considerada “tendência”), e mesmo após uma revisão humana, o moderador o removeu, mesmo que o conteúdo não violasse política alguma da Meta. Como resultado, o Comitê anulou a decisão da Meta. Com relação ao conteúdo gerado no âmbito das parcerias pagas, no Caso *2023-010-IG-MR - Promoção da cetamina para tratamentos não aprovados pela FDA*, o Comitê recomenda não apenas aprimorar seu processo de revisão para garantir que o conteúdo produzido

por meio de uma parceria paga seja submetido a uma avaliação minuciosa em relação a todas as políticas relevantes, mas também garantir que o conteúdo em questão seja direcionado a revisores ou sistemas automatizados equipados e treinados para aplicar as políticas de conteúdos de marca da Meta quando necessário.

Além da recomendação já feita no Caso *2020-004-IG-UA - Sintomas de câncer de mama e nudez*, as auditorias também foram recomendadas pelo Comitê no Caso *2023-007-FB-UA, 2023-008-FB-UA, 2023-009-IG-UA (decisão conjunta) - Disputa política antes das eleições turcas*. Na ocasião, foi reconhecido que os revisores devem ser avisados quando o sistema automatizado identificar um risco maior de erro de aplicação, evitando a remoção antes que seja feito um escalonamento do conteúdo para uma análise contextual mais detalhada. O objetivo das auditorias propostas é verificar as listas de insultos em países com eleições programadas entre 2023 e o início de 2024, com o fim de detectar e eliminar termos erroneamente incluídos nessas listas. Embora o Dynamic Multi-Review (DMR), ou Análise Dinâmica Múltipla em tradução livre, estivesse ativo na época, de um total de oito revisores humanos que avaliaram as três publicações semelhantes, todas anteriores às revisões adicionais de verificação cruzada, apenas dois revisores (um para cada publicação) determinaram que essas publicações não violavam a política de discurso de ódio. Isso levou a um bloqueio excessivo, removendo conteúdo que não deveria ser removido em primeiro lugar, e por esse motivo o Comitê anulou a decisão de remoção tomada originalmente pela Meta.

Diversas decisões demonstram a dificuldade de reconhecer contextos na moderação de conteúdo, especialmente devido à pluralidade linguística, tanto em termos de detecção pela ferramenta automatizada quanto de avaliação humana após a detecção automatizada. Por exemplo, o Caso *023-032-IG-UA Mulher iraniana confrontada na rua* destaca as dificuldades de moderar o discurso figurativo. Essa decisão de caso demonstra uma relação entre a repetição de erros, pois a precisão da detecção automatizada de conteúdo infringente é afetada pela qualidade dos dados de treinamento fornecidos pelos moderadores humanos. Se os moderadores humanos bloquearem excessivamente as declarações figurativas, é provável que esse erro seja repetido e intensificado pela remoção automática.

No Caso *2021-014-FB-UA Alegação de crimes em Raya Kobo*, os recursos linguísticos da Meta foram questionados, embora uma equipe específica de revisão do amárico tenha sido criada e avaliado o conteúdo gerado pelo usuário desse caso.

De acordo com o Comitê, deve ser feita uma avaliação para analisar os recursos linguísticos da Meta na Etiópia, a fim de verificar se são adequados para proteger os direitos de seus usuários. Outro caso que envolve a equipe de análise do amárico o *2022-006-FB-MR - Gabinete de Comunicações da Região do Tigré*, no qual a análise de especialistas ligada ao Integrity Product Operations Centre (IPOC), ou Centro de Operações de Produtos de Integridade, em tradução livre, para a Etiópia foi confirmada pelo Comitê, enquanto duas avaliações da equipe de análise do amárico não foram confirmadas. Na época da decisão de 2022 mencionada, a Meta estava operando por um curto período (dias ou semanas) um IPOC para a Etiópia, a fim de melhorar a moderação em situações de alto risco, com base no monitoramento de especialistas para avaliar e corrigir qualquer abuso na moderação de conteúdo.

