http://dx.doi.org/ 10.18593/ejjl.v19i2.16496

DISCURSO DE ÓDIO PELO FACEBOOK: TRANSPARÊNCIA E PROCEDIMENTOS DE CONTENÇÃO

HATE SPEECH ON FACEBOOK: TRANSPARENCY AND CONTAINMENT PROCEDURES

Maria Edelvacy Pinto Marinho[1]
Stella Regina Coeli de Souza[2]

Resumo: No presente artigo teve-se por escopo investigar quão transparentes se apresentam os procedimentos pelos quais passa um usuário do Facebook denunciado por propagação de discurso de ódio na referida plataforma, até que, eventualmente, receba uma sanção por tal comportamento. Para tanto, foram analisados documentos disponibilizados publicamente pela aludida rede social em sua página a fim de se buscarem dados acerca da qualificação dos moderadores, sanções aplicáveis especificamente àqueles que incorrem em cometimento de hate speech e o procedimento de análise das denúncias recebidas pelos examinadores. A pesquisa é de base bibliográfica documental e foi desenvolvida a partir do raciocínio indutivo. Concluiu-se que os mecanismos disponibilizados pelo Facebook para controle da propagação do discurso de ódio em sua rede necessitam de aperfeiçoamento tanto quanto ao grau e clareza dos critérios definidores de discurso de ódio aplicados aos usuários, quanto ao grau de transparência relativo ao julgamento pela própria rede no que toca à referida prática e à política de exclusão da conta do usuário.
Palavras-chave: Liberdade de expressão. Discurso de ódio. Facebook. Sanções. Transparência.

Abstract: The purpose of this article was to investigate how transparent the procedures are presented to a Facebook user denounced for the propagation of hate speech on the platform, until he eventually receives a sanction for such behavior. To reach this goal, documents publicly disclosed by the aforementioned social network on its page were analyzed in order to seek data on the qualification of the moderators, sanctions applicable specifically to those who incur in hate speech, and the procedure for analyzing the complaints received by the examiners. The research is based on documentary bibliography and was developed from the inductive reasoning. It was concluded that the mechanisms provided by Facebook to control the spread of hate speech in its network need to be improved concerning both the length and clarity of hate speech criteria applied to users and the length of transparency regarding the judgment by the social network related to this practice and the policy of termination of the user's account.
Keywords: Freedom of speech. Hate speech. Facebook. Sanctions. Transparency.

[1] Doutora em Direito pela Universidade Paris 1- Panthéon Sorbonne; Mestre em Direito pelo Centro Universitário de Brasília; Professora no Curso de Direito da Universidade Presbiteriana Mackenzie; mariaedelvacy@gmail.com; https://orcid.org/0000-0002-6957-3099

[2] Mestre e Graduada em Direito pelo Centro Universitário de Brasília; Pesquisadora-Discente no Centro Universitário de Brasília; SEPN, 707/907 Via W 5 Norte, Asa Norte, 70790-075, Brasília, Distrito Federal, Brasil; stellaregina@gmail.com; https://orcid.org/0000-0002-4269-7024

Introdução

M. S., colunista da revista Vice, divulgou no blog da publicação um artigo no qual descrevia como o Facebook o havia bloqueado em razão de uma postagem na aludida rede social, a qual continha o vocábulo "bicha" (*faggot*, na língua inglesa). Causou estranheza em M. S. tal acontecimento, uma vez que se declara homossexual e utilizador assíduo da referida palavra, "assim como negros utilizam o termo que começa com N e ativistas gays diziam *queer* no início dos anos 90." (SUNDERLAND, 2014). Descreveu que, à ocasião que levou à punição em comento, se referiu a um amigo, também homossexual, como "bicha", tendo o indivíduo "curtido" o comentário. Algumas horas depois, o Facebook sancionou M.S. com a pena de impossibilidade de realização de postagens por sete dias.

Incomodado com a situação, M.S. tentou contato com os administradores da rede social em comento a fim de solicitar o desbloqueio – frisando que o site sequer informa aos usuários por quanto tempo a punição se alonga –, sem que obtivesse resposta.[3]

Em outra circunstância, um usuário do Facebook, R.C., utiliza a Comunidade de Ajuda para buscar entender o motivo pelo qual os administradores da rede consideraram o uso da palavra "estúpido" como ato apto a provocar um bloqueio de sua conta por 30 dias – mesmo que ele próprio já tenha sido ignorado ao denunciar diversas páginas por conteúdos que contrariavam os Termos de Uso da comunidade, segundo conta (FACEBOOK, 2017i). Não houve, ao menos na mencionada página, resposta por parte do Facebook para a demanda em questão.

Os cenários descritos até aqui costumam causar intensa perturbação aos usuários em razão da experiência proporcionada pelo Facebook ao longo de sua existência: afinal, criou-se ali uma plataforma para interação entre utilizadores e diversos grupos de que façam ou desejem fazer parte – os quais podem incluir amigos de agora ou outrora, familiares, colegas de trabalho, conhecidos ou desconhecidos que partilhem interesses em comum ou que estejam em busca de aprendizado e debate acerca de temas diversos –, cujos corpo e alma são compostos pela liberdade de expressão. É esse o contexto no qual, dia após dia, ano após ano, os utilizadores constroem as mais diversas redes de relacionamentos interpessoais e registram as próprias histórias.[4]

Em uma configuração assim, é de se verificar, também, o uso do Facebook como veículo disseminador de discursos de ódio. Os relatos descritos no início desta pesquisa emergem de tentativas de contenção de tal mau uso da plataforma – o qual guarda em si o potencial de, no mundo off-line, causar dano ao alvo do ataque, tanto pela via da violência psicológica, quanto pela imposição de agressões físicas. Embora se trate, indubitavelmente, de uma postura necessária à proteção dos indivíduos e da própria comunidade, observa-se, pelo teor das aludidas histórias, que

[3] *"Facebook never explains how long they will block someone, so I messaged the site to explain my situation and appeal the decision. 'I am a gay man, yet you have blocked me from saying faggot,' I wrote. 'You post HIV ads on my timeline because I'm gay, which is, you know, offensive, but you won't let me say faggot, which is gay men's word to use. Please unblock me'."* (SUNDERLAND, 2014).

[4] Acerca de efeitos ditos devastadores sobre a vida profissional e pessoal de uma pessoa que tem a conta do Facebook bloqueada, veja-se York (2016).

os gestores do Facebook ainda enfrentam problemas no que toca à diferenciação entre discursos de ódio e interações pacíficas entre os usuários, acabando por punir sujeitos a quem talvez sequer tenham sido concedidas explicações ou o direito à promoção de suas defesas por meio do exercício do contraditório.

O dano causado aos usuários e a responsabilidade do Facebook em desenvolver ferramentas que reduzam a propagação de discurso de ódio só podem ser compreendidos a partir do grau de penetração que essa rede social alcança entre os membros da sociedade. Um bom parâmetro consiste no número de usuários cadastrados na rede. No dia 02 de junho de 2017, o Facebook anunciou que havia atingido dois bilhões de usuários (FACEBOOK, 2017a). O grau de penetração da referida rede social e seu uso como um dos principais canais de comunicação entre seus usuários torna seu papel na concretização do direito à liberdade de expressão relevante.

Tendo essa conjuntura como pano de fundo de análise, no presente estudo tem-se por escopo descrever o atual estado da arte em que se encontra o Facebook no que diz respeito à transparência com que divulga detalhes acerca do processo pelo qual o usuário passa até que receba dada punição quando incorre no cometimento de discurso de ódio: quem realiza tais decisões e com base em quais critérios? Quais são as possibilidades de contraditório oferecidas ao indivíduo? Quais são as punições possíveis? O que define, de fato, a medida de tais sanções?

Para tanto, far-se-á a análise dos Termos de Uso publicamente fornecidos pela rede social, bem como outros informes e documentos por ela expedidos no sentido de aclarar detalhes sobre o procedimento utilizado para apenar aqueles que sejam considerados como propagadores de discurso de ódio na plataforma. Parte-se da premissa de que o Facebook se constitui em um sistema que segue uma lógica autônoma, o qual se desenvolve de acordo com seu próprio *modus operandi*, de maneira que tende a "definir suas próprias condições de operação, produzindo suas próprias regras e procedimentos normativos, forjando suas próprias racionalidades e estabelecendo seus próprios códigos comunicativos." (FARIA, 2010, p. 33). O que se pretende aqui é, portanto, verificar quais são os procedimentos relativos à análise e à punição a discursos de ódio expressos nos documentos expedidos pelo Facebook, de maneira a compreender, sob a ótica da transparência, sua real efetividade quanto ao combate a ataque a grupos vulneráveis pela via do *hate speech*. Para atingir tal desiderato, faz-se necessário entender que espaço o Facebook tem ocupado como definidor dos limites à liberdade de expressão no ambiente digital, conhecer as recentes declarações do Facebook sobre o que entende ser discurso de ódio, passando-se, a seguir, ao exame propriamente dito dos ritos em comento.

