Inteligência Artificial, moderação de conteúdos no YouTube e a proteção de direitos: características, problemas e impactos políticos[†]

Artificial Intelligence, content moderation on YouTube and the rights protection: characteristics, problems and political impacts

Sivaldo Pereira da Silva [a, b, *]

Daniel Jorge Teixeira Cesar [b]

RESUMO: Este artigo tem como principal objetivo caracterizar o papel de sistemas de inteligência artificial na moderação de conteúdos de usuários no YouTube, seus problemas e impactos políticos, especialmente no horizonte da proteção de direitos individuais. A pesquisa utilizou o método de rastreamento de processo, baseado na coleta e análise de dados qualitativos para estabelecer relações causais e compreender como ocorre a moderação de conteúdos na plataforma. Foram analisados 79 textos dentre notas e informações publicadas pela plataforma em seu blog corporativo; documentos de políticas de moderação de conteúdos e relatórios de transparência da empresa. Os resultados demonstram que há uma crescente centralização da Inteligência Artificial (IA) no processo de moderação de conteúdo, transformando a moderação humana em um dispositivo do sistema automatizado. Isso tem gerado um crescimento substancial do número de remoções de conteúdo potencialmente nocivo que traz, por outro lado, efeitos colaterais como o aumento das violações de direitos individuais pela plataforma. Opacidade; acirramento no problema da escala; moderação guiada por princípios comerciais; falhas na captação de contexto; fragilidade nos processos de participação e *accountability* são outros problemas também identificados.

Palavras-chave: Moderação de Conteúdos; Inteligência Artificial; Mídias Sociais; YouTube; Governança Algorítmica.

ABSTRACT: The main objective of this paper is to characterize the part artificial intelligence systems play in moderating user generated content on YouTube, it's problems and political impacts particularly on the matter of protection of individual rights. This research used process tracing methods based on the gathering and analysis of qualitative data to establish causal relations and to understand how the platform moderates content. A total of 79 documents were analyzed, between notes and information publicized by the platform's corporate blog, documents describing content moderation policies and the company's transparency reports. Results indicate there is a growing centralization of artificial intelligence (AI) in the process of moderating content, turning human moderation into a device of the automated system. This has generated substantial increases in the number of potentially harmful contents removed and causes, on the other hand, side effects such as the increase of individual rights violations by the platform. Opacity; aggravation of the scaling problem; moderation guided by commercial principles; failure to grasp context; fragility in participation processes and accountability are other problems also identified.

Keywords: Content Moderation; Artificial Intelligence; Social Media; YouTube; Algorithmic Governance.

[a] Faculdade de Comunicação, Universidade de Brasília, Brasília, DF, Brasil.

[b] Programa de Pós-Graduação em Comunicação, Universidade de Brasília, Brasília, DF, Brasil.

[*] Correspondência para/*Correspondence to*: Sivaldo Pereira da Silva. E-mail: sivaldop@unb.br.

[†] Este trabalho foi produzido durante a pesquisa de pós-doutorado de Sivaldo Pereira da Silva como professor visitante sênior na Technische Universität Dortmund (2021-2022), com bolsa do edital CAPES/PROBRAL nº 14/2019, no âmbito do projeto "Communication and Democracy".

Recebido em/*Received*: 15/08/2022; Aprovado em/*Approved*: 23/11/2022.

INTRODUÇÃO

A publicação de vídeos, textos ou imagens por grupos ou indivíduos em mídias sociais é hoje objeto de intenso monitoramento. A moderação de conteúdos assumiu um lugar importante no processo de produção da opinião pública contemporânea. Plataformas digitais se tornaram *players* fundamentais, determinando regras de visibilidade, removendo conteúdo e banindo perfis quando esses violam os seus termos de serviço. Ainda que tais corporações não sejam produtoras de conteúdo elas possuem hoje um papel cada vez mais relevante nos processos de formação da opinião pública, indo bem além daquilo que foram os tradicionais *gatekeepers* que preponderaram no século passado (como a TV, o rádio e o jornal no passado).

Com o avanço do processo de plataformização da opinião pública, que tem a internet como sua principal arena, e diante do aumento do fluxo de desinformação e discurso de ódio que assolam a rede, a moderação se tornou um elemento inevitável. Passou a ser uma necessidade do próprio modelo de negócio das empresas digitais. Porém, se por um lado a moderação age para evitar a difusão de conteúdos nocivos e violação de direitos, isso também implica em mais concentração de poder, e também outras violações de direitos como privacidade e liberdade de expressão.

Este problema se tornou ainda maior com volume vertiginoso de conteúdos *online* publicados a cada segundo. Para dar conta deste cenário, plataformas como o YouTube, um dos maiores e mais populares repositórios mundiais de conteúdo audiovisual, passaram a empregar sistemas de Inteligência Artificial (IA) no processo de moderação. Em seus documentos e diretrizes, a plataforma ressalta a eficiência tecnológica da automação e a importância de seu uso para garantir o efetivo monitoramento e identificar conteúdos nocivos em meio aos bilhões de vídeos publicados.

Porém, este cenário não é pacífico e levanta diversos questionamentos. Qual o papel dos sistemas de IA na moderação de conteúdo? O que dizem documentos e justificativas de empresas como o YouTube sobre isso e quais os possíveis impactos políticos de sistemas automatizados na produção da opinião pública e no exercício de direitos, como a liberdade de expressão?

Diante dessas questões, o objetivo principal deste artigo é caracterizar o papel de sistemas de Inteligência Artificial (IA) na moderação de conteúdo no YouTube e seus impactos políticos, sobretudo no horizonte dos direitos individuais. Para tanto, o texto segue dividido em duas partes. Primeiramente, será realizada uma análise sintetizando os principais elementos que caracterizam hoje o processo de moderação de conteúdo no YouTube e qual o lugar que os sistemas de Inteligência Artificial estão assumindo. Na segunda parte, serão identificados os principais problemas e impactos políticos que este cenário acarreta.

MODERAÇÃO DE CONTEÚDO NO YOUTUBE: A CENTRALIDADE DOS SISTEMAS AUTÔNOMOS

Moderação de conteúdo é a atividade de intervenção das plataformas para filtrar informações postadas por usuários impedindo a sua publicação ou deteriorando o alcance daqueles que violam os termos de uso e políticas da empresa. Embora não seja um processo aberto ao escrutínio público, plataformas como o YouTube têm sido pressionadas a responder, esclarecer e justificar esta atividade. Os documentos publicados pela empresa devem ser considerados um aporte importante deste fenômeno pois diz respeito tanto ao discurso público dessas corporações na forma de resposta para o problema dos efeitos danosos de conteúdos nocivos, como também o resultado mais visível das pressões por *accountability* ao qual as plataformas estão sendo cada vez mais submetidas.

Para compreender como o YouTube tem posicionado os sistemas de Inteligência Artificial na sua política de moderação de conteúdo, este estudo observou documentos e diretrizes oficiais da plataforma. Para isso utilizou-se a metodologia de rastreamento de processo, que se baseia na coleta e análise de dados qualitativos para estabelecer relações causais que expliquem como um processo se desenvolve ao longo de um período de tempo. Essa metodologia combina técnicas de análise documental e de estudo de caso para analisar uma situação específica e tem sua origem na psicologia na década de 1970, no estudo dos processos cognitivos de tomada de decisão por parte de indivíduos. Na década de 1980 passa a ser incorporada pelas ciências políticas como ferramenta de análise do processo de formação de políticas públicas. Mais recentemente autores como Beach e Pedersen (2013) e Falleti (2016) ampliam o rigor científico na coleta de evidências e as possibilidades de utilizar o método para formular ou testar hipóteses.

