Contribuições aos Princípios de Santa Clara

Comentários do IRIS à chamada pública

INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE

CONTRIBUIÇÕES AOS PRINCÍPIOS DE SANTA CLARA SOBRE TRANSPARÊNCIA E ACCOUNTABILITY EM MODERAÇÃO DE CONTEÚDO

AUTORIA
Gustavo Ramos Rodrigues
Lahis Pasquali Kurtz
Luiza Couto Chaves Brandão

PROJETO GRÁFICO, CAPA E DIAGRAMAÇÃO
Felipe Duarte

INSTITUTO
DE REFERÊNCIA
EM INTERNET
E SOCIEDADE

Contribuições aos Princípios de Santa Clara sobre Transparência e Accountability em Moderação de Conteúdo

Os Princípios de Santa Clara sobre Transparência e Accountability em Moderação de Conteúdo são um instrumento internacional, resultado do trabalho de algumas organizações que buscam estabelecer critérios e orientações para a moderação de conteúdo, transparência e responsabilidade de plataformas sobre o conteúdo de terceiros. É, assim, um mecanismo de *soft law,* que representa um guia, de padrão internacional, para a matéria.

Considerando os temas de pesquisa do Instituto de Referência em Internet e Sociedade - IRIS[1], respondemos à chamada de consulta pública sobre os Princípios de Santa Clara, suas definições, limitações e potencialidades, a partir das 15 perguntas a seguir[2]. A iniciativa de fortalecer os princípios a partir da integração de distintas perspectivas e abertura para considerações de atores a nível global é bem-vinda no atual cenário, no qual as discussões sobre o conteúdo online tomam cada vez mais força. Nesse sentido seguem as contribuições do IRIS, disponíveis em formato de eletrônico, para consulta do público em geral e fortalecimento desse debate no cenário brasileiro.

Belo Horizonte, 30 de junho de 2020.

1. **Atualmente, os Princípios de Santa Clara estão focados na necessidade de obter números, de estabelecer avisos prévios e recursos em torno da moderação de conteúdo. Este conjunto de questões vai discutir se estas categorias deveriam ser expandidas, aperfeiçoadas ou revisitadas.**

   a. A primeira categoria estabelece, como padrão, que as empresas devem publicar os números de posts removidos e de contas suspensas, temporária ou permanentemente, em razão de violações às suas diretrizes de conteúdo. Por favor, indique qualquer recomendação ou componentes específicos desta categoria que deveriam ser revisitados ou expandidos.

   b. A segunda categoria estabelece, como padrão, que as empresas devem avisar cada usuário cujo conteúdo for removido ou que tiver sua conta suspensa acerca das razões para remoção ou suspensão. Por favor, indique qualquer recomendação ou componentes específicos desta categoria que deveriam ser revisitados ou expandidos.

   c. A terceira categoria estabelece, como padrão, que as empresas devem fornecer oportunidades significativas para recursos tempestivos acerca de qualquer remoção de conteúdo ou suspensão de conta. Por favor, indique qualquer recomendação ou componentes específicos desta categoria que deveriam ser revisitados ou expandidos.

1 O Instituto de Referência em Internet e Sociedade (IRIS) é uma organização sem fins lucrativos, constituída em 2015, como instituto independente que não está vinculado a nenhuma outra instituição, pública ou privada. Sua missão é explorar, investigar e entender os desdobramentos da Internet e novas tecnologias sobre a sociedade contemporânea. Mais informações em: www.irisbh.com.br.

2 A consulta ficou disponível até o dia 30 de junho de 2020. A fim de facilitar a compreensão das respostas oferecidas, reproduzimos aqui, em negrito, as perguntas da página de submissões em portugês: https://santaclaraprinciples.org/pt/cfp/

2. **Você acredita que os Princípios de Santa Clara devem ser expandidos ou emendados para incluir recomendações específicas acerca da transparência em torno do uso de ferramentas automatizadas e do processo de tomada de decisão (o que inclui, por exemplo, o contexto em que estas ferramentas são utilizadas e a extensão em que as decisões são tomadas sem seres humanos), em alguma das seguintes áreas:**

   Moderação de conteúdo (uso de inteligência artificial para revisar conteúdos e contas para determinar se devem ou não ser removidos; processos utilizados para conduzir as revisões quando conteúdos são sinalizados como impróprios por usuários ou outros)