1 Liberdade de expressão no ambiente digital: a construção jurídica de seus limites através do reconhecimento do discurso de ódio

A limitação ao direito à liberdade de expressão é tratada como exceção à regra em sociedades democráticas. No caso do discurso de ódio, essa limitação ganha contornos mais visíveis a partir da

Segunda Guerra Mundial (KÜBLER, 1998, p. 335-336) – fato esse explicado pelo papel da propaganda nazista na disseminação de suas ideias baseadas na supremacia de uma raça em detrimento às demais. Se no período da Guerra o discurso de ódio era disseminado pelos próprios Estados, no pós-Guerra os seus propagadores são representados por um grupo de indivíduos. A mensagem preconceituosa não aparece de maneira tão direta: por vezes vem mascarada por justificativas baseadas em estatísticas retiradas de um contexto específico e generalizadas como verdade universal (ROSENFELD, 2003, p. 6).

As limitações à liberdade de expressão são construções jurisprudenciais, refletidas e dosadas a partir do caso concreto. Considerada um direito humano pela Declaração Universal de Direito do Homem (art. 19),[5] a liberdade de expressão ganha contornos próprios a depender do país que a aplique. Apesar do reconhecimento dos direitos expressos na referida Declaração ser compartilhado entre os países signatários, a interpretação sobre seu conteúdo pode divergir. Do mesmo modo, o tratamento dado ao discurso de ódio pelos Estados também é tema de acordos internacionais; citam-se como exemplos: Convenção Internacional Sobre a Eliminação de Todas as Formas de Discriminação Racial (artigo 4º), Pacto Internacional dos Direitos Civis e Políticos (artigo 20), Convenção Americana sobre Direitos Humanos (art. 13.5) e Convenção Europeia dos Direitos do Homem (art. 10.2).

Os direitos humanos podem ser considerados universais, mas o grau de integração normativa entre os Estados-membros não garante a unificação do seu conteúdo. O sentido dado a cada direito estaria protegido por uma "margem nacional de apreciação"[6] que cada Estado, a partir de sua cultura e sistema jurídicos, estaria apto e legitimado a definir. Esse é um traço comum dos direitos humanos. Se seu próprio conteúdo não é uniforme, as limitações ao seu exercício não poderiam ser diferentes.

Nos EUA, a liberdade de expressão é protegida pela Primeira Emenda à Constituição. O objetivo do dispositivo é proibir a criação de leis que visem a limitar seu exercício. A visão é a de que o Estado seria a maior ameaça à liberdade dos indivíduos, e o papel da Constituição seria proteger o cidadão, restringindo a atuação do Estado. Não há, contudo, leis infraconstitucionais que regulem o discurso de ódio. Quando o discurso parte de um particular, a interpretação dada pela jurisprudência americana tem valorizado a liberdade de expressão em detrimento a outros direitos, como a igualdade. Por consequência, essa concepção libertária acaba por permitir a difusão de discursos que estigmatizam minorias – seja por sua raça, etnia, religião, gênero ou orientação sexual (KÜBLER, 1998, p. 335-336). Tal posicionamento pode ser explicado pela valorização do individualismo na construção da sociedade e cultura americana (SARMENTO, 2006), que eleva a proteção das liberdades, e dentre elas a de expressão, a um patamar superior ao concedido a outros direitos.

Na Alemanha, o conflito entre liberdade de expressão e os direitos da personalidade dos ofendidos é resolvido através da técnica da ponderação de interesses, tendo como referência a

[5] "Todo o indivíduo tem direito à liberdade de opinião e de expressão, o que implica o direito de não ser inquietado pelas suas opiniões e o de procurar, receber e difundir, sem consideração de fronteiras, informações e ideais por qualquer meio de expressão." (ORGANIZAÇÃO DAS NAÇÕES UNIDAS, 1948).

[6] Sobre o tema, ver Delmas-Marty (2006).

supremacia da dignidade da pessoa humana (SARMENTO, 2006). É analisado o conteúdo do discurso questionado, o quanto o debate proporcionado é de interesse público e o quanto é ofensivo aos direitos da personalidade de minorias (SARMENTO, 2006). A judicialização quanto à constitucionalidade dos discursos negacionistas do Holocausto foi decisiva para a construção do tratamento jurídico do discurso de ódio pela Suprema Corte Alemã e nos dispositivos normativos infraconstitucionais (KÜBLER, 1998, p. 342). Além do tratamento constitucional, o discurso de ódio é regulado pelos sistemas administrativo, civil e penal (KÜBLER, 1998, p. 342).

Na França, também há dispositivos precisos de repressão à promoção do discurso de ódio. Errera (1992, p. 144-159) divide a história do tratamento jurídico do discurso de ódio na França em três períodos: 1939, 1945-1972 e 1972-1992. Em 1939 se aprova um decreto legislativo em que se acrescenta um outro tipo de difamação: aquela proferida contra grupos em razão de sua raça ou origem que tenham por objetivo incitar o ódio. Em razão da eclosão da Segunda Guerra Mundial, o dispositivo foi pouco efetivo para os objetivos que se propôs. Nas palavras de Errera (1992, p. 144), "muito pouco, muito tarde." No período pós-Segunda Guerra, as iniciativas governamentais de punição do discurso de ódio podem ser consideradas brandas. Empregavam-se, ainda, os dispositivos criados em 1939. Para ser aplicada à norma dever-se-ia provar a intenção de provocar ódio entre grupos, o que dificultava seu uso (ERRERA, 1992, p. 146). No período de 1972-1992 o Estado Francês passa a reconhecer a gravidade da conduta e modifica sua legislação de modo a respeitar as obrigações advindas da assinatura da Convenção internacional sobre a eliminação de todas as formas de discriminação racial. A lei sobre a luta contra o racismo, de 01 de julho de 1972, modificou estatutos em vigência, como a lei de impressa e de associações e dispositivos que passaram a permitir que associações postulassem em casos envolvendo discurso de ódio. Assim como na Alemanha, há previsão legal de instrumentos na esfera administrativa, civil e penal contra a propagação do discurso de ódio (HOFMANN, 1992, p. 159-170).

Com o desenvolvimento do processo de integração entre os países europeus, o tema passou a fazer parte de uma agenda comum, o que permitiu a harmonização, em certo grau, do tratamento jurídico das limitações à liberdade de expressão. Este pode ser compreendido como resultante de um conflito de direitos fundamentais. Em razão da Segunda Guerra Mundial e das ideologias propagadas durante o período, há o reconhecimento pelas instituições europeias do perigo da disseminação do discurso de ódio entre nacionais dos Estados-membros. Em 1997, o Comitê de Ministros emitiu uma recomendação aos Estados com linhas diretrizes sobre o tratamento a ser conferido às mídias no que se refere à propagação do discurso de ódio. Esse documento traz um conceito sobre o que vem a ser discurso de ódio e que pode servir de referência futura:

> o termo "discurso de ódio" deve ser entendido como abrangendo todas as formas de expressão que propagam, incitam, promovem ou justificam o ódio racial, a xenofobia, o antissemitismo ou outras formas de ódio baseadas na intolerância, incluindo a intolerância expressa na forma de nacionalismo agressivo e

> etnocentrismo, discriminação e hostilidade em relação a minorias, imigrantes e pessoas de origem imigrante. (COUNCIL OF EUROPE, 1997).

Busca-se, por meio dessa recomendação, encontrar "um equilíbrio entre a luta contra o racismo e a intolerância, com a necessidade de proteger a liberdade de expressão, a fim de evitar o risco de minar a democracia na tentativa de defendê-la." (COUNCIL OF EUROPE, 1997). Os Estados, dentro da margem de apreciação que lhes cabe, buscaram introduzir em suas legislações nacionais dispositivos no sentido de reconhecer o discurso de ódio, quando assim qualificado, como uma limitação legítima à liberdade de expressão. A década de 1990 é marcada pela aprovação de uma série de resoluções que visam a combater o racismo e a xenofobia no bloco.[7] Em 2000 é aprovada a diretiva relativa à aplicação do princípio da igualdade de tratamento entre pessoas sem distinção de origem racial ou étnica (CONSELHO DA UNIÃO EUROPEIA, 2000), e em 2009 entra em vigor a Carta dos Direitos Fundamentais da União Europeia, que em seu artigo 21 proíbe qualquer tipo de discriminação baseada em "sexo, raça, cor ou origem étnica ou social, características genéticas, língua, religião ou convicções, opiniões políticas ou outras, pertença a uma minoria nacional, riqueza, nascimento, deficiência, idade ou orientação sexual."