O rastreamento de processos é uma ferramenta para o estudo da criação e aplicação de políticas, no caso, de moderação de conteúdo de uma grande plataforma de comunicação. Para além da descrição, a proposta é desenvolver análises levando em conta as inferências causais sobre o funcionamento de processos a partir de dados qualitativos, relacionando os resultados das políticas de moderação de conteúdo e com os efeitos que produzem. Dessa forma, a partir das evidências e dos dados disponíveis, se procura estabelecer teorias ou hipóteses sobre as relações causais e compreender como funciona a moderação de conteúdo da plataforma.

Foram coletadas evidências dentro de um universo de dados da pesquisa composto por um total de 79 documentos públicos da plataforma, sendo 4 relatórios de transparência; 34 páginas de suporte e ajuda, contendo as políticas e documentos explicativos sobre as regras e sua aplicação; e 41 postagens do *blog* corporativo publicadas entre 2006 e 2021 com informações sobre procedimentos e medidas especificamente sobre moderação de conteúdo. Deste universo de documentos, 1 relatório trata do uso de sistemas automatizados, 6 documentos regulatórios abordam uso de sistemas de aprendizado de máquina e 10 documentos fazem referências a processos automatizados. Os demais trazem temas gerais sobre moderação e

abordam questões como sinalização de conteúdo por usuários e atualizações sobre o funcionamento da moderação.

A autorregulação a partir do trabalho de empregados dedicados à remoção de informação publicada é um dos aspectos da moderação comercial de conteúdos que tem como objetivo principal para identificar as práticas de regulação de conteúdos gerados por usuários em plataformas (Roberts 2019; Gorwa, Binns, Katzenbach 2020). Roberts (2019) destaca que essas plataformas, que baseiam seu modelo de negócios em conteúdo gerado pelos usuários, recorrem ao trabalho de moderadores humanos para manter os espaços comunicacionais livres de conteúdos ilegais, ofensivos, abusivos ou violentos. Esses empregados, em geral de empresas terceirizadas em regime de contrato com salários baixos e longos períodos de trabalho, são expostos ao pior que essas redes podem abrigar, realizam o trabalho de limpar as redes e tornar os espaços comercialmente viáveis para o grande público, possibilitando que marcas possam divulgar seus produtos em anúncios intercalados aos conteúdos gerados pelos usuários. O trabalho realizado por essa categoria de empregados das plataformas é invisibilizado pelas empresas, que evitam expor informações sobre as regras e procedimentos de moderação de conteúdos para proteger o segredo comercial de seu negócio e evitar que usuários possam burlar as regras e encontrar brechas para postar conteúdos proibidos. Quanto aos moderadores é comum que desenvolvam problemas psicológicos relacionados ao stress causado pelo trabalho e pelo tipo de conteúdo revisado. Por se tratar de empresas terceirizadas, muitas delas estão fora do país de origem da empresa e, portanto, pode ocorrer de moderar conteúdos a partir de uma barreira cultural e de linguagem, o que pode interferir no processo de revisão realizado pelas plataformas.

A partir de 2006, com a compra pelo Google e devido ao crescente volume de publicações e a impossibilidade de regular tudo que é enviado pelos usuários, o YouTube adotou o sistema de sinalização ou *flags*[1], um tipo de ferramenta bastante difundida também em outras plataformas (Crawford, Gillespie 2016). Esse sistema estabelece um canal de comunicação direto com a plataforma para denunciar conteúdos ilícitos e solicitar que sejam retirados do ar e não mais sejam visualizados por outros usuários. Previamente a isso a plataforma não possuía um sistema próprio de moderação, que se torna uma necessidade pelo aumento do número de usuários e as questões que envolvem o tipo de conteúdo postado por eles. O sistema de sinalizações passou a permitir ao usuário denunciar conteúdos que não se adequem às Diretrizes da Comunidade. Após ser sinalizado o conteúdo era revisado por um moderador humano para determinar se havia infração das normas e aplicar a sanção devida. Neste período, pelo menos até 2016, o sistema estava baseado justamente no trabalho dos moderadores humanos que faziam a avaliação e remoção de conteúdos potencialmente nocivos. O uso de sistemas automatizados para detectar este tipo de publicação já existia, mas como um sistema secundário e complementar.

[1] Tradução própria do original em inglês publicado em <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/>. Acesso em: 04/07/2022.

Essa estrutura se demonstrou ineficiente e insuficiente para dar conta do volume cada vez maior de conteúdo nocivo publicado na plataforma e isso começou a impactar no próprio modelo de negócios do YouTube com escândalos que envolviam a difusão de conteúdos nocivos na plataforma. Dois casos são emblemáticos neste sentido. O primeiro deles, conhecidos como "Adapocalypse" ocorreu em abril de 2017 quando uma investigação jornalística revelou que propagandas de grandes empresas estavam sendo veiculadas regularmente em vídeos e canais do YouTube de organizações terroristas, antissemitas e grupos neonazistas. A divulgação desta notícia gerou um grande impacto comercial para o YouTube quando mais de 250 grandes anunciantes retiraram seus anúncios da plataforma. No mesmo ano, um jornalista revelou que vídeos aparentemente voltados para crianças usavam personagens infantis de filmes ou animações inoculando conteúdo adulto, violento, sexual, obsceno ou inapropriado para crianças como drogas e consumo de bebidas alcoólicas. Com milhões de visualizações e veiculação de propaganda, o escândalo ficou conhecido como "Elsagate" (em uma referência à personagem Elsa, de animação da Disney, que era comumente utilizada nesses vídeos), também elevou a pressão de empresas que não queriam suas marcas vinculadas à conteúdo considerado perigoso para crianças (Ma, Kou 2021; Tarvin, Stanfill 2022).

Diante desses recorrentes problemas e o impacto financeiro que isso significa, em 2017 a estrutura baseada em *flags* e moderadores humanos sofreu duas mudanças importantes nos anos seguintes. A primeira delas, em 2017 quando houve um primeiro movimento de dar maior protagonismo para os sistemas de IA. Esses, passaram a ser centrais no sistema de *flags*:

> Em 2017, expandimos nosso uso da tecnologia de aprendizado de máquina para ajudar a detectar conteúdo potencialmente violador e enviá-lo para análise humana. O aprendizado de máquina é adequado para detectar padrões, o que nos ajuda a encontrar conteúdo semelhante (mas não exatamente igual) a outro conteúdo que já removemos, mesmo antes de ser visualizado. Esses sistemas são particularmente eficazes para sinalizar conteúdo que geralmente parece o mesmo, como spam ou conteúdo adulto. As máquinas também podem ajudar a sinalizar discursos de ódio e outros conteúdos violadores, mas essas categorias são altamente dependentes do contexto e destacam a importância da revisão humana para tomar decisões diferenciadas. Ainda assim, mais de 87% dos 9 milhões de vídeos que removemos no segundo trimestre de 2019 foram sinalizados pela primeira vez por nossos sistemas automatizados[2].

