Transparência sobre moderação de conteúdo

em políticas de comunidade

iris

Transparência sobre moderação de conteúdo em políticas de comunidade

AUTORIA
Gustavo Ramos Rodrigues
Lahis Pasquali Kurtz

REVISÃO
Luíza Couto Chaves Brandão

PROJETO GRÁFICO, CAPA, DIAGRAMAÇÃO E FINALIZAÇÃO
Felipe Duarte

PRODUÇÃO EDITORIAL
Instituto de Referência em Internet e Sociedade

COMO CITAR EM ABNT
RODRIGUES, Gustavo; KURTZ, Lahis. **Transparência sobre moderação de conteúdo em políticas de comunidade**. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2020. Disponível em: <https://bit.ly/3nUbXYh>. Acesso em: dd mmm aaaa.

Os autores agradecem as contribuiçoes valiosas de Heloisa Massaro (InternetLab) e Jamila Venturini (Derechos Digitales)

INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE

SUMÁRIO

Resumo

Plataformas de conteúdo gerado pelo usuário marcam uma categoria de serviços na internet que adota proativamente medidas interventivas sobre as informações circuladas em seu âmbito. Este paper pretende analisar a transparência da moderação de conteúdo gerado pelo usuário, uma vez que o conhecimento sobre as práticas adotadas permite a conciliação entre o limite ao conteúdo circulado e o exercício democrático da liberdade de expressão. Este estudo avalia em que medida as políticas de comunidade das 8 plataformas online mais usadas no Brasil são instrumentos de transparência que permitem accountability. Para isso, é realizada uma revisão de literatura em que são considerados os contextos histórico, jurídico e cultural que conformam a elaboração de políticas de comunidade, a partir de suas finalidades, e que papel elas exercem na estratégia de transparência das plataformas. Em seguida, por meio de análise de conteúdo, observa-se como esses textos atendem à demanda de tornar públicos os meios de detecção e critérios de análise de conteúdo, bem como quais as medidas interventivas adotadas, como são comunicadas ao usuário e quais as condições de revisão das medidas. Nenhuma das políticas analisadas foi considerada um modelo ideal. Todas apresentaram opacidades, algumas na forma de lacunas na informação apresentada e outras na forma de uma transparência opaca, que oferece informações, dando a aparência de prestar contas, mas essas são inadequadas ou insuficientes para subsidiar a efetiva avaliação das medidas adotadas. A partir do cotejo dos resultados encontrados, são apontadas recomendações de transparência para políticas de comunidade, que, se implementadas, poderão agregar legitimidade às práticas de moderação e possibilitar a construção participativa de aprimoramentos.

Palavras-chave

Conteúdo gerado por usuário; Moderação de conteúdo; Políticas de comunidade; Plataformas; Transparência.

Sumário executivo

- As plataformas que veiculam conteúdo gerado por usuário, na sua crescente importância para a troca e publicação de informação, exercem cada vez mais um papel central nas políticas regulatórias que afetam a liberdade de expressão.

- A atividade de moderação foi profissionalizada devido à demanda por um tratamento padronizado e que controle o conteúdo danoso, a fim de que as plataformas se mantenham como espaços atraentes para anúncios publicitários e sustentem, assim, seu modelo de negócios.

- O modelo de negócios das plataformas também demanda compromisso público com a livre expressão dos usuários, a fim de mantê-los engajados com o conteúdo e, consequentemente, favorecer a visualização dos anúncios.

- Como o ideal sustentado publicamente pelas plataformas é de um espaço livre de censura governamental e de busca da inovação, fortemente influenciado por políticas e leis dos Estados Unidos da América sobre livre expressão e tecnologia, também existe interesse delas em manter a autorregulação sobre a moderação de conteúdo como regra.

- Organizações e comunidades internacionais elaboraram documentos de padrões a serem observados no ambiente regulatório sobre conteúdo. São referências os relatórios da ONU sobre Liberdade de Expressão, os Princípios de Manilla e os Princípios de Santa Clara.

- Os padrões internacionais sobre moderação de conteúdo carregam uma preocupação com o potencial discriminatório e de violação à liberdade de expressão dessas práticas, assim como a necessidade de equilibrar os poderes e obrigações atribuídos a intermediários online - para prevenir cenários de restrição por padrão ou de descontrole sobre conteúdo danoso.

- Situações de abuso impulsionadas por agentes privados e públicos alavancaram a necessidade de responder a demandas e pressões crescentes dos segmentos dos usuários (especialmente representados pela sociedade civil organizada), dos anunciantes e dos reguladores. Esses setores demandam regras balanceadas e democráticas de moderação de conteúdo leva as plataformas a adotarem medidas de transparência quanto a suas políticas.

- Um dos principais canais de transparência disponibilizado pelas plataformas sobre medidas interventivas em conteúdo é o das políticas de comunidade - junto aos Termos de Uso - oferecem algumas referências sobre conteúdos restritos ou indesejáveis e práticas de moderação aplicáveis em caso de

violações.

- Entretanto, a transparência proporcionada por esses canais nem sempre se reflete na democratização da moderação de conteúdo. É possível que se reverta em um tipo de opacidade velada, em que o usuário dispõe de informações insuficientes para compreender os critérios das medidas interventivas e avaliá-las criticamente.

- A análise de políticas de comunidade do YouTube, Facebook, Snapchat, Pinterest, Twitter, Instagram, LinkedIn, TikTok - 8 plataformas de conteúdo gerado por usuário representativas perante o público brasileiro - permitiu identificar diversos déficits de transparência nos quesitos de formas de detecção, meios e critérios de avaliação de conteúdo infringente, assim como de exemplificação de conteúdos proibidos, delimitação de exceções, e ainda de que medidas interventivas são aplicáveis, como se dá a notificação quando realizada e possibilidade de recurso sobre medida aplicada.

- As plataformas majoritariamente apontam meios reativos de detecção de conteúdo potencialmente infringente (mediante denúncia), mas não apontam quais meios proativos são adotados. Isso gera preocupação quanto a possível omissão sobre uso de mecanismos automatizados, viés algorítmico e remoção preventiva de conteúdo. Também não há informação completa sobre os mecanismos de denúncia.

- Foi constatada omissão quanto aos meios de avaliar conteúdo potencialmente infringente, etapa prévia à adoção de medida interventiva. Também é problemático que o autor do conteúdo não seja contatado nesse processo de análise, o que seria benéfico para contextualizá-la.

- Quanto aos critérios de avaliação, as políticas não permitem concluir, em sua maioria, o que é considerado no momento de adotar medida de moderação de conteúdo. Ainda, quando presente alguma menção a critérios, eles não são concretos, nem permitem prever ou prevenir uma situação de intervenção.

- Outro déficit foi a carência de exemplos em algumas políticas de comunidade, que citam apenas um ou nenhum exemplo para alguma categoria moderada. Via de regra, não existe um conjunto de decisões público que auxilie na compreensão de quais casos são aceitáveis ou não, embora esse fosse o cenário ideal. Dessa forma, ainda que não supram essa necessidade, exemplos ajudariam a estabelecer esses contornos. No caso das exceções, elas também são indefinidas ou genéricas em alguns textos, o que gera opacidade sobre a extensão das proibições e incompreensão das normas.

- A especificação de quais os conteúdos infringentes também contém lacunas.

Estas ocorrem pelo uso de termos inespecíficos ou de sentido amplo, como imperativo ou verbos que indicam possibilidade, mas não a certeza, de proibição ou de aplicação de medida interventiva naqueles casos. Essa mesma incerteza permeia a previsão de medidas interventivas, que são descritas com termos ambíguos sobre quando serão aplicadas.

- Quanto aos mecanismos de contestação e notificação sobre medida interventiva para o autor do conteúdo, há políticas que não informam se realizam notificação ao autor, ou como se pode recorrer de uma medida interventiva quando se acredita que o conteúdo não é infringente.

- A fim de que haja maior transparência sobre a moderação de conteúdo, de modo a possibilitar o exercício de direitos fundamentais como liberdade de expressão e acesso à informação, que gera a possibilidade de conhecer e demandar por respostas institucionais, são apresentadas recomendações abaixo.

RECOMENDAÇÕES

Às plataformas:

1. Comunicar de forma expressa, visível e pública quais os meios e critérios empregados na detecção e avaliação de conteúdo passível de moderação, especificando sempre que sistemas automatizados forem utilizados.

2. Indicar conteúdos que são proibidos por meio de termos que denotam proibição expressamente, como “não é permitido” ou “é proibido”.

3. Eliminar termos e construções frasais marcadamente ambíguas da redação de seus padrões de comunidade, em especial o uso do modo imperativo e do verbo “poder”.

4. Indicar quais os critérios ou fatores que influenciam na determinação da medida tomada sempre que múltiplas medidas forem cabíveis.

5. Apresentar múltiplos exemplos e contraexemplos de conteúdo infrator e/ou sujeito à intervenção após cada norma.

6. Delimitar os critérios ou categorias de conteúdos exceptuados de forma específica, eliminando o uso de categorias finalísticas excessivamente amplas (“conteúdo de relevância pública”, por

exemplo).

7. Garantir o acesso público permanente a todas as versões anteriores sempre que as políticas de comunidade forem atualizadas. Sinalizar a data em que a política foi atualizada.

8. Incluir dados quantitativos sobre todos os conteúdos que sofreram algum tipo de intervenção em seus relatórios de transparência, não apenas publicações removidas e contas suspensas.

9. Informar ao usuário nas políticas de comunidade sobre seu direito à notificação e à contestação na ocasião de alguma intervenção de moderação.

10. Notificar o usuário sempre que este for alvo de alguma intervenção. A notificação deve incluir, no mínimo, as seguintes informações: URL da publicação, trecho do conteúdo que causou a intervenção (ou dados adicionais que possibilitem sua identificação), cláusula violada dos padrões de comunidade, meios de detecção e intervenção sobre o conteúdo. Ainda, deve ser fornecida em formato duradouro e deve permanecer disponível mesmo que a conta do usuário seja suspensa ou indisponibilizada.

11. Instituir um sistema de contestação robusto, que inclua, no mínimo: revisão humana por um ou mais indivíduos que não estiveram envolvidos na decisão inicial, oferta ao usuário da oportunidade de apresentar informações adicionais a serem consideradas na revisão, notificação dos resultados da revisão e uma declaração de motivos suficiente para que o usuário compreenda a decisão.

12. Quando utilizada detecção automatizada de conteúdo a ser moderado, explicitar as formas de identificação proativa e os canais para comunicação de terceiros sobre falhas, vieses e discriminações potencializadas ou criadas pelo algoritmo.

13. Desenvolver mecanismos que assegurem ampla participação social na elaboração das políticas de comunidade da plataforma.

14. Apoiar - mediante auxílio financeiro, fornecimento de dados, disponibilização de especialistas, etc. - pesquisa científica multidisciplinar destinada a compreender as diferentes dimensões da circulação de conteúdo em escala massiva, bem como os efeitos

das decisões relativas ao processo de moderação.

Ao setor governamental:

15. Condicionar a efetuação de qualquer alteração no regime de responsabilização de intermediários, especialmente por conteúdos gerado por terceiros, a amplo debate público prévio, voltado à promoção da transparência e participação dos usuários, da sociedade civil organizada e da comunidade científica - tendo em consideração os riscos de censura colateral e da importância da internet como ferramenta de expressão.

16. Desenvolver soluções regulatórias, fundamentadas em direitos humanos e amplo debate público, que obriguem as plataformas a implementar um robusto regime de transparência e de procedimentos democráticos previamente estipulados na moderação de conteúdo.

17. Apoiar a pesquisa científica multidisciplinar sobre moderação de conteúdo e expressão na internet, apta a fundamentar factualmente as soluções legislativas e decisões judiciais adotadas.

18. Compatibilizar quaisquer medidas de transparência adotadas com as normas de privacidade e proteção de dados pessoais, a fim de garantir a segurança jurídica e a proteção do direito do usuário à autodeterminação informativa.

19. Disponibilizar publicamente dados sobre pedidos de moderação de conteúdo feito às plataformas a partir de instituições ou autoridades públicas, a fim de permitir o exame de compatibilidade com os relatórios atualmente disponibilizados unicamente pelas plataformas.

À sociedade civil organizada e à academia:

20. Acompanhar as propostas legislativas e a jurisprudência relativas à transparência, moderação de conteúdo e responsabilidade de intermediários.

21. Conduzir pesquisa científica capaz de embasar soluções protetivas dos direitos fundamentais dos usuários para os desafios

relacionados à governança de plataformas.

22. Demandar canais de participação, junto ao setor governamental e às plataformas, na elaboração e avaliação de políticas de comunidade.

23. Estabelecer parâmetros para avaliação de práticas de moderação de conteúdo e seu nível de accountability.

24. Participar dos fóruns e espaços de debate multissetorial, contribuindo com o debate público sobre moderação de conteúdo e transparência.

25. Denunciar violações de direitos humanos realizadas por atores públicos ou privados no âmbito das práticas de moderação de conteúdo - sejam elas motivadas por políticas de comunidade ou por ordens de autoridades governamentais.

1. Introdução

A liberdade de publicação, circulação e acesso a conteúdo é um valor intrínseco à internet e à cultura informacional que a rede promove. Sob a promessa da informação global ao alcance do público, as ferramentas de comunicação online são, em teoria, agentes que otimizam a formação de agendas de discussão, a socialização de notícias, a construção de conhecimento e a conexão entre pessoas. Na prática, elas vão além disso e lidam diariamente com um aspecto indesejado do triplo papel de consumidor, agente de circulação e produtor de conteúdo oportunizado a cada usuário: o conteúdo danoso ou indesejável, que origina conflitos e demandas de usuários às plataformas. As soluções são as mais diversas e compreendem a imposição de limites à liberdade idealizada. Nesse sentido, essas plataformas não equivalem meramente a um condutor indiferente na medida em que arquitetam os algoritmos mediadores da circulação e a interface de exibição do conteúdo, bem como moderam conteúdos específicos.

Com a intensificação dessa atividade, cresce a estrutura que serve de suporte às medidas interventivas. Consequentemente, essas empresas passam a exercer papéis responsivos às novas situações e pressões por uma postura mais ativa. Assim, elaboram e aplicam regras sobre condutas e situações indesejadas e impondo crescentes restrições à expressão online. Também desenvolvem canais de comunicação por meio dos quais usuários podem reportar violações, recorrer de decisões ou buscar apoio. Essa atuação deixa evidente a necessidade de accountability[1] sobre as medidas adotadas, a fim de garantir uma delimitação democrática do que é conteúdo danoso e uma resposta proporcional.Também é necessário o conhecimento acerca das limitações impostas à expressão, ou seja, do tipo de conteúdo restrito, suspenso ou excluído de circulação. Usuários e informação devem ter tratamentos equitativos e baseados em princípios e normas consolidados na doutrina e jurisprudência de direitos humanos para a proteção da liberdade de expressão. Para isso, considerando o tamanho poder que as plataformas exercem sobre os fluxos de conteúdo online, as regras e medidas que elas adotam devem ser públicas e sujeitas a escrutínio.

Esta pesquisa se fundamentou na importância da transparência para uma accountability democrática sobre plataformas de conteúdo gerado por usuário. Busca-se aferir se as informações disponibilizadas por esses serviços viabilizam o questionamento, pressão e demanda do usuário por explicações. O estudo foi realizado a partir de uma contextualização histórica e social do papel desses serviços. Em seguida, foi realizada análise de conteúdo sobre as diretrizes publicadas por elas sobre as atividades de moderação e condutas esperadas dos usuários.

1 O significado deste termo, embora de difícil tradução, será, para os fins desta pesquisa, a capacidade de demandar por respostas institucionais, que pressupõe a disponibilidade de informações suficientes e meios de comunicação e responsividade institucional.

Assim, o trabalho está dividido em mais cinco seções além desta primeira, de introdução. A segunda consiste na delimitação contextual dos moderadores de conteúdo aqui estudados - as plataformas de conteúdo gerado por usuário - e o reconhecimento dessa atividade como sujeita a controle social e, portanto, a práticas de transparência, destacadas as políticas de comunidade como principal documento oferecido ao usuário. A terceira é a descrição da metodologia de análise de conteúdo desses documentos, com a justificativa de escolha da amostra, critérios e meios de análise adotados e descrição dos procedimentos. Em seguida, na quarta são apresentados os resultados, na forma de relatorias e quadro-síntese. Na quinta, é realizada a discussão sobre como a expectativa de transparência se traduz nas políticas de comunidade e quais as lacunas encontradas. Por fim, na sexta são apresentadas as conclusões sobre as práticas de transparência da moderação de conteúdo pelas plataformas e recomendações para o aprimoramento dos instrumentos analisados.

2. Como a transparência foi incorporada na moderação de conteúdo

Usuários de internet têm, à sua disposição, ferramentas para encontrar e para publicar conteúdo de maneira direta e instantânea. Na posição de intermediárias, as empresas de serviços de conteúdo gerado por usuários exercem papel ativo sobre o que pode ser publicado, compartilhado ou restringido na rede. Com o incremento da importância dessa atividade, a transparência sobre ela se torna uma demanda latente. Emergem questionamentos públicos sobre o que as empresas responsáveis pelos meios de circulação de informação fazem a respeito de conteúdo danoso. Em 2019, por exemplo, o jornal britânico *The Guardian* noticiou os danos sofridos comunidades LGBT e de mulheres não-brancas devido à redução do alcance de suas publicações sem qualquer notificação pelo Instagram[2]. Em 2020, a plataforma TikTok foi denunciada por instruir seus moderadores a reduzir o alcance dos conteúdos com usuários considerados "pouco atraentes" e de casas contendo paredes cuja pintura estivesse danificada[3].

O debate público foi, portanto, crescentemente pautado pelo gradual reconhecimento de que compete às plataformas intervir para coibir a circulação de conteúdo danoso, por um lado. Por outro, reconhece-se cada vez mais que tais

2 JOSEPH, Chanté. Instagram's murky 'shadow bans' just serve to censor marginalised communities. **The Guardian**, Londres, 8 nov. 2019. Disponível em: https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive. Acesso em: 07 out. 2020.

3 BIDDLE, Sam; RIBEIRO, Paulo Victor; DIAS, Tatiana. TikTok escondeu "feios" e favelas para atrair novos usuários e censurou posts políticos. **The Intercept Brasil**, 16 mar. 2020. Disponível em: https://theintercept.com/2020/03/16/tiktok-censurou-rostos-feios-e-favelas-para-atrair-novos-usuarios/ Acesso em: 16 mar. 2020.

intervenções podem representar, elas próprias, riscos a direitos e liberdades fundamentais, sobretudo quando realizadas de forma opaca. Nesse cenário, considera-se um valor democrático e um direito do usuário conhecer os limites e valores do ambiente informativo de que participa, e portanto, a forma como são fixadas regras sobre o tráfego de conteúdo. Por isso, esta seção apresenta o contexto de valorização pública da transparência das plataformas e os canais pelos quais elas buscam atender a essa demanda.

2.1. O papel das plataformas na moderação de conteúdo

Moderação de conteúdo é uma prática que se intensificou e foi alterada em decorrência das transformações no tipo e escala de interação e compartilhamento de informação pelas quais a internet tem passado. Parte importante do fluxo de dados que trafega e é moderado na rede pode ser atribuído às comunidades online. Conforme o público aumentou, o conteúdo e as interações entre usuários cresceram, assim como os tipos de comunidade possíveis, com adaptações daquelas já existentes. As comunidades mais antigas eram majoritariamente fóruns autorregulados, ou então regulados por usuários com poderes especiais (os administradores). Esses espaços contavam com um perfil mais ou menos homogêneo de usuários e com conteúdo majoritariamente não-comercial, caseiro e independente. Entre 1996 e 1998, entretanto, grandes jornais impressos começaram a disponibilizar conteúdo online, o que deu à internet um caráter informativo, e aos provedores de plataformas, em adição a seu papel de provedor de software, cada vez mais papéis semelhantes aos de editores[4].

Hoje, a predominância é das plataformas; nelas, convivem os mais variados grupos de pessoas, com objetivos e perfis diversos. Em algumas, circula grande volume de informação comercial, no formato de anúncios e publicidade por usuários influentes. É o sucesso na promoção desse tipo de conteúdo que interfere na forma como as postagens não-comerciais, a ele associadas, são disponibilizadas aos usuários. Assim, uma das principais atividades de uma plataforma é decidir que conteúdo será priorizado, a quem será mostrado, de quem será ocultado ou quando será removido. O conteúdo mostrado a cada indivíduo é personalizado por meio de algoritmos, que analisam o que mantém a atenção de determinado usuário[5]. Ainda, esse cenário é agravado devido à elevada concentração do setor num grupo

4 KLONICK, Kate. The new governors: The people, rules, and processes governing online speech. **Harv. L. Rev.**, v. 131, p. 1598, 2017. p. 1.618

5 BRITO CRUZ, Francisco (coord.); MASSARO, Heloisa; OLIVA, Thiago; BORGES, Ester. **Internet e eleições no Brasil**: diagnósticos e recomendações. InternetLab, São Paulo, 2019. Disponível em: Acesso em: https://www.internetlab.org.br/wp-content/uploads/2019/09/policy-infopol-26919_4.pdf 18 ago. 2020.

de pouquíssimas empresas, o que as reveste de imenso poder econômico do qual podem valer para influenciar o debate público sobre o tema de acordo com seus interesses comerciais. [6]

Diversos serviços online são chamados de "plataformas", mas existem distintas funcionalidades incluídas nesse termo, dependendo da intenção do interlocutor: serviços que abrigam comunidades no formato mais tradicional de fórum, como Reddit, os que ofertam conteúdo pago, como Netflix, os que facilitam o encontro entre quem busca e quem oferta algo, como Uber, os de pagamento, como PicPay, e os de construção colaborativa, como Wikipedia. Plataforma não é um termo nitidamente definido, mesmo na literatura especializada. Em termos técnicos, aponta Gillespie[7] que a palavra é usada como sinônimo de infraestrutura computacional, significado muito diferente daquele usualmente adotado pela mídia. O significado adotado por esta pesquisa é sintetizado por Valente[8], segundo o qual:

> *As plataformas digitais são sistemas tecnológicos que funcionam como mediadores ativos de interações, comunicações e transações entre indivíduos e organizações operando sobre uma base tecnológica digital conectada, especialmente no âmbito da Internet, provendo serviços calcados nessas conexões, fortemente lastreados na coleta e processamento de dados e marcados por efeitos de rede.*

Nesse conceito, muitos tipos de serviços podem ser identificados, sempre relativos ao tratamento de informação[9]. Entre eles, a importância da moderação de conteúdo é maior em plataformas de conteúdo gerado por usuário. Nelas, quem se encarrega de gerir e elaborar políticas precisa intervir diretamente na forma como usuários interagem com outros e com o conteúdo compartilhado. Isto é, a informação, o tipo de postagem, comentário, compartilhamento que pode ser feito - e acessado - pelos usuários é definido por regras delineadas pela empresa, com base no perfil de usuário e interação que ela deseja associar a seus anunciantes. Esse tipo de plataforma é definido por Klonick[10] como aquelas que "hospedam, publicam e moderam conteúdo gerado por usuários", exemplificadas pela autora por Twitter,

6 VALENTE, J. C. L. Plataformas digitais e concentração na internet. In: Encontro Anual da Rede de Pesquisa em Governança da Internet, III, 2019, Manaus., **Anais...** [S.I.]: Rede de Pesquisa em Governança da Internet, 2020. p. 1-25.

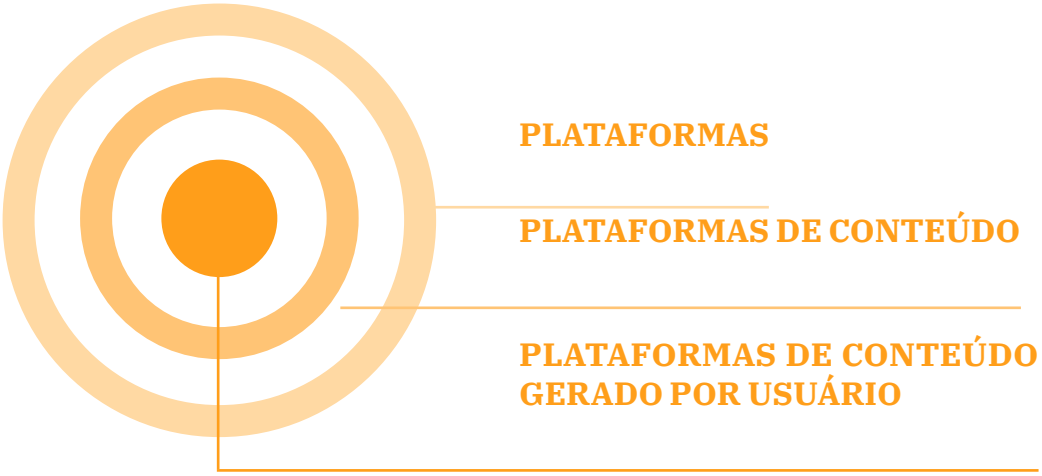
7 GILLESPIE, Tarleton. **Custodians of the Internet:** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018. p. 36

8 VALENTE, J. C. L. **Tecnologia, informação e poder**: das plataformas online aos monopólios digitais. Tese (Doutorado em Sociologia). Universidade de Brasília, Brasília, 2019. p. 170

9 CARMO, Paloma; DUARTE, Felipe; GOMES, Ana Bárbara. **Glossário da Inclusão Digital** - Volume II. Instituto de Referência em Internet e Sociedade: Belo Horizonte, 2020. Disponível em: http://bit.ly/3aqUlfP. Acesso em: 07 ago. 2020. p. 30
AAAA

10 KLONICK, Kate. The new governors: The people, rules, and processes governing online speech.

YouTube e Facebook. Gillespie[11] adota essa concepção e acrescenta mais duas características: 1. elas não produzem nem patrocinam a produção da maior parte daquele conteúdo; e 2. possuem uma infraestrutura de processamento dos dados para serviço aos clientes, publicidade e lucro.



Nesse sentido, plataformas de conteúdo gerado por uuário são uma gama ampla de comunidades e serviços, que incluem tanto redes sociais quanto ferramentas de compartilhamento/disponibilização de vídeos, comentários etc. Diferenciam-se, ainda, de outros tipos de plataforma online, como as de *streaming*, de compartilhamento de bens e serviços, ou de notícias, por exemplo, pelo caráter de autonomia de cada usuário. Este, além de consumir, pode produzir conteúdo para acesso dos demais, sem haver uma curadoria prévia ou contrato comercial entre os produtores de conteúdo e a plataforma que o disponibiliza. Além disso, pode interferir na dinâmica de priorização de conteúdo, compartilhando-o e interagindo com ele de forma a torná-lo mais relevante para a plataforma. Os contratos comerciais, quando ocorrem, são entre a plataforma e os anunciantes que negociam a divulgação e o direcionamento de sua publicidade para os usuários, gerando renda para a plataforma.

O modelo de negócios dessas empresas não rompe completamente com aquele presente nas mídias tradicionais, como rádio e televisão, pois ainda se baseia essencialmente na venda de espaço publicitário. No entanto, as novas mediações introduzidas pela internet e pelo ambiente plataformizado atualizam esse modelo de formas significativas. O tratamento dos dados pessoais de seus usuários em escala massiva possibilita o direcionamento de anúncios específicos para públicos específicos a partir da noção de "relevância". Como observa Gillespie[12], essa noção não tem um significado óbvio e os sentidos que ela adquire são o produto de disputas políticas e culturais pela definição do que é relevante, similarmente ao que ocorre com termos como "popular" ou "digno de notícia" (*newsworthy),* cuja

Harv. L. Rev., v. 131, p. 1598, 2017. p. 1617

11 GILLESPIE, Tarleton. **Custodians of the Internet:** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018. p. 15

12 GILLESPIE, Tarleton. A relevância dos algoritmos. **Revista Parágrafo**, v.6, n.1, p.95-121, 2018.

fluidez vem sendo discutida há décadas por midiálogos.