As normas aqui apresentadas são objeto de controle quanto à intepretação do equilíbrio entre liberdade de expressão e promoção da manifestação de ódio pelo Tribunal Europeu de Direitos Humanos. Este tem se manifestado em favor da limitação à liberdade de expressão em casos de discursos de ódio, como se pode apreender da decisão do caso Erbakan c. Turquia:

> [...] em princípio, pode ser considerado necessário em sociedades democráticas, sancionar ou mesmo prevenir todas as formas de expressão que propagam, incitam, promovem ou justifiquem o ódio baseado na intolerância (incluindo a intolerância religiosa), garantindo que as "formalidades", "condições", "restrições" ou "sanções" impostas são proporcionais ao legítimo objetivo perseguido (no que diz respeito ao discurso do ódio e apologia a violência.[8]

O direito à liberdade de expressão pode ser fundamentado na ideia de igualdade e no quanto esse tipo de discurso reduz e denigre um determinado grupo (FISS, 2005, p. 40). Além desse primeiro impacto à ideia de igualdade, o discurso de ódio tem efeitos no próprio exercício da liberdade e da democracia (FISS, 2005, p. 47). Há um efeito silenciador como resultado da propagação do discurso

[7] Dentre os documentos, citam-se: Resolução do Conselho e dos Representantes dos Governos dos Estados-membros, reunidos no Conselho, de 05 de outubro de 1995, relativa à luta contra o racismo e a xenofobia em matéria de emprego e assuntos sociais; Resolução de 23 de outubro de 1995, sobre a resposta do sistema educativo aos problemas do racismo e da xenofobia; Resolução do Conselho e dos representantes dos governos dos Estados-membros, reunidos no Conselho de 23 de Julho de 1996, relativa ao Ano europeu contra o racismo.

[8] De acordo com a Corte no caso Erbakan c. Turquia nº 59405/00, 06/07/2006: «*La présente affaire se caractérise notamment par le fait que le requérant a été sanctionné pour des déclarations qualifiées par les juridictions internes de 'discours de haine'. A cet égard, la Cour souligne que la tolérance et le respect de l'égale dignité de tous les êtres humains constituent le fondement d'une société démocratique et pluraliste. Il en résulte qu'en principe on peut juger nécessaire, dans les sociétés démocratiques, de sanctionner voire de prévenir toutes les formes d'expression qui propagent, incitent à, promeuvent ou justifient la haine fondée sur l'intolérance (y compris l'intolérance religieuse), si l'on veille à ce que les 'formalités', 'conditions', 'restrictions' ou 'sanctions' imposées soient proportionnées au but légitime poursuivi (en ce qui concerne le discours de haine et l'apologie de la violence, voir, mutatis mutandis, Sürek c. Turquie (no 1) [GC], no 26682/95, § 62, CEDH 1999-IV, et, notamment, Gündüz, précité, § 40).*» (EUROPEAN COURT OF HUMAN RIGHTS, 2006).

de ódio. Os grupos atacados se sentem menos propensos a expor suas opiniões, e mesmo quando o fazem, o ambiente no qual essa opinião será veiculada retira ou diminui a credibilidade de suas palavras (FISS, 2005, p. 47).

Como se pode notar pelos dados expostos, a construção das limitações à liberdade de expressão no que concerne à propagação do discurso de ódio difere entre os países, principalmente quando se compara as visões europeia e norte-americana. Trata-se de um tema sensível, objeto de apreciação judicial de cortes constitucionais e cujo equilíbrio é definido e analisado a partir de critérios objetivos e subjetivos e de acordo com elementos de natureza cultural.

O objetivo de se apresentarem tais diferenças neste artigo é apenas reforçar a tese, aqui apresentada, quanto à legitimidade e ao papel que o Facebook e outras redes sociais têm desempenhado como órgão julgador dos critérios definidores desse equilíbrio e os impactos que a visão desses grupos pode apresentar para o conceito de democracia, já que pode exercer um efeito silenciador quanto ao conteúdo e à forma de apresentação dos interesses e ideias de minorias.

O fenômeno aqui estudado reflete uma mudança quanto à gestão do exercício da liberdade de expressão no que concerne ao grau de unificação promovido pelas redes sociais, em especial o Facebook. A forma como as redes sociais têm realizado o controle sobre o conteúdo postado não toma os devidos cuidados com a identificação real do sentido dessa comunicação e tampouco se observa uma relação clara entre a "ofensa" e a punição aplicada. Como o reconhecimento do que seria discurso de ódio é feito em grande parte por meio da identificação de palavras-chave, perde-se o sentido real que pode ser compreendido apenas pelo contexto. Entende-se que a solução ofertada pelo sistema jurídico, via processo judicial, não é adaptada ao contexto no qual o Facebook fornece seus serviços. Tal dificuldade de adaptação é evidente quando se toma como exemplo o elemento tempo. Entre a apresentação da demanda, seu exame e a execução da decisão judicial, a propagação do discurso de ódio atinge, via rede social, uma escala que torna impossível a sua exclusão do ambiente digital. Compreende-se que o volume de interações também torna utópico um controle sobre cada contexto no qual o suposto conteúdo de ódio foi postado e a definição caso a caso sobre a medida adequada e proporcional à ofensa.

O exame é, portanto, feito por softwares, única maneira de se viabilizar o controle com a velocidade e escala necessárias. Tais softwares utilizam palavras-chave e expressões pré-determinadas como referência para definição se as palavras postadas violam ou não a política de não propagação de discurso de ódio em suas plataformas. O controle pode ser feito também pelos próprios usuários através de um sistema de denúncias. Nos dois casos, faz-se necessário entender qual o espaço destinado à defesa do propagador do discurso e os critérios utilizados para sua censura.

O que chama atenção é que as diretrizes sobre o que venha a ser discurso de ódio, algo que tem sido objeto de construção jurisprudencial há décadas, venham sendo monopolizadas por uma organização privada que concentra a prestação de seus serviços de comunicação, tendo mais de dois bilhões de usuários de diferentes nacionalidades, sem que haja uma discussão pública sobre como o

discurso de ódio deverá ser tratado no ambiente digital. Na seção a seguir será examinado como o Facebook tem construído o referido conceito.

2 O conceito de discurso de ódio segundo o Facebook e o risco de subversão da regra tornada pública

Recentemente, surgiu como pronunciamento oficial do Facebook, em sua *Newsroom* (FACEBOOK, 2017e), um manifesto assinado por Richard Allan,[9] no qual foram abordados diversos itens atinentes ao discurso de ódio e aos desafios enfrentados pelos gestores da rede social no que se refere ao seu combate.

Merece destaque, de início, a tentativa de se buscar conceituar o que o Facebook considera *hate speech*: trata-se de ataques diretos a indivíduos que sejam motivados pelas "características protegidas" – raça, etnia, nacionalidade de origem, afiliação religiosa, orientação sexual, sexo, gênero, identidade de gênero, doenças ou incapacidades físicas.[10] Menciona-se, além disso, que não há uma definição universalmente válida a esse respeito.[11]

Outro ponto de relevo alude ao número de *posts* deletados por terem sido reportados como *hate speech*: em um período de dois meses, foram contabilizadas cerca de 66.000 exclusões por semana, ou 288.000 por mês.[12] Reforça-se que incitações diretas à violência contra características protegidas ou degradações e "desumanizações" contra outrem são tidas como discurso de ódio e levam à supressão das postagens com esse teor.[13]

Por outro lado, admite-se que há situações em que não há consenso sobre a ocorrência, ou não, de *hate speech*, em decorrência da ambiguidade das palavras empregadas, do desconhecimento acerca da intenção da utilização de tais termos e da falta de clareza quanto ao contexto em que o fato analisado ocorreu.[14]

No que toca à análise contextual de postagens reportadas como *hate speech*, diz-se que palavras que são comumente utilizadas em um país, sem que tenham conteúdo ofensivo, podem ser manejadas por outros sujeitos que tenham por fito perpetrarem ofensas a determinados grupos –

[9] Vice-presidente de Políticas Públicas do Facebook – Europa, Oriente Médio e África.

[10] *"Our current definition of hate speech is anything that directly attacks people based on what are known as their 'protected characteristics' — race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, gender identity, or serious disability or disease."* (FACEBOOK, 2017e).

[11] *"There is no universally accepted answer for when something crosses the line. Although a number of countries have laws against hate speech, their definitions of it vary significantly."* (FACEBOOK, 2017e).

[12] *"Over the last two months, on average, we deleted around 66,000 posts reported as hate speech per week — that's around 288,000 posts a month globally. (This includes posts that may have been reported for hate speech but deleted for other reasons, although it doesn't include posts reported for other reasons but deleted for hate speech)."* (FACEBOOK, 2017e).

[13] *"But it's clear we're not perfect when it comes to enforcing our policy. Often there are close calls — and too often we get it wrong. Sometimes, it's obvious that something is hate speech and should be removed – because it includes the direct incitement of violence against protected characteristics, or degrades or dehumanizes people. If we identify credible threats of imminent violence against anyone, including threats based on a protected characteristic, we also escalate that to local law enforcement."* (FACEBOOK, 2017e).