Dados no relatório de transparência do YouTube publicado pelo Google[3] apresentam informação sobre a origem de detecção de vídeos removidos entre 2017 e 2021. É

[2] Ver em <https://blog.youtube/inside-youtube/responsible-policy-enforcement-during-covid-19/> e <https://transparencyreport.google.com/youtube-policy/removals?hl=na>. Acesso em: 04/07/2022.
[3] Ver em <https://transparencyreport.google.com/youtube-policy/removals?hl=na>. Acesso em: 06/07/2022.

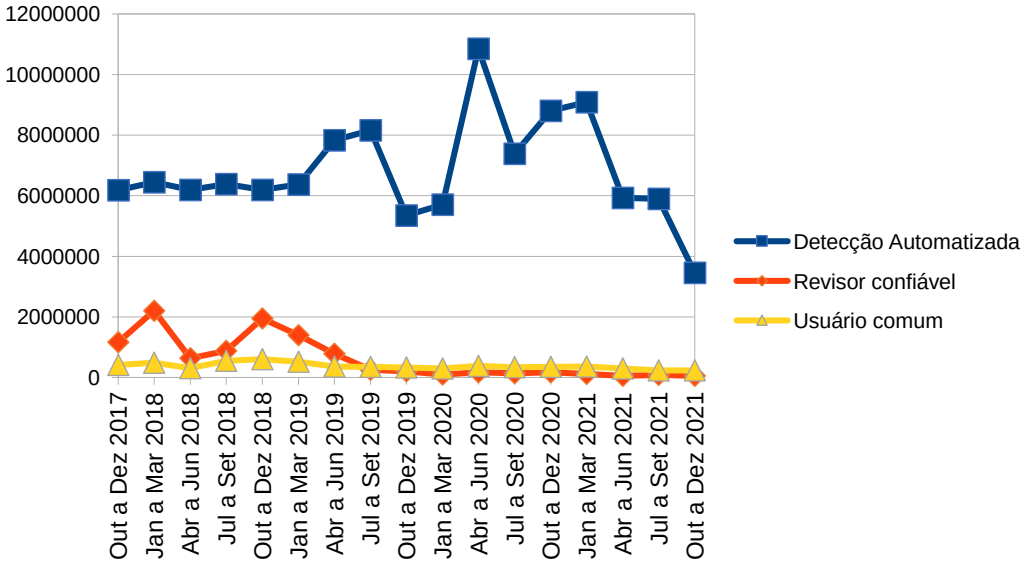
importante notar que a empresa não disponibiliza dados anteriores a esse período, que corresponderia apenas ao total de vídeos detectados por moderadores humanos ou sinalizados por usuários. A tabela 1 e o gráfico 1 apresentam o total de vídeos encontrados por sistemas automatizados, por revisores confiáveis, parte de um programa introduzido pelo YouTube em parceria com órgãos governamentais e ONGs para localizar conteúdos ilícitos; e usuários regulares.

Tabela 1. Vídeos removidos segundo origem de detecção.

	Detecção automatizada	**Revisor confiável**[4]	**Usuário comum**[5]	**Total**
Out a Dez 2017	6182263	1159861	409747	7751871
Jan a Mar 2018	6441778	2203306	491318	9136402
Abr a Jun 2018	6194346	635745	302939	7133030
Jul a Set 2018	6387658	878273	547142	7813073
Out a Dez 2018	6190148	1942913	603696	8736757
Jan a Mar 2019	6372936	1396945	520430	8290311
Abr a Jun 2019	7833449	783228	365339	8982016
Jul a Set 2019	8163067	253178	345435	8761680
Out a Dez 2019	5344863	202105	327876	5874844
Jan a Mar 2020	5711586	91203	300407	6103196
Abr a Jun 2020	10849634	167318	382499	11399451
Jul a Set 2020	7390963	139408	340694	7871065
Out a Dez 2020	8800082	165180	354429	9319691
Jan a Mar 2021	9091315	114061	362110	9567486
Abr a Jun 2021	5927201	54339	296454	6277994
Jul a Set 2021	5901241	85791	233349	6220381
Out a Dez 2021	3451691	59240	232737	3743668

[4] Dados representam o total de vídeos sinalizados por integrantes do programa de revisores confiáveis da plataforma, descrito mais adiante no texto.
[5] Valores referentes ao total de vídeos sinalizados por indivíduos que utilizam a plataforma cotidianamente como consumidores, sem relação direta com a empresa.

Figura 1. Relação dos fenômenos observados no estudo.



Os dados apontam a prevalência dos sistemas automatizados para detectar vídeos removidos do YouTube, com destaque para o crescimento no período inicial da pandemia de Covid-19, quando houve maior rigor da plataforma no uso de IA para remover vídeos. Nota-se a queda acentuada nas detecções automatizadas no período de 2021, sem explicação oficial, e também a diminuição da importância do programa de revisores confiáveis a partir do final de 2019.

O sistema de sinalização e o trabalho de moderadores humanos serviu de base para estabelecer mecanismos automatizados de detecção de conteúdos. Baseado em dados de milhões de solicitações e remoções para treinar e com o avanço da implementação de sistemas de IA, os algoritmos de automatização passaram não apenas a sinalizar, mas também realizar a remoção automatizada de conteúdos antes que pudessem ser vistos por usuários ou moderadores humanos.

Um segundo movimento que reforçou a centralidade dos sistemas automatizados na moderação de conteúdo do YouTube se deu em 2020, devido à crise causada pela pandemia de Covid-19 e a redução de pessoal para moderação de conteúdo. O YouTube mudou a estrutura de funcionamento da moderação, passando a utilizar mais *machine learning*. Tal mudança ocorreu em um cenário de aumento do tráfego de conteúdo na plataforma, com mais conteúdos nocivos circulando (especialmente negacionismo anti-vacina ou desinformação envolvendo o Coronavirus). Isso impactou em mais remoção de vídeos segundo dados do relatório do segundo trimestre de 2020. Entre as opções de diminuir a capacidade de moderação ou expandir, a empresa preferiu a última consciente de que os efeitos poderiam causar a remoção de vídeos mesmo que não infringissem normas e que não haveria moderação humana para validar o trabalho dos sistemas de IA[6].

[6] Tradução própria do original em inglês disponível em <https://blog.youtube/inside-youtube/responsible-policy-enforcement-during-covid-19/>. Acesso em:04/07/2022.

Em termos de volume desses recursos humanos, o YouTube apresenta em seus documentos dados que constam que aproximadamente 10.000 pessoas trabalham com moderação de conteúdo, sem especificar nos documentos publicados se são funcionários contratados da empresa ou empregados terceirizados contratados especificamente para moderar conteúdos. Não é possível indicar se esse é o número total de funcionários dedicados apenas ao YouTube ou para o Google como um todo, bem como a empresa terceirizada responsável pelos contratos e por realizar as remoções e o treinamento dos mecanismos automatizados de detecção e remoção. A ausência desse tipo de informação reforça a opacidade do sistema e a dificuldade em estabelecer formas de *accountability* sobre os dados apresentados e as capacidades tanto dos moderadores humanos quanto dos sistemas automatizados.