No contexto dos algoritmos de direcionamento de conteúdo, a relevância em questão é a relevância presumida de determinado conteúdo para um usuário, sendo esta determinada a partir dos dados gerados por seus comportamentos prévios. Assim, diferentemente do que ocorre na televisão e no rádio, o conteúdo publicitário exibido é direcionado de maneira altamente personalizada, o que agrega valor substancial para os anunciantes. Do ponto de vista das plataformas, é vantajoso alegar que esse tratamento consiste meramente em conectar, de forma mais rápida e eficiente, consumidores a fornecedores de bens e serviços. No entanto, essa narrativa tornou-se objeto de diversos questionamentos na última década. Críticos[13][14][15] ponderam que ao expor continuamente o usuário a conteúdos que reafirmam e alimentam seus gostos, preferências, ideias e interesses prévios, a arquitetura das plataformas modula, ela própria, a subjetividade que alega estar meramente descrevendo e prevendo - mesmo porque essa subjetividade não é construída num vácuo, mas em interação com o ambiente plataformizado.

No empreendimento baseado em direcionamento de anúncios, os clientes da plataforma são os anunciantes e parte substancial do ferramental do serviço ofertado são os dados e a atenção dos usuários. Por isso, as plataformas assumem um papel mais proativo em relação ao conteúdo que elas suportam. O interesse comercial as leva a não relegar apenas à moderação voluntária de usuários o controle sobre conteúdo ofensivo. Como modelo de negócio, parte de sua garantia de qualidade passa pelo controle do conteúdo associado a seu ambiente, em que também circula publicidade contratada por anunciantes.

Segundo Seering, Yoon e Kaufman[16], nesses casos uma equipe contratada pela plataforma ou terceirizada é encarregada profissionalmente da tarefa, serviço que se convencionou chamar de "moderação comercial de conteúdo". Isso diferencia essas plataformas daquelas cuja moderação é autogerida e coordenada por usuários, como nos fóruns do Reddit[17], Wikipedia, Twitch, etc. Com a mudança de escala das interações e o perfil mais comercial e empresarial dos serviços de

13 BRUNO, F. Monitoramento, classificação e controle nos dispositivos de vigilância digital. **Revista FAMECOS**, v. 36, p. 1-7, 2008.

14 ZUBOFF, Shoshana. **The Age of Surveillance Capitalism**: The Fight for a Human Future at the New Frontier of Power. New York: Public Affairs, 2019.

15 ROUVROY, Antoinette; BERNS, Thomas. Governamentalidade algorítmica e perspectivas de emancipação: o díspar como condição de individuação pela relação?. **Revista Eco Pós**, vol. 18, n. 2, 2015, p. 35-56.

16 SEERING, Joseph et al. Moderator engagement and community development in the age of algorithms. **New Media & Society**, v. 21, n. 7, p. 1417-1443, 2019. p. 1419

17 Plataforma de fóruns, divididos em categorias, que dentro dos fóruns adota moderação voluntária e regras próprias gerenciadas pelos usuários de cada comunidade, mas na comunidade em geral conta com moderação comercial para aplicação das políticas de comunidade. Ver: REDDIT. **Content Policy**. Disponível em: https://www.redditinc.com/policies/content-policy Acesso em: 31 jul. 2020.

plataformas, novas formas de moderação de conteúdo passaram a ser exigidas. Dessa forma, aquelas que veiculam anúncios publicitários não contam apenas com o formato proativo, pelos usuários, de moderação de conteúdo[18].

2.2. Conteúdo gerado por usuário e responsabilidade de intermediários

O caminho entre a autogestão do conteúdo e a moderação comercial, entretanto, não é trilhado de maneira lógica e pública. É possível observar a formação de uma contradição entre discursos e práticas no desenrolar dessas políticas. Como narram diversos autores que tratam de plataformas[19] [20], muitas delas foram concebidas sob a visão de que a internet seria um local de livre expressão, abertura a todas as manifestações de pensamento, uma verdadeira "democracia" no sentido utópico de fórum aberto de ideias. Contudo, tão logo elas começaram a atuar, esse conceito mostrou-se uma fantasia. As empresas que gerenciam essas plataformas são grandes agentes econômicos, que assumem o papel de definir como o conteúdo é priorizado e compartilhado com mais ou menos facilidade, de impor regras e de lidar administrativamente com conflitos e situações potencialmente danosas aos usuários. A moderação de conteúdo tornou-se uma grande parte do que as plataformas fazem, independentemente de seu interesse em assumir publicamente este papel[21].

O perfil de aparente "isenção" de quem oferece esses espaços interativos pode ser associado à localização da sede, nos Estados Unidos da América, das maiores empresas[22] provedoras das plataformas que predominam na internet. A questão cultural de valorização da liberdade de expressão e de não-responsabilização de indivíduos por conteúdo de manifestações de pensamento ou ideias está presente no direito estadunidense[23] [24]. Inclusive, abarca frequentemente temas polêmicos, como discurso de ódio. Assim, o direito que geralmente rege as sedes das empresas

18 GILLESPIE, Tarleton. Governance of and by platforms. **SAGE handbook of social media**, p. 254-278, 2017. p. 16

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

20 CITRON, Danielle Keats. Extremist speech, compelled conformity, and censorship creep. **Notre Dame L. Rev.**, v. 93, p. 1035, 2017.

21 GILLESPIE, Tarleton. **Custodians of the Internet:** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018. p. 15

22 SOWELL, Jesse H. **Evaluating competition in the Internet's infrastructure**: a view of GAFAM from the Internet exchanges. Journal of Cyber Policy, v. 5, n. 1, p. 107-139, 2020.

23 ROSENFELD, Michel. Hate speech in constitutional jurisprudence: a comparative analysis. **Cardozo L. Rev.**, v. 24, p. 1523, 2002.

24 RIDER, Karina. The privacy paradox: how market privacy facilitates government surveillance. **Information, Communication & Society**, v. 21, n. 10, p. 1369-1385, 2018.

responsáveis por moderar conteúdo preza por uma liberdade de expressão ampla e que encontra limites somente em casos como ameaças de violência física, crimes e alguns aspectos da privacidade[25].

Aliado a isso, existe uma cultura de liberdade negocial e incentivo à inovação tecnológica que considera não ser ônus do empreendimento arcar com danos por mau uso de seus serviços e nega intervenção estatal para responsabilização de intermediários[26]- termo amplamente usado pela mídia americana para descrever serviços online[27]. Isso pode ser evidenciado na narrativa de Klonick[28] acerca da emergência das primeiras políticas de comunidade de plataformas. Segundo a autora, a liberdade conferida a intermediários da internet para lidar com expressão online tem seu fundamento na seção 230 da Lei de Decência das Comunicações (Communications Decency Act - CDA)[29] dos Estados Unidos da América. A lei estabelece imunidade dos provedores de "serviços interativos por computador" em relação a conteúdo gerado por usuário. Nota-se que o CDA foi discutido à época como uma resposta a preocupações dos reguladores com material pornográfico e/ou reproduzido sem autorização, que ocupavam um espaço simbólico no debate público sobre internet similar àquele atualmente preenchido por questões como desinformação e crimes graves[30].

Essa interpretação dos intermediários como isentos de responsabilidade por conteúdo gerado por terceiros é concebida a partir do caso Zeran vs. America Online (AOL), em que a AOL foi processada por um usuário em razão de conteúdo publicado por outra pessoa na plataforma. No caso, a corte reconheceu a cláusula do "Bom Samaritano", que corresponde ao item C da seção 230 do CDA, e é composta de dois trechos: o primeiro, também conhecido como cláusula do "porto seguro"[31], equiparou as plataformas à descrição da seção 230 para serviços de telecomunicações que não criam, mas circulam conteúdo e, portanto, não podem

25 Conceito cujas referências paradigmáticas são juristas americanos. Ver: WARREN, Samuel; BRANDEIS, Louis. The right to privacy. **civilistica. com: revista eletrônica de direito civil**, v. 2, n. 3, p. 1-22, 2013.

26 BARLOW, John Perry **Declaração de Independência do Ciberespaço**. (trad. DH net). Disponível em: http://www.dhnet.org.br/ciber/textos/barlow.htm. Acesso em: 29 jun 2020.

27 GILLESPIE, Tarleton. Governance of and by platforms. **SAGE handbook of social media**, p. 254-278, 2017. p. 2

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

29 O CDA está contido no Título V Lei de Telecomunicações de 1996, que, na Seção 509, sob o assunto "Empoderamento familiar online", inseriu a Seção 230 na Lei de Telecomunicações já existente, de 1934. Ver: EUA - Estados Unidos da América. **Communications Decency Act**. Sec. 509. Online Family Empowerment. Disponível em: https://www.govinfo.gov/content/pkg/PLAW-104publ104/pdf/PLAW-104publ104.pdf Acesso em:29 jun. 2020.

30 GILLESPIE, Tarleton. **Custodians of the Internet:** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018. p. 27.

31 GILLESPIE, Tarleton. Governance of and by platforms. **SAGE handbook of social media**, p. 254-278, 2017. p. 260

ser responsabilizadas por ele. Isso conferiu às plataformas alguma liberdade e poder, pois não são obrigadas a monitorar conteúdo e também não são legalmente responsabilizadas caso o façam, perante a lei do país-sede da maioria delas. Existia nelas a preocupação de manter-se como meios livres de censura ou controle prévio de circulação de informação e, ao mesmo tempo, não responder judicialmente por qualquer limitação imposta aos usuários.

A segunda parte da seção 230 dá imunidade ao provedor de serviço que de boa-fé limita ou restringe o acesso a determinado conteúdo. Estaria implícito o propósito da regulação em encorajar que os provedores promovam autorregulação da disseminação de material ofensivo em seus serviços. A não-responsabilização justificava-se contra uma possível "censura colateral" das plataformas, para evitar condenação em juízo.

A decisão judicial do caso Zeran foi importante para o modelo de negócios desses empreendimentos - e também para os direitos dos usuários, pois a responsabilização poderia implicar em um ecossistema digital marcado pelo fenômeno da censura colateral, que ocorre quando conteúdos são removidos preventivamente pela plataforma, a fim de evitar custos associados à responsabilização por eventuais danos causados por conteúdos não removidos. Isso também um incentivaria o monitoramento ativo e retirada de conteúdos infringentes sem a mediação do poder judiciário, assim como geraria risco de determinação judicial determinação judicial de bloqueio de plataformas inteiras e, portanto, conteúdos legítimos. A partir desse apanhado histórico é que Klonick elenca os dois motivos que, apesar dessa posição, parecem impulsionar as plataformas à formulação de regras e iniciativas de moderação de conteúdo: a responsabilidade social e a viabilidade econômica[32]. O primeiro motivo é vinculado à segurança dos usuários e seu encorajamento a usar e conectar-se à plataforma. O segundo tem ligação com a veiculação de anúncios nas plataformas ao lado do conteúdo gerado por usuários e a ameaça que o conteúdo indesejado representa para essa renda.

Moderar o conteúdo para que ele seja agradável leva os usuários a interagirem mais e torna a plataforma mais atrativa para anunciantes. No entanto, a remoção excessiva de conteúdos poderia levar à perda de confiança dos usuários e diminuição dessas interações e do apelo comercial, como aponta Gillespie[33]. Para ele, ainda há outro fator nessa ambivalência das plataformas em posicionarem-se como moderadoras de conteúdo.Além da "neutralidade" dos serviços na internet como apenas meios por onde a informação trafega ser uma crença de seus fundadores, é também uma maneira das plataformas evitarem responsabilização judicial. Gillespie ainda ressalta que **a expansão de plataformas de mídias sociais para fora dos EUA**

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

33 GILLESPIE, Tarleton. **Custodians of the Internet:** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018. p. 32

evidenciou um mundo onde a Primeira Emenda é apenas uma regra local. Além disso, potencializou desafios já enfrentados por equipes de funcionários majoritariamente brancas, masculinas, educadas formalmente, tecnológicas, liberais ou libertárias ao lidar com grupos minoritários e diversidade[34].

No Brasil, a legislação que rege esse tópico, o Marco Civil da Internet[35], em seu artigo 19, seguiu a tendência de não responsabilizar os intermediários. Segundo o dispositivo, apenas mediante ordem judicial é possível obrigar uma plataforma a tomar medidas em relação a conteúdo online e aplicar medidas coercitivas caso ela não o faça. Entretanto, situações cada vez mais recorrentes expõem os reflexos da circulação, na internet, de discursos de ódio e extremistas, além de campanhas de desinformação com fins políticos, o que fortalece o discurso de insuficiência das medidas de autorregulação. Emerge desse contexto uma mobilização por regular de maneira mais incisiva o ambiente online. A discussão sobre o regime de responsabilidades dos intermediários de internet e os reflexos da moderação de conteúdo sobre a livre expressão e a democratização do acesso à informação é reaberta, uma vez que as plataformas moldam, de maneiras diferentes mas significativas, as formas como as pessoas se comunicam[36].

Outro fator de debate é que muitos países a partir dos quais usuários acessam essas plataformas têm regulações conflitantes e expectativas distintas daquela dos Estados Unidos da América, onde a maioria está constituída. Estabelecem, nesse sentido responsabilidade mais estrita tanto para proteção dos direitos dos usuários quanto por uma visão de que a autoridade estatal deve preponderar nesses meios[37][38]. Nesse contexto, a preocupação em manter a internet como ambiente democrático e em garantir direitos dos usuários leva os critérios e abordagens para moderação de conteúdo a escrutínio internacional. Por isso, são cada vez mais discutidos por diversos grupos, autoridades, acadêmicos e membros da comunidade técnica.

34 GILLESPIE, Tarleton. **Custodians of the Internet:** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018. p. 18-25

35 BRASIL. Lei Nº 12.965, de 23 de Abril de 2014. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Brasília, DF, Disponível em: <http://bit.ly/32WiEPA>

36 SUZOR, Nicolas; VAN GEELEN, Tess; MYERS WEST, Sarah. **Evaluating the legitimacy of platform governance:** A review of research and a shared research agenda. International Communication Gazette, v. 80, n. 4, p. 385-400, 2018. p. 836.

37 GILLESPIE, Tarleton. Governance of and by platforms. **SAGE handbook of social media**, p. 254-278, 2017. p. 8

38 ONU - Organização das Nações Unidas. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression.** 22 abr. 2020. Disponível em: https://undocs.org/A/HRC/44/49 Acesso em: 30 jun. 2020. p. 7

2.3. Preocupações institucionais com moderação de conteúdo

Diversas situações criam preocupações com a padronização e criação de princípios para moderação de conteúdo. Os direitos humanos sofrem riscos com propostas regulatórias que responsabilizam plataformas por conteúdo gerado por terceiros, gerando risco de "censura colateral". Práticas de moderação que desconsideram a diversidade de contextos ferem a liberdade de expressão, em especial de grupos minoritários, e o acesso à informação. São relatados acordos entre plataformas e governos com perfil autoritário para remoção de conteúdo, majoritariamente aquele de teor crítico a autoridades políticas[39]. Esse tipo de intervenção favorece um ambiente antidemocrático, que fere a livre manifestação do pensamento.

Frente ao conjunto de problemáticas apontadas, grupos de especialistas em direitos humanos e organizações internacionais têm movido esforços para promover referências sobre moderação de conteúdo. A seguir, são apresentados instrumentos que formalizam a demanda por medidas justas e por transparência das plataformas nesse âmbito e os canais que são estruturados e aprimorados pelas plataformas como maneira de resposta a essas pressões

2.3.1. Práticas, princípios e padrões de transparência

No que toca à transparência, as interações entre plataformas e governos têm recebido especial atenção, com enfoque nos relatórios de pedidos governamentais (administrativos ou judiciais) de remoção de conteúdo[40]. Nesse contexto, duas grandes lacunas, ao menos, podem ser observadas na transparência dos relatórios

39 ONU - Organização das Nações Unidas. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression**. Symbol A/HRC/38/35. 6 abr. 2018. Disponível em: https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35. Acesso em: 30 jun. 2020. p. 7

40 O acesso à informação é considerado um dos pilares dos direitos civis, internacionalmente protegido pela Convenção de Direitos Civis e Políticos. Existe expresso reconhecimento, em documentos da ONU, da necessidade de sua preservação contra autoritarismo estatal, mesmo em contextos de emergência, ou principalmente nesses contextos, por ser essencial à adoção e suporte de políticas públicas adequadas e informadas. Assim, qualquer medida estatal de controle de informação, especialmente por bloqueio ou remoção de conteúdo, deve ser amplamente fundamentada e limitada a combater uma ameaça específica e determinada. Isso coíbe práticas generalistas de moderação de conteúdo via regulações estatais amplas e genéricas, bem como impõe uma transparência por proporcionalidade, necessidade e legalidade de qualquer intervenção governamental no acesso a informações. Ver: ONU - Organização das Nações Unidas. **Freedom of Expression and Opinion**. Annual Reports. Disponível em: https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/Annual.aspx#:~:text=Annually%20the%20Special%20Rapporteur%20is,and%20expression%20in%20all%20its Acesso em: 30 jun. 2020. Em especial, ver: Idem, **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression**. Disease pandemics and the freedom of opinion and expression. 23 abr. 2020. Disponível em: https://undocs.org/A/HRC/44/49 Acesso em: 30 jun. 2020. p. 6

na forma como se apresentam. A primeira é que, em geral, eles são fornecidos pelas próprias plataformas, que relatam unilateralmente quantos pedidos foram realizados pelas autoridades. Não há um relatório das autoridades indicando a quais plataformas os agentes estatais dirigiram ordens de remoção ou restrição de conteúdo. Isso dificulta a verificação dos dados que as plataformas divulgam de maneira fragmentada sobre pedidos de autoridades de cada país. A segunda lacuna é que os dados apenas quantificam de maneira genérica as solicitações de remoção e as ordens cumpridas em cada país[41].

Essa forma de transparência sofre críticas, por ser considerada unilateral (uma vez que o dado é fornecido pela própria plataforma e não há como verificá-lo) e por não permitir um diagnóstico adequado das providências que as plataformas adotam para garantir que medidas interventivas no conteúdo respeitem princípios democráticos. Outro limite ao formato atual dos relatórios de transparência é que não se dedicam, de maneira geral, a retratar a moderação proativa pela plataforma. Esse seria um aspecto essencial ao considerar que elas também impactam o debate público e direitos dos indivíduos[42] [43] [44].

Para além disso, relatórios não são as únicas medidas consideradas no cenário internacional como relevantes à transparência na moderação de conteúdo. Outras medidas e instrumentos são, por exemplo: 1. a comunicação ou notificação do usuário sobre medidas adotadas em relação a conteúdo; 2. a possibilidade de revisão dessas medidas; 3. informação sobre como é feita a moderação de conteúdo e como é o trabalho dos moderadores; além de 4. publicidade das regras que delineiam medidas interventivas.

No âmbito da sociedade civil organizada, dois documentos sobre práticas a serem observadas pelas plataformas se destacam: 1. os Princípios de Manilla sobre responsabilidade civil de intermediários[45] e 2. os Princípios de Santa Clara sobre transparência e *accountability* na moderação de conteúdo[46]. O último dos seis

41 LOSEY, James. Surveillance of communications: A legitimization crisis and the need for transparency. **International Journal of Communication**, v. 9, p. 3450-3459, 2015. p. 3452-3453

42 ONU - Organização das Nações Unidas. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression**. Symbol A/HRC/38/35. 6 abr. 2018. Disponível em: https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35 Acesso em: 30 jun. 2020. p. 14

43 Uma exceção reportada é o relatório de implementação de políticas de comunidade do Facebook, divulgado pela primeira vez em 2018. Ver: RDR - Ranking Digital Rights. **2019 RDR Accountability Index**. Maio 2019. Disponível em: https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf Acesso em: 09 jul. 2020. p. 20.

44 FACEBOOK. **Relatório de aplicação dos padrões de comunidade**. Disponível em: https://transparency.facebook.com/community-standards-enforcement Acesso em: 20 jul. 2020.

45 EFF et al. Princípios de Manilla sobre responsabilidade civil de intermediários. Disponível em: https://www.manilaprinciples.org/pt-br Acesso em: 09 jul. 2020.

46 EFF et al. **Santa Clara Principles on transparency and accountability in content moderation**. Disponível em: https://santaclaraprinciples.org/ Acesso em: 09 jul. 2020.

Princípios de Manilla trata da transparência em práticas de restrição de conteúdo. Recomenda, entre outras medidas, a publicação, pelos intermediários, de políticas de restrição de conteúdos, a divulgação de relatórios que informem sobre restrições realizadas, notificação clara ao usuário sobre motivos para medidas interventivas e indicação pública de que conteúdo foi removido e da razão. Os Princípios de Santa Clara compreendem três seções de recomendações: a primeira, sobre divulgação de quantitativos de remoções; a segunda, sobre notificação ao usuário dos motivos para remoção, suspensão ou restrição de conteúdo; e a terceira sobre meios de revisar decisões de remoção ou restrição de conteúdo.

Já no contexto das organizações internacionais, são pertinentes ao tema os relatórios anuais do Relator Especial para a Liberdade de Expressão e de Opinião[47], da Organização das Nações Unidas - ONU, que demonstram preocupação com a elaboração de padrões internacionais de direitos humanos em relação a moderação de conteúdo online. A elaboração dessas diretrizes tanto possibilitaria que plataformas evitassem abuso dos Estados em pedidos excessivos de remoção de conteúdo quanto estabeleceria limites às empresas. Dessa forma, a definição de obrigações a ambas as partes, potencializaria a fiscalização mútua. Um dos relatórios da ONU constata opacidade sobre a maneira como as plataformas interpretam suas próprias regras. Frente a isso, recomenda maior transparência desde o processo de regramento até a implementação das políticas, a fim de viabilizar a fiscalização e engajamento da sociedade civil nos processos[48][49].

A Organização para a Cooperação e o Desenvolvimento Econômico (OCDE) também já demonstrou preocupação com o tema. Seu Conselho, em 2011, reconheceu a importância em limitar a responsabilidade de intermediários na internet. Essa limitação é recomendada como princípio, ao lado de outros, como transparência, devido processo, *accountability*. Recomenda, ainda, um processo de elaboração de políticas de internet multissetorial e inclusivo[50]. No centro da preocupação do Conselho, encontram-se os direitos humanos, o estado de direito, que permita a atuação democrática e a livre expressão. Em especial quanto à limitação da responsabilidade de intermediários, a OCDE reconheceu o papel que eles desempenham na coibição de atividades ilícitas e que sua responsabilização ilimitada pode atingir negativamente seu potencial criativo, de livre fluxo de

47 ONU https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/Annual.aspx#:~:text=Annually%20the%20Special%20Rapporteur%20is,and%20expression%20in%20all%20its

48 ONU http://daccess-ods.un.org/access.nsf/Get?Open&DS=A/HRC/38/35&Lang=E p. 20

49 A mesma recomendação é feita em declaração conjunta do mesmo Relator Especial com OEA, OAS, OSCA, CADHP. Ver: OEA et al. **Declaração conjunta do vigésimo aniversário: desafios para a liberdade de expressão na próxima década**. 2019. Disponível em: https://www.oas.org/pt/cidh/expressao/showarticle.asp?artID=1146&lID=4 Acesso em: 8 jul. 2020.

50 OCDE - Organização para a Cooperação e o Desenvolvimento Econômico. **OECD Council Recommendation on Principles for Internet Policy Making**. 2011. Disponível em: http://www.oecd.org/internet/ieconomy/49258588.pdf Acesso em: jul. 2020.

informação e inovação. Ainda, observou que as limitações podem estimular a cooperação entre diferentes atores para manutenção dos direitos dentro de seus serviços.

Como documento comunitário de referência, destaca-se o Guia dos Direitos Humanos para os Utilizadores de Internet[51], do Conselho da Europa. A publicação de 2014 ressalta que direitos de liberdade de expressão e de acesso à informação são direitos também aplicáveis na internet, ressaltando sua conexão ao tratar dos temas no mesmo tópico. O Guia ainda menciona a importância de que as restrições a essas liberdades sejam fundamentadas em finalidades legítimas orientadas por direitos humanos previstos na Convenção Europeia dos Direitos do Homem.

A transparência e *accountability* sobre moderação de conteúdo não se concretizam unicamente por meio de quantitativos divulgados posteriormente à adoção de medidas. Também é necessária a transparência contínua durante a elaboração e implementação dos procedimentos e políticas de moderação, enquanto eles são realizados. Nesse sentido, outros dois canais comumente são disponibilizados pelas plataformas: os termos de uso e as políticas ou diretrizes de comunidade.

2.3.2. O papel dos termos de uso e das políticas de comunidade

Os termos de uso têm caráter predominantemente contratual, ou seja, geram parâmetros para eventual discussão judicial sobre matérias envolvendo um usuário específico e a plataforma. Eles definem o direito da plataforma, enquanto prestadora de serviço comercial, de exigir determinado comportamento do usuário[52]. Sua finalidade é contratual, estabelece limites e atribuições de responsabilidade por atos e escolhas do usuário em relação à empresa e ao serviço utilizado e vice-versa. Por isso, adotam uma linguagem predominantemente jurídica[53], pois precisam atender ao rigor exigido perante órgãos judiciais[54]. No entanto, por serem um contrato, ainda que relevante para o universo da moderação de conteúdo, como

51 CE - Conselho da Europa. **Guia dos Direitos Humanos para os Utilizadores de Internet**. 2014. Disponível em: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806a0532. Acesso em: 08 out. 2020.

52 MACKINNON, Rebecca et al. **Fostering freedom online: The role of internet intermediaries**. UNESCO Publishing, 2015. p. 20.

53 VENTURINI, Jamila et al. **Terms of service and human rights**: An analysis of online platform contracts. 2016. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/18231 Acesso em: 21 Jul. 2020. p. 24.

54 Há debate acadêmico quanto à necessidade de esses termos também serem traduzidos para linguajar menos técnico, inclusive para cumprir com exigências de que qualquer contrato seja feito de forma livre e esclarecida pelas pessoas que o compõem. Ver: WAUTERS, Ellen; LIEVENS, Eva; VALCKE, Peggy. Towards a better protection of social media users: A legal perspective on the terms of use of social networking sites. **International Journal of Law and Information Technology**, v. 22, n. 3, p. 254-294, 2014.

destacado por West[55], os termos de uso não desempenham esse papel educativo.

As políticas de comunidade, por sua vez, são análogas a uma espécie de regulação, com reflexos diretos e que independem de uma corte decisória ou de uma autoridade externa para serem implementadas. Ainda assim, há casos em que formam vínculo jurídico contratual, por serem citadas nos Termos de Uso da plataforma na maioria dos casos. É por meio delas que a plataforma pode manifestar observância da maior parte das recomendações sobre transparência que lhe são feitas nos documentos de diretrizes. São, ainda, instrumentos com uma finalidade educativa e informativa, que comunicam o tipo de conteúdo aceitável ou não e de que forma a plataforma lida com isso. Elas servem de controle sobre o que se pode esperar como consequência de determinados comportamentos naquele ambiente. Portanto, a transparência sobre as regras, critérios de avaliação e medidas aplicáveis aos conteúdos tem, como principal canal documental, as políticas de comunidade. Elas complementam o conjunto de canais de transparência que pode compreender, mas não é limitado a, relatórios sobre pedidos e intervenções de moderação, termos de uso, centrais de ajuda, etc[56].