[14] *"But sometimes, there isn't a clear consensus — because the words themselves are ambiguous, the intent behind them is unknown or the context around them is unclear. Language also continues to evolve, and a word that was not a slur yesterday may become one today."* (FACEBOOK, 2017e).

exemplo citado no documento é o da palavra "kalar", que em Mianmar possui significação histórica e utilização rotineira, ao passo em que é empregada por budistas nacionalistas contra mulçumanos em discursos de ódio.[15]

Ainda acerca da contextualização, aduz-se que situações específicas podem deflagrar procedimentos internos de análise e adaptação das políticas de uso da rede social – a exemplo do que ocorreu quando houve aumento do influxo de imigrantes na Alemanha. Nesse cenário os gestores do Facebook incluíram no rol de discursos de ódio puníveis toda manifestação de violência ou desumanização contra tais indivíduos, embora tenham buscado preservar o espaço de debate acerca da imigração em si, enfatizando-se que a plataforma é comprometida com discussões factualmente legítimas.[16]

Admite-se, ainda, que o contexto pode indicar também a intenção do utilizador na postagem analisada – a qual pode ter sido construída com termos que denotassem discurso de ódio, mas que se trata, apenas, de uma piada autodepreciativa ou de uma letra de música.[17] Ponto de específico interesse para o presente artigo, a esse respeito, aduz à manifestação, por parte do Facebook, de que o julgamento é feito tendo por enfoque apenas os textos ou imagens marcadas como ofensivas às políticas de uso, sem que o cenário como um todo seja considerado.[18]

A análise do exposto leva a algumas reflexões relevantes para a presente pesquisa. A primeira delas alude à facilidade com que indivíduos ou grupos de sujeitos que desejem propagar o discurso de ódio podem se utilizar da regra posta pelos gestores do Facebook a fim de, ironicamente, transgredi-las. Ora, os próprios gestores reconhecem que dentro do universo de postagens *denunciadas*, nem sempre é possível realizar-se o efetivo controle de contextos e intenções, conforme mencionado alhures. Imagine-se, então, quão dificultosa se torna a moderação de conteúdo em um conjunto de 1,3 milhões de inserções *por minuto* diariamente (SIMPSON, 2017), dentro do qual seja possível que, por diversos mecanismos de alteração da linguagem utilizada, se torne inviável a identificação do *hate speech*.

O retrato descrito se aproxima com o que foi reportado pelo portal do jornal *The Guardian* acerca do combate ao terrorismo por parte do Facebook. Uma fonte familiar às políticas antiterroristas

---

15 *"Often the most difficult edge cases involve language that seems designed to provoke strong feelings, making the discussion even more heated — and a dispassionate look at the context (like country of speaker or audience) more important. Regional and linguistic context is often critical, as is the need to take geopolitical events into account. In Myanmar, for example, the word "kalar" has benign historic roots, and is still used innocuously across many related Burmese words. The term can however also be used as an inflammatory slur, including as an attack by Buddhist nationalists against Muslims."* (FACEBOOK, 2017e).

16 *"When the influx of migrants arriving in Germany increased in recent years, we received feedback that some posts on Facebook were directly threatening refugees or migrants. We investigated how this material appeared globally and decided to develop new guidelines to remove calls for violence against migrants or dehumanizing references to them — such as comparisons to animals, to filth or to trash. But we have left in place the ability for people to express their views on immigration itself. And we are deeply committed to making sure Facebook remains a place for legitimate debate."* (FACEBOOK, 2017e).

17 *"There are times someone might share something that would otherwise be considered hate speech but for non-hateful reasons, such as making a self-deprecating joke or quoting lyrics from a song."* (FACEBOOK, 2017e).

18 *"People's posts on Facebook exist in the larger context of their social relationships with friends. When a post is flagged for violating our policies on hate speech, we don't have that context, so we can only judge it based on the specific text or images shared. But the context can indicate a person's intent, which can come into play when something is reported as hate speech."* (FACEBOOK, 2017e).

da rede social afirmou que tal controle é uma "missão impossível" em face da quantidade de material recebido pelos moderadores – volume esse que dá aos controladores cerca de 10 segundos para ignorarem ou excluírem o conteúdo denunciado (HOPKINS, 2017). Essa impossibilidade – ou limitadíssima possibilidade – de efetiva análise e controle conferiria aos líderes de organizações terroristas a possibilidade de subverterem as políticas de uso da rede a fim de dela se utilizarem para propagarem seus discursos – o que ocorreia, por exemplo, pela postagem de links ("o que torna a moderação extremamente difícil e o conteúdo 'muito resistente à censura'.") (HOPKINS, 2017). Não é difícil concluir-se que o mesmo raciocínio – e, portanto, o mesmo desafio – se aplica à proliferação de discursos de ódio *lato sensu*.

Abra-se, aqui, um parêntese quanto à recente divulgação de documentos, também pelo *The Guardian*, que contêm especificações direcionadas aos moderadores acerca da maneira como devem realizar o exame das postagens denunciadas (HOPKINS, 2017), a qual pode constituir-se em fato simultaneamente benéfico e maléfico. Dentre os efeitos benefícios, destaque-se a possibilidade de debate sobre tal regramento pela comunidade que a ela se submete – o que inclui, além dos usuários individualmente considerados, veículos de comunicação, grupos políticos e ONGs que tenham por escopo a luta pela proteção a direitos individuais e coletivos.

Tal discussão é inserida em um contexto de globalização, no qual se anuncia uma crise de funcionalidade e eficácia do direito positivo. Associado a isso, tem-se o surgimento de um novo modelo jurídico, que se destaca por suas feições pluralistas e é "pactuado por diferentes atores – empresas, fundações, associações comunitárias, entidades de classe, órgãos de representação corporativa e organizações não-governamentais." (FARIA, 2010, p. 21).

É exatamente nesse cenário de necessidade de interação de diversos sujeitos interessados na efetiva proteção dos indivíduos contra os ataques realizados via discursos de ódio que se verifica benéfico o acesso aos informes divulgados pelo *The Guardian*, no sentido de buscar-se a compreensão do procedimento a fim de aperfeiçoá-lo mediante o debate empreendido nesse sentido. Isso se torna ainda mais relevante ao atentar-se para o engajamento do Facebook na construção de uma "comunidade global" (ZUCKEBERG, 2017): em sua busca por regras aceitáveis em níveis mundiais, faz-se necessário que permita que os atores que compõem tal coletividade tenham voz ativa na sinalização de falhas que podem levar à subproteção, pela rede social, de determinados grupos. A via do vazamento de informações, entretanto, não é o caminho mais desejável para isso, competindo ao Facebook, na medida do possível, disponibilizar à sociedade tais procedimentos de forma pública em seus próprios canais.

Voltando-se ao problema da subversão das regras tornadas conhecidas pelo mencionado vazamento, verifica-se que a publicidade conferida aos procedimentos de análise de *hate speech*, proporcionada pelo vazamento ora comentado, tornou possível propagar-se o discurso de ódio sem que as especificações impostas pelo Facebook sejam quebradas. Isso porque o exame em questão se baseia nas seguintes combinações:

Categoria protegida + categoria protegida = Categoria Protegida

Categoria protegida + categoria não protegida = Categoria Não Protegida

As ditas categorias protegidas são aquelas defendidas pelo Facebook, contra as quais, objetivamente, proíbe o direcionamento dos discursos de ódio, baseadas, como já mencionado, em sexo, raça, identidade de gênero, afiliação religiosa, nacionalidade, etnia, orientação sexual, e na defesa a portadores de doenças e incapacidades graves (HATE SPEECH..., 2017). Dentre as categorias não protegidas estão classe social, aparência, religiões, ideologias políticas e países isoladamente considerados (uma vez que a proteção é direcionada ao povo, aos nacionais) (HATE SPEECH..., 2017).

Ilustrando-se exemplificativamente, tem-se que "homem branco" é categoria protegida, pois combina "homem" (categoria protegida) com "branco" (categoria protegida). Entretanto, tem-se que "crianças negras" se trata de grupo desprotegido, uma vez que "crianças" é categoria não protegida. Assim, ao combinar-se um termo incluso nas categorias não protegidas, traz-se um grupo originariamente protegido ao status de não proteção. Em termos práticos, torna-se possível aos propagadores de discursos de ódio que postem conteúdos materialmente abusivos, mas que não serão excluídos em razão da maneira como é realizado o procedimento de análise: descontextualizado e indiferente, via de regra, às intenções do autor, por força do pouco tempo despendido em tais exames em decorrência do volume de submissões realizadas (ANGWIN, 2017). A esse respeito, menciona Dave Willner, ex-integrante da equipe de moderação do Facebook: "em razão do volume de decisões – milhões por dia – a abordagem é muito mais utilitarista do que aquilo a que estamos acostumados em nosso sistema de justiça." (ANGWIN, 2017).