A amplitude de terminologia e a impossibilidade de proibi-la sem análise contextual foram apontadas na Opinião Consultiva *PAO-2023-01 Uso do termo "shaheed" em referência a indivíduos designados como perigosos*. As dificuldades na identificação de conteúdo retórico, seja por ferramenta automatizada ou por humanos, também são apontadas no Caso *2023-011-IG-UA, 2023-012-FB-UA, 2023-013-FB-UA - Postagens dos Estados Unidos que discutem o aborto (decisão conjunta),* destacando a possibilidade de riscos sistemáticos causados por erros na precisão da aplicação da política de Violência e Incitação da Meta. No Caso *2023-014-IG-UA - Convocação das mulheres para protesto em Cuba*, de acordo com o Comitê, tanto os sistemas automatizados da Meta quanto os revisores de conteúdo devem incorporar informações contextuais em seu processo de tomada de decisão.

As informações contextuais são relevantes especialmente quando os usuários tentam evitar a detecção automática de termos proibidos. No *Caso 2020-003-FB-UA - Armênios no Azerbaijão*, embora não tenha sido expressamente mencionado na decisão se o conteúdo foi de fato detectado automaticamente, com base nos fatos no fundo da controvérsia, ficou demonstrado que o usuário manipulou recursos (*in casu*, o uso de pontuação entre caracteres) para evitar a detecção automática de termos proibidos considerados discurso de ódio. De acordo com o Comitê, essa ação sugere que o usuário estava ciente do uso de linguagem proibida e, por esse motivo, a decisão para remover o conteúdo da plataforma foi mantida pelo Comitê.

Considerando o discorrido até aqui, é possível agrupar as recomendações do Comitê em dois grandes grupos: (i) melhorias na transparência; (ii) melhorias procedimentais.

As recomendações sobre melhorias na transparência são aqui subclassificadas em três grupos: (i) transparência para os usuários, em geral; (ii) transparência para o usuário, individualmente; (iii) transparência para o Comitê, de acordo com a Tabela 1. A título de exemplo, conforme mostram o Relatório Anual do Comitê de Supervisão 2023 (OVERSIGHT BOARD, 2024) e o Rastreador de Recomendações (RECOMMENDATION TRACKER, 2024), a implementação da recomendação para auditar listas de insultos para encontrar e remover termos adicionados erroneamente foi demonstrada por informações públicas divulgadas pela Meta. No entanto, a Meta ainda não progrediu quanto à implementação de um procedimento de auditoria interna para aprender com os erros de aplicação (OVERSIGHT BOARD, 2024).

As recomendações do Comitê convergem em pelo menos um aspecto fundamental: a necessidade de garantir o direito de acesso à informação online: por meio da proteção de conteúdos que aumentem a conscientização e por meio da criação de mecanismos de transparência processual na moderação de conteúdo. Reconhece-se que os erros não se limitam a uma modalidade de moderação: é provável que ocorram erros no uso de análises automatizadas e humanas.

**Tabela 1: Recomendações – Melhorias quanto à transparência**

| Usuários, em geral | Usuários, individualmente | Comitê |
|---|---|---|
| Implementar procedimento de auditoria interna para aprender com os erros de aplicação. ♦ | Informar os usuários, de forma descritiva, quando a automação for usada para tomar medidas coercitivas contra seu conteúdo. ♦ | Fornecer os dados ao Comitê para evitar, entre outros, erros decorrentes de problemas sistêmicos na aplicação. ○ |
| Incluir nos Relatórios de Transparência o número de decisões por meio de remoção automatizadas e a proporção dessas decisões posteriormente revertidas após análise humana. ♦ | | |
| Incluir taxas de erro quanto a conteúdo incluído erroneamente nos bancos de MMS, divididos pela política de conteúdo, nos relatórios de transparência.◊ | | |
| Auditar listas de insultos para localizar e remover termos adicionados por engano. ■ | | |

♦ 2020-004-IG-UA – Sintomas de câncer de mama e nudez
◊ 2022-004-FB-UA – Ilustração retratando policiais colombianos
■ 2023-007-FB-UA e outros – Disputa política antes das eleições turcas
○ 2023-011-IG-UA e outros – Postagens dos Estados Unidos que discutem o aborto

Fonte: elaboração própria.

Um desafio adicional envolve os inúmeros conflitos de interesse associados a uma empresa como a Meta e como eles influenciam as plataformas que a companhia gerencia. Entre os diversos interesses da Meta, o bem-estar do usuário é apenas um aspecto, e é difícil determinar definitivamente se essa é a principal força que impulsiona as decisões da empresa, como a de criar o Comitê de Supervisão. Do ponto de vista jurídico, o Comitê, que se originou da autorregulação e foi implementado dentro do escopo da supervisão externa, não está imune a mal-entendidos e não é – e não deve ser – visto como uma panaceia para todos os desafios apresentados pela moderação de conteúdo.