2.4. Políticas de comunidade como canal de transparência sobre conteúdo moderado

As discussões judiciais do caso Zeran, encerradas em 1998, impulsionaram a organização e implementação de políticas de comunidade pelas plataformas. Elas foram importantes para revelar a conexão entre o regime de não-responsabilização das plataformas e a garantia de livre expressão online. Nesse contexto, influenciaram na contratação de advogados especialistas na Primeira Emenda Constitucional dos EUA, referente à livre expressão, como encarregados de políticas de comunidade das principais aplicações online, como narra Klonick[57]. Ela defende que isso marca as políticas com uma forte crença nesse direito dos usuários face a autoridades governamentais e censura colateral.

Quando a Google adquiriu o YouTube, em 2006, contratou a advogada Nicole Wong para formular diretrizes sobre o conteúdo permitido na plataforma. Foi, então, adotada a regra de que nenhum conteúdo lícito seria removido, a não ser

55 WEST, Sarah Myers. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. **New Media & Society**, v. 20, n. 11, p. 4366-4383, 2018.

56 Uma crítica à diversidade de canais sobre os procedimentos e regras que orientam a moderação de conteúdo das plataformas é encontrada em: BELLI, Luca; VENTURINI, Jamila. Private ordering and the rise of terms of service as cyber-regulation. **Internet Policy Review**. v. 5, n. 4. 2016. p. 9. Disponível em: https://www.econstor.eu/bitstream/10419/214032/1/IntPolRev-2016-4-441.pdf Acesso em:08 out. 2020

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

que violasse aqueles padrões. Em 2009, Jud Hoffman e outras seis pessoas, entre elas o advogado David Willner, foram contratadas pelo Facebook para compor um conjunto de normas com os padrões de comunidade para assegurar transparência em relação aos usuários. No mesmo ano, o Twitter, plataforma que adotava política de ampla livre expressão e não-monitoramento sobre o que os usuários postam, contratou o advogado Alexander Macgillivray para o conselho geral. Ele desempenhou forte resistência contra pedidos governamentais de remoção de conteúdo durante seus quatro anos nesta posição. Esses são exemplos de uma tendência das plataformas em, inicialmente, como resposta às implicações do Caso Zeran, criar esforços para moderar conteúdo para evitar ofensas e danos a usuários e, ao mesmo tempo, sinalizar preocupação com a liberdade de expressão perante o público.

A intenção, com as políticas, é promover a imagem da plataforma como ambiente de respeito e abertura ao usuário. Assim, elas traçam linhas gerais sobre conteúdo indesejado, a fim de informar quanto ao que é ou não chancelado e de que forma é possível esperar que ela aja frente a violações. A fuga do ideal autoritário e a promoção da cultura de liberdade de expressão vincula-se a uma busca por transparência, que precisa ser demonstrada, segundo Gillespie, a três segmentos. São eles: **os usuários**, sobre a proteção a sua expressão e contra abusos, **os anunciantes**, sobre o ambiente amigável, **e os legisladores**, sobre a diligência da plataforma e a desnecessidade de outro tipo de regulação[58].

O último segmento, dos legisladores, é evidenciado em estudos como o mais afetado, na prática, pela publicação deste tipo de documento. Isso fica evidente na medida em que muitos dos termos de uso incluem, em suas cláusulas, a obrigação de respeito às políticas de comunidade. Esses documentos são, assim, inseridos na relação contratual entre o usuário e a plataforma, e demonstram que há um conjunto de regras governando as situações dela decorrentes em relação ao conteúdo circulado. Isso reforça a manifestação, para o setor estatal, de que essas empresas já estão elaborando soluções para lidar com potenciais conflitos, e que seria, portanto, desnecessário elaborar leis a respeito[59].

Em um relatório publicado pela Electronic Frontier Foundation, são elencadas três vantagens na publicação de políticas de comunidade: 1. a maior facilidade de percepção e documentação sobre mudanças de postura da plataforma; 2. a disponibilidade para que sejam analisadas e sejam discutidos publicamente quais padrões deveriam ser observados na moderação de conteúdo; e 3. a possibilidade de revisão de políticas pelas plataformas entre si, servindo de modelo de boas práticas para negócios iniciantes[60].

58 GILLESPIE, Tarleton. **Custodians of the Internet:** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018. p. 74

59 VENTURINI, Jamila et al. **Terms of service and human rights**: An analysis of online platform contracts. 2016. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/18231 Acesso em: 21 Jul. 2020.

60 CROCKER, Andrew et al. **Who has your back?** Censorship edition 2019. EFF - Electronic

A transparência também é retratada positivamente em estudos de referência sobre gestão de comunidades, como uma estratégia de abordagem de regras em relação aos usuários. Kraut e Resnick[61], em suas recomendações sobre o tema, consideram medidas interventivas sobre o conteúdo como atitudes extremas. Apontam ainda que meios comportamentais e brandos de moderação, como esclarecimento de diretrizes e envolvimento da comunidade como um todo no processo de moderação, são recomendáveis para dar legitimidade àquele conjunto de regras. Nesse sentido, Centivany[62] salienta a importância da participação do usuário no processo de elaboração de políticas de comunidade. Uma das formas de aproximar o usuário desse processo é a transparência oferecida pelas plataformas sobre seu conjunto de regras aplicáveis.

Embora aparentemente aliadas à transparência, as políticas de comunidade podem não necessariamente gerar a abertura prometida ou o ambiente regulado. Isto é, podem não capacitar o público a verificar o que e como, na prática, é moderado. Os ideais de livre expressão e manifestação envolvem uma democratização, ou seja, uma possibilidade de participação da comunidade nas políticas que a afetam. Nesse sentido, a ideia de accountability, segundo Fox[63], é complementar à de transparência e possui duas dimensões: a capacidade ou direito de demandar soluções e a capacidade de aplicar sanções. A *accountability* diz respeito à possibilidade de o usuário exigir determinada atitude das plataformas ou demandar das autoridades providências. E a transparência, por sua vez, não necessariamente implica nessa possibilidade.

Frontier Foundation. 2019. Disponível em: https://www.eff.org/wp/who-has-your-back-2019 Acesso em 21 Jul. 2020. p. 11

61 KRAUT, Robert E.; RESNICK, Paul. **Building successful online communities: Evidence-based social design**. Mit Press, 2012.

62 CENTIVANY, Alissa. Values, ethics and participatory policymaking in online communities. **Proceedings of the Association for Information Science and Technology**, v. 53, n. 1, p. 1-10, 2016.

63 FOX, Jonathan. The uncertain relationship between transparency and accountability. **Development in practice**, v. 17, n. 4-5, p. 663-671, 2007. p. 247

Fox alerta para **iniciativas de "transparência opaca", que blindam o ator de responsabilização efetiva enquanto geram uma fachada de preocupação social e política**. A transparência está associada à disseminação e acesso à informação, enquanto a *accountability* envolve possibilidade concreta de sanção ou resposta. A responsividade de uma instituição está vinculada a ambos os conceitos, entretanto, um não decorre automaticamente do outro. Isto é, transparência não necessariamente gera *accountability*[64]. As políticas de comunidade, dependendo do tipo de informação e ferramentas disponibilizadas, podem estar mais próximas de um ou de outro modelo.

Nesse sentido, alguns tipos de transparência podem, até mesmo, ser prejudiciais à garantia de direitos. Isso porque resultariam na imagem de que ações institucionais são adotadas para apresentar soluções, sem haver iniciativas reais para lidar com as situações necessárias. Como pontua Roberts[65], a experiência dos usuários pode ser afetada de maneiras diversas por políticas de comunidade opacas e indefinidas. Dessa forma, um usuário pode ter a impressão de absoluto respeito à liberdade de expressão em determinada plataforma e encarar o conteúdo que aparece como a verdadeira agregação de todas as ideias de todos os usuários. Outros, enquanto isso, podem considerar as regras de remoção e medidas interventivas como restritivas demais. A autora considera que a própria falta de nitidez das regras pode servir à plataforma. Isso pois a ausência de posicionamento aparente, como ato de despolitização, dá a impressão de objetividade e tecnicidade às medidas de moderação de conteúdo, apenas guiada pela atração de usuários e seu direcionamento a anunciantes.

Por sua vez, as políticas de comunidade vão muito além da tarefa de regular a moderação de conteúdo potencialmente ilícito ou considerado violador de regras de autoridades nacionais ou internacionais. Por meio delas, as plataformas criam conjuntos de regras e restrições sobre conteúdos não necessariamente ilícitos. As seções de políticas de comunidade geralmente são organizadas em tópicos que sofrem moderação. Abrangem venda de produtos ilícitos, falsidade ideológica, violação de propriedade intelectual, apologia a crimes e outras atividades que representam ilicitude, ao lado de segmentos temáticos como nudez, spam, comportamentos de risco, que são conteúdos controversos, porém cujo compartilhamento não necessariamente é ilícito. Isto é, elas delimitam o que é conteúdo violador, com base em seus próprios parâmetros. E, como os espaços de expressão por ela oferecidos são considerados por estudiosos como "quase-públicos"[66], devido a seu papel enquanto disseminadoras de informação *online*,

64 FOX, Jonathan. The uncertain relationship between transparency and accountability. **Development in practice**, v. 17, n. 4-5, p. 663-671, 2007. p. 250-252

65 ROBERTS, Sarah T. Digital detritus:'Error' and the logic of opacity in social media content moderation. **First Monday**, 2018. p. 4

66 A expressão é usada em ANANNY, Mike; GILLESPIE, Tarleton. Public Platforms: Beyond the Cycle of Shocks and Exceptions. In: **The Internet, Policy and Politics Conferences**. Oxford Internet

há implícita a obrigação de verificar o quanto essas normas privadas interferem em direitos humanos quando aplicadas[67].

As políticas de comunidade são, antes de tudo, declarações públicas de intenções e valores. A publicização de regras, como lembra Gillespie, é uma declaração não só do que é ou não permitido, mas do porquê[68]. Com a forte influência do direito americano por estarem sediadas, majoritariamente, nos Estados Unidos, e resguardadas pela seção 230 do DCA enquanto intermediárias, as plataformas são autorreguladas nesse ponto. Entretanto, isso não significa que seus interesses comerciais sejam o único fator influente sobre o caráter de maior ou menor transparência que decidem adotar em relação a suas políticas e medidas de moderação de conteúdo.

A transparência, no âmbito das plataformas online, é especialmente relevante. A declaração de justificativa dos princípios de Manila reconhece que, se uma autoridade pública só pode limitar a expressão em casos específicos, como segurança nacional ou proteção de direitos de terceiros, os intermediários online podem limitá-la conforme suas próprias regras - desde que elas sejam transparentes[69]. Entretanto, pode-se ir além, pois essa disponibilização ao público das regras presta-se à sua avaliação e à aferição de respeito a princípios democráticos em relação ao tratamento do conteúdo. A possibilidade de criar as próprias restrições de circulação de informação não é uma carta branca às plataformas. Elas ainda são sujeitas ao sistema internacional de proteção aos direitos humanos e, portanto, devem respeitar os princípios desenvolvidos para guiar as ações de empresas, conhecidos como Princípios Ruggie e mencionados nos relatórios especiais da ONU, por David Kaye[70]. Ele enfatiza que negócios

Institute, 2016. p. 15. A ideia de transposição da barreira entre público e privado pelo serviço prestado por plataformas também é presente em WEST, Sarah Myers. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. **New Media & Society**, v. 20, n. 11, p. 4366-4383, 2018. p. 4367; KLONICK, Kate. The new governors: The people, rules, and processes governing online speech. **Harv. L. Rev.**, v. 131, p. 1598, 2017 p. 1611; GORWA, Robert. What is platform governance?. **Information, Communication & Society**, v. 22, n. 6, p. 854-871, 2019. p. 862

67 MACKINNON, Rebecca et al. **Fostering freedom online**: The role of internet intermediaries. UNESCO Publishing, 2015. p. 55

68 GILLESPIE, Tarleton. **Custodians of the Internet:** Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018. p. 72

69 Esse aspecto é reconhecido no estudo que deu origem aos Princípios de Manilla, nos quais contribuíram representantes das seguintes organizações não-governamentais: Electronic Frontier Foundation, o Centre for Internet and Society (CIS, Índia), a Artigo 19 (Reino Unido), KICTANET (Quênia),a Derechos Digitales (Chile), a Asociación por los Derechos Civiles (ADC, Argentina) e a Open Net (Coréia do Sul). Ver: EFF - Electronic Frontier Foundation. **The Manila Principles on Intermediary Liability Background Paper**. 1. v. online, maio 2015. Disponível em: https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf#page=49 Acesso em: 20 ago. 2020. p. 48

70 ONU - Organização das Nações Unidas. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression**. Symbol A/HRC/38/35. 6 abr. 2018. Disponível em: https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35 Acesso em: 30 jun. 2020. p. 5-6

teriam um mínimo de responsabilidades no meio digital, como 1) evitar causar ou contribuir para efeitos colaterais em direitos humanos, 2) comprometer-se, em suas políticas, com proteção de alto nível aos direitos humanos de seus usuários, 3) conduzir devida diligência e análise de risco de suas atividades para direitos humanos, 4) adotar estratégias de mitigação e prevenção que priorizem direitos humanos reconhecidos internacionalmente ao máximo frente a regulações locais, 5) conduzir avaliações e revisões de suas políticas e práticas em relação a direitos humanos, consultando partes interessadas, 6) oferecer reparação apropriada e mecanismos de reclamação ao usuário. Pode-se argumentar que a elaboração de regras de restrição de conteúdo por uma plataforma privada só é justificada mediante a sujeição dessas a algum tipo de controle social efetivo.

O ambiente online exige uma coordenação que vai além do âmbito nacional, de maneira que os instrumentos existentes atualmente para nortear boas práticas em relação a comunidades online são códigos de conduta. Estes são corregulações construídas em diálogo das autoridades com as entidades afetadas, bem como princípios elaborados pela comunidade acadêmica e pela sociedade civil em busca de garantias democráticas para os usuários. Nenhum desses instrumentos tem força de lei, isto é, não podem ser aplicadas sanções às empresas responsáveis pelas plataformas se elas não seguirem os princípios ou códigos de conduta estabelecidos por esse tipo de dispositivo. Entretanto, eles representam instrumentos de *soft law*[71], contribuem com o debate sobre a transparência na moderação de conteúdo e apontam critérios para avaliação de como uma plataforma lida com suas políticas.

Com base na função de garantidoras da transparência e liberdade de expressão sob a qual as políticas de comunidade são concebidas, esta pesquisa buscou analisar em que medida elas observam os padrões internacionalmente estabelecidos. Para isso, foi desenhada uma coleta dos textos e realizada uma análise de conteúdo, conforme a metodologia descrita na seção seguinte.

3. Metodologia de análise de políticas de comunidade

A proposta desta pesquisa foi analisar a transparência promovida pelas políticas de comunidade e como esse canal informativo possibilita conhecer os critérios,

71 *Soft law* são padrões de orientação sobre procedimentos para determinado setor, normalmente estabelecidos por entidades relevantes naquela área, mas que não possuem teor juridicamente vinculante. Por isso, seu descumprimento não resulta na aplicação de sanções formais e não se pode exigir seu cumprimento perante tribunais. Devido a isso, a *soft law* não é considerada direito, mas exerce influência sobre condutas, com potencial de gerar consequências para quem não a observa. A não-observância da *soft law* pode ter efeitos indiretos, aplicados pela comunidade comprometida com aqueles parâmetros, geralmente se refletindo na reputação e na capacidade negocial do agente implicado. Não existe uma tradução do termo para o português, inclusive por ser uma categoria reconhecida essencialmente no âmbito das relações internacionais.

medidas e atitudes interventivas adotadas pelas plataformas de conteúdo gerado por usuário. Para isso, foi realizada uma coleta, codificação e análise dos textos que compõem essas políticas. Essa parte da pesquisa compreendeu as etapas de: 1. definição da amostra, 2. elaboração de critérios de análise, 3. coleta dos documentos, 4. codificação e 5. análise do material, que são melhor articuladas a seguir.

3.1. Definição da amostra de plataformas

A análise de conteúdo foi considerada a metodologia adequada para a investigação dos aspectos específicos das políticas de comunidade. Segundo Bardin[72], a prática consiste em um conjunto de técnicas de análise das comunicações com caráter objetivo e sistemático, que permite inferências sobre as condições de produção e recepção das mensagens (sua origem, seu contexto ou seus efeitos).

Para a coleta dos textos, foi necessário delimitar quais plataformas seriam objeto da pesquisa pretendida. A fim de garantir uma amostra representativa[73] das políticas utilizadas para moderar conteúdo de potencial grande de alcance, definiu-se o critério de quantidade de usuários. Reconhecendo ainda que a presente pesquisa é um recorte do ponto de vista de pesquisadores brasileiros, buscaram-se dados sobre as plataformas mais utilizadas no país.

Para saber quais as plataformas com mais usuários, recorreu-se à pesquisa "Digital 2019", realizada pelas agências de marketing online Hootsuite e We Are Social[74], na qual consta o ranking de plataformas de mídias sociais mais ativas no Brasil. Como não foi possível localizar essa informação em bases de pesquisa acadêmica, recorreu-se às estatísticas de uso de aplicativos no Brasil divulgadas por agências[75]. Dessa forma, buscou-se apoiar a confiabilidade da pesquisa verificando que seus dados são usados por sites especializados[76], que a citam como fonte. Ela também é coerente com outras pesquisas, como o relatório da SensorTower[77], agência de

72 BARDIN, Laurence. **Análise de conteúdo**. Trad. Luís Antero Reto e Augusto Pinheiro. Lisboa: 70, 1977. p. 42

73 BARDIN, Laurence. **Análise de conteúdo**. Trad. Luís Antero Reto e Augusto Pinheiro. Lisboa: 70, 1977.p. 97

74 O relatório Digital 2019 foi publicado no site Datareportal, o qual reúne dados sobre mídias sociais para área de marketing e publicidade. O material foi produzido a partir de diversas fontes, que podem ser conferidas em: https://datareportal.com/data-sources

75 Uma vez que essa informação é produzida por atores do setor privado, há possibilidade de conflito de interesses com os fins da pesquisa científica. Por esse motivo, os dados devem ser considerados com cautela.

76 https://www.techtudo.com.br/noticias/2019/02/conheca-as-redes-sociais-mais-usadas-no-brasil-e-no-mundo-em-2018.ghtml

77 SENSORTOWER. **Q4 2019**. Store Intelligence Data Digest. Disponível em: https://go.sensortower.com/Q4-2019-Data-Digest.html?src=blog Acesso em: 20 ago. 2020

análise de dados, que divulgou ranking dos 10 aplicativos mais instalados em aparelhos de celular em 2019 e contava com 5 das plataformas de maior uso pelos brasileiros segundo a Digital 2019.

Na pesquisa que serviu de fonte para a escolha, apresentava-se um gráfico que distinguia as plataformas entre redes sociais e aplicativos mensageiros. Esses últimos foram excluídos da amostra por não comportarem moderação de conteúdo gerado por usuário nos moldes aqui narrados. Com objetivos e funcionalidades distintas, não permitem a publicação de conteúdo em geral, mas voltam-se ao seu envio como forma de comunicação direta com a pessoa ou o grupo interessado. Assim, em geral, esses serviços não dispõem de ferramentas de filtragem de conteúdo, visto que oferecem o sigilo das comunicações como um diferencial[78], e, via de regra, suas políticas de comunidade são mais enxutas.

Observado esse critério, foram incluídas as sete plataformas mais utilizadas pelo público brasileiro. Isso considera a grande adesão dos usuários a esses serviços e marcas, enquanto as demais contam com menos de um quinto dos usuários. Por essa razão, não haveria um equilíbrio entre os esforços desprendidos para sua análise e a representatividade delas no cenário estudado.

Adicionalmente, a plataforma TikTok foi incluída na amostra devido a seu crescimento. A rede social rapidamente ocupou uma posição de destaque e veio a constar no ranking de 2020 da mesma pesquisa usada como referencial para o ranking de plataformas[79]. Essa ascensão também foi acompanhada de cobertura midiática[80] de situações problemáticas envolvendo políticas de moderação de conteúdo implementadas inicialmente pela plataforma. Isso a tornou relevante objeto de estudo para os fins aqui pretendidos.

Portanto, a amostra final que compõe as plataformas nas quais foram coletadas políticas de comunidade para análise é a lista das 8 aplicações a seguir:

78 Embora esses serviços podem contar com moderação no nível de controle de metadados, como o número de usuários ao qual uma mensagem será encaminhada, quantitativo máximo de usuários em determinado grupo, possibilidade de bloqueio de contas, etc. Entretanto, esses aspectos de moderação contam com um tipo de intervenção distinta (por exemplo, em regra não há remoção de conteúdo) e, comumente, as políticas que as norteiam não são direcionadas para o tipo/mérito do conteúdo, mas para a forma de sua propagação.

79 HOOTSUITE; WE ARE SOCIAL. **Digital 2020**: Brazil - DataReportal Global Digital Insights. Disponível em: https://datareportal.com/reports/digital-2020-brazil?rq=brazil Acesso em: 20 ago. 2020. p. 43

80 BIDDLE, Sam; RIBEIRO, Paulo Victor; DIAS, Tatiana. TikTok escondeu “feios” e favelas para atrair novos usuários e censurou posts políticos. **The Intercept Brasil**. 16 mar. 2020. Disponível em: https://theintercept.com/2020/03/16/tiktok-censurou-rostos-feios-e-favelas-para-atrair-novos-usuarios/ Acesso em: 16 mar. 2020.

1. Youtube
2. Facebook
3. Instagram
4. Twitter
5. LinkedIn
6. Pinterest
7. Snapchat
8. TikTok

Ainda, devido à localização da equipe de pesquisa, explicita-se que as políticas de comunidade analisadas foram as disponíveis para acesso a partir do Brasil, de forma que todas se encontravam em língua portuguesa.

3.2. Elaboração de critérios de análise

Definidos o objeto e a amostra, foi necessário firmar categorias de análise do conteúdo das políticas, que permitissem identificar iniciativas e lacunas na transparência oferecida por esses documentos. A partir dos padrões democráticos internacionais para a atividade de moderação, foram elaborados pontos a serem observados no texto das políticas, sobre seu conteúdo informativo.

A primeira categoria, “Fundamentação da política”, diz respeito aos posicionamentos da plataforma em relação a seu papel ou ao tipo de conteúdo ou interação incentivado. Inicialmente, foram destacados trechos sobre valores e fundamentos das políticas. A segunda e a terceira categorias básicas criadas foram as de ”Proibição” e ”Exceção”. Elas são trechos que permitam ao usuário conhecer os limites estabelecidos pela plataforma ao conteúdo, identificar potenciais violações e saber como adequar o conteúdo ao contexto de exceção à regra, conforme o caso. Também foi criada a categoria “Recomendação de Conteúdo”, referente a frases que não representavam proibições de conteúdo, mas eram importantes para identificar como a política lida com limites ao conteúdo e às interações esperadas pela plataforma.

Também foram criadas categorias “Meio de análise” e “Critério de análise”, para os trechos que indicam os meios de triagem de conteúdos a serem moderados e os critérios que levam a decidir sobre aplicação ou não de medida interventiva.

Essas informações permitiriam ao usuário questionar ferramentas de triagem e automatização de moderação de conteúdo, quando houvesse desequilíbrio em relação a liberdade de expressão. Ainda possibilitariam ao usuário que sofre medida interventiva identificar se houve uma decisão adequada e que observou os critérios prefixados, de forma a desafiar medidas aplicadas de maneira ilegítima. Além disso, na categoria "Apoio ao usuário", foram identificados trechos que servem menos de orientação sobre a atividade de moderação em si e mais de apoio ao usuário. Eles foram considerados informações adicionais sobre as políticas e seus procedimentos, maneiras de contactar a plataforma quando constatada uma violação e formas de contestar medidas interventivas aplicáveis.

Todas essas categorias integram elementos a serem observados para a construção de políticas de comunidade que comuniquem, de fato, o sistema de regras implementado e a ele integrem o usuário. Kraut e Resnick[81] recomendam que, para uma boa gestão de comunidades, o usuário deve ter as regras em mente ao tomar decisões. Os autores afirmam, nessa linha, que há 3 formas de conhecer as regras de uma comunidade: observação do comportamento de seus membros; 2. leitura de códigos de conduta ou generalizações; ou 3. tendo retorno sobre seu comportamento. As duas últimas maneiras possuem reflexos nas políticas de comunidade e foram incorporadas na análise aqui realizada.

Para Kraut e Resnick, instruções diretas são mais eficazes do que exemplos ou conselhos indiretos. Além disso, consideram que avisos ao violador das normas e chances de reverter a conduta demonstram confiança em sua boa-fé. São formas, assim, de incentivar os usuários a cumprirem mais as regras. Essa prática está relacionada à nitidez de proibições, exceções, apoio ao usuário sobre contestação e notificação de medidas interventivas. Também consideram que é preciso ter meios eficazes de detecção das violações, como, por exemplo, sinalização pelo usuário. Essa prática é incorporada nas políticas pela menção aos meios de detecção de conteúdo potencialmente infringente. Ainda, sustentam que é mais importante as sanções serem aplicadas com regularidade, ainda que sejam leves, do que serem severas e incertas. Nesse caso, a incorporação é verificada nas políticas pela análise de informações sobre medidas interventivas aplicáveis.

As grandes categorias manejadas foram, assim: 1) fundamentação da política; 2) proibição; 3) exceção; 4) recomendações de conteúdo; 5) medida interventiva; 6) meio de análise; 7) critério de análise; e 8) apoio ao usuário.

Com exceção da primeira categoria, todas foram subdivididas em códigos específicos. Esses códigos eram aplicados a trechos do texto, que variavam entre frases ou parágrafos, denominados "citações". Com eles, buscou-se analisar a especificidade da proibição ou da exceção, para identificar se o texto permite

81 KRAUT, Robert E.; RESNICK, Paul. **Building successful online communities**: Evidence-based social design. Mit Press, 2012.

ao usuário saber que situações ensejam violação da política ou não. A fim de determinar se o conteúdo violador seria sempre nitidamente proibido ou se estaria indicado com sutileza em alguns trechos. Quando os termos usados não denotavam expressamente proibição (“é proibido”, ”não é permitido”, etc.), o conteúdo foi enquadrado não como proibição, mas como conteúdo desencorajado. Os parâmetros também consideram se o usuário teria acesso aos meios de que a plataforma lança mão para localizar conteúdo a ser moderado e à forma de definição da intervenção e sua eventual aplicação. Um relato detalhado dos códigos construídos e do significado de cada um pode ser encontrado no Apêndice A, ao final deste trabalho.

Fundamentação da política

Proibição::

Proibição:: Específica

Proibição:: Genérica

Exceção::

Exceção: Específica

Exceção: Genérica

Recomendações de conteúdo::

Recomendações de conteúdo:: Desencorajado

Recomendações de conteúdo:: Encorajado

Medida interventiva::

Medida interventiva: Específica

Medida interventiva: Genérica

Meio de análise::

Meio de análise:: Determinado

Meio de análise:: Indeterminado

Critério de análise::

> Critério de análise:: Indeterminado
>
> Critério de análise:: Determinado
>
> Apoio ao usuário::
>
> Apoio ao usuário:: Denúncia
>
> Apoio ao usuário:: Contestação
>
> Apoio ao usuário:: Informações

Tendo definidos as categorias e códigos, procedeu-se à documentação e codificação dos textos das políticas.