   Ranking e downranking de conteúdo (uso de inteligência artificial para promover certos conteúdos a despeito de outros em rankings de ferramentas de busca, por exemplo, e para promover o downranking de conteúdos como desinformação ou clickbaits)

   Direcionamento de publicidade (uso de inteligência artificial para segmentar e mirar grupos específicos de usuários e enviar publicidade a eles)

   Recomendações de conteúdo e auto-complete (uso de inteligência artificial para recomendar conteúdos como vídeos, posts e palavras-chave com base nos perfis e no comportamento passado dos usuários)

O crescente emprego de ferramentas de análise automatizada de conteúdo em massa por parte das plataformas suscita preocupações sérias relativas à não-discriminação, haja vista que sistemas algorítmicos que moderam conteúdo frequentemente apresentam vieses de classe, raça, gênero, orientação sexual e de outros marcadores sociais. Somam-se a essas questões o potencial que os sistemas de direcionamento automatizado de publicidade apresentam para a disseminação de desinformação, o que afeta a segurança e a saúde públicas, bem como a estabilidade dos sistemas democráticos.

Nesse cenário, a demanda por transparência de decisões automatizadas tem sido de progressiva importância em âmbito internacional, sendo reconhecida, inclusive, como condição para a efetivação do direito à autodeterminação informativa. Um exemplo disso é o Regulamento Geral de Proteção de Dados da União Europeia, que reconhece o direito do usuário de ser informado sobre a lógica de processos de decisões automatizadas. Assim sendo, conclui-se ser adequado que o tema seja abordado de forma específica, posto que os sistemas automatizados constituem um dos meios principais pelos quais todo o tratamento de conteúdo pelas plataformas ocorre.

3. **Você acredita que os atuais Princípios de Santa Clara fornecem a estrutura adequada ou poderiam ser aplicados para intermediar restrições (tais como age-gating, adicionar avisos a determinados conteúdos e adicionar informações de qualificação a determinados conteúdos)? Se não, nós deveríamos procurar incluir estas categorias em um processo de revisão dos princípios ou um conjunto separado de princípios seria melhor para abarcar estas questões?**

Do ponto de vista da demanda por transparência, o tratamento do conteúdo pelas plataformas deve ser entendido a partir de suas implicações para os direitos

fundamentais dos usuários, em especial para a liberdade de expressão. Dessa perspectiva, práticas de moderação potencialmente restritivas da liberdade de expressão não se limitam a remoção de publicações e suspensão ou banimento de contas, mas incluem, também, intervenções que reduzem seu alcance sem removê-lo necessariamente.

Dada a curadoria algorítmica de conteúdos que molda o fluxo de publicações exibido aos usuários, as recomendações de conteúdo e os resultados de buscas por palavras-chave na plataforma, a prática de *downranking* pode ter efeitos tão severos sobre a capacidade do usuário de se comunicar quanto a remoção. Esses efeitos são profundamente agravados pela opacidade que caracteriza a prática, pois, diferentemente da remoção, o usuário sequer pode ter ciência de que está sofrendo *downranking*, o que compromete sua capacidade de acionar mecanismos de contestação. Similarmente, o *age-gating* e a adição de avisos podem obstaculizar o acesso do conteúdo a segmentos demográficos inteiros, o que também suscita preocupações severas relativas a transparência, liberdade de expressão e não-discriminação.

Ademais, as normas que determinam a aplicação de restrições intermediárias a conteúdos específicos usualmente são parte das mesmas políticas de comunidade que regulam a remoção de conteúdo. Presumivelmente, os sistemas e equipes que aplicam tais normas também são os mesmos empregados pelas plataformas para a detecção e análise dos conteúdos passíveis de remoção. As preocupações com violações de direitos favorecidas pela opacidade desses sistemas são, portanto, igualmente aplicáveis às restrições intermediárias. Por esses motivos, entendemos que a estrutura atual dos princípios seja aplicada não apenas a suspensões de contas e remoções de publicações, mas sejam estendidas para abordar restrições intermediárias.

4. De que maneiras você já utilizou os Princípios de Santa Clara como ferramenta ou recurso de advocacy? Se você estiver confortável em compartilhar, por favor inclua os links para quaisquer exemplos que você tiver.