Dados do relatório de transparência da empresa reforçam essa diretriz de centralidade gradual das máquinas, que passaram a ter maior papel na remoção de conteúdos. De todo modo, o trabalho de moderadores humanos continua existindo, mas agora como um dispositivo do sistema (invertendo a lógica inicial) tendo agora a função de alimentar bancos de dados de sistemas automatizados que ajudem a realizar as remoções; fazer a revisão de determinados conteúdos que foram sinalizados ou removidos considerados controversos; e agir no sistema de apelações, isto é, avaliando críticas ou pedidos de reavaliação de usuários que tiveram seus conteúdos afetados ou sinalizados.

> Normalmente contamos com uma combinação de pessoas e tecnologia para fazer cumprir nossas políticas. O aprendizado de máquina ajuda a detectar conteúdo potencialmente prejudicial e o envia a revisores humanos para avaliação. A revisão humana não é apenas necessária para treinar nossos sistemas de aprendizado de máquina, mas também serve como uma verificação, fornecendo *feedback* que melhora a precisão de nossos sistemas ao longo do tempo[7].

Atualmente, o processo de moderação no YouTube adota diversas formas de sanções e, em todas elas, os sistemas de IA tem um papel relevante a desempenhar:

Strikes – O YouTube implementou em 2019 o sistema de *strikes* que utiliza atualmente[8], que serve como advertência inicial de ocorrências de violações das Diretrizes da Comunidade, com efeitos imediatos para o usuário, caráter cumulativo e válidos por um período de 90 dias. O criador é inicialmente notificado sobre o conteúdo referido, a política violada e como afeta o canal, passando a seguir para o sistema de *strikes* em caso de nova violação. O sistema de *strikes* regula por períodos de tempo determinados a inclusão de novas postagens e funcionalidades do usuário como transmissões ao vivo e edição nos conteúdos já publicados. O primeiro *strike* congela postagens e funções por uma semana, o segundo por duas e o terceiro deleta a conta[9].

[7] Ver em <https://blog.youtube/news-and-events/making-our-strikes-system-clear-and/> Acesso em: 29/07/2022.
[8] Ver em <https://support.google.com/youtube/answer/2802032>. Acesso em: 29/07/2022.
[9] Ver em <https://support.google.com/transparencyreport/answer/9209072>. Acesso em: 29/07/2022.

O usuário pode apelar a qualquer momento para uma nova revisão por moderador humano[10].

Degradação do alcance – A diminuição de alcance ocorre nos casos em que não há infração das normas, mas vídeos considerados impróprios por conter desinformação ou conteúdo considerado ofensivo ou inapropriado para determinados públicos podem ter seu alcance reduzido ao ser removido do sistema de recomendações. Isso envolve também restrição por idade e limitação de recursos, neste caso, "esses vídeos permanecerão disponíveis no YouTube, mas apresentarão uma mensagem de aviso no início e alguns recursos serão desativados, incluindo compartilhamento, comentários e posicionamento entre os vídeos sugeridos." [11]. O algoritmo que determina o sistema de recomendações mede dados sobre audiência e interesse de usuários para ranquear sugestões de vídeos sobre temas relacionados ao vídeo acessado e retira das recomendações conteúdos considerados no limite daquilo que é aceitável segundo as Diretrizes da Comunidade.

Remoção de conteúdo – Neste caso, vídeos que são considerados violações da política da plataforma ficam indisponíveis para o público (só podem ser acessados internamente, na conta do usuário que postou) mas não possuem mais visibilidade externa, mesmo que o *link* anterior seja compartilhado (aparecerá mensagem informando que o conteúdo está indisponível)[12]. Um conteúdo é removido do YouTube somente após ser sinalizado por usuário ou pela detecção automatizada e revisado por um moderador humano, que deleta o vídeo com base nas Diretrizes da Comunidade, ou pelo sistema de *machine learning* treinado para realizar remoções automaticamente em caso de infração de direito autoral, conteúdo de violência e terrorismo ou de abuso infantil e conteúdos já revisados por moderadores humanos.[13] Os moderadores treinados removem conteúdos violadores, exceto em casos em que há contexto educacional, científico, documental ou artístico identificado no conteúdo e nos metadados apresentados pelo usuário.[14]

Desabilitação de comentários – O YouTube utiliza sistemas automatizados e sinalização de usuários para detectar e remover comentários ofensivos ou abusivos, especialmente em casos de conteúdo voltado para público infantil, chegando até mesmo a desativar a possibilidade de comentários caso necessário.[15] A empresa também oferece ao criador de conteúdo a possibilidade de limitar comentários nos

[10] Ver em <https://support.google.com/transparencyreport/answer/9209072>. Acesso em: 29/07/2022.
[11] Ver em <https://support.google.com/transparencyreport/answer/9209072>. Acesso em: 29/07/2022.
[12] Ver em <https://support.google.com/transparencyreport/answer/9209072>. Aesso em: 29/07/2022.
[13] Ver em <https://www.youtube.com/howyoutubeworks/policies/community-guidelines>. Acesso em: 29/07/2022.
[14] Ver em <https://blog.youtube/news-and-events/5-ways-were-toughening-our-approach-to/>. Acesso em: 29/07/2022.
[15] Ver em <https://blog.youtube/news-and-events/faster-removals-and-tackling-comments/>. Acesso em: 29/07/2022.

vídeos publicados e utiliza sistemas automatizados para detectar e remover a maior parte dos conteúdos de spam na plataforma.[16]

Restrições de funcionalidades – Em casos determinados o YouTube pode desativar funções de compartilhamento e comentários para conteúdos no limite das Diretrizes da Comunidade e restringir transmissões ao vivo quando o usuário apresenta *strikes* por violação das normas. É possível também que vídeos com metadados enganosos, que descrevam contexto diferente do que apresenta o conteúdo, possam ser limitados como privados, inacessíveis ao público.[17]

Desmonetização – Nos casos em que não há violação direta das normas com relação a violência e terrorismo, mas o conteúdo apresenta controversa com relação a religião ou supremacista, uma das possíveis sanções além da desativação de comentários e funções é a desativação da captação financeira para o criador de conteúdo.[18]

Banimento do usuário – Como exposto no sistema de *strikes*, quando um usuário acumula três violações em um período de 90 dias o YouTube deleta o canal. O usuário pode ser banido e sua conta deletada após repetidas violações das Diretrizes da Comunidade ou Termo de Uso, ou abuso grave das normas como spam, pornografia, comportamento predatório ou por infração de direito autoral[19]. Nos casos de violação de direito autoral o usuário é recomendado a procurar o setor responsável pela área no YouTube. Em todos os outros é possível apelar por meio de um formulário disponibilizado pela plataforma.