3.3. Coleta e codificação dos documentos

Para registrar e documentar as políticas de comunidade objeto de análise da pesquisa, foram acessadas as páginas de cada uma das plataformas, na seção correspondente às políticas ou diretrizes de comunidade. Em seguida, criou-se um arquivo de texto para cada uma, na qual foi copiado o conteúdo, mantida a formatação básica original. Esta coleta foi realizada no dia 22 de abril de 2020, após a realização de testes e refinamento da codificação, para garantir uma análise temporalizada e o mais atualizada possível de todas as políticas até aquele momento, ainda que possa haver modificação posterior. Também foram realizadas, no dia 24 de junho de 2020, capturas de tela das páginas, a fim de ilustrar a disposição do conteúdo e a diagramação na qual ele é fornecido ao usuário que acessa as políticas.

Os arquivos de texto foram, sucessivamente, codificados de acordo com as categorias estabelecidas anteriormente. Ainda, foram criados códigos personalizados para cada política de comunidade, com o nome da plataforma e o título da seção. Eles foram aplicados conjuntamente ao código de análise, a fim de facilitar a identificação da estrutura de apresentação do conteúdo dentro de cada seção e os padrões observados por cada plataforma. Também havia a intenção de verificar se, internamente à política, ocorria padronização da apresentação das informações sobre moderação para cada segmento moderado, o que indicaria maior preocupação com transparência sistemática. Ainda, houve a percepção de que os formatos dos documentos não são padronizados entre si, de maneira que não formam um corpus homogêneo para análise; a codificação das seções buscou evidenciar essa diversidade de formatos e subdivisões temáticas dos tópicos tratados pelas políticas.

Para essa etapa, a codificação foi realizada com apoio do software Atlas.ti Cloud[82], uma ferramenta de suporte a análise qualitativa e quantitativa de conteúdo, que permite a criação de códigos e dispõe de meios de tratamento estatístico automatizado de dados qualitativos. O software permitiu a realização coletiva da codificação, organizada da seguinte maneira: os 8 documentos foram inseridos em um projeto conjunto dos pesquisadores na plataforma e foram distribuídos entre eles. Cada um foi responsável por codificar 4 textos e revisar a codificação dos outros 4, realizada pelo colega. Nos casos de dúvida, houve aplicação de um código com essa sinalização, e posteriormente foi realizada análise conjunta, o que permitiu melhor nitidez dos critérios e delimitações de cada código. Destaca-se que o software em questão não realiza análises de forma autônoma e que a interpretação dos dados para a produção do conhecimento resultante requer o trabalho interpretativo dos pesquisadores, que o fazem suportados por sua fundamentação teórica e conceitual.[83].

3.4. Análise do material codificado

Feita a codificação, foi possível identificar quais os trechos relevantes das políticas e de que forma os critérios de transparência eram abordados: 1. de maneira completa; 2. incompleta; ou 3. estavam ausentes. A partir desse material, a transparência na política de cada uma das plataformas foi individualmente relatada. Em seguida,compararam-se as diferentes formas pelas quais os critérios elencados na codificação são observados.

As relatorias estão organizadas de acordo com os critérios de codificação e têm caráter descritivo, com apontamentos sobre a informação proporcionada pelo texto lido e também sobre eventuais dúvidas suscitadas pela forma como as informações são estruturadas e o seu contexto.

Os comparativos não buscam estabelecer um ranking, pois esta tarefa exigiria a atribuição de relevâncias, pesos e parâmetros classificatórios aos critérios de análise, o que não integra o escopo desta pesquisa. Seu objetivo, na verdade, é fornecer um panorama de continuidades e rupturas entre as abordagens realizadas por cada política de comunidade no aspecto da transparência. Assim, estão organizados no formato de quadros: o primeiro mais ilustrativo e objetivo e o segundo mais informativo, com exemplificações para uma imagem mais concreta do material discutido.

Essa análise compreendeu a interpretação e definição de parâmetros para leitura

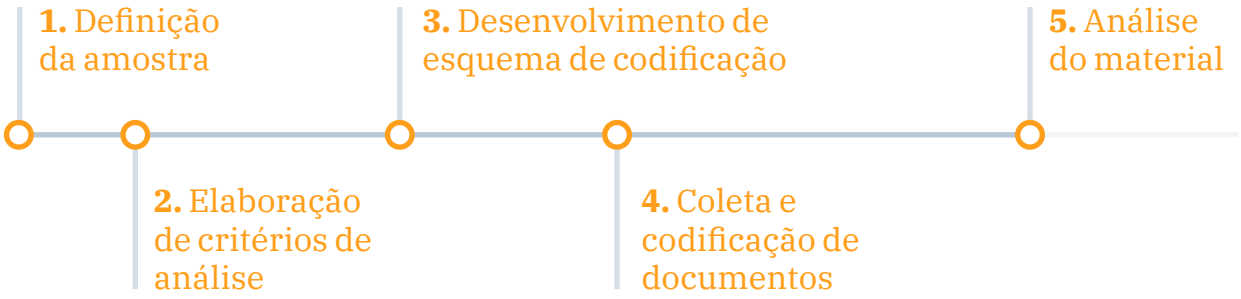
82 Disponível em: https://atlasti.com/cloud/

83 SILVA JUNIOR, Luiz Alberto; LEAO, Marcelo B. C. O software Atlas.ti como recurso para a análise de conteúdo: analisando a robótica no Ensino de Ciências em teses brasileiras. **Ciênc. educ. (Bauru)**, Bauru , v. 24, n. 3, p. 715-728, set. 2018.

das citações (trechos codificados), bem como a constatação de presença ou ausência de citações sob determinado código nos documentos. Em algumas seções (critérios de avaliação, exceções), preservamos o conteúdo específico das citações (por exemplo: o critério é o histórico do usuário, a exceção é para fins educacionais). Em outras, por outro lado, relatamos somente se o elemento analisado é especificado de forma precisa ou imprecisa (por exemplo: se há ou não exemplos nas proibições, se as medidas interventivas são determinadas ou indeterminadas).

Adotar a primeira abordagem mostrou-se inviável na medida em que há variações demais no conteúdo das proibições. Além disso, saber o conteúdo desses elementos não parece especialmente relevante para a aferição de transparência dos padrões de comunidade. Tendo em vista esse objetivo, o que importa é saber se uma proibição é expressa de forma inequívoca, por exemplo, e não o que ela proíbe especificamente. Por outro lado, os dados levantados acerca do conteúdo que é moderado podem ser interessantes do ponto de vista ilustrativo, motivo pelo qual são citados exemplificativamente.

A organização dos resultados pretende ser um retrato de quais os déficits e boas práticas das plataformas em relação à transparência proporcionada pelas políticas de comunidade. Busca, portanto, servir de suporte às recomendações que são realizadas ao final desta pesquisa.



4. Resultados

Esta seção apresenta os resultados da análise. Inicialmente, descrevemos os critérios que compuseram o modelo de relatoria utilizado e as questões que orientaram o preenchimento da seção correspondente ao critério. A seguir, são apresentadas as relatorias individuais de cada caso. Elas incluem uma breve descrição da plataforma e o relato das respostas às questões orientadoras a partir da codificação dos padrões de comunidade. Por fim, discutem-se os resultados de cada critério da relatoria, com especial atenção a práticas que reduzem a transparência de cada critério.

4.1. Critérios de relatoria

Da estrutura geral da política. Como está estruturada a exposição dos padrões de comunidade? Trata-se de uma página única e centralizada ou de diversas páginas relativas a normas específicas, por exemplo?

Da detecção de conteúdo potencialmente infringente. Os padrões de comunidade informam ao usuário como a plataforma ganha ciência de que algum conteúdo está potencialmente violando suas políticas? Por exemplo, se tal detecção se dá por meio de análise proativa massiva ou exclusivamente por meio de denúncias.

Dos meios de avaliação de conteúdo potencialmente infringente. Uma vez que conteúdo potencialmente infrator foi detectado, os padrões de comunidade informam ao usuário os meios empregados pela plataforma para avaliar se aquele conteúdo efetivamente infringe suas políticas? Por exemplo, se essa análise é feita por moderadores humanos ou se é feita por algoritmos.

Dos critérios de avaliação de conteúdo potencialmente infringente. No contexto da avaliação de conteúdo potencialmente infrator, os padrões de comunidade informam ao usuário os critérios empregados pela plataforma para avaliar 1. se aquele conteúdo efetivamente infringe suas políticas e 2. quais as medidas cabíveis?

Dos exemplos. Na ocasião da afirmação de uma proibição qualquer, os padrões de comunidade oferecem exemplos e contraexemplos de conteúdos considerados infratores da proibição?

Das exceções. Os padrões de comunidade informam ao usuário sobre exceções para proibições específicas? Em caso afirmativo, quais são essas exceções?

Da especificação não-ambígua de conteúdos infringentes. Os padrões de comunidade informam ao usuário de forma não-ambígua quais são os conteúdos proibidos na plataforma?

Das medidas interventivas aplicáveis. Os padrões de comunidade especificam de forma não-ambígua ao usuário quais medidas são aplicáveis a quais infrações?

Da contestação. Os padrões de comunidade informam ao usuário que ele possui direito a contestar a medida interventiva sofrida? Em caso afirmativo, oferecem indicação de como o usuário pode efetuar a contestação?

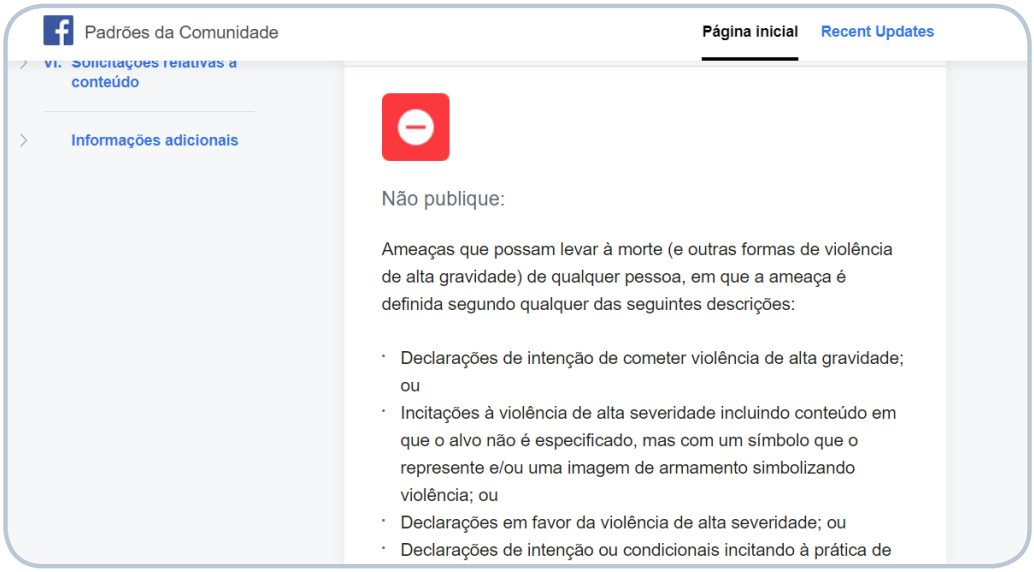
Da notificação. Os padrões de comunidade informam ao usuário sobre como ele será notificado em caso de violação das políticas de comunidade?

4.2. Facebook

Descrição: Trata-se de uma plataforma voltada ao compartilhamento de conteúdo em múltiplos formatos (visual, textual, audiovisual, etc). Os conteúdos podem ser disponibilizados na forma de postagens duradouras, publicáveis nos perfis dos usuários, em grupos ou páginas, ou efêmeras, publicáveis nos perfis dos usuários. É possível seguir usuários, reagir, compartilhar ou comentar suas publicações e enviar mensagens de forma privada. O conteúdo dos perfis, páginas e grupos que o usuário segue é exibido em forma de fluxo de publicações em sua página inicial a partir de uma curadoria algorítmica.

Da estrutura geral da política. Uma página centralizada apresenta a introdução e a lista de seções, cada uma contendo subseções para políticas temáticas específicas. Cada subseção temática inclui um ou dois parágrafos intitulados "fundamentação da política" e o endereço de um segmento da página que dispõe de mais detalhes sobre a norma.

Página da política de comunidade do Facebook



Exemplo de sinalização de proibição no Facebook

Da detecção de conteúdo potencialmente infringente. A plataforma indica a denúncia como meio de detecção generalizado. Além disso, especificamente em sua política de notícias falsas, afirma que utiliza diversos recursos, incluindo feedback da comunidade, para informar um modelo de aprendizado de máquina que detecta de forma proativa conteúdo potencialmente desinformativo. Ainda, em sua política de exploração sexual de adultos, também alude a mecanismo automatizado de detecção, oferecendo o endereço de uma página para mais informações sobre tal tecnologia.

Dos meios de avaliação de conteúdo potencialmente infringente. Além dos casos supracitados, não há citações que permitam determinar de que modo os conteúdos potencialmente infringentes são analisados.

Dos critérios de avaliação de conteúdo potencialmente infringente. Há numerosas referências explícitas à análise do contexto em diversas políticas e critérios específicos são apontados em determinadas normas: a avaliação de potencial caso de exploração sexual de adultos considera a evidência de falta de consentimento; a política de violência e incitação fatora a presença de indícios críveis de dano físico ou ameaça à segurança pública, bem como o quão direto foi o ataque, a visibilidade da pessoa afetada e a intenção do usuário. Estes dois últimos critérios também são observados na política de bullying e assédio.

Dos exemplos. Diversos exemplos de violações são oferecidos em distintas políticas.

Das exceções. Algumas políticas estabelecem exceções para conteúdo educativo, artístico, satírico, religioso, documental, médico, de empoderamento ou de interesse público. A política de nudez adulta e atividades sexuais excetua amamentação, conteúdo relacionados a parto e puerpério e conteúdo em que a atividade sexual esteja implícita ou as imagens não apresentarem detalhes suficientes, com apenas formas e contornos corporais visíveis. São exceções à política de produtos controlados: 1. a promoção de itens disponíveis para venda fora da plataforma no âmbito da política de produtos controlados, desde que obedecidas as normas pertinentes; 2. os debates sobre a venda de armas e peças de armas; e 3. a discussão sobre a regulação dos tópicos.

Da especificação não-ambígua de conteúdos infringentes. A plataforma faz uso de linguagem visual que alude à proibição no início de seus segmentos contendo mais detalhes sobre cada norma. Isso sugere que os conteúdos ali situados seriam proibidos, porém o uso intercambiável dos termos “comportamentos proibidos”, “não publique” e, simplesmente, “não” dificulta a identificação inequívoca de que conteúdos efetivamente infringem as políticas e quais deles não o fazem. Adicionalmente, esses segmentos incluem conteúdos que não são efetivamente proibidos, mas cuja visualização está sujeita a restrições etárias ou à sinalização

de conteúdo sensível em algumas políticas (produtos controlados e automutilação, por exemplo).

Das medidas interventivas aplicáveis. Na seção inicial, a plataforma afirma que considera a gravidade da violação e o histórico do usuário na determinação de qual medida será aplicada. Os conteúdos sujeitos à sinalização e restrição etária são listados de maneira específica e pormenorizada. De modo geral, contudo, não é possível precisar de que maneira os diferentes critérios influenciam a determinação das medidas aplicáveis à maioria das infrações. Adicionalmente, o recurso frequente ao verbo poder ("poderemos remover") e a advérbios de incerteza ("geralmente removemos") amplia a insegurança sobre a resposta suscitada por cada infração.

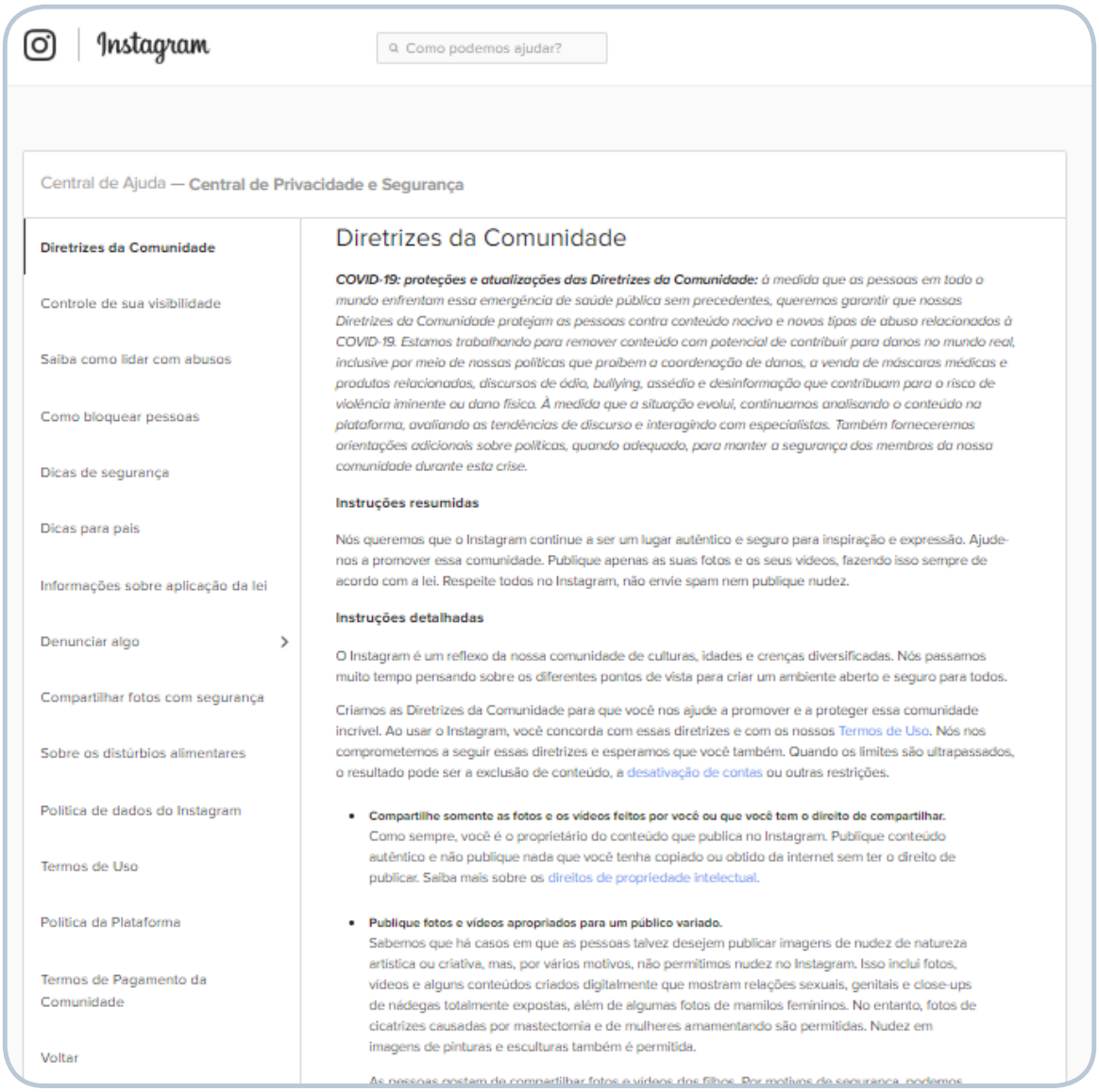
Da contestação. Não há quaisquer referências a mecanismos de contestação disponíveis ao usuário que sofreu medida interventiva.

Da notificação. Não há quaisquer referências à notificação enviada ao usuário na ocasião de seu conteúdo ou conta ser alvo de medida interventiva.

4.3. Instagram

Descrição: Consiste em plataforma focada em compartilhamento de conteúdo visual (imagens) e audiovisual (vídeos). É possível que o usuário publique fotografias, desenhos, vídeos curtos em dois canais: seu *feed*, na forma de postagens compostas de imagem e legenda, que ficam associadas ao perfil do usuário; e seus *stories*, que é uma ferramenta de compartilhamento temporário de conteúdo de forma sequencial e efêmera. Nesse caso, a plataforma oculta aquelas imagens depois de um certo período em horas. É possível seguir usuários para visualizar suas atualizações, compartilhar conteúdo do *feed* ou dos *stories*, "curtir" fotos do *feed* de outros usuários, bem como comentar na postagem, curtir comentários de outros usuários, reagir aos *stories* de outros usuários e enviar mensagens de forma privada.

Da estrutura geral da política. O Instagram tem uma central de ajuda estruturada na forma de menu de tópicos, na qual está contida a página de diretrizes de comunidade. Esta página possui um texto que conta com introdução, parágrafo de instruções resumidas e depois uma lista de itens intitulada "instruções detalhadas", em que cada item corresponde a um tipo de conduta desejada em relação ao conteúdo. Esses itens são parágrafos de pouco mais de 10 linhas.



Página inicial das políticas de comunidade do Instagram - visão geral

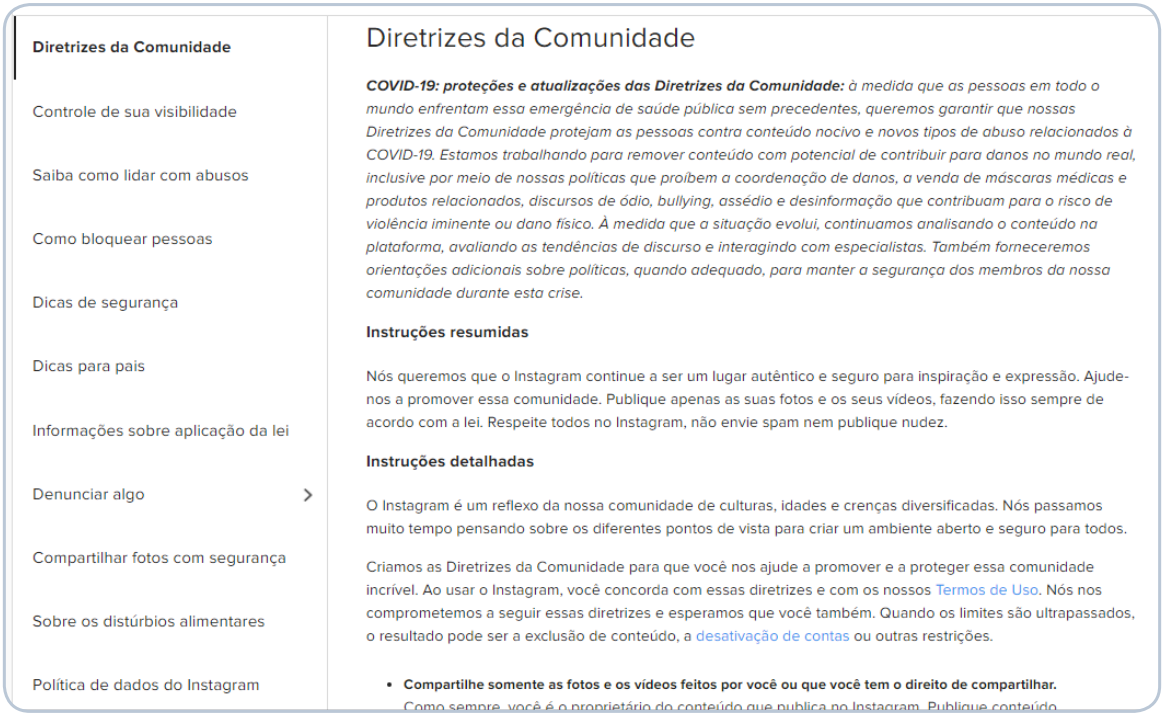


Imagem aproximada com tópicos e texto de abertura das políticas

Da detecção de conteúdo infringente. Apenas menciona que há análise de denúncias. Há instruções sobre a forma de realizar denúncia, menção à ferramenta de denúncia e aconselhamento sobre em que casos é cabível usá-la.

Dos meios de avaliação de conteúdo potencialmente infringente. Menciona que tem "uma equipe global que analisa as denúncias e trabalha o mais rápido possível para remover o conteúdo que não segue as nossas diretrizes".

Dos critérios de avaliação de conteúdo potencialmente infringente. "Analisamos cuidadosamente as denúncias de ameaças e consideramos várias questões para determinar se uma ameaça é real". Não são explícitas as questões consideradas na maioria dos casos. Há uma referência à visibilidade do usuário como critério para a aplicação das políticas - a plataforma afirma que permite "discussões mais acaloradas" sobre pessoas de elevada visibilidade, embora não explicite o que isso significa.

Dos exemplos. De nove citações proibitivas, duas incluem múltiplos exemplos. As outras consistem na mera enunciação da regra ou citam um único exemplo.

Das exceções. Há algumas exceções relativas ao teor do conteúdo em normas específicas: 1. conteúdo compartilhado para fins condenatórios ou de conscientização (discurso de ódio, violência explícita); e 2. imagens relativas a mastectomia e amamentação, bem como pinturas e esculturas (nudez). Há, também, exceções relacionadas à situação da conta que publicou o conteúdo: anúncios de vendas de animais físicos são permitidos quando realizados por lojas, mas não por pessoas físicas. Similarmente, anúncios de compra e venda de produtos controlados são permitidos desde que com a permissão prévia da plataforma. O uso do verbo "poder" prejudica a determinação plena da exceção quanto a discurso de ódio.

Da especificação não-ambígua de conteúdos infringentes. Existe um nível de ambiguidade entre os conteúdos infringentes e meramente desencorajados. Existem citações que contém termos expressos, como "não é permitido", "nunca é aceitável", "nós não toleramos", "o instagram não aceita" e "podemos remover". Também há, contudo, citações como "não publique", "não + verbo no imperativo", ou "evite", que podem significar conteúdo desencorajado, mas não necessariamente são proibições explícitas.

Das medidas interventivas aplicáveis. De maneira geral, não há definição sobre em que casos são aplicadas, de fato, medidas interventivas. Diversos trechos apresentam a expressão "podemos remover", que indica de mera possibilidade. Também há trechos com medidas alternativas entre si ("removeremos o conteúdo ou desativaremos as contas", "o resultado pode ser a exclusão de conteúdo, a desativação de contas ou outras restrições"), sem especificar em que contexto

uma ou outra é tomada.