Assim, diante da extrema limitação para efetiva monitoração do mau uso da plataforma, é forçoso que se admita que trazer à tona um suposto conjunto de regras que venham a publicizar os contornos do que é considerado discurso de ódio para o Facebook não se traduz em medida efetiva em seu combate, nem significa que se estaria, necessariamente, lidando com transparência para com o usuário. Isso porque a problemática acaba sempre girando em torno de um mesmo eixo: a dificuldade diante das limitações (humanas e algorítmicas) de realizar-se o efetivo controle de todo o conteúdo postado na rede – ainda que só se considerasse o universo de postagens denunciadas como impróprias.

A próxima reflexão é resultado da anterior: em termos de transparência, ao invés de investir-se na delimitação pormenorizada do conceito de discurso de ódio e sua publicização, parece ser mais urgente, nessa linha, que sejam disponibilizados os procedimentos pelos quais um usuário pode vir a passar acaso uma postagem por ele realizada seja tida como *hate speech*. O que se defende aqui, em outras palavras, é que em termos de *accountability* é mais interessante ao usuário que tenha prévio conhecimento de dados como possíveis sanções e informes sobre os moderadores do que, efetivamente, do conceito de *hate speech*. Afinal, se por um lado o conceito de discurso de ódio guarda particularidades em sua aplicação necessárias à amplitude que se espera de uma concepção "universalizável", por outro, a relação utilizador-rede social é fixa, e não arbitrária – no sentido de ser a todos aplicável de modo igual e previsível –, a partir do momento em que se sabe exatamente como

será o trato dado ao utilizador caso incorra em discurso de ódio. Isso, inclusive, reforçaria os laços necessários à existência de uma comunidade, como o próprio Facebook se autodenomina.[19]

Nesse sentido, impende registrar-se que o combate ao discurso de ódio foi o centro das atenções, também, quando da assinatura de acordo, com tal intento, entre Facebook, Twitter, Youtube e Microsoft e Comissão Europeia. O objetivo central era assegurar que plataformas on-line não oferecessem oportunidade para propagação viral de *hate speech*.[20]

Esse tipo de "parceria normativa" entre grandes empresas e Estados apenas reconhece um fenômeno já identificado por juristas: a fragmentação do Direito.[21] Este pode ser entendido como é resultado do processo de globalização, do aumento do poder das empresas multinacionais, do enfraquecimento do Estado como único produtor de fontes normativas e da necessidade de um ambiente que se garanta um grau mínimo de harmonização normativa para viabilização das relações comerciais. A existência de fontes normativas diversas daquelas originadas via processo estatal caracteriza o cenário no qual o Facebook tem ditado as regras sobre o exercício da liberdade de expressão em sua plataforma. Em razão do processo de fragmentação, a produção normativa se torna cada vez mais especializada e setorizada (CAFAGGI, 2010).

Assumindo-se que as interações ocorridas em plataformas on-line como o Facebook cada vez mais têm repercussões concretas no "mundo não virtual", os Estados, sozinhos, não podem acolher reivindicações internas relativas ao discurso de ódio propagado em redes sociais, fazendo-se necessário o diálogo que resulta em compromissos como o ora em comento (FARIA, 2010, p. 37).

Reconhecendo-se, destarte, a impossibilidade de conter os efeitos do discurso de ódio utilizando-se apenas das vias jurídicas tradicionais, Estados e instituições que defendem interesses estatais se veem mobilizados a buscar, junto às companhias tecnológicas, consensos entre direitos que pretendem amparar e formas de fazê-lo em conjunto com tais empresas – uma vez que o direito positivo, sozinho, já não é capaz de lidar com tal realidade de maneira efetiva.[22]

---

[19] *"The company now has 2 billion users, which it calls its 'community'—an interesting choice of term, considering its rules and their enforcement actually serve to further divide, by creating different standards for different people."* (ONLINE CENSORSHIP TEAM, 2017).

[20] *"The IT Companies support the European Commission and EU Member States in the effort to respond to the challenge of ensuring that online platforms do not offer opportunities for illegal online hate speech to spread virally."* (EUROPEAN COMMISSION, 2016, tradução nossa).

[21] Para maior aprofundamento no assunto, ver: Teubner (1996, p. 3-28), Fisher-Lescano e Teubner (2004, p. 999-1046) e Koskenniemi e Leino (2002, p. 553-579).

[22] Neste sentido ver: "Com [...] o advento de matérias e situações novas e não padronizáveis pelos paradigmas jurídicos vigentes, o alcance do direito positivo tende a ser cada vez mais reduzido e a eficácia de suas normas a ficar cada vez mais frágil, limitada e volátil, como tem sido evidenciado, por exemplo, por sua incapacidade cada vez mais flagrante de limitar o uso da internet e regulamentar a comunicação virtual, que constituem um espaço essencialmente não-estatal e transterritorial. [...] limitação estrutural do direito positivo e de suas instituições judiciais diz respeito à discrepância entre seu perfil arquitetônico e a crescente complexidade do mundo contemporâneo. Suas normas tradicionalmente padronizadoras, com sequencias lógicas e binárias, editadas com base nos princípios da impessoalidade, da generalidade, da abstração e do rigor semântico e organizadas sob a forma de um sistema fechado, coerente e postulado como isento de lacunas e antinomias, são singelas demais tanto para apreender quanto para dar conta de uma pluralidade de situações sociais, econômicas, políticas e culturais cada vez mais funcionalmente diferenciadas." (FARIA, 2010, p. 47).

O exemplo referente ao acordo em comento traduz-se em uma solução não totalmente distanciada do direito positivo, mas que também não é nele completamente centrada: deseja-se o reforço da legislação estatal que hostiliza o *hate speech* nos ambientes on-line e off-line, o que ocorre por ações que não estão mais com fulcro em medidas puramente estatais, mas, sim, no intercâmbio complementar entre tais ações e aquelas que devem partir das empresas-parte no aludido Código de Conduta.[23] Busca-se conciliar o combate ao mau uso das redes sociais e a proteção aos alvos de discursos de ódio por meio do compromisso firmado entre os Estados representados pela Comissão Europeia – e as legislações aplicáveis ao combate ao *hate speech* – e os gestores das empresas de tecnologia – que se propõem, de acordo com o Código, a manter processos claros e efetivos de análise de denúncias relativas ao cometimento de tais discursos, de maneira a eliminarem conteúdos a isso afetos.[24]

Nessa fase inicial de aproximação entre Estados e redes sociais há o reconhecimento recíproco da necessidade de cooperação. Por se tratar do exercício de direitos fundamentais dos usuários das redes sociais, o grau de cooperação deve ultrapassar a primeira barreira do reconhecimento e da indicação de princípios e diretrizes para a criação de um espaço contínuo de elaboração normativa e fiscalização conjunta e colaborativa. Assim, o próximo passo da presente pesquisa é verificar quais são os procedimentos disponibilizados pelo Facebook para a supressão do discurso de ódio em sua plataforma, bem como quão transparentes são. É o que será explorado na seção a seguir.

3 Das penas possíveis e do procedimento pelo qual passa o usuário até a sanção: o que dizem os documentos expedidos pelo Facebook

O retrato apresentado no tópico anterior dá conta de que o direito positivo emanado pelo Estado não é suficiente para, sozinho, combater as práticas de discursos de ódio possíveis em uma plataforma como o Facebook. Esse, por sua vez, possui um conjunto de regras que constituem uma espécie de ordenamento "legal" próprio, o qual visa a estabelecer padrões aplicáveis globalmente.[25]

Isso inclui, certamente, procedimentos próprios, os quais visam a atender expectativas internas de construção de uma comunidade segura, nas quais se alicerça o Facebook. Apesar disso, conforme pontuado noutro momento, tais "ritos" são realizados sob intensa pressão, a qual se relaciona à quantidade de denúncias, prescindindo, muitas vezes, de exames satisfatórios no que se refere ao contexto e à intenção do autor.

[23] "*In order to prevent the spread of illegal hate speech, it is essential to ensure that relevant national laws transposing the Council Framework Decision 2008/913/JHA are fully enforced by Member States in the online as well as the in the offline environment. While the effective application of provisions criminalising hate speech is dependent on a robust system of enforcement of criminal law sanctions against the individual perpetrators of hate speech, this work must be complemented with actions geared at ensuring that illegal hate speech online is expeditiously acted upon by online intermediaries and social media platformsi*". (EUROPEAN COMMISSION, 2016).

[24] "*The IT Companies to have in place clear and effective processes to review notifications regarding illegal hate speech on their services so they can remove or disable access to such content.*" (EUROPEAN COMMISSION, 2016).

[25] "*But Facebook says its goal is different — to apply consistent standards worldwide. 'The policies do not always lead to perfect outcomes,' said Monika Bickert, head of global policy management at Facebook. 'That is the reality of having policies that apply to a global community where people around the world are going to have very different ideas about what is OK to share'.*" (ANGWIN, 2017).