Desde 2019, o IRIS desenvolve um projeto de pesquisa sobre transparência em moderação de conteúdo em plataformas. O projeto atende a uma demanda crescente por maior compreensão dos padrões de comunidade desses serviços, seus termos e políticas de uso, sobre o conteúdo gerado por seus usuários. Nesse sentido, os Princípios de Santa Clara são usados como fonte de referência para padrões internacionais nesse tema, assim como o Relatório da ONU sobre Liberdade de Expressão (2018).

Atualmente, o Brasil discute o projeto de lei 2630 para tratar de desinformação, como uma problema amplamente reconhecido no país. Existem propostas diversas, mas a mais debatida e possivelmente votada pelo Senado no dia 30/06 é denominada “Lei da Liberdade, Transparência, Responsabilidade na Internet”, do senador Alessandro Vieira. Já existem variadas versões deste projeto e, por isso, optamos por comentar os assuntos em vez dos artigos. Nossa contribuição está disponível aqui os Princípios de Santa Clara baseiam algumas de nossas observações sobre transparência e a responsabilidade dos intermediários sobre o conteúdo de desinformação e sua remoção das plataformas.

5. Como os Princípios de Santa Clara podem ser mais úteis na sua atividade de advocacy em torno destas questões daqui para frente?

Seria interessante que os princípios tivessem mais alcance para que servissem não apenas de parâmetros para empresas, mas também para reguladores. Como

nossas atividades envolvem o acompanhamento de iniciativas legislativas que incluem moderação de conteúdo, seria interessante que os princípios também se direcionassem à regulação dessa matéria. Para tanto, acreditamos ser necessário mais divulgação, especialmente em português, dos princípios e oportunidades de capacitação de tomadores de decisão estatais sobre as direções por eles estabelecidas.

6. Você acredita que os Princípios de Santa Clara deveriam ser aplicáveis à moderação de propaganda, para além da moderação do conteúdo não-comercial gerado por usuários? Se sim, você acredita que eles devem ser aplicados em todo ou apenas em parte?

Como conteúdos publicitários, de maneira geral, contam com um agente econômico como responsável, não parecem ser equiparáveis a conteúdo gerado por usuário, e sua moderação pode atender a critérios distintos, sendo mais factível a responsabilização do anunciante em caso de dano. Como os anunciantes desempenham papel diferente do usuário nas plataformas, sendo que os anúncios são rastreáveis e distinguíveis de outros tipos de conteúdo, bem como os criadores podem ser localizados e identificados, a moderação de publicidade pode respeitar as regras (regulação ou autorregulação nacional, estadual, municipal, local) existentes nos locais onde é veiculada.

A moderação de publicidade, no Brasil, é delegada ao CONAR (Conselho Nacional de Autorregulamentação Publicitária) e está sujeita a autorregulação interna à área, motivo pelo qual as plataformas não podem ser responsabilizadas por moderar/remover conteúdo publicitário de anúncios. O procedimento para denunciar ou questionar a suspensão de circulação de conteúdo publicitário que não seja gerado pelo usuário é estabelecido por essa entidade pelo Código Brasileiro de Autorregulação Publicitária, baseado em moderação reativa, e não proativa, de forma a evitar controle prévio sobre anúncios.

7. Há alguma parte dos Princípios de Santa Clara que você entende como pouco clara ou difícil de entender?

- “discrete post” é uma expressão que parece ter sido usada no seu significado técnico. Se significa “número inteiro”, como em uma expressão matemática, sugerimos substituir por “quantidade exata” ou alguma expressão menos técnica.

- “category of rule violated” é uma expressão aberta e que não seguiu o padrão de outras expressões nos princípios, de exemplificar o universo possível. Poderia haver exemplos entre parênteses indicando o que é considerado uma categoria.

- a divulgação de “locations of flaggers and impacted users” não fornece contexto e qual o parâmetro de divulgação. Poderia estar explícita a importância de saber a localização de usuários impactados ou “flaggers”, ademais de ser necessário indicar a especificidade dessa localização (país, estado, região...).

- “independent external review processes” é um termo que pode se referir a múltiplos agentes, como empresas de auditoria, tribunais ou verificadores de fatos; seria interessante especificar ou exemplificar.

8. Existem riscos específicos aos direitos humanos que os Princípios de Santa Clara poderiam ajudar a mitigar ao encorajar que as empresas forneçam outros tipos de dados adicionais? (Por exemplo, existe algum tipo de campanha maldosa de sinalização de conteúdo que não estaria visível nos dados exigidos pelos Princípios de Santa Clara, mas que estariam visíveis se adicionássemos uma nova coluna para tratar deste tema?)

Os princípios deveriam incluir de forma explícita a proteção da privacidade e dados pessoais dos usuários das plataformas em medidas de transparência. Além disso, deveriam recomendar, no que se refere à regulação, processos democráticos de tomada de decisão e que seja evitada a abordagem criminal em matéria de moderação de conteúdo, a fim de não deflagrar autocensura, entre outras violações à liberdade de expressão.

9. Você entende que existem considerações regionais, nacionais ou culturais que não estão refletidas nos Princípios de Santa Clara, mas que deveriam estar?

A revisão humana de decisões deve levar em conta a diversidade do corpo revisor, tendo em consideração a diversidade de países, culturas e legislações que permeiam os contextos dos usuários de plataformas online. É necessário reforçar a importância do domínio das línguas do conteúdo revisado, considerando as granularidades do discurso e a possibilidade de identificar usos sociais da linguagem, como ironia, sarcasmo, humor, etc.

10. Você entende que existem considerações para empresas de pequeno e médio porte que não estão refletidas nos Princípios de Santa Clara, mas que deveriam estar?

Os princípios 1 e 2 podem ser mais facilmente implementáveis na medida em que seus custos principais se relacionam a alterações no software e podem, mesmo, ser incluídos ao desenvolvimento das plataformas na fase inicial. Os custos de recursos humanos relacionados à implementação do princípio 3, por outro lado, podem ser significativamente maiores, pois exigem o provimento de uma segunda instância de revisão humana do conteúdo. Os custos de manutenção dessa segunda instância podem variar sensivelmente de acordo com o volume de usuários da plataforma. Assim sendo, é possível que a capacidade econômica de implementação do princípio 3 varie de acordo com os volumes de usuários, des publicações, de contestações enviadas à plataforma e de acordo com seu faturamento. Desse modo, há possibilidade de que esse princípio impacte, de forma significativa, o ambiente concorrencial em que a atividade se desenvolve. Por isso, entendemos que possibilidade de proporcionalidade do princípio para empresas pequenas e médias ou de definição de padrões mínimos deve ser estudada, e tanto a decisão de flexibilizá-lo ou não quanto os possíveis critérios de definição do que conta como pequena e média empresa no contexto dos princípios devem se fundamentar em análises científicas sobre os impactos econômicos de medidas dessa natureza.

11. Você teria recomendações para garantir que os Princípios de Santa Clara permaneçam viáveis, factíveis e relevantes no longo prazo?

Consideramos interessante a iniciativa de receber comentários e sugestões periódicas, como a presente consulta, a fim de viabilizar a atualização dos princípios. Dessa forma, seria recomendável realizar chamadas para opinião da comunidade

de forma anual. Também sugerimos que os princípios abarquem diretrizes sobre conteúdo de políticas de comunidade, de modo que seja possível visualizar sua adesão, ou não, pelas plataformas, por meio de uma ferramenta de comparação.

12. Quem você recomendaria que participasse em consultas futuras sobre os Princípios de Santa Clara? Se possível, por favor compartilhe seus nomes e endereços de e-mail conosco.

Recomendaríamos entidades em destaque no acompanhamento e aconselhamento do cenário regulatório e social envolvendo internet em países externos ao contexto norte-americano e europeu, que são pouco representadas nas discussões sobre expressão e conteúdo online. Em específico no Brasil, indicamos a Coalizão Direitos na Rede. Ela é uma articulação composta por mais de trinta e cinco entidades comprometidas com o avanço dos direitos digitais no Brasil, entre elas o instituto de pesquisa de que fazemos parte, Instituto de Referência em Internet e Sociedade. Sua atuação é pautada pela defesa da liberdade de expressão, da privacidade, da transparência e de uma internet segura, universal e aberta. O contato pode ser feito com Fabricio Solagna, secretário-geral da CDR. contato@direitosnarede.org.br O IRIS é membro da Coalizão. Link: https://direitosnarede.org.br/p/declaracao-cdr/

13. Se os Princípios de Santa Clara exigissem a divulgação de informações sobre o background cultural ou de treinamento dos moderadores de conteúdo empregados por uma plataforma, o que você gostaria que as plataformas dissessem a este respeito? (Por exemplo: divulgar o percentual de moderadores que passou por um teste de idiomas, de acordo com os idiomas em que a plataforma estiver moderando conteúdo, ou divulgar que todos os moderadores passaram por um tipo de treinamento específico.)

O recente escândalo de supressão discriminatória de publicações na moderação da plataforma TikTok reafirmou a importância da transparência no processo de treinamento dos moderadores de conteúdo. Naquele caso, ficou evidenciada a disparidade entre o que era afirmado nas políticas de comunidade que a empresa divulgava ao público e o que os moderadores eram ensinados em documentos internos e instrumentos pedagógicos da companhia, assim como relatado acerca de outras plataformas por Sarah West (Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366-4383.) e por Sarah Roberts (Roberts ST (2014) *Behind the screen: the hidden digital labor of commercial content moderation*. PhD Thesis, University of Illinois, Chicago, IL.). Para reduzir essa lacuna e maximizar a *accountability* sobre as práticas de moderação das empresas, entendemos que o processo de educação e treinamento dos moderadores deve ser tão público quanto possível, o que inclui a divulgação detalhada de todas as etapas do processo e o compartilhamento dos documentos e materiais didáticos empregados nele. Dados quantitativos referentes ao perfil dos moderadores, resguardadas sua privacidade e proteção de seus dados pessoais, também devem ser divulgados, inclusive a porcentagem dos aprovados no teste de proficiência no idioma moderado.

14. Você teria alguma sugestão adicional?

Sugerimos que os comentários recebidos nessas chamadas para contribuições sejam publicados, ainda que de forma anonimizada, para que haja transparência no processo de consideração dos apontamentos. Além disso, a disponibilização do material que inclui diversas perspectivas seria interessante para acesso das partes interessadas,

inclusive, como fonte de pesquisa. Outra sugestão é a abertura de chamada pública para a participação em eventos de debate sobre os princípios, a exemplo do COMO Summit.

15. Eventos como o COVID-19 aumentaram sua consciência sobre necessidades específicas de transparência e responsabilidade ou deficiências dos Princípios de Santa Clara?

A pandemia da COVID-19 desencadeou, no Brasil, uma mobilização por leis de combate à desinformação. O problema ganhou destaque midiático e político devido às desinformações que circulam na internet sobre curas, medicamentos e formas eficazes de prevenção ou de distanciamento social. O legislativo brasileiro está elaborando propostas de lei para lidar com esse tipo de conteúdo na internet e há a tentativa de uma tramitação rápida e sem abertura participativa ou a devida publicidade. O texto de um dos projetos (PL 2630) foi levado a votação por três vezes sem discussão de seu conteúdo, com urgência em sua tramitação, dada por esse contexto. Por isso, sofreu numerosas propostas de emendas e ainda não há nitidez sobre o conteúdo efetivamente em debate ou que será votado. Outras propostas estão sendo discutidas em paralelo, com outras numerações. Entretanto, diversas delas são de controle proativo, pelas plataformas, de alguns tipos de conteúdo online, ou então de intervenção de autoridades governamentais na definição da veracidade do conteúdo que circula pelas plataformas online, bem como medidas de divulgação de dados pessoais ou exposição de usuários infratores, ou diminuição do acesso de usuários a serviços como redes sociais por meio de sua burocratização, o que gera perigos e insegurança sobre a liberdade de expressão, abrindo a possibilidade de vigilância massiva.

Esses projetos de lei, por sua vez, ganham força frente à insatisfação com a atual forma de resposta e controle de campanhas de desinformação e a incerteza quanto às medidas efetivamente adotadas pelas plataformas para lidar com essa questão. Assim, ficou evidenciada a necessidade de maior transparência quanto a critérios, métodos, ferramentas e garantias aos usuários sobre as medidas interventivas adotadas em relação a conteúdo online. Medidas que permitam identificar facilmente abusos de autoridades em pedidos de remoção de conteúdo, ou erros da própria plataforma, permitissem ao usuário recorrer e ter acesso ao devido processo quando seu conteúdo é questionado, e que tivessem abertura para questionamento da sociedade, são relevantes para contrapor propostas que, a fim de lidar com essas incertezas, impõem monitoramento proativo e responsabilização de terceiros por conteúdo online. A falta de transparência gera incerteza sobre as medidas adotadas e deixa margem para propostas de regulação que tolhem direitos dos usuários e arrisca restringir o potencial da internet, ainda mais em um cenário no qual os reflexos da desinformação são quantificáveis e sentidos de forma rápida, como o da pandemia.