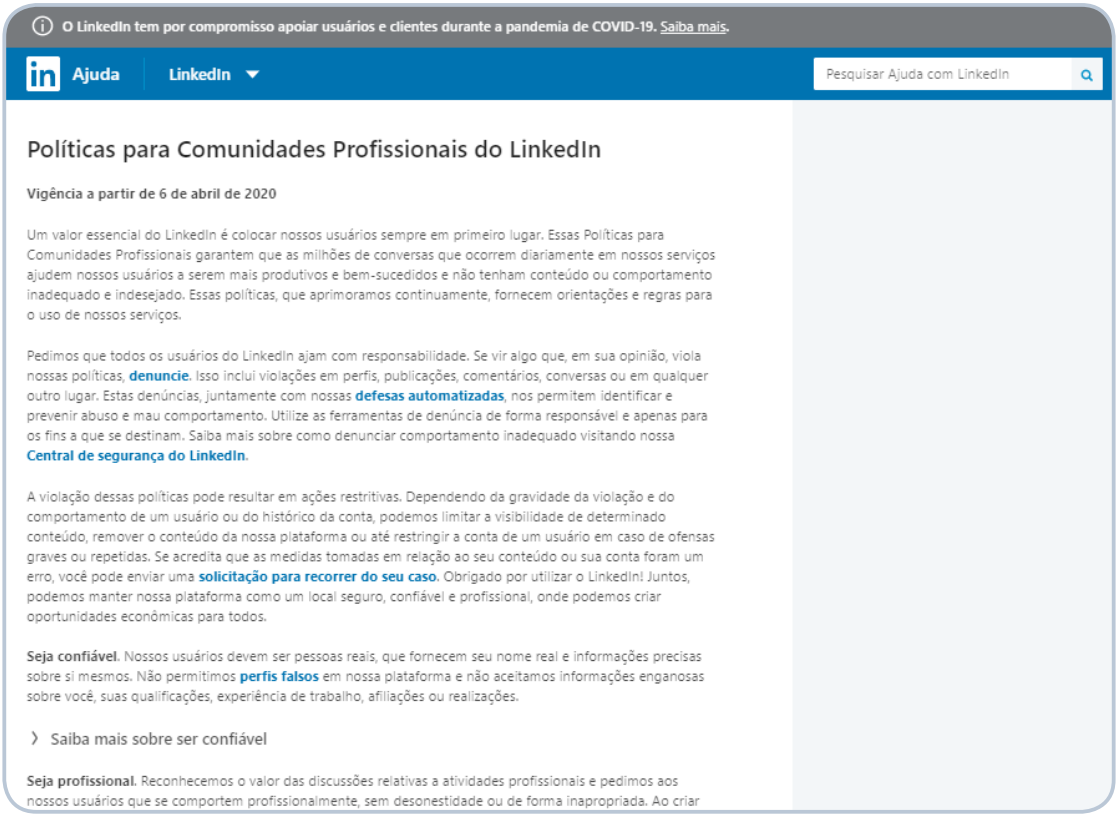
Da contestação. Não há menção à possibilidade de contestar uma denúncia nas políticas.

Da notificação. Não há menção à notificação do usuário quando sofre uma denúncia ou medida interventiva.

4.4. LinkedIn

Descrição: É uma plataforma voltada a profissionais e suas atividades nas respectivas áreas, compartilhamento de currículos e dicas sobre mercado de trabalho e gestão. A plataforma possibilita ao usuário criar um perfil detalhado no formato de currículo, compartilhar conteúdo de outras plataformas da internet (como postagens em blogs) e também textos e imagens no formato de postagem. É possível seguir usuários para visualizar suas atualizações, bem como visualizar postagens com as quais eles interagiram, recomendar competências presentes no currículo dos outros usuários, reagir e comentar em postagens e enviar mensagens privadas.

Da estrutura geral da política. Toda a política é em uma página, dividida em itens/condutas que se expandem na mesma página ao clicar em "saiba mais", em cada um deles.


O LinkedIn tem por compromisso apoiar usuários e clientes durante a pandemia de COVID-19. Saiba mais.

Ajuda | LinkedIn

Pesquisar Ajuda com LinkedIn

Políticas para Comunidades Profissionais do LinkedIn

Vigência a partir de 6 de abril de 2020

Um valor essencial do LinkedIn é colocar nossos usuários sempre em primeiro lugar. Essas Políticas para Comunidades Profissionais garantem que as milhões de conversas que ocorrem diariamente em nossos serviços ajudem nossos usuários a serem mais produtivos e bem-sucedidos e não tenham conteúdo ou comportamento inadequado e indesejado. Essas políticas, que aprimoramos continuamente, fornecem orientações e regras para o uso de nossos serviços.

Pedimos que todos os usuários do LinkedIn ajam com responsabilidade. Se vir algo que, em sua opinião, viola nossas políticas, **denuncie**. Isso inclui violações em perfis, publicações, comentários, conversas ou em qualquer outro lugar. Estas denúncias, juntamente com nossas **defesas automatizadas**, nos permitem identificar e prevenir abuso e mau comportamento. Utilize as ferramentas de denúncia de forma responsável e apenas para os fins a que se destinam. Saiba mais sobre como denunciar comportamento inadequado visitando nossa **Central de segurança do LinkedIn**.

A violação dessas políticas pode resultar em ações restritivas. Dependendo da gravidade da violação e do comportamento de um usuário ou do histórico da conta, podemos limitar a visibilidade de determinado conteúdo, remover o conteúdo da nossa plataforma ou até restringir a conta de um usuário em caso de ofensas graves ou repetidas. Se acredita que as medidas tomadas em relação ao seu conteúdo ou sua conta foram um erro, você pode enviar uma **solicitação para recorrer do seu caso**. Obrigado por utilizar o LinkedIn! Juntos, podemos manter nossa plataforma como um local seguro, confiável e profissional, onde podemos criar oportunidades econômicas para todos.

Seja confiável. Nossos usuários devem ser pessoas reais, que fornecem seu nome real e informações precisas sobre si mesmos. Não permitimos **perfis falsos** em nossa plataforma e não aceitamos informações enganosas sobre você, suas qualificações, experiência de trabalho, afiliações ou realizações.

› Saiba mais sobre ser confiável

Seja profissional. Reconhecemos o valor das discussões relativas a atividades profissionais e pedimos aos nossos usuários que se comportem profissionalmente, sem desonestidade ou de forma inapropriada. Ao criar

Página de políticas de comunidade do LinkedIn

Da detecção de conteúdo infringente. A única forma de detecção de conteúdo infringente mencionada no texto da política é a denúncia pelos usuários.

Dos meios de avaliação de conteúdo potencialmente infringente. Não há indicativo de como é feita a análise do conteúdo potencialmente infringente.

Dos critérios de avaliação de conteúdo potencialmente infringente. Não há critérios explícitos para avaliar se o conteúdo é infringente ou não, exceto por uma referência à intenção do usuário na avaliação de publicações contendo conteúdo adulto.

Dos exemplos. As proibições são majoritariamente descritas na forma de regras, com enunciados abstratos, sem exemplos. Entretanto, alguns tópicos incluem exemplos, como no trecho "Não permitimos nenhuma ameaça de violência contra um indivíduo ou um grupo em nossa plataforma. Isso inclui declarações de intenção de matar ou infligir danos físicos graves", ou "você não pode usar os serviços para enviar propostas sexuais não solicitadas, participar ou promover conteúdo sexualmente explícito não consensual (por exemplo, pornografia por vingança),[...]".

Das exceções. Há uma exceção quanto ao teor nos casos de conteúdo adulto: educacional, médico, científico ou artístico e não gratuitamente explícito. Também há previsão de exceção quanto à proibição de contas fictícias, desde que com a permissão prévia da plataforma.

Da especificação não-ambígua de conteúdos infringentes. As citações expõem, em sua maioria, conteúdo infringente. Este se reflete em expressões como "não permitimos", "não aceitamos", "nunca admitimos", "proibimos", "você não pode". Há também citações que se referem a conteúdo encorajado, como em "pedimos aos usuários que se comportem profissionalmente", ou desencorajado, como em "não publique", "não devem ser utilizados".

Das medidas interventivas aplicáveis. As medidas interventivas são aplicáveis dependendo do contexto em que o conteúdo causa violação e também de acordo com o comportamento/histórico do usuário. As intervenções previstas são: 1."limitar a visibilidade de determinado conteúdo"; 2."remover o conteúdo"; ou 3. "restringir a conta de um usuário", esta "em caso de ofensas graves ou repetidas". Não há especificidades sobre quando se caracterizam casos aos quais se aplicam limitação de visibilidade ou remoção.

Da contestação. O mecanismo de contestação é descrito em uma citação, que indica que o usuário pode enviar uma solicitação para recorrer do caso.

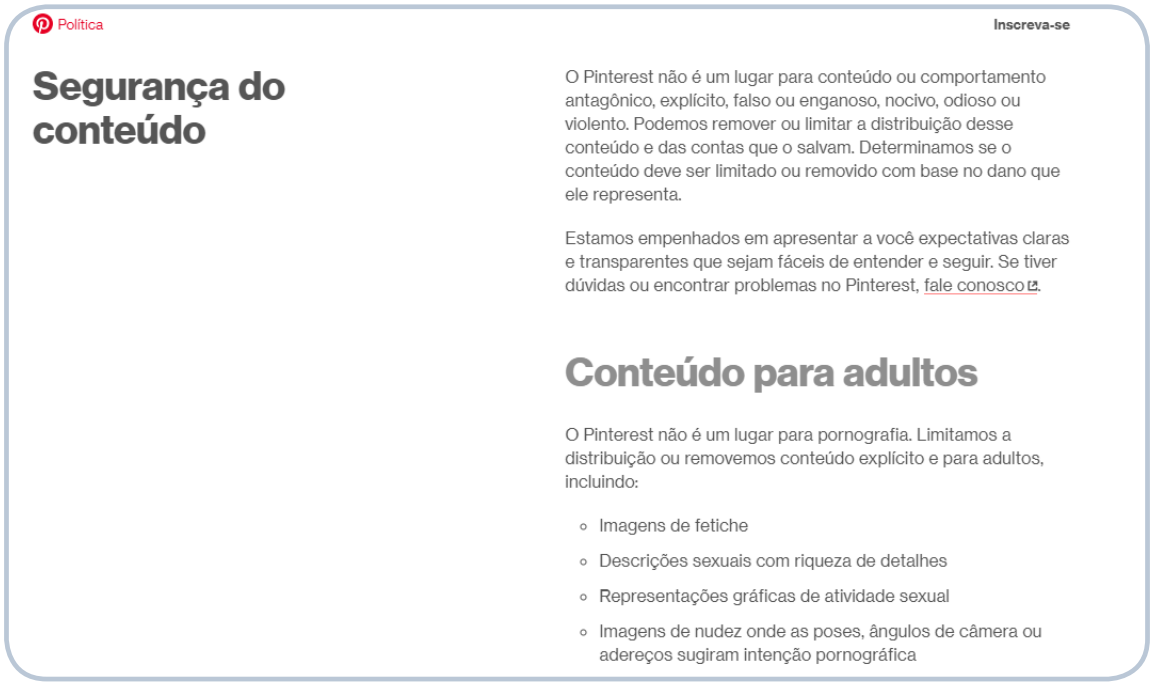
Da notificação. Não há menção a meios de notificação do usuário sobre denúncias ou medidas interventivas.

4.5. Pinterest

Descrição: Trata-se de uma plataforma voltada a organização e compartilhamento de conteúdo no formato visual (fotos, desenhos, gifs e vídeos curtos). Volta-se para arte ou imagens que sirvam de base para atividades criativas, como indica a citação "a missão do Pinterest é trazer inspiração para as pessoas criarem a vida que amam". Não há um recorte de faixa etária ou público-alvo. É possível compartilhar imagens e seguir usuários ou pastas alimentadas com imagens de determinado tipo, bem como "salvar" imagens específicas para sua coleção e comentar em imagens postadas.

Da estrutura geral da política. A política está disposta em uma página, sob o link "comunidade", separada de diversos outros links que aparecem no topo. Entre eles estão os termos privacidade, comerciante, publicidade, desenvolvedores, copyright, marca comercial. A página de políticas de comunidade é composta de uma breve introdução sobre a plataforma, seguida de itens sobre o conteúdo moderado, explicados de maneira sucinta.

Cabeçalho da página de políticas de comunidade do Pinterest


Política

Inscreva-se

Segurança do conteúdo

O Pinterest não é um lugar para conteúdo ou comportamento antagônico, explícito, falso ou enganoso, nocivo, odioso ou violento. Podemos remover ou limitar a distribuição desse conteúdo e das contas que o salvam. Determinamos se o conteúdo deve ser limitado ou removido com base no dano que ele representa.

Estamos empenhados em apresentar a você expectativas claras e transparentes que sejam fáceis de entender e seguir. Se tiver dúvidas ou encontrar problemas no Pinterest, fale conosco.

Conteúdo para adultos

O Pinterest não é um lugar para pornografia. Limitamos a distribuição ou removemos conteúdo explícito e para adultos, incluindo:

- Imagens de fetiche
- Descrições sexuais com riqueza de detalhes
- Representações gráficas de atividade sexual
- Imagens de nudez onde as poses, ângulos de câmera ou adereços sugiram intenção pornográfica

Exemplo de tópico de conteúdo moderado nas políticas

Da detecção de conteúdo infringente. A política menciona apenas a denúncia, indicando categorias existentes para o conteúdo ("Experiências irrelevantes, Informações privadas, Mensagens de corrente não solicitadas, Críticas usando linguagem ofensiva, Comentários e solicitações de natureza sexual, Promoção de produtos e negócios não solicitados ou irrelevantes, Observações odiosas ou constrangedoras como comentários sobre aparência física"). Também fornece o link para uma Central de Ajuda, em que é possível denunciar uma violação de política. No trecho sobre organizações e indivíduos perigosos, informa que a plataforma trabalha "com especialistas em segurança, do setor e do governo para nos ajudar a identificar esses grupos".

Dos meios de avaliação de conteúdo potencialmente infringente. Não há menção à forma como o conteúdo potencialmente infringente é avaliado.

Dos critérios de avaliação de conteúdo potencialmente infringente. Os critérios de avaliação de conteúdo denunciado não são explícitos, há apenas uma lista de proibições ao longo da política. Na parte sobre segurança do conteúdo, há uma menção sutil ao critério usado para remover conteúdo, que é o dano que ele representa, na citação: "determinamos se o conteúdo deve ser limitado ou removido com base no dano que ele representa."

Dos exemplos. Múltiplos exemplos são oferecidos na maior parte das proibições.

Das exceções. A única menção a exceções é no trecho sobre violência, de forma genérica. Afirma-se que "em alguns casos, permitimos que sejam salvas imagens perturbadoras para fins de lembrança e defesa, mas limitamos a distribuição desse

conteúdo em partes públicas da plataforma".

Da especificação não-ambígua de conteúdos infringentes. A distinção não é tão nítida, pois há conteúdos sinalizados como proibidos por meio da expressão "o Pinterest não é um lugar para", seguido de uma lista de exemplos de conteúdo que é removido ou limitado. Por sua vez, os conteúdos desencorajados são no modo de imperativo negativo ("não publique", "não faça", "não use") e direcionados a atitudes de usuários que geram conteúdo indesejado, sem especificação de alguma medida interventiva adotada para aquelas condutas.

Das medidas interventivas aplicáveis. Em diversos tópicos de conteúdos moderados, há indicação de que a distribuição do conteúdo pode ser limitada ou ele pode ser removido. Evidências sugerem que a regra sobre se uma ou outra medida é aplicada é o dano que ele pode causar, mencionada na seção sobre segurança do conteúdo.

Da contestação. Não há indicativo, na página de políticas, de que há meios para contestar uma medida interventiva ou uma denúncia. Entretanto, há um link para a página de Central de Ajuda (fora das políticas), que contém essas informações.

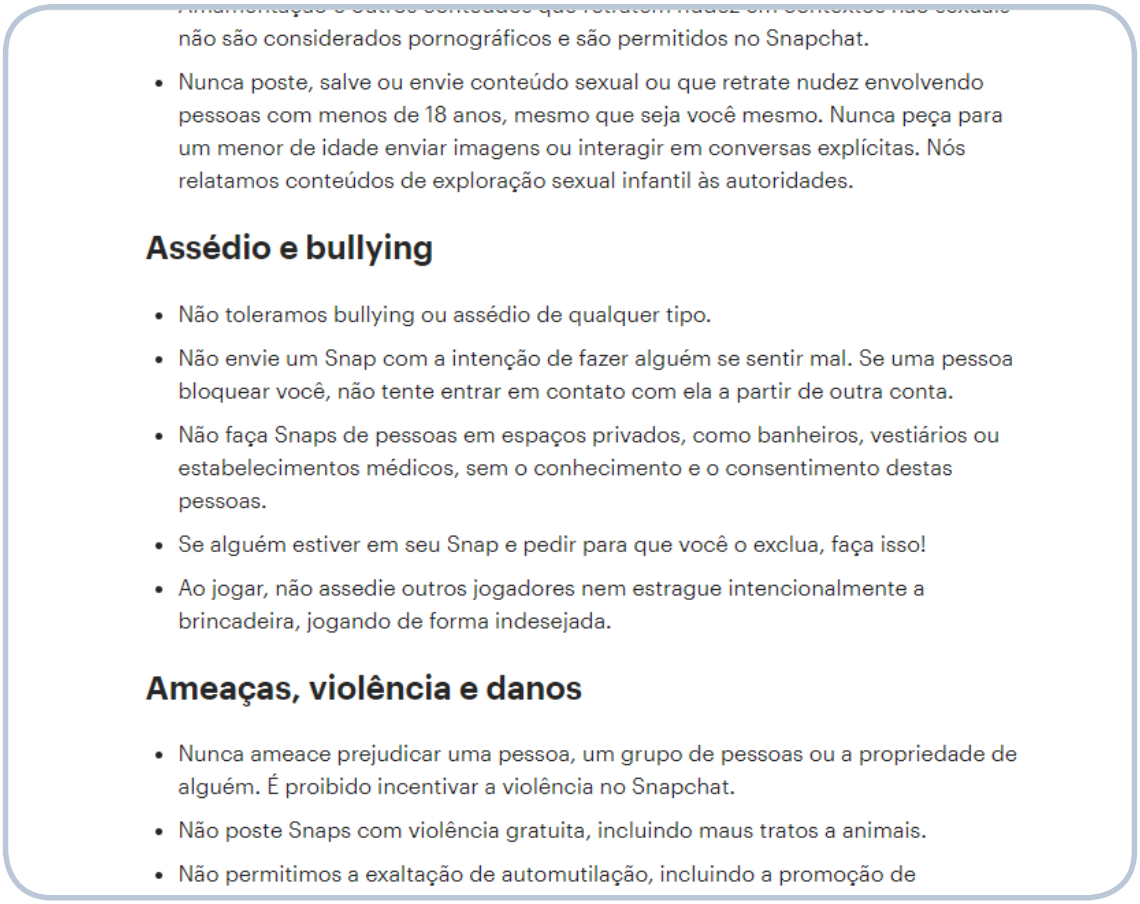
Da notificação. Não há indicativo de que o usuário é notificado quando uma medida interventiva é aplicada ou quando seu conteúdo foi denunciado.

4.6. Snapchat

Descrição: A aplicação é voltada para o compartilhamento de mídias em foto ou vídeo, que podem ser editados e incorporar desenhos ou figuras. Os usuários costumam utilizar os filtros disponíveis na galeria do aplicativo que alteram as imagens, voz, cores e outros elementos de forma customizada. As mídias criadas podem ser compartilhadas com toda a lista de contatos do usuário por 24 horas ou podem ser destinadas apenas a listas de usuários específicos, selecionados de forma privada. Isso permite somente uma exibição e uma repetição e, após a sua visualização, o material se autodestrói. Há também o *chat* escrito e a aba que permite explorar outras contas que estão produzindo conteúdo na rede.

Da estrutura geral. O documento consiste em uma página única. Dois parágrafos introdutórios apresentam fundamentos e escopo de aplicação. As diretrizes se dividem, então, tematicamente: conteúdo sexualmente explícito; assédio e bullying; ameaças, violência e danos; falsificação e spam; discurso de ódio e informações falsas; conteúdo ilegal; e terrorismo. Ao fim, dois parágrafos tratam das denúncias e mais informações e das medidas cabíveis em caso de infração.

Cabeçalho da página de políticas de comunidade do Snapchat



Exemplo de tópico de conteúdo moderado

Da detecção de conteúdo potencialmente infringente. O penúltimo parágrafo da política afirma que a empresa investiga denúncias para determinar se houve violação e se há necessidade de intervenção. Não há mais referências a mecanismos utilizados para a detecção de conteúdo infrator.

Dos meios de avaliação de conteúdo potencialmente infringente. Não há quaisquer referências aos meios empregados na avaliação de conteúdo denunciado. Não é possível determinar se a análise é automatizada ou é conduzida por moderadores humanos.

Dos critérios de avaliação de conteúdo potencialmente infringente. Nenhum critério explícito de avaliação é apresentado ao longo do texto. O último parágrafo inclui a alegação genérica: "tentaremos fazer o que acharmos que melhor reflete estes valores [consistência e justiça] em cada situação, a nosso critério."

Dos exemplos. De quatorze citações proibitivas, três incluem mais de um exemplo: os espaços considerados privados, da exaltação à automutilação e de atividades ilegais. As outras incluem apenas um exemplo ou nenhum.

Das exceções. Amamentação e nudez não sexual são excetuadas da proibição de conteúdo sexualmente explícito. O segundo parágrafo introdutório afirma que em certos casos a plataforma se isentará de tomar medidas sobre conteúdo "com relevância midiática que seja sobre alguma questão política, social ou outra preocupação geral da nossa comunidade", mas não oferece maiores especificações.

Da especificação não-ambígua de conteúdos infringentes. A distinção entre conteúdos meramente desencorajados e aqueles efetivamente proibidos é ambígua em metade das prescrições negativas devido ao recurso frequente ao modo imperativo ("não publique, não poste"). Há sete instâncias de citações com proibição expressa.

Das medidas interventivas aplicáveis. O último parágrafo informa o usuário de que na ocasião de violação das diretrizes, a plataforma poderá "remover o conteúdo ofensivo, excluir sua conta e/ou notificar as autoridades". Não há, todavia, especificação de quais violações estão sujeitas a quais medidas, exceto no caso de conteúdo de exploração sexual infantil, que tem previsão de ser notificado às autoridades. Não há qualquer especificação dos fatores considerados na definição de quais serão as medidas aplicáveis a cada caso.

Da contestação. Não há quaisquer referências a mecanismos de contestação disponíveis ao usuário que sofreu medida interventiva.

Da notificação. Não há quaisquer referências à notificação enviada ao usuário na ocasião de seu conteúdo ou conta ser alvo de medida interventiva.

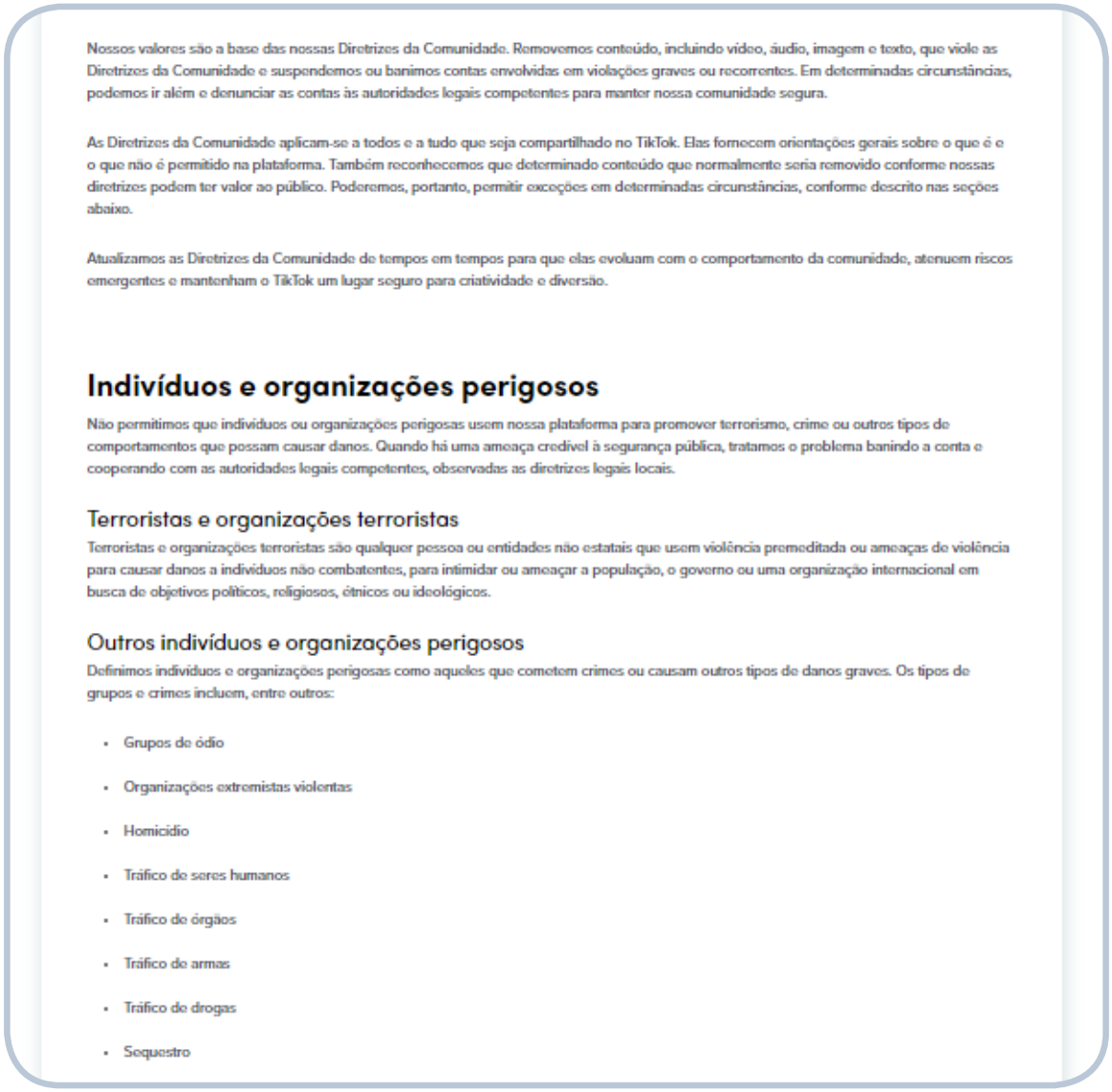
4.7. TikTok

Descrição: É uma rede social de compartilhamento de vídeos de até 15 segundos, frequentemente utilizada pelos usuários para a produção e o compartilhamento de conteúdo de teor artístico ou jocoso. A plataforma oferece uma série de efeitos visuais e sonoros que podem ser incorporados aos vídeos. Os usuários podem seguir uns aos outros, comentar ou reagir aos vídeos alheios e buscar por novos perfis. A página inicial exibe um fluxo contínuo de vídeos dos criadores de conteúdo seguidos pelo usuário.

Da estrutura geral da política. O documento consiste em uma página única. Dois parágrafos introdutórios apresentam fundamentos e escopo de aplicação. As diretrizes se dividem, então, tematicamente: conteúdo sexualmente explícito; assédio e bullying; ameaças, violência e danos; falsificação e spam; discurso de ódio e informações falsas; conteúdo ilegal; e terrorismo. Ao fim, dois parágrafos tratam das denúncias e mais informações e das medidas cabíveis em caso de infração.

Cabeçalho da página de políticas de comunidade do TikTok

Cabeçalho da página de políticas de comunidade do TikTok



Nossos valores são a base das nossas Diretrizes da Comunidade. Removemos conteúdo, incluindo vídeo, áudio, imagem e texto, que viole as Diretrizes da Comunidade e suspendemos ou banimos contas envolvidas em violações graves ou recorrentes. Em determinadas circunstâncias, podemos ir além e denunciar as contas às autoridades legais competentes para manter nossa comunidade segura.

As Diretrizes da Comunidade aplicam-se a todos e a tudo que seja compartilhado no TikTok. Elas fornecem orientações gerais sobre o que é e o que não é permitido na plataforma. Também reconhecemos que determinado conteúdo que normalmente seria removido conforme nossas diretrizes podem ter valor ao público. Poderemos, portanto, permitir exceções em determinadas circunstâncias, conforme descrito nas seções abaixo.