A relevância da declaração que alude à precariedade ínsita à análise das supostas postagens revestidas de discurso de ódio reside na exata preocupação que esta pesquisa pretende expor: afinal, o pouco que se sabe sobre a avaliação de postagens que supostamente ofendem as políticas de uso do Facebook aponta para a realização de inspeções que, embora feitas de forma inconsistente e superficial, podem levar a sanções de exclusão da postagem, bloqueio e até desativação do perfil. Reconhecer a possibilidade de erro não torna tal relação menos arbitrária do que aquela que se apresenta hoje, conforme o verificado ao longo da pesquisa que originou o presente artigo. Os motivos para isso estão na ausência de regramento que apresente ao usuário quem realiza tais análises, bem como a relação de causa e efeito existente entre o mau uso da plataforma e as punições aplicáveis. Sequer existe a publicização das sanções possíveis e suas gradações.

Tem-se que a transparência quanto aos procedimentos e às sanções relativas ao cometimento de discurso de ódio tem importância para a relação usuário-rede social no sentido de efetiva possibilidade de autorregulação, sem que se faça necessária a intervenção estatal de conflitos que podem ser resolvidos pelas vias internas ao próprio Facebook.[26]

A relevância de cada fator mencionado para a transparência da relação usuário-rede social está descrita nas subseções a seguir, bem como o nível de clareza de informações disponibilizadas sobre tais aspectos.

3.1 **O fator humano**

No pronunciamento a respeito das medidas que visam a coibir o discurso de ódio de que já se tratou em outro momento deste artigo, Richard Allan manifestou-se acerca da quantidade de moderadores responsáveis pelo exame das postagens reportadas como abusivas: ao longo deste ano, 3.000 empregados serão adicionados ao time que já conta com 4.500 analistas.[27] Para que se tenha uma noção de proporcionalidade, as medidas antiterroristas contam com uma equipe de 150 especialistas que se dedicam exclusivamente ao tema.[28]

Além de tal numerário, pouco se sabe acerca do grupo ou da maneira como lhes são distribuídas as demandas. Não existem manifestações por parte do Facebook acerca de como são selecionadas tais pessoas, que tipo de formação possuem e que medidas são tomadas para que a

[26] "Essa é mais uma das facetas paradoxais da metamorfose que o Estado e suas instituições jurídicas vêm sofrendo ao longo das últimas décadas. Desregulamentação e deslegalização não significam menos direito. Significam, isto sim, menos direito positivo e menos mediação das instituições políticas na produção de regras, em benefício de uma normatividade emanada de diferentes formas de contrato e da tendência dos diferentes setores da vida social e econômica à autorregulação e autocomposição de conflitos." (FARIA, 2010, p. 60).

[27] *"We're building up these teams that deal with reported content: over the next year, we'll add 3,000 people to our community operations team around the world, on top of the 4,500 we have today."* (FACEBOOK, 2017e).

[28] *"At Facebook, more than 150 people are exclusively or primarily focused on countering terrorism as their core responsibility. This includes academic experts on counterterrorism, former prosecutors, former law enforcement agents and analysts, and engineers."* (FACEBOOK, 2017f).

análise dos *reports* não seja enviesada em razão de ideologias ou visões étnico-culturais próprias de tais indivíduos.

A necessidade de transparência quanto ao fator humano, nos termos anteriormente expostos, reside na certeza aos usuários de que as denúncias realizadas efetivamente chegam a *alguém* e que tal pessoa está preparada para *analisar* a questão com precisão e pouca margem para erros. Nesse cenário, mesmo a sensação de liberdade de que gozam indivíduos e grupos que desejem utilizar a plataforma para propagar discursos de ódio seria minorada pela soma da vigilância exercida pela comunidade – responsável por realizar as delações – e por um incremento na certeza de que atitudes serão tomadas em *todos* os casos.

No estado em que se encontra, é de se questionar acerca da real capacidade de análise do conteúdo abusivo – lembremo-nos das questões de intenção e contexto reportadas no tópico anterior – por menos de 10 mil moderadores em um universo no qual as métricas indicam números da ordem de quatro mil *uploads* de fotos por segundo (ou 350 milhões por dia) (ASLAM, 2017) e 1,3 milhão de postagens escritas por minuto (ASLAM, 2017). Ainda que, evidentemente, apenas uma parcela desse montante se trate de conteúdo sujeito à análise por abusos eventualmente cometidos, trata-se de uma proporção questionável. A inexistência desse tipo de dado torna difícil chegar-se a uma conclusão que não seja meramente especulativa.

Outro ponto levantado no manifesto em comento diz respeito ao uso de inteligência artificial (IA).[29] Segundo Richard Allan, o emprego de tecnologia é voltado para o trato conferido a palavras utilizadas em comentários e que, "obviamente", estejam ligadas ao cometimento de discursos de ódio. Admite, ainda, que há um longo caminho a ser percorrido até que tais mecanismos possam lidar com toda a complexidade inerente às análises de conteúdo abusivo. Reconhece-se, assim, aquilo que Ash (2016) já pontuava em sua obra acerca da liberdade de expressão ao mencionar que mesmo o mais sofisticado algoritmo não é capaz de realizar o julgamento individual das condições de momento, local, forma e contexto em que certas expressões são utilizadas, de modo que se pudesse realizar a diferenciação entre a caracterização de discurso de ódio ou meros pronunciamentos rudes, ou mesmo quando o discurso de ódio se torna ameaçador.[30]

---

[29] "*People often ask: can't artificial intelligence solve this? Technology will continue to be an important part of how we try to improve. We are, for example, experimenting with ways to filter the most obviously toxic language in comments so they are hidden from posts. But while we're continuing to invest in these promising advances, we're a long way from being able to rely on machine learning and AI to handle the complexity involved in assessing hate speech.*
That's why we rely so heavily on our community to identify and report potential hate speech. With billions of posts on our platform — and with the need for context in order to assess the meaning and intent of reported posts — there's not yet a perfect tool or system that can reliably find and distinguish posts that cross the line from expressive opinion into unacceptable hate speech. Our model builds on the eyes and ears of everyone on platform — the people who vigilantly report millions of posts to us each week for all sorts of potential violations. We then have our teams of reviewers, who have broad language expertise and work 24 hours a day across time zones, to apply our hate speech policies." (FACEBOOK, 2017e).

[30] "*Automated filtering designed by clever engineers can detect nudity and block much spam. But even the world's most sophisticated algorithm cannot make the individual judgement of time, place, manner and context which determines what is hate speech, as opposed to merely rude speech, and when hate speech becomes dangerous speech.*" (ASH, 2016).

Em razão de tais limitações, aparentemente não há que se falar, ainda, em um cenário de perfeita possibilidade de complementação entre inteligência artificial e o fator humano. Nesse momento, o último prepondera. Apesar disso, sob a ótica da necessidade de prover à comunidade mais informes a respeito dos procedimentos pelos quais uma denúncia de conteúdo abusivo passa até uma eventual sanção, seria desejável que o Facebook divulgasse a proporção de uso real da IA no combate ao *hate speech*: em que casos seu uso tem se mostrado eficaz? Há auxílio do componente humano no momento decisório? Se não há, há que se pesar entre a necessidade de existência de medidas que permitam a manifestação do usuário antes da aplicação da punição ou a mera possibilidade de envio de recursos pelo usuário (como já ocorre) – e que tudo isso seja explicitado e acessível ao usuário.

O fator humano, portanto, resta inteiramente comprometido quando a lente que se utiliza para sua análise é a da problemática da transparência – o que conduz à sensação de que o Facebook, em sua pretensão de universalização de regras de conduta que visem a "proteger" o usuário, cria um ambiente no qual o *hater* possui, ainda, grande espaço de atuação, na medida em que a comunidade não possui informes que lhes indiquem qualidade e efetividade quando da investigação de todo o conteúdo abusivo por ela reportado.

3.2 **Das penalidades**

Em termos de transparência, a questão das sanções aplicáveis ao *hate speech* não oferece maior alento se comparada à problemática do fator humano.

Os Padrões da Comunidade restringem-se a informar ao usuário que conteúdos considerados como discurso de ódio serão removidos:

> Discurso de ódio
> O Facebook *remove* discursos de ódio, o que inclui conteúdos que ataquem diretamente as pessoas com base em:
> raça, etnia, nacionalidade, religião, orientação sexual, gênero ou identidade de gênero, ou deficiências graves ou doenças. (FACEBOOK, 2017h).

Não há, em nenhuma das demais seções do documento, menção a qualquer tipo de penalidade que não seja a de exclusão de conteúdo.

A situação não muda quando se investiga o teor da Declaração de Direitos e Responsabilidades,[31] na qual apenas se lê, em sua terceira parte, que "você não publicará conteúdos que contenham discurso de ódio, sejam ameaçadores ou pornográficos; incitem violência; ou contenham nudez ou violência gratuita ou gráfica." (FACEBOOK 2017d). Não há qualquer abordagem acerca de sanções previstas para aqueles que infrinjam tal mandamento (ou os demais).