AUTOMATIZAÇÃO, PROBLEMAS E DESAFIOS

Os documentos publicados pela plataforma frequentemente destacam números gerais de remoção de conteúdo buscando demonstrar eficiência tecnológica do sistema, principalmente no que diz respeito à adoção e ampliação do uso de Inteligência Artificial, especialmente a partir de 2017. Porém, os relatórios e cifras escondem problemas e dimensões políticas que afetam direitos e demonstram que a plataforma enfrenta ainda muitos desafios e fragilidades nesses processos. Podemos enumerar seis questões mais proeminentes que a análise dos documentos demonstraram e que merecem especial atenção na construção de uma política de moderação de conteúdo mais efetiva:

a) Mais opacidade, menos transparência. Um dos problemas mais recorrentemente mencionado por diversos analistas se refere à precariedade da transparência sobre o uso de IA na moderação de conteúdo (Suzor et al 2019; Ananny, Crawford 2018; Puddephatt 2021). Após muita pressão externa, relatórios de transparência, textos

[16] Ver em <https://support.google.com/transparencyreport/answer/9209072>. Acesso em: 29/07/2022.
[17] Ver em <https://blog.youtube/news-and-events/an-update-on-our-commitment-to-fight-terror/>. Acesso em: 29/07/2022.
[18] Ver em <https://support.google.com/youtube/answer/2802168>. Acesso em 29/07/2022.
[19] Tradução própria do original em inglês, disponível em <https://blog.youtube/inside-youtube/responsible-policy-enforcement-during-covid-19/>. Acesso em: 04/07/2022.

explicativos em *blogs* e outros documentos passaram a ser publicados regularmente pela plataforma. A partir dos dados oferecidos pelo YouTube em seu relatório de transparência e nas comunicações públicas da empresa sobre suas práticas, anúncios de medidas e resultados, não é possível determinar com precisão se os sistemas de IA utilizados pela empresa realizam o trabalho esperado de remover conteúdos infratores mantendo conteúdos lícitos. Além disso, a opacidade também ocorre no processo de justificação para aqueles que sofreram penalidades, como demonstra o estudo de Ma e Kou (2021):

> Algorithmic opacity in moderation refers to the situations where YouTubers who experienced algorithmic punishments felt confused and had no clues of how algorithms made decisions. Even if the adjudication of a moderation case is clear-cut to most people, the YouTuber who received the penalty could experience it differently and feel confused. We found that YouTubers experienced algorithmic opacity at multiple layers. Moderation decisions could sometimes puzzle the ordinary audience (MA; Kou, 2021 p. 9).

Se antes, num cenário de predominância da moderação humana, já havia um cenário de pouca transparência, com a intensificação da presença de sistemas de IA essa insuficiência aumentou pois tornou o processo de moderação ainda mais opaco, adicionando novas áreas obtusas e novos problemas de invisibilidade. Neste sentido, podemos sintetizar dois problemas principais.

Primeiro, falta de clareza sobre a relação entre moderação humana e sistemas de IA. A empresa aponta que o total de vídeos removidos demandaria o trabalho de 180.000 pessoas em regime de 40 horas por semana. Apesar das afirmações de sucesso o YouTube não apresenta números para comparar, apenas destaca que aplicará a mesma tecnologia usada para detectar extremismo e violência para detectar e remover conteúdos de discurso de ódio e abuso infantil. Como analisa Roberts (2019), a maior parte do trabalho dos moderadores humanos é invisibilizado pelas empresas que não são transparentes sobre as práticas de moderação, seja na criação e aplicação das normas, total de empregados e empresas que realizam o trabalho e dados sobre os conteúdos removidos. Ao mesmo tempo, a plataforma tem destacado em seus documentos a importância e sinergia da combinação de pessoas e máquinas, porém, não explica em detalhes como este relacionamento funciona em termos de hierarquia; fluxograma.

Segundo, ocorrem incompletude e superficialidades nos dados sobre como a Inteligência Artificial realmente funciona. A plataforma tem optado por uma linguagem simplificada e com poucos dados técnicos claramente direcionada para ser mais fácil de ser assimilada pelo usuário médio. Mas isso esconde a falta de transparência sobre o real funcionamento da plataforma sobre questões relacionadas aos algoritmos. Isso também é reforçado por outros estudos:

> The YouTube company announced that "we had been working on an even more effective classifier, that will identify and remove

> predatory comments. [...] We accelerated its launch and now have a new comments classifier in place that is more sweeping in scope, and will detect and remove 2X more individual comments" (YouTube Creator Blog, 2019). This may or may not have been intentionally opaque, but nevertheless uses the term "classifier," which as a technical term referring to an algorithmic or machine learning process does not appear in standard dictionaries like Merriam-Webster, without explaining what it means.[...] What it doesn't do is specify what will be classified (usernames associated with that behavior in other comments, word choice in comments, use of time stamps to call attention to particular parts of the video, something else?), using what criteria (with 51% certainty? 90%?), nor indeed what number it is they claim to have doubled (Tarvin; Stanfill, 2022, p. 823).

Ao mesmo tempo, a plataforma publica parcialmente as informações ao não produzir dados globais sobre conteúdo que beira a infração das normas. Possui apenas dados de mercados de países anglófonos.

b) Aumento de violações de direitos pela plataforma. Dados apresentados no relatório de transparência da empresa, que possui informações apenas posteriores à implementação de sistemas de *machine learning* para detecção e remoção de conteúdos, apontam os sistemas automatizados são responsáveis pela maior parte da detecção de vídeos, muitos deles removidos antes sequer de ser visualizados por outro usuário. Entre 2017 e 2021 a plataforma removeu 83 milhões de vídeos e 7 bilhões de comentários por violar as regras. Cerca de 94% dos vídeos foram removidos por IA, 75% desse total antes que tivesse 10 visualizações. Embora uma boa parte desse volume trate de conteúdos nocivos, a plataforma também admite que há efeitos colaterais devido aos erros no processo de moderação que afeta também usuários e conteúdos inocentes:

> Para algumas áreas de política delicadas, como extremismo violento e segurança infantil, aceitamos um nível mais baixo de precisão para garantir a remoção do maior número possível de conteúdos violadores. Isso também significa que, nessas áreas especificamente, uma quantidade maior de conteúdo que não viola nossas políticas também foi removida. A decisão de aplicar em excesso nessas áreas de política – com muita cautela – levou a um aumento de mais de 3 vezes nas remoções de conteúdo que nossos sistemas suspeitavam estar ligado ao extremismo violento ou potencialmente prejudicial às crianças. Isso inclui desafios, desafios ou outros conteúdos postados inocentemente que possam colocar menores em risco[20].

Esses dados demostram que a plataforma coloca os erros de moderação como algo natural, sem dar a devida importância e não disponibiliza detalhes sobre esses erros que são, na verdade, violações da plataforma sobre a liberdade de expressão. Não é possível determinar se conteúdos que não violem as regras são removidos e em que

[20] Tradução própria do original em inglês disponível em <https://blog.youtube/inside-youtube/look-how-we-treat-educational-documentary-scientific-and-artistic-content-youtube/>. Acesso em: 04/07/2022.

escala isso ocorre, apesar dos números indicarem que há um grande volume de vídeos e canais deletados, não oferece informações sobre como funciona o sistema de *Flagging*, ou como ocorre a comunicação como usuário sobre o conteúdo sinalizado, se removido e a razão da decisão.

Não por acaso, dados dos relatórios também apontam aumento substancial nas apelações e na reintegração de vídeos removidos, o dobro do trimestre anterior. As apelações podem ser vistas como um indicador que aponta potencialmente para o aumento de violações dos direitos dos usuários pela plataforma.

A falta de informações de contexto não permite elaborar análises mais profundas para identificar questões como motivo da remoção ou comparar com outros dados mas os dados apontam que há um evidente aumento dos casos de falsos positivos (Bright et al 2021).