Atualizamos as Diretrizes da Comunidade de tempos em tempos para que elas evoluam com o comportamento da comunidade, atenuem riscos emergentes e mantenham o TikTok um lugar seguro para criatividade e diversão.

Indivíduos e organizações perigosos

Não permitimos que indivíduos ou organizações perigosas usem nossa plataforma para promover terrorismo, crime ou outros tipos de comportamentos que possam causar danos. Quando há uma ameaça credível à segurança pública, tratamos o problema banindo a conta e cooperando com as autoridades legais competentes, observadas as diretrizes legais locais.

Terroristas e organizações terroristas

Terroristas e organizações terroristas são qualquer pessoa ou entidades não estatais que usem violência premeditada ou ameaças de violência para causar danos a indivíduos não combatentes, para intimidar ou ameaçar a população, o governo ou uma organização internacional em busca de objetivos políticos, religiosos, étnicos ou ideológicos.

Outros indivíduos e organizações perigosos

Definimos indivíduos e organizações perigosas como aqueles que cometem crimes ou causam outros tipos de danos graves. Os tipos de grupos e crimes incluem, entre outros:

- Grupos de ódio
- Organizações extremistas violentas
- Homicídio
- Tráfico de seres humanos
- Tráfico de órgãos
- Tráfico de armas
- Tráfico de drogas
- Sequestro

Exemplo de trecho contendo tópico de conteúdo moderado

Da detecção de conteúdo potencialmente infringente. A seção sobre personificação afirma que a plataforma toma medidas após "confirmar os relatos de imitação". O último parágrafo do documento pede que os usuários avisem à plataforma sobre conteúdo potencialmente infrator, a fim de que ela possa "verificar e tomar as medidas cabíveis". Não há mais referências aos meios utilizados para detecção desses conteúdos.

Dos meios de avaliação de conteúdo potencialmente infringente. Além das referências supracitadas sobre os meios de detecção, não há menções adicionais dos mecanismos empregados para a apuração do mérito de denúncias de conteúdo potencialmente infringente.

Dos critérios de avaliação de conteúdo potencialmente infringente. Além das referências supracitadas sobre os meios de detecção, não há menções adicionais aos critérios utilizados na avaliação de conteúdo potencialmente infringente.

Dos exemplos. Múltiplos exemplos de conteúdos encobertos por cada categoria proibida são oferecidos.

Das exceções. Exceções são estabelecidas de forma expressa e específica, sendo as mais frequentes para conteúdo educativo, histórico, satírico, artístico, científico, jornalístico ou com finalidade de conscientização.

Da especificação não-ambígua de conteúdos infringentes. A plataforma define de forma não-ambígua quais conteúdos são proibidos por meio de termos como "não permitimos", "não toleramos" ou "proibimos". Conteúdos desencorajados não são especificados, há somente a determinação de conteúdos proibidos e permitidos.

Das medidas interventivas aplicáveis. De vinte citações referentes a medidas aplicáveis, quatorze estabelecem de forma específica as medidas que a plataforma tomará em caso de violação das normas. As restantes apontam medidas possíveis, porém indeterminam se e em quais circunstâncias a plataforma as tomará ("em certas circunstâncias, podemos ir além e denunciar as contas às autoridades [...]"). Os fatores mais frequentemente citados para a definição da medida são a existência de "risco genuíno de violência e ameaça crível à segurança pública", bem como o histórico do infrator.

Da contestação. Não há quaisquer referências a mecanismos de contestação disponíveis ao usuário que sofreu medida interventiva.

Da notificação. Não há quaisquer referências à notificação enviada ao usuário na ocasião de seu conteúdo ou conta ser alvo de medida interventiva.

4.8. Twitter

Descrição: É uma plataforma voltada primariamente ao compartilhamento de publicações textuais de até 280 caracteres. É possível seguir usuários, reagir, compartilhar ou comentar suas publicações e enviar mensagens de forma privada. Os assuntos mais comentados são exibidos em um menu acessível aos usuários. A plataforma também oferece as possibilidades de publicar conteúdo visual efêmero ou publicar gravações de áudio. O conteúdo dos perfis, páginas e grupos que o usuário segue é exibido em forma de fluxo de publicações em sua página inicial a partir de uma curadoria algorítmica.

Da estrutura geral da política. Uma página central, denominada "As Regras do Twitter", oferece um resumo de cada política temática, do mecanismo de apelação e da publicidade de terceiros em conteúdo de vídeo. A página inclui os endereços das páginas referentes a cada política, à apelação e à abordagem de elaboração e a filosofia de aplicação das medidas corretivas.

Exemplo de tópicos moderados nas políticas do Twitter

Trecho de política de um tópico específico moderado

Da detecção de conteúdo potencialmente infringente. Há referências ao uso de detecção proativa de conteúdo potencialmente infrator para violações às políticas de produtos e serviços legais ou regulamentados, spam e manipulação da plataforma, mídias sintéticas ou manipuladas e para a identificação de comportamento abusivo. A política de falsa identidade inclui a afirmação expressa de que a plataforma não faz detecção proativa. Em todas as políticas, há menção ao uso do mecanismo de denúncia para a identificação.

Dos meios de avaliação de conteúdo potencialmente infringente. Além do mecanismo de denúncia repetidamente indicado, a plataforma afirma recorrer ao contato com o potencial infrator ou com a pessoa afetada pela infração para avaliar casos de violações a políticas específicas, como as de nudez não consensual e de informações privadas. Segundo a plataforma, uma "equipe interfuncional" decide sobre cada caso após avaliação.

Dos critérios de avaliação de conteúdo potencialmente infringente. A plataforma menciona repetidamente alguns critérios utilizados para avaliar os conteúdos potencialmente infringentes, entre eles: a intenção do usuário, a gravidade da violação, a autoria da denúncia, o histórico do usuário e a existência de interesse público no conteúdo.

Dos exemplos. Diversos exemplos e contraexemplos são oferecidos na maioria das políticas, usualmente em seções destinadas especificamente a isso.

Das exceções. Algumas políticas estabelecem exceções para conteúdo educativo, artístico, satírico, religioso, documental, médico, de empoderamento ou de interesse público. A política de glorificação de violência também excetua violência cometida por membros do Estado, desde que o alvo não seja um grupo protegido. Há, ainda, exceções referentes à variação cultural quanto a informações privadas.A política contra terrorismo e extremismo violento inclui "exceções limitadas" para grupos reformados, eleitos ou em processo de resolução pacífica.

Da especificação não-ambígua de conteúdos infringentes. A plataforma define de forma não-ambígua quais conteúdos são proibidos por meio de termos como "não permitimos", "não toleramos" ou "proibimos". Conteúdos desencorajados não são especificados, há somente a determinação de conteúdos proibidos e permitidos.

Das medidas interventivas aplicáveis. Em algumas políticas, a plataforma informa que a repetição da infração levará a sanções mais graves (da suspensão à remoção permanente da conta, por exemplo). Em outros casos, afirma que a gravidade da violação levará a medidas mais contundentes desde a primeira violação. O recurso frequente ao verbo poder ("poderemos remover") amplia a insegurança sobre a resposta suscitada por cada infração.

Da contestação. Quase todas as políticas temáticas incluem indicação de formulário de contestação para usuários que entendam que seu conteúdo foi moderado injustamente.

Da notificação. Somente a política de direitos autorais trata do processo de notificação. Nela, afirma-se que na ocasião da decisão de remoção do conteúdo, o usuário será notificado com informações de contato do denunciante, uma cópia completa da reclamação e como instruções sobre como contranotificá-lo[84].

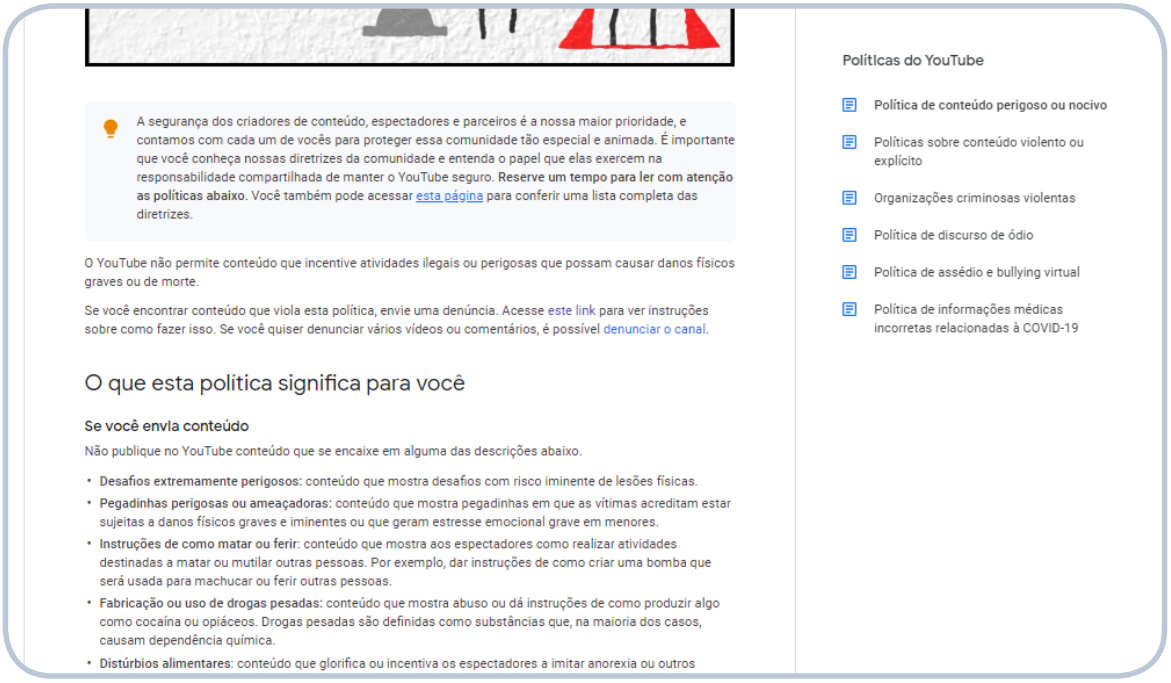
84 Isso é coerente com o Digital Millennium Copyright Act (DMCA), legislação estadunidense de direitos autorais que adota uma política mais restritiva a conteúdo que viola direito autoral, indicando procedimentos para tratar esses casos. Ver: EUA - Estados Unidos da América. **The Digital Millennium Copyright Act**. Dez. 1998. Disponível em: https://www.copyright.gov/legislation/dmca.pdf Acesso em: 08 out. 2020.

4.9. Youtube

Descrição: O Youtube é uma plataforma de compartilhamento de vídeos, que possui dois segmentos: um para vídeos disponibilizados de forma gratuita por usuários e outro de vídeos pagos (YouTube Movies), comercializados pela plataforma mas produzidos por terceiros com interesse comercial direto. É uma plataforma que encampa a ideia de renda por meio de anúncios, que são inseridos em meio aos vídeos disponibilizados. Ao mesmo tempo, permite ao usuário criador de conteúdo receber uma parte dessa renda proporcional ao número de visualizações de seu vídeo. Não possui público-alvo específico, porém em suas políticas é possível identificar diversas medidas voltadas à segurança infantil. Além disso, um segmento da plataforma de vídeos gratuitos é para crianças (Youtube Kids), o que sugere presença significativa de usuários dessa faixa etária na plataforma. No entanto, não é o único segmento de público relevante, pois a plataforma segmenta o conteúdo declaradamente infantil do conteúdo em geral. É possível assinar canais de usuários para receber as atualizações, "curtir" ou "não curtir" vídeos e comentar. Também existe a modalidade de tornar-se membro de um canal, que é uma forma de contribuir financeiramente de maneira periódica e ter acesso a outras vantagens oferecidas pelo usuário criador.

Da estrutura geral da política. O Youtube tem uma central de ajuda com diversas informações, que abarca a política de comunidade. Ela é robusta e separada em tópicos (assuntos objeto de moderação) que são estruturados de forma padronizada entre si. Existem 3 níveis de informação nas políticas: o texto introdutório da página central, os cartões-resumo dos tópicos moderados, que estão nessa página, e as páginas específicas de cada tópico, abertas a partir do link "saiba mais" nesses cartões.

Exemplo dos cartões-resumo na página inicial de políticas do Youtube



Trecho de página específica de tópico sobre conteúdo moderado, com menu de outros tópicos à esquerda

Da detecção de conteúdo infringente. Na introdução da página inicial, é indicada a possibilidade de denunciar conteúdo inadequado, por meio da sinalização para análise da equipe do YouTube. A política indica que a equipe faz um trabalho diuturno de analisar conteúdo sinalizado pelos usuários. Nos resumos nos cartões de tópicos, apenas naquele referente à privacidade há menção à possibilidade de denúncia pelo usuário. O monitoramento proativo da plataforma é mencionado em um dos cartões, que trata da segurança infantil. Ele explica que a plataforma denuncia comportamento irregular às autoridades e trabalha em conjunto com elas. Nas páginas por tópico, o início do texto de cada uma delas indica a possibilidade do usuário denunciar conteúdo irregular/violador. Não há menção a outros meios de detecção de conteúdos inadequados, como filtragem automatizada de publicações. A política apresenta ao usuário os meios de denúncia, com links e instruções sobre como denunciar violações.

Dos meios de avaliação de conteúdo potencialmente infringente. A política menciona uma equipe de análise do conteúdo potencialmente violador/denunciado.

Dos critérios de avaliação de conteúdo potencialmente infringente. A plataforma apresenta critérios explícitos de análise contextual para configuração ou não da violação para todos os conteúdos, exceto direitos autorais. Por exemplo, na política de privacidade, “a pessoa precisa ser claramente identificável no vídeo”, ou na política de assédio e bullying, “quando o criador de conteúdo: incentiva repetidamente sua audiência a ter comportamento abusivo[...]”. Ainda, na de conteúdo explícito ou violento: “se imagens violentas ou sangrentas são o foco do vídeo, como se concentrar somente na parte com violência explícita de um filme ou videogame”.

Dos exemplos. Nas páginas dos tópicos moderados, uma vasta gama de exemplos de conteúdo proibido ou indesejado é explorada, com o aviso de que a lista não é exaustiva.

Das exceções. Há exemplos de exceções, que incluem conteúdo educativo, de protesto, com fins artísticos ou históricos.

Da distinção entre conteúdos infringentes e desencorajados. Conteúdos infringentes de maneira geral são indicados como proibidos nitidamente, exceto por algumas citações em que não fica expressa a proibição. Nelas, é indicado apenas que alguns assuntos "são levados muito a sério", ou, em meio a uma descrição de conteúdo proibido, é usado imperativo negativo "não poste", "não crie", "não estimule". Também há conteúdo desencorajado, que não está em meio às proibições, mas é tratado de maneira ambígua, com termos como "não é legal", "não é apropriado".

Das medidas interventivas aplicáveis. Há critérios contextuais, exemplificados em cada tópico, que indicam quais as medidas aplicadas a cada caso (restrição de idade ou remoção do conteúdo). Por exemplo, a seção sobre segurança infantil especifica que "podemos adicionar uma restrição de idade ao conteúdo que incluir qualquer um dos itens a seguir: Atos nocivos ou perigosos que possam ser imitados por menores: conteúdo que apresenta adultos participando de atividades perigosas que podem ser facilmente imitadas por menores". Na seção sobre falsificação de identidade dispõe que "contas criadas com o intuito de se passar por outro canal ou outra pessoa podem ser removidas de acordo com nossa política de falsificação de identidade".Já na seção sobre assédio e bullying, define que "em alguns casos raros, podemos remover o conteúdo ou aplicar outras penalidades quando o criador de conteúdo: incentiva repetidamente sua audiência a ter comportamento abusivo; faz envio recorrente de conteúdos que insultam e assediam um indivíduo identificável com base nas características intrínsecas dele [...]". Também são indicadas características da conduta do usuário que podem levar ou não à aplicação de medidas interventivas (por exemplo, a possibilidade de encerramento do canal após 3 avisos).

Da contestação. A única seção que explicita meios de recurso frente a medidas interventivas é a de direitos autorais, que tem procedimento específico para notificação e contranotificação[85]. Apenas em seções externas à política de comunidade, que são alcançadas a partir do acesso ao link da ferramenta de denúncia (que abre em uma nova central, com diversos outros links e informações), há instruções sobre contestação de medidas interventivas. Não há esse link para a ferramenta de denúncia nas políticas de comunidade, somente sendo possível acessá-lo na página principal da central de ajuda..

85 Novamente, vale ressaltar que este trecho é influenciado pelo DMCA, já mencionado. Ver: EUA - Estados Unidos da América. **The Digital Millennium Copyright Act**. Dez. 1998. Disponível em: https://www.copyright.gov/legislation/dmca.pdf Acesso em: 08 out. 2020.

Da notificação. Sempre que o conteúdo é removido, há notificação do usuário por e-mail. Não há menção à notificação em caso de restrição de alcance, porém há indicação de que o usuário recebe uma advertência ao violar as políticas.

Quadro 1 - Comparação de relatorias

	FACEBOOK	INSTAGRAM	LINKEDIN	PINTEREST	SNAPCHAT	TIKTOK	TWITTER	YOUTUBE
ESTRUTURA GERAL	Página centralizada com menu de tópicos que carregam na mesma página	Página centralizada com menu de tópicos que carregam na mesma página	Página única centralizada e dividida em seções	Página única centralizada e dividida em seções	Página única centralizada e dividida em seções	Página única centralizada e dividida em seções	Página centralizada com tópicos resumidos e links para páginas específicas	Página centralizada com tópicos resumidos e links para páginas específicas
MEIOS DE DETECÇÃO	Denúncias; Detecção proativa automatizada (exploração sexual de adultos; notícias falsas)	Denúncias	Denúncias	Denúncias; Consultoria externa.	Denúncias	Denúncias	Denúncias; Detecção proativa automatizada (produtos e serviços legais ou regulamentados, spam e manipulação, midias sintéticas e manipuladas, comportamento abusivo).	Denúncias; Monitoramento proativo (segurança infantil).
MEIOS DE AVALIAÇÃO	*	Análise humana	Não há informações	*	*	*	Análise humana; Contato com o potencial infrator ou com a pessoa afetada	Análise humana

*sem informações

	FACEBOOK	INSTAGRAM	LINKEDIN	PINTEREST	SNAPCHAT	TIKTOK	TWITTER	YOUTUBE
CRITÉRIOS DE AVALIAÇÃO	Análise contextual; Evidência de falta de consentimento (exploração sexual de adultos); Indícios de dano ou ameaça à segurança pública; Visibilidade do afetado; Intenção do usuário	Visibilidade do afetado.	Intenção do usuário (conteúdo adulto)	Dano potencial	*	*	Intenção do usuário, gravidade da violação, a autoria da denúncia, o histórico do usuário e a existência de interesse público no conteúdo	A análise contextual; Identificabilidade da pessoa retratada (privacidade); Foco do vídeo (conteúdo explícito ou violento)
EXEMPLOS	Múltiplos exemplos ao longo da maioria das proibições	Apenas um ou nenhum exemplo na maioria das proibições	Apenas um ou nenhum exemplo na maioria das proibições	Múltiplos exemplos ao longo da maioria das proibições	Apenas um ou nenhum exemplo na maioria das proibições	Múltiplos exemplos ao longo da maioria das proibições	Múltiplos exemplos ao longo da maioria das proibições	Múltiplos exemplos ao longo da maioria das proibições
EXCEÇÕES	Finalidade educativa, artística, satírica, religiosa, documental, médica, de empoderamento ou de interesse público. Outras exceções específicas	Finalidade condenatória ou de conscientização (discurso de ódio e violencia explícita); Imagens referentes a mastectomia, amamentação, pinturas e esculturas (nudez); Outras específicas.	Finalidade educacional, médica, científica ou artística (conteúdo adulto); Contas fictícias são permitidas se autorizadas previamente pela plataforma.	Finalidades de lembrança e defesa (imagens perturbadoras)	Relevância midiática e tema político ou social.	Finalidade educativa, histórica, satírica, artística, científica, jornalística ou de conscientização	Finalidade educativa, artística, satírica, religiosa, documental, médica, de empoderamento ou de interesse público. Outras exceções especificadas.	Finalidade educativa, de protesto, artística ou histórica

*sem informações

	FACEBOOK	INSTAGRAM	LINKEDIN	PINTEREST	SNAPCHAT	TIKTOK	TWITTER	YOUTUBE
ESPECIFICAÇÃO DE CONTEÚDOS PROIBIDOS	Prejudicada pelo recurso frequente ao modo imperativo ou a termos ambíguos	Predominantemente determinada, porém prejudicada pelo recurso frequente ao modo imperativo.	Predominantemente determinada, porém prejudicada pelo uso de termos ambíguos	Prejudicada pelo recurso frequente ao modo imperativo ou a termos ambíguos	Prejudicada pelo recurso frequente ao imperativo	Completamente determinada	Completamente determinada	Predominantemente determinada, porém prejudicada pelo recurso frequente ao modo imperativo ou a termos ambíguos
MEDIDAS INTERVENTIVAS	Determinadas somente nos casos de restrição etária e de sinalização. Critérios de gravidade da infração e histórico do usuário são considerados.	Determinação é prejudicada pelo uso do verbo poder e pela apresentação de mais de uma medida cabível sem indicação dos critérios de avaliação.	Indicação frequente de múltiplas medidas cabíveis. Critérios de contexto da violação e histórico do usuário são considerados.	Indicação frequente de múltiplas medidas cabíveis. Critério de dano é considerado.	Determinação é prejudicada pela apresentação de mais de uma medida cabível sem indicação dos critérios de avaliação.	Determinação é prejudicada pelo uso do verbo poder. Critérios de histórico do usuário e existência de risco de violência são considerados.	Determinação é prejudicada pelo uso do verbo poder. Critérios de histórico do usuário e gravidade da infração são considerados.	Indicação frequente de múltiplas medidas cabíveis. Critério de histórico do usuário é considerado. Outros critérios são indicados para infrações específicas.
CONTESTAÇÃO	*	*	Menção sem detalhamento.	*	*	*	Menção sem detalhamento.	Restrita a direitos autorais
NOTIFICAÇÃO	*	*	*	*	*	*	Restrita a direitos autorais.	Menção de notificação por e-mail no caso de remoção. Indicação de que o usuário recebe advertência ao cometer uma infração.

*sem informações

Fonte: autoria própria

Vale ressaltar que as relatorias se ativeram à amostra e aos documentos coletados e selecionados como foco da pesquisa, ou seja, os textos das políticas de comunidade. Isso significa que podem haver outros problemas de transparência não passíveis de identificação pela análise desse material, e que portanto não constam no quadro-síntese. A partir das relatorias elaboradas, foi possível traçar um panorama de lacunas a serem superadas em cada um dos critérios de transparência nas políticas de comunidade. Ele é elaborado na discussão, a seguir.

5. Discussão dos critérios de transparência analisados

As relatorias demonstram que a transparência é incorporada nas políticas de comunidade de maneira incompleta e desigual entre as políticas e mesmo em relação a diferentes assuntos moderados ou seções. A seguir, são discutidos alguns riscos identificados e possíveis pontos a serem aprimorados para que esses canais exerçam o papel de informar e possibilitar ao público que sejam avaliadas e discutidas as práticas de moderação de conteúdo adotadas, desde a detecção até a adoção de medidas interventivas.

5.1. Da detecção de conteúdo potencialmente infringente

Dado que o volume total de conteúdos publicados nas grandes plataformas provavelmente é imensamente superior ao volume de conteúdos moderados, a detecção se configura como etapa inicial da moderação. Ela é crucial na medida em que decisões relativas ao uso ou desuso de mecanismos específicos podem implicar na inclusão ou exclusão indevidas de conteúdos específicos no universo de publicações a serem avaliadas como potencialmente infringentes pela plataforma.

A possibilidade de uso de sistemas automatizados de detecção de conteúdo potencialmente infringente chama especial atenção e levanta uma série de preocupações[86]. Entre elas, as relativas à autodeterminação do usuário sobre seus dados pessoais, bem como ao risco desses dados serem utilizados para embasar uma decisão automatizada de teor discriminatório. Por exemplo, nuances culturais ou linguísticas na língua inglesa poderiam ser captadas mais facilmente por um algoritmo que aquelas em outros idiomas, o que poderia resultar em mais

86 ONU - Organização das Nações Unidas. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression**. Symbol A/HRC/38/35. 6 abr. 2018. Disponível em: https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35 Acesso em: 30 jun. 2020. p. 12, 18

remoções nesses idiomas devido a uma análise contextual de menor qualidade.

Um estudo realizado pelo centro de pesquisa InternetLab[87] ilustra os riscos de discriminação associados a esse tipo de sistema. O trabalho analisou o Perspective, desenvolvido pelo grupo Alphabet (conglomerado proprietário de Google e Youtube), um sistema de inteligência artificial que se propõe a identificar o nível de "toxicidade"[88] de conteúdo textual. A pesquisa comparou as avaliações feitas pelo Perspective sobre os perfis de 80 *drag queens* com aquelas feitas sobre amostras de controle que incluíam desde perfis com conteúdo considerado saudável até outros, pertencentes a supremacistas brancos e a Donald Trump, considerados controversos ou "muito tóxicos". Seus resultados indicaram que "um número significativo de perfis das drag queens no Twitter foram considerados como potencialmente mais tóxicos que o perfil de Donald Trump e de supremacistas brancos". Além disso, palavras como *gay*, *lesbian*, *queer* e *transvestite* foram consideradas significativamente tóxicas. Mais elevados ainda foram os níveis de toxicidade atribuídos a certos termos que, ainda que frequentemente sejam utilizados de forma ofensiva, são comumente reivindicados positivamente pelas comunidades afetadas, como *fag* e *sissy*.

Assim, dadas as problemáticas associadas a diferentes meios de detecção utilizados, especialmente no contexto de decisões automatizadas, a transparência é relevante tanto para que cada usuário individual tenha condições plenas de recorrer em caso de intervenções indevidas quanto para que os diversos setores afetados possam exercer controle social sobre as plataformas.

A leitura dos padrões de comunidade sugere que o principal método declaradamente empregado pelas plataformas permanece sendo a denúncia individual - meio citado em todos os casos analisados. Além dela, há referências ao monitoramento proativo para violações específicas nos casos do Facebook, do Twitter e do Youtube. As duas primeiras caracterizam esse monitoramento expressamente como automatizado. Ainda, o Youtube e o Pinterest afirmam recorrer ao contato com especialistas e autoridades externas, embora não especifiquem precisamente como isso se dá.

Quando as plataformas de fato utilizam exclusivamente a denúncia como mecanismo de detecção, é positivo que elas informem aos usuários a esse respeito. Na ocasião do emprego de meios adicionais, a omissão em relação ao teor desses meios resulta em transparência incompleta.Isso pode gerar, inclusive, incompreensão dos usuários, cuja percepção pode ser de que estão sujeitos apenas aos meios citados pela plataforma. Se a detecção é realizada proativamente, cabe

87 GOMES, A.; ANTONIALLI, D.; OLIVA, T. Drag queens e Inteligência Artificial: computadores devem decidir o que é 'tóxico' na internet? **InternetLab**, São Paulo, 28 mai. 2019. Disponível em: https://www.internetlab.org.br/pt/liberdade-de-expressao/drag-queens-e-inteligencia-artificial-computadores-devem-decidir-o-que-e-toxico-na-internet/. Acesso em: 08/10/2020.

88 O sistema define como tóxico um comentário que é "rude, desrespeitoso ou não-razoável que provavelmente fará com que você abandone uma discussão".