[31] A qual é "baseada nos Princípios do Facebook e representa os termos de serviço que regem nosso relacionamento com os usuários e outras pessoas que interagem com o Facebook, bem como marcas, produtos e serviços do Facebook que não possuam termos separados ou que estejam vinculados a estes termos, que chamamos de 'Serviços do Facebook' ou 'Serviços'." (FACEBOOK, 2017d).

É apenas na Central de Ajuda que são encontradas alusões a medidas punitivas outras que não a remoção de conteúdo: bloqueios e desativação de contas. Acerca dos primeiros, veja-se:

> O Facebook tem políticas para impedir comportamentos que outras pessoas podem achar impróprios ou abusivos. Se sua conta estiver bloqueada, você ainda poderá fazer *login* no Facebook, mas não poderá usar os recursos.
> O Facebook tem políticas estabelecidas para evitar comportamentos que outras pessoas podem achar irritantes ou abusivos. Determinamos que você utilizou um recurso de maneira que pode ser considerada abusiva, mesmo que não tenha sido de propósito.
> Lembre-se de algumas coisas sobre seu bloqueio:
> Bloqueios são temporários e podem durar de algumas horas a alguns dias;[32]
> Não podemos cancelar o seu bloqueio por motivo algum.
> Para evitar novos bloqueios, reduza ou interrompa esse comportamento. Por exemplo, envie menos mensagens por dia, marque as pessoas somente se elas estiverem nas fotos e convide apenas as pessoas conhecidas que possam ter interesse em participar dos eventos. Caso contrário, sua conta será permanentemente desativada. (FACEBOOK, 2017b).

Destaque-se, portanto, que o bloqueio visa a impedir, temporariamente, a continuidade de comportamentos abusivos, impróprios ou irritantes: condutas repetitivas, marcações indevidas, perturbação de desconhecidos são alguns exemplos. Extrai-se do texto, ainda, que o abuso prescinde de intenção – o que denota não haver possibilidade de contraditório –, e que os bloqueios podem durar de horas a dias – sem que haja a fixação de períodos mínimos ou máximos. Não há também a possibilidade de se cancelar o bloqueio uma vez que tenha sido imposto – o que reforça a impossibilidade de contraditório. A reiteração de comportamentos que levem ao bloqueio pode fazer com que o usuário tenha sua conta *permanentemente* desativada – mesmo que, repise-se, o utilizador não aja de tal modo propositadamente.

Acerca da desativação de contas, dispõe-se que:

> Se a sua conta estiver desativada, não será possível acessá-la. Lembre-se de que há diversos motivos para uma conta ser desativada e lidamos com cada um desses casos de maneira diferente.
> Se sua conta do Facebook foi desativada, você verá uma mensagem especial ao tentar entrar.
> Desativamos contas do Facebook que não seguem os Termos e Condições do Facebook. Alguns exemplos incluem:
> *Publicar conteúdo que não segue os Termos e Condições do Facebook;*
> Usar um nome falso;
> Fingir ser outra pessoa;
> *Comportamento contínuo que não é permitido no Facebook por violar os nossos Padrões da Comunidade;*
> Contatar outras pessoas com o propósito de assédio, propaganda, promoção, encontro amoroso ou outras condutas não permitidas;
> Se você acredita que sua conta tenha sido desativada por engano, envie um pedido de recurso. (FACEBOOK, 2017c).

[32] Registre-se que o site Urban Dictionary (2017) traz um rol de progressão atinente à quantidade de horas e dias a que o usuário pode permanecer bloqueado, a saber: *"24h/48h/72h/3 days/7 days/14 days/30 days. Then if you just don't get the hint, Facebook will eventually permanently delete your profile."* Apesar disso, não foram encontradas tais informações em documentos expedidos pelo próprio Facebook.

Por se tratar da sanção mais grave de que se tem notícia, fala-se expressamente na possibilidade de recurso. As hipóteses para que tal pena seja aplicada são várias, dentre as quais, a título exemplificativo, destaca-se a publicação de conteúdo que não segue os Termos e Condições do Facebook, bem como o comportamento proibido pelos Padrões da Comunidade. Sabendo-se que o discurso de ódio integra ambos os róis, conclui-se pela possibilidade de imposição de tal pena a indivíduo que proceda dessa forma. Anteriormente, registrou-se que há a previsão, nos Padrões da Comunidade, de exclusão de conteúdos tidos como *hate speech*. Depreende-se, assim, que as sanções aplicáveis ao aludido discurso variam entre a supressão da postagem e a desativação da conta.

A interpretação da primeira circunstância pontuada no rol exemplificativo – *"Publicar conteúdo que não segue os Termos e Condições do Facebook"* – leva à compreensão de que não é preciso se incorrer reiteradamente no *hate speech* para que dada conta seja desativada. Basta que o moderador entenda ser o caso de subsumir tal sanção ao caso. Não há qualquer sinalização acerca da gravidade ínsita ao comportamento, ou a qualquer fator que diferencie uma ocorrência apta a conduzir à desativação da conta de outra que leve à simples exclusão de conteúdo. Assim, a penalização do utilizador se trata de resultado de um julgamento de alta carga subjetiva, a qual tende a variar entre os diversos componentes da equipe de moderadores – já que, conforme se pontuou anteriormente, não existe uma previsão expressa que garanta um mínimo de isenção por parte dos referidos analistas, tratando-se o exame das denúncias de um momento em que ideologias podem sobrepor a lógica e, assim, levar a penalizações significativamente injustas – formal ou materialmente.

Ainda acerca da desativação das contas, depreende-se do trecho a seguir a possibilidade de aviso anterior à desativação:

> Não recebi um aviso antes de minha conta ser desabilitada.
> Em alguns casos, não enviamos aviso antes de desabilitar sua conta. Além disso, nós não restauramos contas que foram desativadas por violações graves dos Padrões da comunidade do Facebook. (FACEBOOK, 2017g).

Sob o enfoque da transparência, veja-se que não há qualquer esclarecimento sobre os "alguns casos" passíveis de desativação sem aviso, não se sabe se referido comunicado é regra, exceção ou se também está condicionado ao subjetivismo decisional dos moderadores. Além disso, não há possibilidade de restauração das contas que afrontem gravemente os Padrões da Comunidade – sem que exista, em tal documento, qualquer informação acerca de que comportamentos são tidos por mais ou menos intensos para os gestores da rede social em estudo.

Conclusões

O discurso de ódio tem sido considerado uma possível limitação ao exercício da liberdade de expressão. Essa restrição tem sido construída no direito positivo por meio da jurisprudência,

fruto do exame caso a caso. Isso porque limitar o discurso em uma sociedade democrática é uma medida excepcional justificada pela necessidade de equilíbrio com outros direitos fundamentais como a igualdade. O efeito silenciador do discurso pelo Estado tem impactos na definição e exercício da democracia. Esse sistema de controle estatal era eficaz, em certa medida, quando a produção da informação era centralizada em alguns grupos. O cenário era de poucos produtores e muitos receptores de informação. Com a internet e as redes sociais esse cenário se altera, já que consumidores são também produtores de informação. Por consequência, o controle via judiciário da propagação do discurso de ódio se torna pouco efetivo: o sistema estatal em vigor não foi pensado para lidar com a velocidade e o volume das demandas.

Diante desse espaço, as redes sociais passam a realizar tal controle. Do exame do que venha a ser discurso de ódio pelo Facebook se conclui que: a plataforma concentra esforços na definição de palavras representativas de tal discurso; há o reconhecimento de que a mesma palavra pode ter sentidos diferentes a depender da cultura e que tal dado é considerado quando se definem quais seriam as palavras-chaves; e a divulgação prévia de quais seriam tais palavras inviabilizaria o controle realizado pela plataforma.

Os recentes vazamentos documentais divulgados pelo *The Guardian*, ainda que por vias questionáveis, lançaram novas luzes quanto à propagação do discurso de ódio no Facebook. Verificou-se não ser difícil realizar combinações que acabem por expor grupos vulneráveis aos referidos ataques, ao mesmo tempo em que a abordagem meramente utilitarista da análise perpetrada pelos moderadores acaba por permitir que tais agressões sejam ignoradas, de modo que não infrinjam os Termos de Uso e importem nas sanções cabíveis.

Acerca da transparência conferida aos procedimentos pelos quais passam as postagens denunciadas como *hate speech*, notou-se que não há clareza acerca das sanções possíveis – tanto em termos de espécies aplicáveis quanto em relação às suas durações –, nem em relação às características e à política de seleção das pessoas que cuidam de tais exames.

Se o direito positivo estatal já não possui capacidade de, sozinho, trazer soluções que efetivamente sirvam ao combate do *hate speech* propagado nas redes sociais, verifica-se necessário que os gestores de tais plataformas se comprometam a garantir à comunidade a existência de procedimentos eficazes nesse sentido. Demais disso, cabe ao Estado ser o fiscal de como tais políticas são criadas e executadas – afinal, trata-se do exercício da liberdade de expressão, um dos pilares da democracia. Quanto mais claros os processos por meio dos quais as redes sociais gerenciam o direito à liberdade de expressão, mais subsídios possui a comunidade – interna e externa ao Facebook – para fiscalizar e aperfeiçoar o exercício da liberdade de expressão no ambiente virtual.

Referências

ANGWIN, Julia. Facebook's secret censorship rules protect white men from hate speech but not black children. *ProPublica*, 28 jun. 2017. Disponível em: <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>. Acesso em: 05 jul. 2017.

ASH, Timothy. Garton. *Free speech*: ten principles for a connected world. Londres: Atlantic Books, 2016.

ASLAM, Salman. Facebook by the numbers: stats, demographics and fun facts. Omnicore, 01 Jan. 2018. Disponível em: <https://www.omnicoreagency.com/facebook-statistics/>. Acesso em: 30 jan. 2018.

CAFAGGI, Fabrizzio. New foundations of transnational private regulation. *EUI Working Papers:* RSCAS 2010/53. Florença, Dec. 2010.

CONSELHO DA UNIÃO EUROPEIA. Directiva 2000/43/CE do Conselho, de 29 de junho de 2000. Aplica o princípio da igualdade de tratamento entre as pessoas, sem distinção de origem racial ou étnica. *Jornal Oficial das Comunidades Europeias*, 17 jul. 2000. Disponível em: <http://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32000L0043&from=PT>. Acesso em: 30 dez. 2017.

COUNCIL OF EUROPE. Comité des ministres. *Recommandation nº R (97)20.* Adoptée par le Comité des Ministres le 30 octobre 1997. Disponível em: <https://www.coe.int/en/web/freedom-expression/committee-of-ministers-adopted-texts/-/asset_publisher/aDXmrol0vvsU/content/recommendation-no-r-97-20-of-the-committee-of-ministers-to--member-states-on-hate-speech-?inheritRedirect=false>. Acesso em: 31 dez. 2017.

DELMAS-MARTY, M. *Le pluralisme ordonné*: les forces imaginantes du droit. Paris: Ed. Du Seuil, 2006. v. 2.

EUROPEAN COMMISSION. *Code of conduct on countering illegal hate speech online*. 2016. Disponível em: <http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf>. Acesso em: 02 jul. 2017.

EUROPEAN COURT OF HUMAN RIGHTS. *Erbakan v. Turkey*. Application n. 59405/00. 06 jul. 2006. Disponível em : <https://webcache.googleusercontent.com/search?q=cache:F5qunROnn4s-J:https://hudoc.echr.coe.int/app/conversion/pdf/%3Flibrary%3DECHR%26id%3D003-1728198-1812055%26filename%3D003-1728198-1812055.pdf+&cd=3&hl=pt-BR&ct=clnk&gl=br>. Acesso em: 30 dez. 2017.

ERRERA, Roger. In defence of civility: racial incitement and group libel in french law. In: Coliver, S. (Ed.). *Striking a Balance*: Hate Speech, Freedom of Expression and Non-Discrimination. London: Article 19 and Human Rights Centre, 1992.

FACEBOOK atinge marca de 2 bilhões de usuários, anuncia Zuckerberg. *Folha de São Paulo*, São Paulo, 27 jun. 2017a. Disponível em: <http://www1.folha.uol.com.br/tec/2017/06/1896428-facebook-atinge-marca-de-2-bilhoes-de-usuarios-anuncia-zuckerberg.shtml>. Acesso em: 31 dez. 2017.

FACEBOOK. *Bloqueios.* Central de Ajuda. 2017b. <https://www.facebook.com/help/174623239336651>. Acesso em: 02 jul. 2017.

FACEBOOK. *Contas desativadas.* 2017c. Disponível em: <https://www.facebook.com/help/185747581553788>. Acesso em: 03 jul. 2017.

FACEBOOK. *Declaração de direitos e responsabilidades*. 2017d. Disponível em: <https://www.facebook.com/legal/terms>. 2017f. Acesso em: jul. 2017.

FACEBOOK. Hard questions: hate speech. *Newsroom*, 27 jun. 2017e. Disponível em: <https://newsroom.fb.com/news/2017/06/hard-questions-hate-speech/>. Acesso em: 02 jul. 2017.

FACEBOOK. Hard questions: how we counter terrorism. *Newsroom*, 15 jun. 2017f. Disponível em: <https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/>. Acesso em: 02 jul. 2017.

FACEBOOK. *Não recebi um aviso antes de minha conta ser desabilitada*. 2017g <https://www.facebook.com/help/228409123842014?helpref=uf_permalink>. Acesso em: 04 jul. 2017.

FACEBOOK. *Padrões da comunidade*. 2017h. Disponível em: <https://www.facebook.com/communitystandards>. Acesso em: 05 jul. 2017.

FACEBOOK. *Unlawfully blocked for 30 days?* 2017i. Disponível em: <https://www.facebook.com/help/community/question/?id=10155849281675214>. Acesso em: 25 jun. 2017.

FARIA, José Eduardo. *Sociologia jurídica*: direito e conjuntura. 2. ed. São Paulo: Saraiva, 2010.

FISS, Owen. *A ironia da liberdade de expressão*: estado, regulação e diversidade na esfera pública. Tradução Gustavo Binenbojm e Caio Mário da Silva Pereira Neto. Rio de Janeiro: Renovar, 2005.

HATE SPEECH and anti-migrant posts: Facebook's rules. The Guardian Online, 24 May 2017. Disponível em: <https://www.theguardian.com/news/gallery/2017/may/24/hate-speech-and-anti-migrant-posts-facebooks-rules#img-2>. Acesso em: 30 jun. 2017.

HOFMANN, Rainer. Incitement to national and racial hatered: the legal situation in Germany. In: COLIVER, Sandra (Ed.). *Striking a balance*: hate speech, freedom of expression and non-discrimination. Londres: Article 19 and Human Rights Centre, 1992. p. 159-170.

HOPKINS, Nick. *Facebook struggles with 'mission impossible' to stop online extremism.* Disponível em: <https://www.theguardian.com/news/2017/may/24/facebook-struggles-with-mission-impossible-to-stop-online-extremism>. Acesso em: 20 jun. 2017.

KÜBLER, Friedrich. How much freedom for racist speech? Transnational aspects of a conflict of human rights. *Hofstra Law Review*, New York, v. 27, i. 2, p. 334-376, 1998.

ONLINE CENSORSHIP TEAM. *Facebook must go further on transparency*. Disponível em: <https://onlinecensorship.org/news-and-analysis/facebook-must-go-further-on-transparency>. Acesso em: 05 jul. 2017.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS. *Declaração Universal dos Direitos Humanos*. 10 dez. 1948. Disponível em: <http://unesdoc.unesco.org/images/0013/001394/139423por.pdf>. Acesso em: 31 dez. 2017.

ROSENFELD, Michel. Hate speech in constitutional jurisprudence: a comparative analysis. *Cardozo Law Review*, New York, v. 24, i. 4, 2003. Disponível em: <https://ssrn.com/abstract=265939>. Acesso em: 30 dez. 2017.

SARMENTO, Daniel. A liberdade de expressão e o problema do "hate speech". *Revista de Direito do Estado*, Rio de Janeiro: Renovar, p. 53-106, 2006.

SIMPSON, Jon. The power of content intelligence in marketing. *Forbes Agency Council*, 10 Jan. 2017. Disponível em: <https://www.forbes.com/sites/forbesagencycouncil/2017/01/10/the-power-of-content-intelligence-in-marketing/#13136bf26dae>. Acesso em: 05 jul. 2017.

SUNDERLAND, Mitchell. *Facebook blocked me because I said 'faggot,' even though I'm gay*. 2014. Disponível em: <https://www.vice.com/en_us/article/bn575w/facebook-wont-let-faggots-say-faggot>. Acesso em: 25 jun. 2017.

URBAN DICIONARY. *Facebook Jail*. Disponível em: <http://www.urbandictionary.com/define.php?term=Facebook%20Jail>. Acesso em: 03 jul. 2017.

YORK, Jilian. Getting banned from Facebook can have unexpected and professionally devastating consequences. Quartz, 31 Mar. 2016. Disponível em: <https://qz.com/651001/getting-banned--from-facebook-can-have-unexpected-and-professionally-devastating-consequences/>. Acesso em: 25 maio 2017.

ZUCKEBERG, Mark. Building global community. *Facebook*, 16 fev. 2017. Disponível em: <https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/>. Acesso em: 30 jun. 2017.

Data da submissão: 07 de fevereiro de 2018
Avaliado em: 13 de maio de 2018 (AVALIADOR A)
Avaliado em: 11 de maio de 2018 (AVALIADOR B)
Aceito em: 25 de junho de 2018