O YouTube admite em seus documentos lidar com vídeos que, apesar de não infringirem as normas e por isso não serem removidos, devem sofrer sanções e, para isso, a plataforma remove determinados vídeos das possíveis sugestões nas recomendações. Trata-se de conteúdo limítrofe, que apesar de não infringir regras, contém desinformação e por isso são sancionados pela redução de alcance. A confiabilidade de uma informação, entre outros fatores, é analisada por moderadores humanos que determinam se um vídeo pode ser recomendado ou removido da plataforma. Vídeos limítrofes são retirados das recomendações, enquanto vídeos que infringem as normas são removidos da plataforma. Diminuir alcance e possibilidade de engajamento ao tornar disponível para uma parcela ínfima de público que pode acessar por outros meios que não pelo sistema de recomendações.

c) Julgamento automatizado e o problema do contexto. Um vídeo com símbolos nazistas e menções à Ku Klux Klan pode ser detectado por sistemas de IA, porém, isso por si só não significa que o conteúdo viola as normas por tratar de racismo. Isso pode ser tanto um vídeo de fato racista postado por grupos supremacistas como também pode ser um documentário ou uma aula sobre os perigos do racismo. Essa diferença só será percebida na análise qualitativa e de contexto no qual o conteúdo está inserido. Os sistemas de IA são ineficazes em captar contextos mais complexos ou elementos como ironia, paródias etc. A plataforma reconhece que este é um desafio para os sistemas automatizados que utiliza:

> Conteúdos semelhantes sinalizados pelo mesmo motivo podem conter sentidos ou apresentar ideias divergentes, de modo que uma poderia ser aceitável segundo as regras da plataforma e mesmo assim poderia ser removida. Para resolver esse problema é preciso alimentar o sistema com dados para facilitar a classificação de algoritmos e melhorar a eficiência do mecanismo. [...] identificação de contexto para determinar se vídeo possui conteúdo educativo, documental, científico ou artístico para determinar se o conteúdo pode permanecer na plataforma. Categorias como discurso de ódio, violência gráfica, violência organizada e desinformação possuem padrões mais elevados para identificar contexto. Conteúdos

> violentos ou que reproduzem perigo a crianças são removidos independentemente de contexto[21].

Para tentar lidar com esse problema, a plataforma chama a atenção dos usuários para a importância de estabelecer contexto nos vídeos, por meio de descrição ou *tags* para evitar sanções. Isto é, procura resolver o problema de contextualização repassando ao usuário a responsabilidade de oferecer explicações em detalhe no momento que um vídeo é enviado ao site. A descrição e outras informações colaboram para oferecer contexto. Na prática, uma medida ineficaz pois conteúdos nocivos podem simular descrições falsas dando ao vídeo metadados capazes de driblar os sistemas automatizados de moderação.

O YouTube não oferece informações sobre como o sistema funciona além das explicações nos relatórios de transparência e publicações do *blog* corporativo, que de maneira vaga e geral anunciam medidas ou informam sobre procedimentos na remoção de conteúdos.

d) Acirramento do efeito escala e seus impactos. Uma das características mais marcantes de sistemas de Inteligência Artificial diz respeito ao poder da repetição. Isto é, a capacidade de replicar estruturas, processos, procedimentos e escalar isso de modo automatizado, o que significa uma nova forma de poder (Silva 2020). Gillespie (2020) ressalta esta característica lembrando que "escala" não significa só "tamanho":

> Too often, when platform representatives point to their scale, they mean little more than the enormous number of users or amount of content. But scale is something more than size. Scale is about how the small can be made to have large effects; or how a process can be proceduralized such that it can be replicated in different contexts, and appear the same (GILLESPIE 2020, p. 2).

Em seus documentos, o YouTube explica que os vídeos analisados e julgados pelos moderadores servem como matéria-prima para o treinamento dos sistemas automatizados, isto é, os algoritmos assimilam os padrões de análise e julgamento dos moderadores humanos e reaplicam esses critérios em análises futuras:

> Nossas equipes revisaram manualmente mais de 2 milhões de vídeos detectados pelos sistemas automatizados que identificam conteúdo extremista. Com isso, coletamos um grande volume de exemplos para treinamento, que ajudam a aprimorar a tecnologia de sinalização por aprendizado de máquina[22].

Como há imprecisões e falhas tanto no processo automatizado quanto no julgamento do moderador humano, os sistemas tendem a herdar vieses e reproduzir erros ao não identificar contextos e causar remoções indevidas em larga escala, ainda que em percentuais aparentemente irrelevantes. Sobretudo porque, quando falamos de

[21] Ver em <https://support.google.com/transparencyreport/answer/9209072>. Acesso em: 29/07/2022.
[22] Ver em <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove>. Acesso em: 04/07/2022.

bilhões de vídeos, qualquer percentual ínfimo de remoção equivocada, significa milhares de direitos afetados.

Para Gillespie (2020) os sistemas de IA podem ser úteis para realizar grande parte do trabalho, como encontrar cópias de postagens já removidas, e deixar casos menos óbvios e que requerem mais cuidado para moderadores humanos, e questiona se a automatização e substituição de humanos na tarefa é a melhor resposta, apontando que sistemas de *machine learning* podem ajudar a realizar a função, mas não devem ser integralmente responsabilizadas.

e) Moderação automatizada guiada por princípios comerciais. A moderação de conteúdos faz parte do modelo de negócios e a maneira como as plataformas se protegem de riscos legais. O YouTube busca uma maneira de equilibrar a liberdade de expressão com o potencial de engajamento do que usuários publicam nas plataformas e circular a informação em um modelo de economia de atenção. Não há uma preponderância dos direitos como princípio norteador. Por isso, os documentos sobre moderação publicados pela plataforma deixam claro que o processo visa manter a comunidade mais segura, porém, sem alterar o modelo de negócios[23].

Outros estudos (Ma, Kou 2021; Tarvin, Stanfill 2022) também têm demonstrado que há uma moderação seletiva quanto ao tratamento das atividades dos perfis de grandes empresas em comparação a pequenos canais, como aponta pesquisa sobre usuários que tiveram conteúdos removidos pela plataforma:

> [...] the YouTuber compared the thread poster's videos with several famous YouTubers' ones and described the observations where algorithmic punishments were imposed unequally between small and large video creators, also resonating with Caplan's findings (MA; KOU, 2021, p. 11).

A importância e defesa da manutenção do modelo de negócios também fica evidente na definição de quatro áreas temáticas prioritárias de ação da moderação automatizada que são frequentemente citadas nos documentos: violência, terrorismo, infância e direitos autorais. Essas áreas priorizadas pela empresa para implementação de sistemas automatizados utilizam recursos como Copyright Match e ContentID[24] para identificar e remover conteúdos protegidos por direito autoral (Gray et al 2020) e PhotoDNA para identificar casos de abuso infantil e terrorismo, como indicam os relatórios da plataforma.

Para esses casos, as máquinas detectam conteúdos com essas violações já no *upload*. Essas áreas não foram definidas por acaso. São resultado da pressão de anunciantes principalmente após os casos de escândalos como o "Adapocalypse" e "Elsagate". A proteção ao modelo de negócios passa pela manutenção dos espaços seguros para

[23] Ver em <https://blog.youtube/news-and-events/how-flagging-works/>. Acesso em: 04/07/2022.
[24] Ver em <https://blog.youtube/intl/pt-br/news-and-events/relatorio-de-transparencia-de-copyright-do-youtube-mostra-como-funciona-nosso-sistema-de-garantia-de-respeito-direitos-autorais/>. Acesso em: 04/07/2022.

usuários e anunciantes. Anunciantes não querem sua publicidade associada a conteúdos violentos ou abusivos. No caso de direitos autorais, este é um dos segmentos mais desenvolvidos no monitoramento automatizado, pois permite que grandes empresas de conteúdo controlem e retirem do ar (elas próprias) conteúdos protegidos por *copyright*.

Os resultados apontados nos relatórios sobre a moderação dessas áreas prioritárias demonstram um volume expressivo de ações e crescentes taxas de sucesso em detectar e remover conteúdo antes que seja sequer visualizado. Porém, não há detalhamento dos dados capazes de possibilitar comparação e compreensão mais abrangente sobre as causas e remoções, não sendo possível afirmar que os números garantem a eficiência ou geram mais violações.

f) Insuficiência nos processos de participação e *accountability*. Sistemas de IA são impactantes e podem afetar a vida de milhares de indivíduos. Por isso, diversos analistas (Cath et al 2018; Donahoe, Metzger 2019; WirtzIRTZ et al 2020; 2021; Puddephat 2021) e organizações (GPAI 2021; ITU 2021; OECD 2022) têm destacado a importância da abertura e participação mais ampla dos vários *stakeholders* para validar e aprimorar tais ferramentas, dando assim mais pluralidade e minimizando seus efeitos colaterais. Ao mesmo tempo, deve-se criar processos contínuos de *accountability* tendo em vista a constante evolução e readaptação dessas ferramentas.

Os documentos analisados não trazem informações capazes de assegurar tal dimensão. Para Tarvin E Stanfill (2022) o YouTube na verdade pratica aquilo que chamam de "*governance-washing*". Isto é, quando uma organização tenta melhorar sua reputação pública divulgando informações parciais sobre suas atividades, selecionando determinados dados positivos e reforçando imagem de uma cultura corporativa comprometida, incluindo a ideia de abertura e participação sem que isso ocorra efetivamente:

> Another way the YouTube company's governance was governance-washing is that they offloaded responsibility for governance onto users. That is, at the same time as emphasizing their actions, especially speed, volume, and technology, the company also asked those who use the platform to contribute to platform governance–not in the sense of involvement in setting policy, but as the eyes and ears and reporting clicks of enforcement. That is, in a playbook common across social media, users are asked to do some of the work of content moderation. (TARVIN; STANFILL, 2022, p. 826-827).

Neste sentido, os documentos do YouTube tentam demonstrar que suas políticas são construídas com especialistas, acadêmicos e organizações de defesa de direitos civis:

> Continuamos a investir na rede de mais de 150 acadêmicos, parceiros governamentais e ONGs que trazem conhecimentos valiosos para nossos sistemas de fiscalização, como o Centro Internacional para o Estudo da Radicalização no King's College London, Liga Anti-Difamação e Instituto de Segurança Online da

> Família. Isso inclui adicionar mais parceiros focados na segurança infantil de todo o mundo, como Childline South Africa , ECPAT Indonesia e Parents' Union on Net da Coréia do Sul.

Este é o máximo de detalhamento que os documentos trazem sobre o quadro mais amplo de inserção de *stakeholders* nos processos de moderação. Apesar da menção de algumas parcerias, não há critérios claros sobre como se dá essa participação, qual seria a lista completa de todos os parceiros envolvidos e qual é exatamente o papel dessas parcerias e como esses atores operam dentro do sistema.

Um bom exemplo da instabilidade e ausência de uma política mais consistente e ampla de participação é o programa de revisores confiáveis. Iniciado pela plataforma em 2012, esta iniciativa foi criada para fornecer ferramentas a usuários regulares com altos índices de precisão na sinalização de vídeos e representantes de organizações não-governamentais com interesse na efetiva aplicação das normas para reportar conteúdo violador e repassar a decisão de remover ao time de moderadores humanos. Nos anos seguintes, e com auge em 2017, o YouTube investiu no crescimento do programa para aumentar a precisão na remoção de vídeos, chegando a dobrar o número de revisores confiáveis para tratar sobre abuso infantil[25] e expandir o número de ONGs e cobertura de atuação para tratar sobre terrorismo e discurso de ódio[26]. Porém, a partir de 2018 houve gradual desativação do programa, acentuada pela diminuição de pessoal devido à pandemia de Covid-19 e culminando na reformulação em 2022. Baseando-se nos avanços nos sistemas automatizados para detecção, o YouTube cortou os revisores confiáveis voluntários e restaram apenas os especialistas de ONGs. Oficialmente, a plataforma não publicou informações referentes aos cortes no programa de revisores confiáveis em seus *sites*, mas há declaração de representante da empresa sobre isso externamente, como no site Tubefilter em abril de 2022[27].

Todo esse quadro demonstra que a plataforma não tem um programa efetivo, transparente e de longo prazo com critérios claros que garantam diversidade e independência de *stakeholders* capazes de contribuir de forma independente e objetiva com a resolução dos problemas e desafios que o processo de moderação enfrenta.

[25] Ver em <https://blog.youtube/news-and-events/5-ways-were-toughening-our-approach-to/04/07/2022> Acesso em 04/07/2022.
[26] Ver em <https://blog.youtube/news-and-events/an-update-on-our-commitment-to-fight-terror/> e <https://blog.youtube/news-and-events/an-update-on-our-commitment-to-fight/>. Acesso em 04/07/2022.
[27] Ver em <https://www.tubefilter.com/2022/04/29/youtube-trusted-flagger-program-individuals-organizations>. Acesso em 12/08/2022.

CONSIDERAÇÕES FINAIS

Este artigo teve como principal objetivo compreender e caracterizar o papel de sistemas de inteligência artificial na moderação de conteúdos no YouTube, tendo como pano de fundo a dimensão política sobretudo no tocante a direitos individuais.

O crescimento da plataforma demanda que cada vez mais sejam necessárias regras e procedimentos para regular os conteúdos publicados pelos usuários de modo a evitar que pornografia, violência, discursos de ódio, desinformação e outras questões de interesse público circulem livremente nas redes. Esse tipo de conteúdo negativo interfere tanto socialmente, sobre o tipo de informação que pode ser acessada nas redes sociais por um grande público, quanto para a plataforma em relação ao modelo de negócios, considerando que os anunciantes não querem ser associados a esse tipo de conteúdo. Ao longo da última década o sistema de moderação do YouTube, originariamente baseado em moderadores humanos agindo sob demanda a partir da sinalização provocada pelos usuários, passou a depender de algoritmos para realizar a maioria das suas funções. O mesmo pode ser dito de outras plataformas de mídias sociais pois a automatização de sistemas de moderação foi a solução encontrada pelas empresas o problema de escala e de eficiência na aplicação de sanções.

A análise dos documentos, políticas e diretrizes publicadas pela plataforma demonstrou que houve nos últimos anos um claro movimento de centralização dos sistemas de IA no processo de moderação de conteúdo. Se antes a moderação humana prevalecia como eixo central com sistemas de IA, enquanto um suporte, esse quadro se inverteu, transformando as equipes de funcionários em dispositivos que passaram a servir ao sistema central automatizado de moderação. Isso ocorreu principalmente a partir de 2017, sobretudo em um contexto de escândalos e pressões que a plataforma sofreu diante da veiculação de marcas de importantes anunciantes a conteúdo nocivo. Assim, algoritmos de *machine learning* passaram a atuar, direta ou indiretamente, em todas as sanções aplicadas pela plataforma como *strikes*, degradação do alcance, remoção de conteúdo, desabilitação de comentários, restrições de funcionalidades, desmonetização e banimento do usuário.

Neste cenário, o estudo identificou os 6 problemas principais que a moderação automatizada tem acarretado. (1) Os sistemas de IA trouxeram mais opacidade ao processo de moderação. (2) A plataforma passou a violar com mais frequência direitos individuais como efeito colateral do aumento das remoções de conteúdos nocivos. (3) O julgamento automatizado tende a ignorar o problema do contexto, penalizando conteúdos inocentes. (4) Houve um acirramento do efeito escala, isso é, a replicação de erros de avaliação ganharam maior dimensão e alcance.(5) Identificou-se uma moderação automatizada centrada em valores comerciais e priorizada por este. (6) Há uma insuficiência nos processos de participação e *accountability* onde a empresa não criou efetivamente mecanismos de abertura para incorporar visões e demandas de *stakeholders*.

Diante da tendência de *datificação* social e digitalização de processos de comunicação a ocorrência de conteúdos *online* nocivos também seguirá crescente. Políticas e

processos de moderação e restrições serão inevitáveis e necessárias. Porém, devido à dimensão pública e ao impacto político que isso acarreta é preciso que haja processos mais maduros e justificados de intervenção. Os efeitos colaterais, que violam direitos individuais, não podem ser tratados como parte normal do processo. Dada a dimensão ampla que o fenômeno da plataformização da opinião pública representa, isso significaria implantar um regime permanente de violações de direitos onde o combate a conteúdos nocivos se tornaria um jogo de soma negativa no qual usuários inocentes seriam punidos e teriam seus direitos frequentemente violados em um grau crescente em termos de escala, através de um sistema repetitivo e automatizado de violações, contraditoriamente, em nome do combate a violações.

Por isso, o uso de Inteligência Artificial não deve ser naturalizado como a parte central do sistema, agindo em escala potencialmente afetando direitos. Sistemas automatizados são ferramentas potentes para auxiliar nesses processos, mas a sua eficiência não pode ser medida apenas em termos quantitativos genéricos. É preciso reposicionar o papel da supervisão humana e possibilitar avaliações mais consistentes em termos qualitativos detalhados. Plataformas como o YouTube precisam de mais abertura, transparência e precisam ser mais *accountable* para enfrentar problemas dessa magnitude.

REFERÊNCIAS

ANANNY, Mike e CRAWFORD, Kate, 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media and Society, vol. 20, no. 3, p. 973–989.

BEACH, Derek, PEDERSEN, Rasmus Brun, 2013. Process-Tracing Methods. Foundations and Guidelines. Ann Arbour, University of Michigan Press.

BUNTAIN, Cody, BONNEAU, Richard, NAGLER, Jonathan e TUCKER, Joshua, 2021. YouTube Recommendations and Effects on Sharing Across Online Social Platforms. Proceedings of the ACM on Human-Computer Interaction. 5. 1-26. 10.1145/3449085. [Acesso em 24 julho 2022]. Disponível em: https://arxiv.org/abs/2003.00970.

BURROUGHS, Benjamin, 2017. YouTube Kids: The App Economy and Mobile Parenting. Social Media + Society, vol. 3 no. 2 p. 1–8.

BRIGHT, J et al. Addressing Hate Speech On Social Media: Contemporary Challenges. Paris: UNESCO, 2021.

CATH, C. et al. Artificial Intelligence and the "Good Society": the US, EU, and UK approach. Science and Engineering Ethics, v. 24, n. 2, 2018, p. 505-528.

CRAWFORD, Kate, GILLESPIE, Tarleton, 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media & Society. 2016 vol.18 no. 33 p. 410-428.

DONAHOE, E.; METZGER, M. M. Artificial intelligence And Human rights. Journal of Democracy, v.30, n. 2, p. 115-126, 2019.

FALLETI, Tulia G, 2016. Process tracing of extensive and intensive processes. New Political Economy, vol. 21, no. 5, p. 455–462.

GILLESPIE, Tarleton, 2018. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. New Haven: Yale University Press.

GILLESPIE, Tarleton, 2020. Content moderation, AI, and the question of scale. Big Data & Society. 2020. Vol 7 , no. 2.

GORWA, Robert, BINNS, Reuben e KATZENBACH,Christian, 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society vol. 3 no. 1 p. 1–15.

GRAY, Joanne E e SUZOR, Nicolas P, 2020. Playing with machines: Using machine learning to understand automated copyright enforcement at scale. Big Data & Society. Janeiro 2020.

GPAI. The Global Partnership on Artificial Intelligence. Responsible AI Working Group ReporT 2021 - GPAI Paris Summit. Disponível em < https://www.gpai.ai/projects/responsible-ai/gpai-responsible-ai-wg-report-november-2021.pdf > Acesso em 17 março 2022.

ITU. International Telecommunication Union. United Nations Activities on Artificial Intelligence (AI)2021. Geneva, 2021. Disponível em < https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2021-PDF-E.pdf > Acesso em 12 julho 2022.

KUMAR, Sangeet, 2019. The algorithmic dance: YouTube's Adpocalypse and the gatekeeping of cultural content on digital platforms. Internet Policy Review, vol. 8 no. 2.

MA, Renkai e KOU, Yubo, 2021. How advertiser-friendly is my video?: YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2).

OECD. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. 2022. Disponível em < https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 > Acesso em 25 julho 2022.

PUDDEPHATT. Andrew, 2021 Letting the Sun Shine In: Transparency and Accountability in the Digital Age. Paris: UNESCO.

ROBERTS, Sarah T, 2019. Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven: Yale University Press.

SILVA, Luiz Rogério Lopes, BOTELHO-FRANCISCO, Rodrigo Eduardo., ALISSON AUGUSTO DE OLIVEIRA, Alisson Augusto e PONTES, Vinícius Ramos, 2019. A gestão do discurso de ódio nas plataformas de redes sociais digitais: um comparativo entre Facebook, Twitter e Youtube. Revista Ibero-Americana De Ciência Da Informação, vol. 12 no 2, p. 470–492.

SILVA, Sivaldo P. da, 2020. Democracia, Inteligência Artificial e desafios regulatórios: direitos, dilemas e poder em sociedades datificadas. Revista Eletrônica do Programa de Pós-Graduação da Câmara dos Deputados, v. 13, p. 226-248.

SUZOR, Nicolas. P. et al, 2019. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. International Journal of Communication, 13, p. 1526–1543.

TARVIN, Emily e STANFILL, Mel, 2022. YouTube's predator problem: Platform moderation as governance-washing, and user resistance. Convergence, vol. 28, no. 3, p. 822–837.

WIRTZ, Bernd W. et al, 2020. The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. International Journal of Public Administration, p. 1-12.