à plataforma explicitar se ela é feita de forma automatizada ou, em caso negativo, de que outra forma.

5.2. Dos meios de avaliação de conteúdo potencialmente infringente

Uma vez realizada a detecção de conteúdo potencialmente passível de moderação, cabe à plataforma avaliar se aquele conteúdo efetivamente é passível de intervenção conforme suas políticas, seja por violar alguma norma ou por se enquadrar em algum caso excepcional previsto que justifique a intervenção (conteúdo sujeito a restrições etárias, por exemplo, que não necessariamente é proibido). A transparência sobre os meios de detecção empregados é relevante pois abre a possibilidade de discutir as insuficiências e consequências de cada meio. O uso de algoritmos pode ser criticado pela sua imprecisão, que resulta em remoção de conteúdo legítimo e potencialização de vieses e discriminações. Já a contratação de moderadores humanos em geral tende a tornar a moderação mais precisa e potencialmente contextualizada[89] - ainda que com problemas nesse campo, dependendo da diversidade e representatividade do grupo de moderação -, mas implica em ser transparente também sobre como é o acompanhamento dos reflexos dessa atividade sobre a saúde e as condições de trabalho dessas pessoas. Além das implicações relativas a controle social e devido processo, esse aspecto levanta questões específicas quanto às condições de trabalho dos moderadores humanos, incluindo assédio moral no ambiente de trabalho, ausência de treinamento adequado e exposição a conteúdo danoso a sua saúde mental sem que a plataforma forneça as proteções adequadas[90].

Nos casos analisados, predomina a opacidade quanto aos meios de avaliação empregados: as diretrizes de cinco empresas (Facebook, LinkedIn, Pinterest, Snapchat e TikTok) não incluem menções aos meios pelos quais cada caso é examinado. Três delas (Instagram, Twitter e Youtube) afirmam lançar mão de análise humana para a realização dessas avaliações. No caso do Twitter, recorre-se também ao contato com o infrator potencial ou com a possível vítima da violação no contexto de políticas específicas. Essa é uma prática positiva na medida em que muitas infrações potenciais exigem averiguação contextual e a possibilidade de que o usuário forneça mais informações sobre a publicação é consistente com

89 Como relata esta coluna jornalística, o algoritmo do YouTube removeu o dobro de conteúdo legítimo após ser adotado massivamente no período de pandemia. Ver: NEWTON, Casey. YouTube gets sued by its moderators. **The Interface**. 22 set. 2020, n. 572 Disponível em: https://www.getrevue.co/profile/caseynewton/issues/youtube-gets-sued-by-its-moderators-280023?utm_campaign=Issue&utm_content=view_in_browser&utm_medium=email&utm_source=The+Interface Acesso em: 08 out. 2020.

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

as expectativas de devido processo. Para poder discutir e exigir a reconsideração de uma medida interventiva aplicada sobre conteúdo, o usuário precisa saber de que forma o material foi avaliado e que tipo de informação complementar pode servir para esse propósito. Também a forma de proceder quanto à identificação de conteúdo realmente danoso pode ser revista, o que torna a política aplicada mais eficaz e legítima, limitada ao que é necessário.

5.3. Dos critérios de avaliação de conteúdo potencialmente infringente

A categoria diz respeito tanto aos critérios empregados pela plataforma com a finalidade de determinar se um conteúdo efetivamente se enquadra em alguma hipótese de intervenção prevista nas políticas quanto de definir qual será a medida. A comunicação desses critérios de maneira nítida e não-ambígua é crucial para a transparência quanto aos valores e princípios que influenciam as decisões relativas a moderar ou não, e sobre *como* moderar. Similarmente ao que se passa com o critério anterior, a opacidade quanto a esses critérios também se torna uma questão mais sensível no contexto do uso de sistemas automatizados de avaliação do conteúdo devido aos riscos de discriminação algorítmica.

Dos oito casos analisados, três (Facebook, Twitter e Youtube) descrevem os múltiplos critérios empregados em sua análise: análise do contexto da publicação, gravidade da violação, histórico do usuário, indícios de dano ou ameaça à segurança pública, intenção do usuário, visibilidade da pessoa afetada. Também há critérios mais específicos para normas determinadas - o Facebook considera “evidência de falta de consentimento” em sua política da exploração sexual de adultos, o Youtube leva em conta a identificabilidade da pessoa retratada ao considerar violações a sua política de privacidade e considera o foco do vídeo ao aplicar suas normas referentes a conteúdo violento.

Nos outros cinco casos, o quadro geral é predominantemente opaco - três plataformas referenciam um único critério de avaliação: dano representado pelo conteúdo (Pinterest), visibilidade do afetado (Instagram) e intenção do usuário (Linkedin - proibição de conteúdo adulto). Quanto ao Snapchat e TikTok, as referências são simplesmente ausentes ou consistem em apelos genéricos à discricionariedade da plataforma. Também é digno de nota o uso da expressão “poderemos permitir” no caso da exceção estabelecida para o Instagram para discursos de ódio, o que introduz indeterminação.

Percebe-se, no conjunto de políticas analisado, ausência de parâmetros concretos ou mensuráveis, que delimitem ou permitam projetar expectativas comuns sobre

que tipo de conteúdo é danoso. Isso pode reafirmar a ideia de Roberts[91], que enxerga as políticas de comunidade como uma resposta quanto a demandas sobre controle de conteúdo danoso que não se compromete com uma posição concreta sobre os limites à expressão decorrentes das soluções adotadas. A ausência de critérios explícitos de avaliação representa riscos. Um deles é o da subjetividade dos indivíduos envolvidos na avaliação dos casos ou do viés algorítmico, quando isso é feito por máquinas. Isso pode desencadear discriminação ou percepções distintas sobre o que consiste em uma violação. Outro risco, ainda, é o do tratamento desigual entre usuários que apresentem a mesma violação, mas não sejam submetidos à mesma medida interventiva. A opacidade também impossibilita o debate sobre os limites à liberdade de expressão e revisão dos repertórios de conteúdo considerado danoso. É por meio desse tipo de transparência procedimental que se podem identificar os tipos de conteúdo considerados infringentes nos casos concretos e, se necessário, rever o escopo das categorias.

5.4. Dos exemplos

Dada a amplitude da maior parte das normas encontradas nos padrões, os exemplos podem facilitar a compreensão dos usuários sobre casos concretos que podem ou não ser englobados pelas proibições - embora não afastem a necessidade de critérios de análise nítidos sobre o conteúdo infringente.. Nos casos analisados, seis das plataformas oferecem exemplos múltiplos e específicos na maior parte de suas publicações: Facebook, Pinterest, TikTok, Twitter e Youtube. O YouTube apresenta listas amplamente exemplificativas de condutas, que permitem ter uma imagem de casos específicos considerados violações. Isso ajuda a entender como as proibições e exceções são interpretadas. O Twitter apresenta, ainda, diversos contraexemplos, casos de condutas às quais às regras não se aplicam, o que pode contribuir para a compreensão do usuário. Nos casos do Snapchat, do Linkedin e do Instagram, a maior parte das proibições cita um único exemplo ou nenhum e se limita à enunciação da regra abstrata.

5.5. Das exceções

É comum que as plataformas excepcionem certos conteúdos de suas políticas. Uma publicação que poderia ser ofensiva ou inadequada em circunstâncias convencionais pode não ser em contexto educativo, artístico ou jornalístico, por exemplo. Na maior parte das vezes, não é razoável esperar que o usuário saiba previamente quais são as exceções admitidas pela plataforma, mesmo porque elas variam entre uma e outra. Desse modo, a explicitação das exceções admitidas para as normas proibitivas das plataformas é essencial para a compreensão adequada

91 ROBERTS, Sarah T. Digital detritus:'Error' and the logic of opacity in social media content moderation. **First Monday**, 2018. p. 4

dessas norma e capacidade contestatória do usuário, e para o controle social. A delimitação precisa das exceções também é importante para determinar o alcance das proibições impostas. Assim, a indefinição das exceções também é uma forma de opacidade quanto às proibições.

Todas as plataformas examinadas estabelecem exceções, em sua maioria com referência ao teor ou à finalidade do conteúdo - se é condenatório, educativo, satírico, religioso, documental, médico, histórico, artístico, científico, jornalístico ou de manifestações políticas. Também são mobilizadas exceções relacionadas à relevância para conteúdos "de interesse público" (Twitter) ou "com relevância midiática" (Pinterest). Ainda, há exceções eventuais quanto à autoria da publicação ou permissão prévia da plataforma (Facebook e Instagram - proibições de venda de produtos controlados, Linkedin - proibição de contas fictícias) e outras, de natureza mais variável e limitada.

É desejável que as plataformas maximizem a transparência nas exceções de suas normas ao estabelecer, de forma determinada e precisa, as condições (finalísticas ou não) pelas quais os conteúdos são excetuados. Ainda, é importante que eliminem hipóteses baseadas na discricionariedade da plataforma. Nesse sentido, o Pinterest, por exemplo, abre exceções para conteúdos que reflitam alguma "preocupação geral da nossa comunidade" sem delimitar quais são essas preocupações. Também seria positivo especificar ou estabelecer diretrizes para a determinação o que significam categorias tão plásticas como conteúdo de "interesse público" e conteúdo "com relevância midiática". Essa parametrização não deve ser realizada apenas de acordo com critérios da plataforma, mas tomando por base os padrões internacionais de direitos humanos. A apresentação de exceções com ausência de demarcação, por uso de termos amplos ou abstratos, pode gerar opacidade e incerteza sobre quais são as regras e limites impostos nas políticas.

5.6. Da especificação não-ambígua de conteúdos infringentes

A moderação de conteúdo não se reduz às intervenções de remoção de postagens e contas. As remoções são, no entanto, o principal meio pelo qual as plataformas respondem à publicação de conteúdo não-autorizado. Por isso, ganham centralidade no debate por seu impacto patente sobre a liberdade de expressão. A delimitação nítida de quais conteúdos são proibidos nas plataformas é necessária tanto para que os usuários individuais compreendam as normas às quais estão sujeitos - para que possam contestar remoções indevidas - quanto para que a sociedade exerça controle sobre essas normas.

Em seis das plataformas analisadas (Facebook, Instagram, LinkedIn, Pinterest,

Snapchat e Youtube), essa delimitação é prejudicada, em diferentes graus, por algum fator introdutor de ambiguidade. No contexto das normas proibitivas, a introdução de ambiguidade ocorre quando a redação de uma norma não explicita se um determinado conteúdo é efetivamente proibido (e) ou meramente desencorajado em função dos valores da plataforma. O primeiro caso atrairia a aplicação de uma intervenção repressiva por parte da plataforma. No segundo, por sua vez, o usuário pode legitimamente discordar dos valores da plataforma e publicar seu conteúdo mesmo assim, desde que não infrinja as normas .

Um dos fatores de ambiguidade é o emprego do modo verbal imperativo na redação das políticas, como “não publique” ou “não poste”.Isso dificulta ao usuário determinar se o conteúdo discutido no trecho é efetivamente proibido ou meramente desencorajado.O mesmo problema ocorre quando termos inespecíficos denotam reprovação a um certo tipo de conteúdo. Alguns exemplos são citações como “O Pinterest não é um lugar para [...]” (Pinterest) e “Não é apropriado [publicar um certo conteúdo]” (Youtube). A leitura dessas citações não informa conclusivamente sobre o conteúdo ser autorizado ou não. Somente duas plataformas (TikTok e Twitter) apresentam a redação inteiramente não-ambígua quanto a suas proibições, com destaque para o recurso a expressões como “não é permitido”, “não é autorizado”, “é proibido” ou “não é tolerado”.

Os trechos apontados ilustram outro meio das plataformas demonstrarem publicamente sua preocupação com o conteúdo. Dessa forma, atendem a pressões sobre direitos e evitam medidas regulatórias, mas sem admitir que há necessariamente intervenção sobre o conteúdo. Ademais, reforçam uma imagem de liberdade de expressão e autonomia do usuário. As políticas, ao utilizarem termos não expressamente proibitivos e enunciarem conteúdos reprovados sem indicar a medida interventiva, apresentam um posicionamento sobre o que é ou não aceitável, sem se comprometer com uma intervenção. A opacidade é gerada na medida em que não se sabe ao certo se - e em quais casos - ocorre moderação.

5.7. Das medidas interventivas aplicáveis

A moderação de conteúdo em plataformas abrange diversas medidas, que vão além da remoção e são adotadas com diversas finalidades. A restrição etária do acesso a um conteúdo e a sinalização de conteúdo sensível, por exemplo, não buscam eliminar conteúdo danoso da plataforma. Na verdade, buscam somente evitar ou prevenir a exposição de certos usuários a conteúdos com os quais não seria adequado que tivessem contato. Esse é o caso, por exemplo, de menores de idade ou usuários que não podem ou não querem entrar em contato com alguns tipos de conteúdo, seja por situações pessoais ou pelo ambiente a partir do qual acessam. Isso não implica que o conteúdo em geral seja danoso.

Além disso, muitas vezes as plataformas lidam com conteúdo danoso por meio de intervenções menos visíveis, mas que ainda assim podem ser restritivas da liberdade de expressão. É o caso, por exemplo, da prática conhecida como *shadow banning*[92], por meio da qual a plataforma reduz o alcance de certos conteúdos sem que os usuários que o publicam tenham ciência da redução. Intervenções de menor visibilidade para o usuário afetado podem, inclusive, ser mais gravosas, na medida em que a possibilidade de recurso é inexistente se o usuário afetado não tem ciência de que está sendo alvo de uma intervenção.

Por esses motivos, a comunicação nítida, por parte das plataformas, em relação a quais medidas são aplicáveis a quais conteúdos publicados é um dos aspectos mais fundamentais da transparência na moderação de conteúdo. Como ideal, o usuário sempre deveria saber a quais medidas está potencialmente sujeito ao publicar um conteúdo específico.Isso, no entanto, nem sempre é possível porque muitas vezes mais de um critério fatora na determinação da medida aplicável. Por exemplo, uma plataforma pode optar por remover a publicação ou suspender a conta tanto em função do teor da publicação quanto do histórico do usuário. Assim, o mínimo necessário é que as plataformas explicitem qual o universo de medidas cabíveis em cada cenário e quais critérios serão considerados na tomada de decisão.

Nos casos analisados, todas as plataformas apresentam algum grau de ambiguidade em relação às medidas aplicáveis a conteúdos passíveis de moderação. Como no critério anterior, essa ambiguidade frequentemente resulta dos usos do modo imperativo (“não publique”, “não poste”) e do verbo poder (“em caso x, a publicação poderá ser removida”). Também se apresenta quando são apontadas alternativas de intervenção cabíveis sem a indicação correlata dos critérios que determinarão qual será a medida aplicada. Cinco das plataformas (Facebook, Pinterest, TikTok, Twitter e Youtube) indicam critérios utilizados para determinar entre uma medida e outra quando mais de uma possibilidade é aplicável. Nesses casos, os critérios são o histórico do usuário, a gravidade da infração e o dano potencial.

5.8. Da contestação

A produção de *accountability* a partir da transparência exige que as informações disponibilizadas pelas plataformas possibilitem tanto a reparação individual em casos de intervenções indevidas quanto o exercício de controle social. Para que a reparação individual seja uma possibilidade efetiva, cabe às plataformas disponibilizar mecanismos pelos quais os usuários que sofreram a aplicação de uma medida interventiva possam contestá-la de maneira significativa. Conforme os Princípios de Santa Clara sobre Transparência e Accountability em Moderação

92 WEST, Sarah Myers. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. **New Media & Society**, v. 20, n. 11, p. 4366-4383, 2018. p. 4374

de Conteúdo[93], um recurso significativo exige, mas não se limita a: 1. revisão humana por um ou mais indivíduos que não estiveram envolvidas na decisão inicial; 2. oportunidade de o usuário apresentar informações adicionais a serem consideradas na revisão; 3. notificação dos resultados da revisão; e 4. uma declaração de motivos suficiente para que o usuário compreenda a decisão.

A análise comparativa sobre os mecanismos presentemente empregados pelas plataformas exigiria outro estudo, que os examinasse de forma detida. Ainda assim, os padrões de comunidade podem desempenhar um papel importante ao dar ciência ao usuário sobre a existência de tais mecanismos e explicar seu funcionamento. Nos casos analisados, a maioria das plataformas (Facebook, Instagram, Pinterest, Snapchat e TikTok) não inclui qualquer referência a mecanismos dessa natureza. O LinkedIn a menciona na introdução de suas políticas e o Youtube cita a possibilidade, porém, a restringe a violações relacionadas a direitos autorais. No caso do Twitter, a menção à possibilidade de contestação é incluída ao fim de quase todas as políticas setoriais.

Seria desejável que todas as plataformas referenciassem de forma expressa o mecanismo recursal do qual o usuário pode dispor em caso de moderação indevida. Além disso, tal mecanismo deve ser indicado ao usuário sempre que este sofrer alguma medida de moderação, não apenas nas políticas de comunidade. Esse mecanismo não deve se restringir a direitos autorais, mas ser estendido a qualquer tipo de intervenção. O enfoque da padronização de mecanismos de contestação somente em políticas de direito autoral demonstra a tendência das políticas serem orientadas também por interesses jurídicos e econômicos. Conforme apontam Belli e Venturini[94], esses mecanismos não dizem respeito apenas a políticas de transparência, mas ao cumprimento estrito do DMCA, lei estadunidense que impõe deveres e responsabilidades sobre intermediários em relação a propriedade intelectual. Se o usuário não tiver conhecimento da existência desse mecanismo, sua capacidade contestatória fica evidentemente prejudicada desde o início.

5.9. Da notificação

Uma vez que a plataforma aplica uma medida interventiva sobre um conteúdo, a primeira etapa dessa aplicação deveria consistir na notificação do usuário de que o conteúdo em questão sofrerá a medida. Assim como a contestação, a análise sobre os mecanismos de notificação em si mesmos exigiria outro estudo, que abordasse o tópico de forma específica. No entanto, é possível que os padrões de

93 EFF et al. **Santa Clara Principles on transparency and accountability in content moderation.** Disponível em: https://santaclaraprinciples.org/ Acesso em: 21 jul. 2020.

94 BELLI, Luca; VENTURINI, Jamila. Private ordering and the rise of terms of service as cyber-regulation. **Internet Policy Review**. v. 5, n. 4. 2016. p. 9. Disponível em: https://www.econstor.eu/bitstream/10419/214032/1/IntPolRev-2016-4-441.pdf Acesso em:08 out. 2020.

comunidade informem ao usuário sobre seu direito a ser notificado, bem como sobre o conteúdo e o formato esperados na notificação. Dessa forma, na ocorrência de uma inconsistência entre as previsões dos padrões de comunidade a esse respeito e a notificação em si, o usuário pode demandar a reparação em caso de intervenção indevida. Conforme os Princípios de Santa Clara[95], a notificação deve incluir, no mínimo, as seguintes informações: 1. URL da publicação; 2. trecho do conteúdo que causou a intervenção (ou dados adicionais que possibilitem sua identificação); 3. cláusula violada dos padrões de comunidade; 4. meios de detecção; e 5. intervenção sobre o conteúdo. Ainda, deve ser fornecida em formato duradouro e permanecer disponível mesmo que a conta do usuário seja suspensa ou indisponibilizada.

Nos casos analisados, seis das plataformas (Facebook, Instagram, LinkedIn, Pinterest, Snapchat, TikTok) não incluem qualquer referência ao mecanismo de notificação. O Twitter limita a menção à política de direito autoral, que detalha os mecanismos de notificação e contranotificação em caso de violação aparente. O Youtube menciona que o usuário recebe um e-mail no caso de uma remoção e indica que ele é advertido ao cometer uma infração.

Trata-se de um cenário indicativo de uma zona de opacidade concernente às informações sobre notificações. A transparência é prejudicada pela ausência quase total de informações sobre notificações ao usuário. A ausência desse tópico nas políticas sinaliza que não há compromisso da plataforma com essa prática. A ausência desse tópico nas políticas sinaliza que não há compromisso da plataforma com essa prática. Sem um conjunto de previsões sobre como a notificação deveria se dar, a capacidade contestatória do usuário no caso de um processo de moderação acompanhado de uma notificação insuficiente é severamente prejudicada.

95 EFF et al. **Santa Clara Principles on transparency and accountability in content moderation.** Disponível em: https://santaclaraprinciples.org/ Acesso em: 21 jul. 2020.

Quadro 2 - Principais riscos, problemas e déficits de transparência

	PLATAFORMAS QUE APRESENTARAM ALGUMA INADEQUAÇÃO	INADEQUAÇÃO	RISCOS E ROBLEMAS	RECOMENDAÇÕES DE TRANSPARÊNCIA
MEIOS DE DETECÇÃO	Youtube	Indicação de detecção proativa sem referência ao teor (automatizado ou não);	Redução da capacidade contestatória do usuário	Citar expressamente quaisquer meios empregados e especificar a quais casos se aplicam; Especificar sempre que sistemas automatizados forem empregados.
MEIOS DE AVALIAÇÃO	Facebook; Linkedin; Pinterest; Snapchat; Tiktok	Omissão dos meios	Injustiça algorítmica; Redução da capacidade contestatória do usuário; Opacidade quanto às condições trabalhistas de moderadores humanos	Citar expressamente quaisquer meios empregados e especificar a quais casos se aplicam; Especificar sempre que sistemas automatizados forem empregados.
CRITÉRIOS DE AVALIAÇÃO	Instagram; Pinterest; Snapchat; TikTok	Omissão dos critérios; Referência à discricionariedade da plataforma; Uso do verbo "poder".	Injustiça algorítmica; Redução da capacidade contestatória do usuário;	Citar expressamente quaisquer critérios empregados ou para treinamento de moderadores; Especificar sempre que sistemas automatizados forem empregados; Eliminar o uso do verbo poder.
EXEMPLOS	Instagram; LinkedIn; Snapchat	Ausência de exemplos; Apresentação de um único exemplo	Indefinição quanto à norma	Apresentar múltiplos exemplos de conteúdo infringente de cada norma proibitiva; Apresentar contraexemplos.
EXCEÇÕES	Twitter, Pinterest	Uso de categorias finalísticas excessivamente amplas; Referência à discricionariedade da plataforma;	Indefinição quanto à norma	Citar especificamente as finalidades particulares ou qualquer outro critério empregado para definir conteúdos exceptuados de proibições; Eliminar o uso de categorias finalísticas excessivamente amplas.

	PLATAFORMAS QUE APRESENTARAM ALGUMA INADEQUAÇÃO	INADEQUAÇÃO	RISCOS E ROBLEMAS	RECOMENDAÇÕES DE TRANSPARÊNCIA
ESPECIFICAÇÃO DE CONTEÚDOS PROIBIDOS	Facebook; Instagram; LinkedIn; Pinterest; Snapchat; Youtube	Terminologia ambígua; Uso do verbo "poder"; Uso do modo imperativo;	Indefinição quanto à norma; Redução da capacidade contestatória do usuário;	Utilizar expressões que denotem proibição de forma inequívoca; Eliminar o uso do modo imperativo e do verbo poder;
MEDIDAS INTERVENTIVAS	Instagram, Snapchat, TikTok e Twitter	Omissão dos critérios; Terminologia inespecífica; Uso do verbo "poder"; Referência a múltiplas medidas aplicáveis sem indicação dos critérios/fatores a serem considerados.	Indefinição quanto à medida interventiva	Citar expressamente quais as medidas cabíveis e a quais casos se aplicam; Eliminar o uso do modo imperativo e do verbo poder; Indicar critérios/fatores a serem considerados na determinação da medida sempre que múltiplas medidas forem apresentadas como aplicáveis ao mesmo conteúdo.
CONTESTAÇÃO	Facebook; Instagram; Pinterest; Snapchat; TikTok; Youtube	Omissão do mecanismo recursal; Citação restrita à política de direitos autorais	Redução da capacidade contestatória ao usuário	Citar expressamente quaisquer meios recursais cabíveis; Detalhar seu funcionamento ou indicar de forma notável e acessível onde o usuário pode aprender a respeito.
NOTIFICAÇÃO	Facebook; Instagram; Linkedin; Pinterest; Snapchat; TikTok; Twitter;	Omissão do mecanismo de notificação; Citação restrita à política de direitos autorais	Redução da capacidade contestatória ao usuário	Citar expressamente o mecanismo de notificação; Detalhar as informações apresentadas na notificação.

Fonte: autoria própria.

Em síntese, os problemas identificados pelo contraste entre todas as relatorias e análise das políticas de comunidade apontam para a necessidade de revisão na forma como elas se apresentam, para que sejam mais transparentes. No geral, percebe-se uma tendência de aprimoramento na quantidade de informações disponíveis ao público sobre a moderação realizada. Entretanto, a filosofia que orienta essas ações das plataformas não mira na accountability, sendo uma transparência ainda instrumental. As informações apresentadas, a linguagem utilizada, os mecanismos disponibilizados e a forma como são construídas as políticas de comunidade não se refletem em uma transparência que viabilize controle social sobre as práticas de moderação de conteúdo.

6. Conclusão

O ecossistema sociotécnico de produção, veiculação e consumo de conteúdos foi profundamente transformado em escala global nas últimas décadas, com um crescente protagonismo das plataformas de conteúdo gerado por usuários nesse meio. Essas aplicações eram comumente concebidas sob o imaginário liberal dominante no debate público sobre a internet durante a década de 1990, que enquadrava a rede como território de realização dos valores de livre expressão, autonomia individual e descentralização. Nesse contexto, essas plataformas não tomavam a moderação das contas e publicações como uma de suas funções centrais. Intervenções dessa natureza eram consideradas um encargo secundário e incidental que se impunha conexo a seu objetivo central de favorecer o aumento no volume e na diversidade dos conteúdos gerados pelos usuários e por elas veiculados. Ainda assim, preocupações relacionadas à difusão de conteúdo danoso, à época identificado sobretudo com material pornográfico e/ou pirateado, começavam a se apresentar como desafios institucionais e suscitavam respostas dos poderes constituídos.

Em diversos países, o resultado dessas disputas iniciais foi a consolidação do modelo de não-responsabilização prévia dos intermediários, paradigma normativo representado nos EUA pela seção 230 do *Communications Decency Act* e no Brasil pelo Art. 19 do Marco Civil da Internet. Esse regime regulatório se fundamenta na premissa de que a responsabilização dos intermediários pelos conteúdos gerados pelos usuários gera um incentivo ao monitoramento e à restrição massiva de conteúdos por parte dos intermediários. Nessa lógica, as plataformas buscariam se autoproteger previamente de sanções, o que poderia ferir a liberdade de expressão e a privacidade dos usuários. Todavia, esse modelo também atribui aos intermediários a capacidade de tomar medidas proativas para a moderação, a fim de oferecer um incentivo a ações de moderação destinadas à redução da circulação de conteúdo danoso (a chamada cláusula do bom samaritano da seção 230).

No século XXI, esse fenômeno se revestiu de maior atenção conforme a internet cresceu em alcance, diversidade e volume de informação trafegada. O crescente protagonismo das plataformas no ambiente digital implicou na ampliação correlata das demandas sociais em torno de seu manejo do conteúdo veiculado. Isso se agrava com o exercício, pelas plataformas, de um protagonismo cada vez maior nas políticas regulatórias da liberdade de expressão.