**Tabela 2: Recomendações – Melhorias procedimentais**

| **Dados de entrada do sistema automatizado** | **Detecção de conteúdo e suas consequências** | **Interação entre o sistema automatizado e a revisão humana** | **Relação entre a detecção automatizada e as políticas da plataforma** |
|---|---|---|---|
| Atualizar classificadores para excluir dados de treinamento de erros de aplicação anteriores. □ | Aprimorar a detecção automática de texto sobreposto em imagens. ♦ | Os revisores humanos devem ser avisados quando os sistemas automatizados identificarem um risco maior de erro. ◘ | Examinar se o sistema HIPO pode efetivamente priorizar possíveis erros na aplicação de exceções. ○ |
| Assegurar que o conteúdo não violador seja rapidamente removido dos bancos de MMS. ■ | Aprimorar a revisão do HIPO para fazer avaliações em todos os idiomas. ○ | A revisão secundária humana deve ser feita por outros revisores que não os que avaliaram o conteúdo anteriormente. ▲ | Aprimorar a revisão automatizada da parceria paga em relação a todas as políticas aplicáveis da plataforma. ◄ |
| Garantir que o conteúdo com altas taxas de apelo e altas taxas de apelo bem-sucedido seja reavaliado para possível remoção dos bancos de MMS. ■ | Suspender "strikes" automáticos de MMS, adicionando uma tela de aviso em vez de remover automaticamente o conteúdo. ▼ | Informar aos moderadores de conteúdo se eles estão envolvidos na revisão secundária e examinar a precisão dos resultados. ◊ | Garantir que o conteúdo seja encaminhado para sistemas automatizados (ou revisores humanos) treinados para aplicar as políticas de conteúdos de marca da Meta. ◄ |
| Aprimorar os sistemas automatizados para incluir contexto. ► | | | |

♦ 2020-004-IG-UA – Sintomas de câncer de mama e nudez
◊ 2021-012-FB-UA – Cinto Wampum
○ 2022-005-FB-UA – Menção do Talibã na divulgação de notícias
◘ 2023-007-FB-UA e outros – Disputa política antes das eleições turcas
■ 2022-004-FB-UA – Ilustração retratando policiais colombianos
□ 2021-006-IG-UA – Isolamento de Öcalan
► 2023-014-IG-UA – Convocação das mulheres para protesto em Cuba
▲ 2023-002-IG-UA e outros – Violência contra mulheres
▼ 2022-011-IG-UA – Vídeo após ataque a igreja na Nigéria, and 2023-050-FB-UA Reféns sequestrados de Israel
◄ 2023-010-IG-MR – Promoção da cetamina para tratamentos não aprovados pela FDA

Fonte: elaboração própria.

## CONCLUSÃO

O Comitê fornece algumas recomendações à Meta para atenuar as ingerências indevidas à liberdade de expressão a partir do uso de sistemas automatizados para detecção de conteúdo infringente. Essas recomendações indicam onde se localizam os problemas dos sistemas automatizados e como eles poderiam ser resolvidos para não remover conteúdo legítimo ou para restaurá-lo o mais rápido possível.

Em termos de estruturação dos dados analisados, as recomendações fornecidas à Meta pelo Comitê para o aprimoramento de sistemas automatizados são categorizados em termos de transparência (para usuários em geral, para usuários individuais e para o Comitê) e aprimoramentos procedimentais (com relação aos dados de entrada, detecção de conteúdo, necessidade de revisão humana e a relação entre sistemas automatizados e suas políticas), atenuando, assim, os impactos negativos que as ferramentas automatizadas possam causar à liberdade de expressão, sem que isso comprometa o exercício de outros direitos fundamentais eventualmente colidentes.

Trata-se de uma “faca de dois gumes”: embora o conteúdo legítimo seja de fato removido indevidamente mediante o uso de ferramentas automatizadas, proibir os provedores de plataformas de usar tais sistemas para detectar conteúdo infringente seria inconstitucional, uma vez que conteúdos ilegítimos continuariam sendo compartilhados em plataformas, sem perspectiva de um bloqueio efetivo. Entretanto, não há garantia de que os conteúdos legítimos removidos indevidamente serão restaurados o mais rápido possível, devendo investir-se no aperfeiçoamento das ferramentas automatizadas a fim de corrigir erros.

A coleção de casos examinada destaca diferentes pontos-chave, incluindo a necessidade de supervisão humana e a importância da transparência em relação ao uso de sistemas automatizados para detectar conteúdo infringente, especialmente no que diz respeito à reversão de decisões humanas. Além disso, persistem desafios em relação à compreensão contextual do conteúdo gerado por usuários, ao potencial do conteúdo para aumentar a conscientização em sociedade e à capacidade geral dos sistemas automatizados e dos moderadores humanos de se alinharem às políticas de moderação de conteúdo da empresa, explorando os riscos e as oportunidades da adoção de mecanismos automatizados como uma solução para encontrar e combater conteúdos infringentes.

## DECLARAÇÃO DE DIVERSIDADE NAS CITAÇÕES

Por meio desta declaração, junta-se a um esforço coletivo para desfazer o apagamento epistemológico estrutural na academia contra mulheres, pessoas não binárias, negres, do Sul global, e de outros grupos sociais, cujas vozes são menos ouvidas devido ao viés encontrado em citações. Acreditamos que a transparência em relação às bibliografias citadas seja fundamental para compreender o presente, e alterar esse quadro estrutural de maneira conjunta e consistente. Compartilha-se que neste artigo, as citações se distribuem da seguinte forma: nomes femininos (28,6%), masculinos (28,6%), feminino-masculino (14,3%), e fontes institucionais (28,6%).

## REFERÊNCIAS


CASE decisions and policy advisory opinions, 2024. Disponível em: www.oversightboard.com/decision. Acesso em: 07 abr. 2024.

CELESTE, Edoardo. Digital constitutionalism: a new systematic theorisation. **International Review of Law, Computers & Technology**, v. 33, n. 1, p. 76-99, 2019.

DOUEK, Evelyn. Facebook's oversight board: move fast with stable infrastructure and humility. **NCJL & Tech.,** v. 21, p. 1, 2019.

DOUEK, Evelyn. The Siren Call of Content Moderation Formalism, New Technologies Of Communication And The First Amendment: The Internet, Social Media And Censorship. Lee Bollinger & Geoffrey Stone eds., 2022 (No prelo).

FINCATO, Denise Pires; GILLET, Sérgio Augusto da Costa. A Pesquisa Jurídica sem Mistérios: do Projeto de Pesquisa à Banca. Porto Alegre: Editora Fi, 2018.

GILLESPIE, Tarleton. **Custodians of the internet.** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018.

HARTMANN, Ivar Alberto Martins; SARLET, Ingo Wolfgang. Direitos fundamentais e direito privado: a proteção da liberdade de expressão nas mídias sociais. **Revista de Direito Público,** v. 16, n. 90, dez. 2019.

KLONICK, Kate. The new governors: The people, rules, and processes governing online speech. **Harv. L. Rev**., v. 131, p. 1598, 2017.

MARTÍNEZ, María Barral. Platform regulation, content moderation, and AI-based filtering tools: Some reflections from the European Union. **JIPITEC**, v. 14, p. 211, 2023.

MENDES, Laura Schertel; MATTIUZZO, Marcela. Discriminação algorítmica: conceito, fundamento legal e tipologia. **RDU**, Porto Alegre, v. 16, n. 90, 2019, p. 39-64, nov./dez. 2019.

OLIVA, Thiago Dias; ANTONIALLI, Dennys Marcelo; GOMES, Alessandra. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. **Sexuality & Culture**, v. 25, p. 700-732, 2021.

OVERSIGHT BOARD 2024 ANNUAL REPORT, 2024. Disponível em: https://www.oversightboard.com/wp-content/uploads/2024/06/Oversight-Board-2024-Annual-Report.pdf. Acesso em: 13/05/2024.

Recommendation Tracker – Oversight Board, June 2024. Disponível em: https://www.oversightboard.com/explore-our-recommendation-tracker/. Acesso em: 21 jun. 2024.

REINHARDT, Jörn. Algorithmizität und Sichtbarkeit. Konflikte um Bilder in den sozialen Medien. *In* **Rechtsästhetik in rechtsphilosophischer Absicht**. Nomos Verlagsgesellschaft mbH & Co. KG, 2020.

TABELA – Casos e opiniões consultivas, 2024. Disponível em: https://docs.google.com/spreadsheets/d/14sDzjrCepXYlg6F2e_7K4k6YQGdeQ4oLZM7Lj6nur9o/edit?usp=sharing. Acesso em: 12 jul. 2024.