Por um lado, exige-se que as plataformas intervenham para coibir conteúdo manifestamente danoso, incluindo aquele que constitui discurso de ódio, desinformação e condutas criminosas. Por outro lado, a restrição da liberdade de expressão mediante a discricionariedade de agentes privados se apresenta como uma ameaça para esse e outros direitos, sobretudo na medida em que suas plataformas ganham relevância para o acesso a informações e o debate público. Consequentemente, um conjunto heterogêneo de soluções regulatórias não-vinculantes tem sido desenvolvido, o qual inclui códigos de conduta, listas de princípios orientadores e documentos que sugerem boas práticas. Além disso, a profissionalização da moderação é favorecida pelas plataformas na medida em que também comunica a seus parceiros comerciais que o ambiente respeitará certos padrões de qualidade.

No setor privado, os padrões, diretrizes ou políticas de comunidade são o principal instrumento dessa natureza. Comumente identificadas como regramentos privados das plataformas sobre as normas e medidas aplicáveis aos conteúdos gerados pelos usuários, a eficácia concreta desses instrumentos historicamente assemelha-se mais à de uma cartilha educativa ou de uma peça publicitária do que de um documento regulatório. O exame das políticas de comunidade das plataformas analisadas neste estudo revela uma série de escolhas de redação que removem a previsibilidade e ampliam incerteza sobre qual é de fato o teor das normas, quais mecanismos e critérios orientam sua aplicação, que medidas serão tomadas pela plataforma em caso de infração e de que direitos dispõe o usuário afetado pela moderação. Exemplos dessas escolhas incluem a opção pelo modo imperativo (“não publique, não poste”) e o uso de linguagem que desobriga a plataforma a se comprometer com um curso de ação específico (“em caso de infração, *poderemos* remover a postagem”).

Essa constatação sugere que as políticas de comunidade presentemente cumprem uma finalidade primariamente publicitária - sinalizar para anunciantes a viabilidade comercial do ambiente e para formuladores de políticas a dispensabilidade de intervenções regulatórias estatais - e pedagógica - ensinar ao usuário sobre que comportamentos são desejáveis ou não do ponto de vista da plataforma. Esse cenário é agravado quando considerado que há casos em que os Termos de Uso obrigam o usuário a observar as políticas de comunidade, transformando-as em aditivos contratuais dotados de eficácia jurídica. Nesses casos, os usuários ficam vinculados a regras contratuais extraordinariamente

imprecisas, elaboradas unilateralmente e alteradas com frequência, ao passo que as plataformas passam a dispor de mais um mecanismo para evadir compromissos legais efetivos.

Mas se do ponto de vista descritivo (referente às funções concretas que cumprem) esse é o estado atual de instrumentos dessa natureza, do ponto de vista normativo (referente às funções que deveriam cumprir) ainda há um longo caminho a se trilhar. O exercício do controle social sobre as plataformas demanda que quaisquer regramentos sobre conteúdos apresentem o máximo de transparência. Ela é necessária tanto no nível das normas e medidas aplicáveis, quanto nos critérios e sistemas que orientam sua aplicação.

Os resultados empíricos desta pesquisa sugerem que ainda há um longo caminho a ser trilhado para que a transparência se efetive nessa seara. Nas plataformas analisadas, a redação dos padrões de comunidade é comumente permeada por construções linguísticas que introduzem ambiguidade e incerteza quanto à aplicabilidade das normas e medidas. Destacam-se, a esse respeito, a carência de exemplos, o uso frequente do modo imperativo ("não publique", "não poste"), o emprego do verbo "poder" ("conteúdo X *poderá* ser removido"), as referências à discricionariedade da plataforma e o uso de terminologia ambígua. Esse tipo de construto não oferece previsibilidade para a moderação de conteúdo. Ao contrário, reafirma a desobrigação das plataformas à aplicação de suas próprias regras de forma consistente, isonômica e transparente. Sob o mesmo ângulo, as omissões e insuficiências quanto à comunicação dos meios e critérios de avaliação dos conteúdos passíveis de moderação impedem que o processo pelo qual os conteúdos são moderados seja conhecido e, por conseguinte, examinado de forma crítica.

As consequências dessa opacidade não são triviais, pois o desconhecimento relativo a tais processos compromete o desenvolvimento de respostas sociais aos riscos representados por violações de direitos ocorridas no exercício da moderação. Tais riscos incluem a imposição de restrições à liberdade de expressão por intervenções indevidas, discriminação por sistemas automatizados de tomada de decisão, violações aos direitos dos trabalhadores contratados para o exercício da moderação e a nulificação da capacidade contestatória de cada usuário submetido a uma intervenção da plataforma. Consequentemente, a opacidade desse tipo de documento se apresenta como um problema social significativo para a defesa dos direitos humanos num ambiente em que as plataformas desempenham papel crescente.

Observa-se, por fim, que a totalidade do debate sobre transparência em moderação de conteúdo não se reduz à transparência em regramentos de natureza similar aos examinados neste estudo. Um regime capaz de garantir essa transparência demanda uma série de garantias de devido processo, que incluem a reavaliação dos procedimentos de notificação e a instituição de sistemas robustos

de contestação. Os únicos fundamentos possíveis para a construção de tal regime são a pesquisa científica continuamente multidisciplinar e atualizada. Igualmente fundamental é o diálogo multissetorial atravessado por ampla participação social e democrática. Somente desse modo será possível avançar rumo à efetiva transparência e *accountability* num contexto de crescente presença das plataformas.

Recomendações

Às plataformas

1. Comunicar de forma expressa, visível e pública quais os meios e critérios empregados na detecção e avaliação de conteúdo passível de moderação, especificando sempre que sistemas automatizados forem utilizados.

2. Indicar conteúdos que são proibidos por meio de termos que denotam proibição expressamente, como “não é permitido” ou “é proibido”.

3. Eliminar termos e construções frasais marcadamente ambíguas da redação de seus padrões de comunidade, em especial o uso do modo imperativo e do verbo “poder”.

4. Indicar quais os critérios ou fatores que influenciam na determinação da medida tomada sempre que múltiplas medidas forem cabíveis.

5. Apresentar múltiplos exemplos e contraexemplos de conteúdo infrator e/ou sujeito à intervenção após cada norma.

6. Delimitar os critérios ou categorias de conteúdos exceptuados de forma específica, eliminando o uso de categorias finalísticas excessivamente amplas (“conteúdo de relevância pública”, por exemplo).

7. Garantir o acesso público permanente a todas as versões anteriores sempre que as políticas de comunidade forem atualizadas. Sinalizar a data em que a política foi atualizada.

8. Incluir dados quantitativos sobre todos os conteúdos que sofreram algum tipo de intervenção em seus relatórios de transparência, não apenas publicações removidas e contas

suspensas.

9. Informar ao usuário nas políticas de comunidade sobre seu direito à notificação e à contestação na ocasião de alguma intervenção de moderação.

10. Notificar o usuário sempre que este for alvo de alguma intervenção. A notificação deve incluir, no mínimo, as seguintes informações: URL da publicação, trecho do conteúdo que causou a intervenção (ou dados adicionais que possibilitem sua identificação), cláusula violada dos padrões de comunidade, meios de detecção e intervenção sobre o conteúdo. Ainda, deve ser fornecida em formato duradouro e deve permanecer disponível mesmo que a conta do usuário seja suspensa ou indisponibilizada.

11. Instituir um sistema de contestação robusto, que inclua, no mínimo: revisão humana por um ou mais indivíduos que não estiveram envolvidos na decisão inicial, oferta ao usuário da oportunidade de apresentar informações adicionais a serem consideradas na revisão, notificação dos resultados da revisão e uma declaração de motivos suficiente para que o usuário compreenda a decisão.

12. Quando utilizada detecção automatizada de conteúdo a ser moderado, explicitar as formas de identificação proativa e os canais para comunicação de terceiros sobre falhas, vieses e discriminações potencializadas ou criadas pelo algoritmo.

13. Desenvolver mecanismos que assegurem ampla participação social na elaboração das políticas de comunidade da plataforma.

14. Apoiar - mediante auxílio financeiro, fornecimento de dados, disponibilização de especialistas, etc. - pesquisa científica multidisciplinar destinada a compreender as diferentes dimensões da circulação de conteúdo em escala massiva, bem como os efeitos das decisões relativas ao processo de moderação.

Ao setor governamental

15. Condicionar a efetuação de qualquer alteração no regime de responsabilização de intermediários, especialmente por conteúdos gerado por terceiros, a amplo debate público prévio, voltado à promoção da transparência e participação dos usuários,

da sociedade civil organizada e da comunidade científica - tendo em consideração os riscos de censura colateral e da importância da internet como ferramenta de expressão.

16. Desenvolver soluções regulatórias, fundamentadas em direitos humanos e amplo debate público, que obriguem as plataformas a implementar um robusto regime de transparência e de procedimentos democráticos previamente estipulados na moderação de conteúdo.

17. Apoiar a pesquisa científica multidisciplinar sobre moderação de conteúdo e expressão na internet, apta a fundamentar factualmente as soluções legislativas e decisões judiciais adotadas.

18. Compatibilizar quaisquer medidas de transparência adotadas com as normas de privacidade e proteção de dados pessoais, a fim de garantir a segurança jurídica e a proteção do direito do usuário à autodeterminação informativa.

19. Disponibilizar publicamente dados sobre pedidos de moderação de conteúdo feito às plataformas a partir de instituições ou autoridades públicas, a fim de permitir o exame de compatibilidade com os relatórios atualmente disponibilizados unicamente pelas plataformas.

À sociedade civil organizada e à academia

20. Acompanhar as propostas legislativas e a jurisprudência relativas à transparência, moderação de conteúdo e responsabilidade de intermediários.

21. Conduzir pesquisa científica capaz de embasar soluções protetivas dos direitos fundamentais dos usuários para os desafios relacionados à governança de plataformas.

22. Demandar canais de participação, junto ao setor governamental e às plataformas, na elaboração e avaliação de políticas de comunidade.

23. Estabelecer parâmetros para avaliação de práticas de moderação de conteúdo e seu nível de accountability.

24. Participar dos fóruns e espaços de debate multissetorial, contribuindo com o debate público sobre moderação de conteúdo e transparência.

25. Denunciar violações de direitos humanos realizadas por atores públicos ou privados no âmbito das práticas de moderação de conteúdo - sejam elas motivadas por políticas de comunidade ou por ordens de autoridades governamentais.

REFERÊNCIAS

ANANNY, Mike; GILLESPIE, Tarleton. Public Platforms: Beyond the Cycle of Shocks and Exceptions. In: **The Internet, Policy and Politics Conferences**. Oxford Internet Institute, 2016.

BARDIN, Laurence. **Análise de conteúdo**. Trad. Luís Antero Reto e Augusto Pinheiro. Lisboa: 70, 1977.

BARLOW, John Perry. **Declaração de Independência do Ciberespaço**. (trad. DH net). Disponível em: http://www.dhnet.org.br/ciber/textos/barlow.htm. Acesso em: 29 jun 2020.

BELLI, Luca; VENTURINI, Jamila. Private ordering and the rise of terms of service as cyber-regulation. **Internet Policy Review**. v. 5, n. 4. 2016. p. 9. Disponível em: https://www.econstor.eu/bitstream/10419/214032/1/IntPolRev-2016-4-441.pdf Acesso em:08 out. 2020.

BIDDLE, Sam; RIBEIRO, Paulo Victor; DIAS, Tatiana. TikTok escondeu "feios" e favelas para atrair novos usuários e censurou posts políticos. **The Intercept Brasil**, 16 mar. 2020. Disponível em: https://theintercept.com/2020/03/16/tiktok-censurou-rostos-feios-e-favelas-para-atrair-novos-usuarios/ Acesso em: 16 mar. 2020.

BRITO CRUZ, Francisco (coord.); MASSARO, Heloisa; OLIVA, Thiago; BORGES, Ester. **Internet e eleições no Brasil**: diagnósticos e recomendações. InternetLab, São Paulo, 2019. Disponível em: Acesso em: https://www.internetlab.org.br/wp-content/uploads/2019/09/policy-infopol-26919_4.pdf 18 ago. 2020.

BRUNO, F. Monitoramento, classificação e controle nos dispositivos de vigilância digital. **Revista FAMECOS**, v. 36, p. 1-7, 2008

CARMO, Paloma; DUARTE, Felipe; GOMES, Ana Bárbara. **Glossário da Inclusão Digital** - Volume II. Instituto de Referência em Internet e Sociedade: Belo Horizonte, 2020. Disponível em: http://bit.ly/3aqUlfP. Acesso em: 07 ago. 2020.

CE - Conselho da Europa. **Guia dos Direitos Humanos para os Utilizadores de Internet**. 2014. Disponível em: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806a0532. Acesso em: 08 out. 2020.

CENTIVANY, Alissa. Values, ethics and participatory policymaking in online communities. **Proceedings of the Association for Information Science and Technology**, v. 53, n. 1, p. 1-10, 2016.

CITRON, Danielle Keats. Extremist speech, compelled conformity, and censorship creep. **Notre Dame L. Rev.**, v. 93, p. 1035, 2017.

CROCKER, Andrew et al. **Who has your back?** Censorship edition 2019. EFF - Electronic Frontier Foundation. 2019. Disponível em: https://www.eff.org/wp/who-has-your-back-2019 Acesso em 21 Jul. 2020

EFF et al. **Princípios de Manilla sobre responsabilidade civil de intermediários.** Disponível em: https://www.manilaprinciples.org/pt-br Acesso em: 21 jul. 2020.

EFF et al. **Santa Clara Principles on transparency and accountability in content moderation.** Disponível em: https://santaclaraprinciples.org/ Acesso em: 21 jul. 2020.

EFF - Electronic Frontier Foundation. **The Manila Principles on Intermediary Liability Background Paper**. 1. v. online, maio 2015. Disponível em: https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf#page=49 Acesso em: 20 ago. 2020.

EUA - Estados Unidos da América. **Communications Decency Act**. Sec. 509. Online Family Empowerment. In: Telecommunications Act of 1996. Disponível em: https://www.govinfo.gov/content/pkg/PLAW-104publ104/pdf/PLAW-104publ104.pdf Acesso em: 29 jun. 2020.

EUA - Estados Unidos da América. **The Digital Millennium Copyright Act**. Dez. 1998. Disponível em: https://www.copyright.gov/legislation/dmca.pdf Acesso em: 08 out. 2020.

FACEBOOK. **Relatório de aplicação dos padrões de comunidade**. Disponível em: https://transparency.facebook.com/community-standards-enforcement Acesso em: 20 jul. 2020.

FIESLER, Casey; et al. Reddit rules! characterizing an ecosystem of governance. In: **Twelfth International AAAI Conference on Web and Social Media**. 2018.

FOX, Jonathan. The uncertain relationship between transparency and accountability. **Development in practice**, v. 17, n. 4-5, p. 663-671, 2007.

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

GILLESPIE, Tarleton. Governance of and by platforms. **SAGE handbook of social media**, p. 254-278, 2017.

GORWA, Robert. What is platform governance?. **Information, Communication & Society**, v. 22, n. 6, p. 854-871, 2019.

HOOTSUITE; WE ARE SOCIAL. **Digital 2020**: Brazil - DataReportal Global Digital Insights. Disponível em: https://datareportal.com/reports/digital-2020-brazil?rq=brazil Acesso em: 20 ago. 2020. p. 43

JOSEPH, Chanté. Instagram's murky 'shadow bans' just serve to censor marginalised communities. **The Guardian**, Londres, 8 nov. 2019. Disponível em: https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive. Acesso em: 07 out. 2020.

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

KRAUT, Robert E.; RESNICK, Paul. **Building successful online communities: Evidence-based social design**. Mit Press, 2012.

LOSEY, James. Surveillance of communications: A legitimization crisis and the need for transparency. **International Journal of Communication**, v. 9, p. 3450-3459, 2015.

MACKINNON, Rebecca et al. **Fostering freedom online**: The role of internet intermediaries. UNESCO Publishing, 2015.

NEWTON, Casey. YouTube gets sued by its moderators. **The Interface**. 22 set. 2020, n. 572 Disponível em: https://www.getrevue.co/profile/caseynewton/issues/youtube-gets-sued-by-its-moderators-280023?utm_campaign=Issue&utm_content=view_in_browser&utm_medium=email&utm_source=The+Interface Acesso em: 08 out. 2020.

OEA et al. **Declaração conjunta do vigésimo aniversário: desafios para a liberdade de expressão na próxima década**. 2019. Disponível em: https://www.oas.org/pt/cidh/expressao/showarticle.asp?artID=1146&lID=4 Acesso em: 8 jul. 2020

ONU - Organização das Nações Unidas. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression**. Symbol A/HRC/38/35. 6 abr. 2018. Disponível em: https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35 Acesso em: 30 jun. 2020.

ONU - Organização das Nações Unidas. **Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression**.

Disease pandemics and the freedom of opinion and expression. 23 abr. 2020. Disponível em: https://undocs.org/A/HRC/44/49 Acesso em: 30 jun. 2020.

RDR - Ranking Digital Rights. **2019 RDR Accountability Index**. Maio 2019. Disponível em: https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf Acesso em: 09 jul. 2020. p. 20

REDDIT. **Content Policy**. Disponível em: https://www.redditinc.com/policies/content-policy Acesso em: 31 jul. 2020.

ROUVROY, Antoinette; BERNS, Thomas. Governamentalidade algorítmica e perspectivas de emancipação: o díspar como condição de individuação pela relação?. **Revista Eco Pós**, vol. 18, n. 2, 2015, p. 35-56.

RIDER, Karina. The privacy paradox: how market privacy facilitates government surveillance. **Information, Communication & Society**, v. 21, n. 10, p. 1369-1385, 2018.

ROBERTS, Sarah T. Digital detritus:'Error'and the logic of opacity in social media content moderation. **First Monday**, 2018.

ROSENFELD, Michel. **Hate speech in constitutional jurisprudence**: a comparative analysis. Cardozo L. Rev., v. 24, p. 1523, 2002.

SEERING, Joseph; KRAUT, Robert; DABBISH, Laura. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In: **Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing**. 2017. p. 111-125.

SENSORTOWER. **Q4 2019**. Store Intelligence Data Digest. Disponível em: https://go.sensortower.com/Q4-2019-Data-Digest.html?src=blog Acesso em: 20 ago. 2020

SILVA JUNIOR, Luiz Alberto; LEAO, Marcelo B. C. O software Atlas.ti como recurso para a análise de conteúdo: analisando a robótica no Ensino de Ciências em teses brasileiras. **Ciênc. educ. (Bauru)**, Bauru , v. 24, n. 3, p. 715-728, set. 2018.

SOWELL, Jesse H. **Evaluating competition in the Internet's infrastructure**: a view of GAFAM from the Internet exchanges. Journal of Cyber Policy, v. 5, n. 1, p. 107-139, 2020.

SUZOR, Nicolas P. et al. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. **International Journal of Communication**, v. 13, p. 18, 2019.

SUZOR, Nicolas; VAN GEELEN, Tess; MYERS WEST, Sarah. **Evaluating the legitimacy of platform governance:** A review of research and a shared research agenda. International Communication Gazette, v. 80, n. 4, p. 385-400, 2018.

VALENTE, J. C. L. **Tecnologia, informação e poder**: das plataformas online aos monopólios digitais. Tese (Doutorado em Sociologia). Universidade de Brasília, Brasília, 2019. p. 170

VALENTE, J. C. L. Plataformas digitais e concentração na internet. In: Encontro Anual da Rede de Pesquisa em Governança da Internet, III, 2019, Manaus., **Anais...** [S.I.]: Rede de Pesquisa em Governança da Internet, 2020. p. 1-25.

VENTURINI, Jamila et al. **Terms of service and human rights**: An analysis of online platform contracts. 2016. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/18231 Acesso em: 21 Jul. 2020.

WEST, Sarah Myers. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. **New Media & Society**, v. 20, n. 11, p. 4366-4383, 2018.

ZUBOFF, Shoshana. **The Age of Surveillance Capitalism**: The Fight for a Human Future at the New Frontier of Power. New York: Public Affairs, 2019.

APÊNDICE A - ESQUEMA DE CODIFICAÇÃO AXIAL

Os códigos associados aos trechos destacados como relevantes nas políticas analisadas foram:

Fundamentação da política

Proibição:

- Proibição: Específica
- Proibição: Genérica

Exceção::

- Exceção: Específica
- Exceção: Genérica

Recomendações de conteúdo:

- Recomendações de conteúdo: Desencorajado
- Recomendações de conteúdo: Encorajado

Medida interventiva:

- Medida interventiva: Específica
- Medida interventiva: Genérica

Meio de análise:

- Meio de análise: Determinado
- Meio de análise: Indeterminado

Critério de análise:

- Critério de análise: Indeterminado

Critério de análise: Determinado

Apoio ao usuário:

Apoio ao usuário: Denúncia

Apoio ao usuário: Contestação

Apoio ao usuário: Informações

Codificação Axial - Descrição

O código “Fundamentação da política” refere-se a citações que contém exposições de valores, missões ou objetivos pela plataforma, bem como àquelas em que esta justifica ou fundamenta suas políticas.

A categoria “Proibição” refere-se a citações que contém proibições explícitas da veiculação de certos tipos de conteúdo na plataforma. Entendemos por proibições explícitas aquelas formuladas utilizando termos como “proibimos”, “não permitimos”, “não toleramos” e suas variações, bem como aquelas em que elementos contextuais (estrutura do documento, uso de imagens, relação com outras frases próximas) sugerem tratar-se de uma interdição. Aplicamos o código “Proibição::Específica”” a proibições que delimitam, por meio de definições ou de múltiplos exemplos, que conteúdos se encaixam no tipo proibido. Na ausência de definição ou de exemplos, aplicamos o código “Proibição::Genérica”.

A categoria “Exceção” refere-se a citações que excetuam subtipos de conteúdos do escopo de alguma política que normalmente seria aplicável a seu tipo. Aplicamos o código “Exceção::Específica” a citações que explicitam as hipóteses excetuadas (por exemplo: “não é permitida nudez, exceto para fins artísticos e educacionais”). Aplicamos o código “Exceção::Genérica” a citações não o fazem ou o fazem de forma inespecífica (por exemplo: “não é permitida nudez na plataforma, porém podemos permitir em alguns casos, a nosso critério”).

A categoria “Recomendações de conteúdo” refere-se a citações que contém conteúdos que não são proibidos, mas buscam conduzir usuários a produzirem conteúdo alinhado com aquele idealizado para o perfil da plataforma, ou a evitar produzir conteúdo que foge a esse perfil.

O código “Recomendações de conteúdo:: Desencorajado” refere-se a citações que contém recomendações aos usuários para que evitem veicular certos tipos de conteúdo na plataforma, porém sem, todavia, proibir sua veiculação. Entendemos por recomendações aquelas formuladas sem o uso dos termos que caracterizam

proibição, sendo comumente escritas com recurso ao modo imperativo (por exemplo: “Não publique conteúdos que possam ofender a terceiros.”).O código “Recomendações de conteúdo:: Encorajado” refere-se a citações que contém recomendações aos usuários para que veiculem certos tipos de conteúdo na plataforma (por exemplo: “Busque publicar conteúdos interessantes em seu feed.”)

A categoria “Medida interventiva” refere-se a citações que contém menções a medidas adotáveis pela plataforma no contexto da moderação de conteúdo (por exemplo: remoção da publicação, suspensão da conta, denúncia às autoridades). Aplicamos o código “Medida interventiva::Específica” a citações que estabelecem as medidas cabíveis e os casos aos quais são aplicáveis (por exemplo: “no evento da publicação de conteúdo contendo incitação à autolesão, a publicação será removida e a conta do usuário será suspensa”). Aplicamos o código “Medida interventiva::Ambígua” a citações que não fixam o universo de medidas cabíveis aos casos (por exemplo: “em caso de violação desta norma, poderemos advertir o usuário, remover a publicação e/ou suspender sua conta”).

A categoria “Meio de análise” refere-se a citações que contém menções aos mecanismos de detecção e/ou avaliação do teor de conteúdos potencialmente infringentes das políticas da plataforma. Entendemos por mecanismos de detecção aqueles que dão ciência à plataforma da existência de conteúdo potencialmente infringente (por exemplo: denúncias de usuários, filtragem automatizada). Entendemos por mecanismos de avaliação aqueles que a plataforma emprega para avaliar o teor do conteúdo potencialmente infringente (por exemplo: moderação por profissionais treinados, inteligência artificial). Aplicamos o código “Meio de análise::Determinado” a citações que informam o mecanismo de detecção e/ou avaliação (por exemplo: “caso uma conta seja denunciada por comportamento abusivo, notificaremos a conta”). Aplicamos o código “Meio de análise:: Indeterminado” a citações que omitem o mecanismo de detecção ou avaliação (“Ao tomarmos ciência de que um usuário está se comportando de forma abusiva, notificaremos a conta”).

A categoria “Critério de análise” refere-se a citações que contém referências aos critérios considerados pela plataforma na avaliação do teor de um conteúdo potencialmente infringente e/ou na definição das medidas interventivas cabíveis (por exemplo: histórico do usuário, gravidade da violação, contexto cultural, intenção do usuário). Aplicamos o código “Critério de análise::Determinado” a citações em que o critério utilizado na avaliação é especificado (Por exemplo: “se o usuário violar esta política novamente, sua conta será suspensa”). Aplicamos o código “Critério de análise::Indeterminado” a citações que fazem referência a uma avaliação na qual o critério não é explicitado (Por exemplo: “A nosso critério, poderemos permitir certas reproduções de nudez na plataforma”).

A categoria “Apoio ao Usuário” refere-se mecanismos de auxílio ao usuário, isto é, recursos e funcionalidades disponibilizados pela plataforma para que usuário se informe, denuncie violações e conteste intervenções relacionadas às políticas de comunidade.

O código “Apoio ao Usuário: Denúncia” refere-se a citações que apontam mecanismos que podem ser utilizados pelo usuário para dar ciência à plataforma de que há uma infração de suas políticas de comunidade ocorrendo.

O código “Apoio ao Usuário::Informações” refere-se a citações que apontam links e outros serviços em que se pode saber mais sobre os conteúdos moderados e os objetivos da plataforma, ou buscar auxílio em relação a alguma situação decorrente desses conteúdos.

O código “Apoio ao Usuário::Contestação” refere-se a citações que indicam os mecanismos disponíveis ao usuário para que este recorra ou se oponha na ocasião de um conteúdo seu ter sofrido uma medida interventiva pela plataforma (por exemplo: formulário de contestação, central de informações).

Para cada documento analisado, foram criados ainda códigos que coocorrem com os acima mencionados, para cada seção da política, iniciado por “[Plataforma]:” e seguido da temática tratada naquele trecho. Por exemplo, em cada uma das citações que estão no trecho sobre indivíduos e organizações perigosos de um documento, foi aplicado, além de um dos códigos acima, o código “Pinterest: indivíduos e organizações perigosos”. Esses códigos de seção foram adicionados concomitantemente a cada citação categorizada com um dos códigos anteriores. Essa sobrecodificação foi feita a fim de se possibilitar a recuperação automática de quais códigos aparecem em cada bloco do texto da política, para identificar sua estrutura (ou seja, saber se uma seção é composta de proibição, recomendações, exceções, etc.). Também será possível analisar, com essa dupla codificação das citações, como cada tipo de conteúdo é tratado pelas plataformas e a maneira como elas abordam as categorias, comparando-as entre si. Então, por exemplo, ao dar o comando de recuperar citações com o código “[Plataforma]: indivíduos e organizações perigosos”, o software retorna todas as citações daquela seção, com o respectivo código complementar, correspondendo a uma das categorias anteriores.

Data de coleta do material: 22/04/2020

INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE