DOI: 10.1590/s0103-4014.202438111.020

Inteligência Artificial explicável para atenuar a falta de transparência e a legitimidade na moderação da Internet

Thomas Palmeira Ferraz [I]
Caio Henrique Dias Duarte [II]
Maria Fernanda Ribeiro [III]
Gabriel Goes Braga Takayanagi [IV]
alexandre alcoforado [V]
Roseli de Deus Lopes [VI]
Mart Susi [VII]

Introdução

Historicamente, as sociedades sempre estabeleceram normas de convivência que impõem limites aos direitos individuais. Era o caso do direito fundamental à liberdade de expressão, que nas democracias modernas tinha claramente estabelecido limites e mecanismos para punir aqueles que não o respeitavam. Entretanto, com o advento das redes sociais, o debate público foi transferido para a internet. Se, por um lado, isso facilita a participação das pessoas; por outro, torna impossível que os mesmos meios convencionais de moderação do debate público sejam aplicados. Para lidar com essa questão, as plataformas digitais criaram corpos de moderadores humanos que atuam na avaliação de reclamações. Nos últimos anos, esses grupos foram substituídos pelo uso integrado de diversos modelos de Inteligência Artificial (IA) que atuam automaticamente na moderação de conteúdo, em muitos casos antes mesmo que haja reclamação de alguma parte. No entanto, o uso massivo de IA neste papel de juiz e o fato de o monopólio desta moderação estar a cargo de entidades privadas levanta uma série de questionamentos éticos e jurídicos que serão explorados neste trabalho.

Do ponto de vista jurídico, há um impasse sobre quem pode moderar o conteúdo. Idealmente seria que ele pudesse seguir os princípios de *legitimidade, transparência, controle e capacidade de execução* (Sander, 2020). Ainda que o Estado seja dotado de legitimidade democrática, um regramento público transparente e poder de ação local, ele possui recursos limitados e pouca capacidade de respon-

der rapidamente às demandas de moderação. A isso se soma o fato de o mesmo não ter jurisdição sobre onde o conteúdo é armazenado (os servidores e centros de dados), geralmente fora de suas fronteiras, sendo incapaz de realizar esta moderação na ponta. Já as plataformas digitais, embora sejam mais capazes de identificar usuários, tenham acesso aos conteúdos com facilidade e conheçam a estrutura de seu portal, na prática, não têm um regramento claro e sua legitimidade como censor público é questionável. Esse cenário nos coloca em um dilema, no qual o Estado não consegue agir sozinho, e as entidades privadas, se agirem sozinhas, podem ser responsabilizadas por ações que pareçam ser excessos ou omissões.

O uso de modelos de inteligência artificial no papel de julgadores, por sua vez, também apresenta importantes contradições éticas a serem abordadas (Nahmias; Perel, 2020). Para além da questão filosófica de máquinas tomando decisões sobre o direito dos humanos, a natureza opaca (ou "caixa-preta") dos modelos de maior complexidade, amplamente utilizados, dificulta a percepção de vieses inconscientes, que podem tratar de forma prejudicial grupos sociais minoritários em comparação com outros, um fenômeno chamado *discriminação algorítmica*. No contexto da moderação do conteúdo na Internet, esses vieses inconscientes podem ser decorrentes de questões locais e culturais, de posicionamento político, de raça, de orientação sexual, entre outros, considerando que cada grupo pode ter um vocabulário particular (Oliva; Antonialli; Gomes, 2021). Tolerar que grupos tenham menos liberdade de expressão que outros fere os fundamentos de um pleno Estado Democrático de Direito.

Essas questões estabelecem, para a sociedade contemporânea, um dilema entre moderar e não moderar: manter a liberdade de expressão inviolável, ainda que dela se abuse, ou combater o abuso virtual e conteúdos potencialmente danosos, correndo o risco de inadvertidamente suprimir um direito fundamental? A evolução das redes sociais inviabiliza que uma moderação seja feita totalmente por humanos, e mesmo que fosse, os humanos também estão sujeitos a vieses, embora sejam mais fáceis de medir e já existam métodos de mitigação. Portanto, o uso de IA na moderação foi, na última década, a escolha pelo "mal menor". Agora sendo o uso de IA plenamente difundido pelas plataformas, é fundamental debatê-lo de modo a refinar este tipo de moderação e achar um ponto de equilíbrio. Este tipo de discussão tem sido feita sobre o uso de inteligência artificial em todos os aspectos da vida humana, e levou a União Europeia a introduzir o direito à explicação de decisões algorítmicas no seu marco legal de tecnologia (GDPR) em 2016 (Goodman; Flaxman, 2017).

Diante desse panorama, tem emergido na pesquisa em Ciência da Computação o campo da *Explainable AI* (XAI), em português IA Explicável, no qual são pesquisadas e desenvolvidas ferramentas que possibilitem tornar interpretável o processo de decisão de modelos já existentes, bem como desenvolver modelos desenhados para serem interpretáveis aos humanos (Adadi; Berrada, 2018; Felzmann et al., 2020). Essa é uma área que avançou muito, incluindo a

obtenção de modelos transparentes com desempenho próximo a modelos opacos para diversas tarefas (Arrieta et al., 2020). No contexto da moderação de conteúdo, esta categoria de IA permite: (*i)* o desenvolvimento de modelos já construídos para seguir um padrão de moderação definido pela sociedade e ser capaz de explicar suas decisões com base neste padrão; e (*ii*) construir modelos que possam atuar em paralelo auditando modelos caixa-preta existentes com base nestes mesmos critérios, explicando porque o modelo de caixa-preta está tomando suas decisões e, assim, dando maior transparência ao processo.

Há vários trabalhos na literatura que abordam os aspectos discutidos até aqui. Susi (2019) foi capaz de formular um instrumento matemático baseado em critérios transparentes, rumo a um padrão de moderação de invasão de privacidade que respeite os direitos humanos fundamentais. Já Reis et al. (2019) foram capazes de definir métricas para auditar o modelo XGBoost para detecção de notícias falsas usando *Explainable AI*. Por outro lado, Mohseni et al. (2021) conseguiram produzir evidências de que critérios e decisões transparentes explicadas ao usuário têm o potencial de reduzir a recorrência do compartilhamento de notícias falsas. Neste trabalho, abordamos aspectos semelhantes aos destes estudos, de uma perspectiva interdisciplinar, mas com foco na definição mais clara dos aspectos do problema e dos atores envolvidos nele, e na construção de um paradigma de moderação transparente que o solucione.

Objetivos e metodologia

O objetivo deste trabalho é analisar o uso da inteligência artificial no contexto de moderação de conteúdo na Internet. Trazemos uma visão holística, porém acessível, do panorama atual desse tema, indicando oportunidades de melhorias, especialmente com a evolução da maturidade tecnológica no campo da *Explainable AI*. Com isso, propomos um novo padrão de moderação que é consistente com as demandas atuais.

Para esse fim, tomamos três pontos de estudo:

• **O atual paradigma da moderação**, considerando seus diversos níveis de decisão, o papel desempenhado pelas plataformas digitais e pelo Estado no processo, bem como os modelos de IA utilizados, seus diferentes níveis de automação, as funções que desempenham no contexto em que são aplicados e suas limitações;

• **A visão ética e jurídica** sob o modelo em vigência na moderação de conteúdo online, explorando como o direito e as ciências sociais observam o sistema atual, e definindo o estado da arte da discussão sobre liberdade de expressão no contexto digital, entendendo quais aspectos devem ser atendidos por um padrão de moderação a ser adotado e que papel cada parte interessada deve desempenhar;

• **O estado da arte da tecnologia**, analisando o panorama da *Explainable AI* e sua maturidade tecnológica atual, compreendendo seu funcionamento, o que é necessário para sua construção e suas limitações, trazendo conclusões so-

bre como os avanços nesta área podem ser aplicados para resolver as dificuldades que se encontram atualmente no modelo vigente de moderação.

Considerando esses três pilares, procuramos sumarizar e relacionar o que foi estudado propondo um novo paradigma de moderação que seja justo, ético e transparente, assim como definir o papel do Estado e das plataformas digitais neste contexto. Discute-se também quais desafios tecnológicos terão que ser enfrentados para sua implementação, bem como seus potenciais benefícios.

Este trabalho, em linhas gerais, procura diagnosticar as deficiências do uso da Inteligência Artificial na moderação automática do conteúdo nas redes sociais e suas ameaças do ponto de vista ético e jurídico a direitos humanos fundamentais como a liberdade de expressão, e a partir disso identificar e desvendar como a *Explainable AI* pode ser útil para mitigar estes efeitos negativos. Na base de dados da Web of Science, há cerca de 4.300 estudos abordando a XAI. Nessa mesma base, existem aproximadamente 2.000 estudos acerca da moderação automática de conteúdo, porém, apenas 21 deles mencionam a XAI, o que demonstra a necessidade de um trabalho que considere os dois aspectos em conjunto. Para essa finalidade, utilizamos uma abordagem interdisciplinar entre Sociologia, Computação e Direito, aliando análises técnicas e matemáticas de dados, a uma revisão de estudos que abordam a sociedade e o Direito.

Primeiramente, realizamos uma revisão exploratória da literatura e dos materiais e dados disponibilizados pelas próprias plataformas Meta (2021) e Google (2021) para caracterizar a forma como as grandes plataformas atualmente realizam a moderação, inclusive buscando formas de medir sua efetividade. Em seguida, aprofundamos os questionamentos relativos ao uso desta tecnologia do ponto de vista ético e jurídico.

Do ponto de vista ético, revisamos trabalhos em Sociologia, Computação Social e Estatística que permitam esclarecer quais características do modelo da "caixa-preta" podem resultar em desigualdade e injustiça no processo de tomada de decisão automático em moderação de conteúdo. Com isso, obtemos lacunas e os desafios éticos motivados pelo uso desse tipo de tecnologia.

Sob um olhar pragmático do Direito, fazemos uma análise comparativa do cenário atual quanto à observância dos princípios da legitimidade, transparência, controle e capacidade de execução, e quanto ao cumprimento da regulamentação jurídica vigente na União Europeia (GDPR) e no Brasil (Marco Civil da Internet e LGPD). Isso nos dá um panorama que permite diagnosticar os objetivos dispostos na regulamentação e na doutrina jurídica que não são atendidos pelo paradigma de moderação adotado atualmente, com especial atenção à legitimidade para moderar a acessibilidade dos critérios de moderação e o controle público sobre o processo.

Com o panorama bem projetado, procuramos definir a *Explainable AI* e desvendar quais de suas características podem mitigar os efeitos nocivos do uso da IA sob os dois aspectos supracitados. Buscamos demonstrar a viabilidade

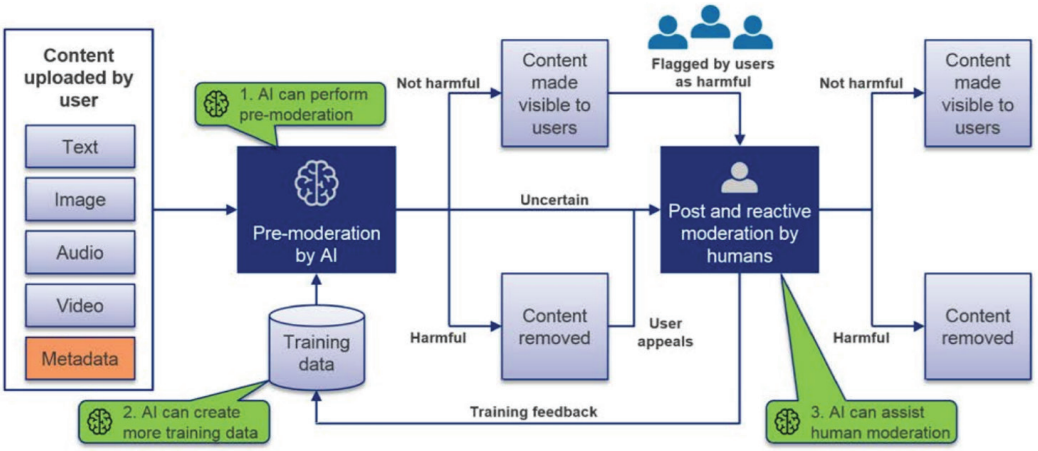
tecnológica da implementação de modelos desse tipo, já que este é um questionamento recorrente. A seguir, delineamos que papéis o Estado e as plataformas digitais devem desempenhar de modo a tornar efetivo o uso desta nova tecnologia, criando um novo paradigma de moderação automática de conteúdo em que a transparência esteja no centro do processo decisório.

Por fim, discutimos avanços que já têm sido feitos no sentido de estabelecer esse novo paradigma, buscando experiências práticas similares. Trazemos evidências de que o uso dessas novas tecnologias tem grande potencial de não só cobrir as lacunas apresentadas neste estudo, mas também tornar a moderação de conteúdo mais efetiva, quando a sociedade é introduzida no processo.

Paradigma atual da Moderação de Conteúdo

Especialmente em razão das pressões da sociedade civil e dos governos de vários países, os principais provedores de conteúdo passaram a desenvolver e implementar um conjunto de normas para lidar com o conteúdo publicado na plataforma que não estava de acordo com os regulamentos locais ou sua visão sobre o que deve ser o ambiente digital – como os Community Standards ("Normas da Comunidade", para Facebook) e as Community Guidelines & Policies ("Políticas e Diretrizes da Comunidade", para YouTube) (Estarque e Achergas 2021). Esses documentos geralmente contêm as principais diretrizes sobre conteúdo e ações que são permitidas ou não na rede social, tais como violência, assédio, discurso de ódio, notícias falsas, nudez e terrorismo. Klonick (2017) relata, no entanto, que o surgimento desses é o resultado de esforços recentes e, no passado, as políticas eram baseadas em diretrizes genéricas. Há ainda esforços recentes para criar comitês de supervisão independentes (Klonick, 2019).

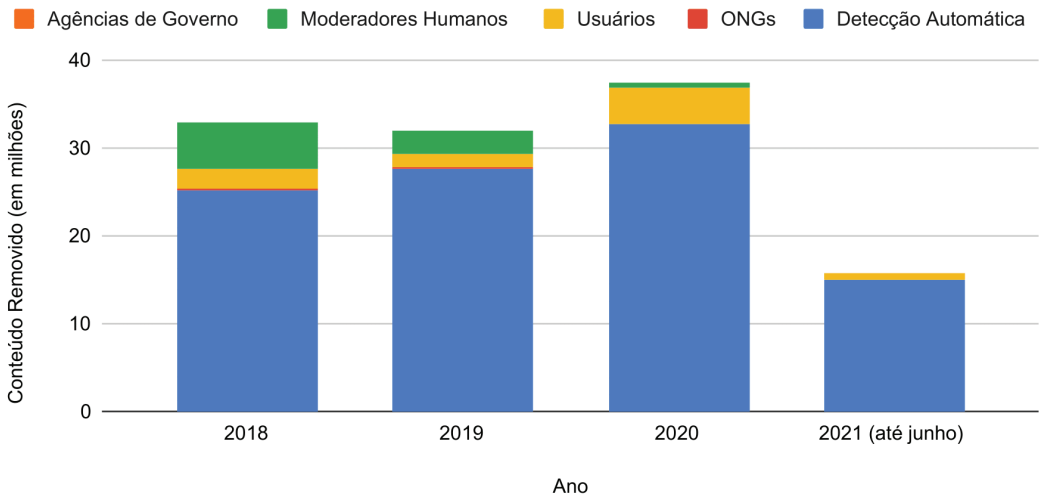
A moderação de conteúdo se aproveita da recente revolução que vem ocorrendo na Inteligência Artificial, com o surgimento de modelos de aprendizagem profunda baseados em redes neurais (LeCun; Bengio; Hinton, 2015) e, mais recentemente, modelos de linguagem pré-treinados (Devlin et al., 2019; Brown et al., 2020; Alcoforado et al., 2022), para fazer cumprir as Normas da Comunidade. A moderação de conteúdo on-line pode ser implementada de várias maneiras, mas geralmente adota uma das seguintes abordagens ou ambas (Winchcomb, 2019; Jiang; Robertson; Wilson, 2020): (i) *Pré-moderação*, quando o conteúdo carregado é moderado antes da publicação, normalmente usando sistemas baseados em IA; (ii) *Pós-moderação* (ou moderação reativa) quando o conteúdo é moderado após sua publicação e foi marcado pelos usuários ou sistemas baseados em IA como prejudicial, ou que foi removido anteriormente, mas requer uma segunda revisão mediante recurso. A Figura 1 ilustra esse caso geral de moderação.



Fonte: Winchcomb (2019).

Figura 1 – Modelo Geral de Moderação de Conteúdo em Mídias Sociais.

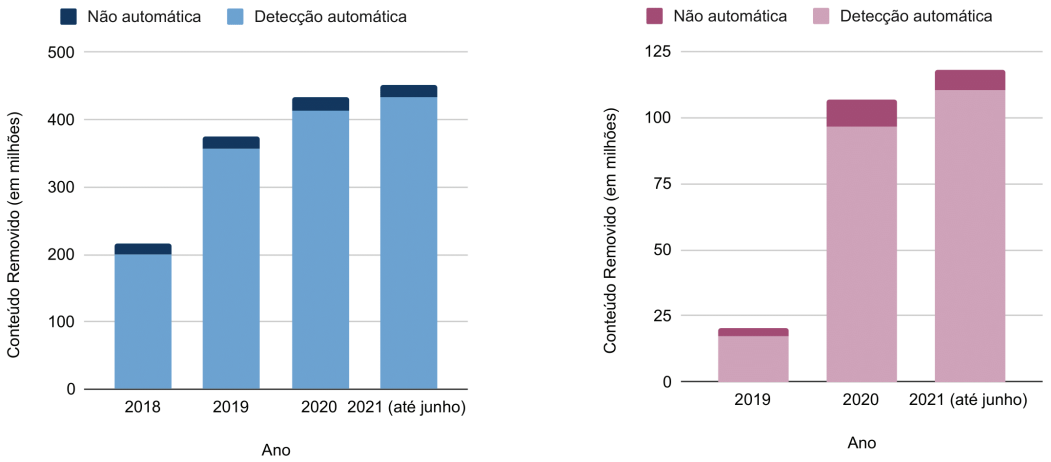
Apesar disso, detectar conteúdo tóxico ou prejudicial pode ser bastante desafiador já que o conteúdo pode aparecer em muitas modalidades diferentes (por exemplo, áudios, imagens, vídeos, GIF, textos, além de combinações multimodais), em formatos diferentes (por exemplo, memes, ou *deepfakes* – falsificações baseadas em redes neurais profundas) (Winchcomb, 2019). Além disso, alguns podem ser transmissões ao vivo, que requerem ação em tempo real. Ainda assim, outros podem depender do contexto em que foram produzidos para serem considerados tóxicos ou prejudiciais. Além disso, a linguagem da Internet pode evoluir com o tempo, e até mesmo os usuários podem aprender técnicas para contornar a moderação de conteúdo usando linguagem proprietária codificando intencionalmente certas palavras através de erros ortográficos, ou *leetspeak* (linguagem informal da internet em que se substituem letras por números ou símbolos, por exemplo, "v1@do", "p*rr*", "p*t@") (Tan et al., 2020). Se esta dificuldade não fosse suficiente, existe uma ampla gama de conteúdos potencialmente nocivos em diferentes níveis, incluindo, mas não limitados a: material de abuso infantil, conteúdo violento e extremo, notícias falsas, discurso de ódio, falsas alegações relacionadas à saúde, conteúdo sexual, material cruel e insensível, e conteúdo de spam. Na prática, o sucesso do uso da Inteligência Artificial para moderação de conteúdo depende do desenvolvimento de vários sistemas especializados para cada categoria, trabalhando em harmonia.



Fonte: Elaboração própria baseada em dados do Google Transparency Report (Google, 2021).

Figura 2 – Número de vídeos (em milhões) removidos pela primeira vez no YouTube (sem considerar o recurso) por ano e pelo autor da remoção. Em azul, a detecção automática é a remoção de IA. Nas outras cores, as remoções são feitas por humanos: Agências governamentais e judiciais, moderadores humanos (que trabalham para o Google), reclamações de usuários, e ONGs (parceiros do YouTube).

Como mostrado na Figura 1, a moderação de conteúdo é tipicamente implementada como um processo híbrido IA-humano, podendo ter diversos níveis de automação. Em um nível inferior de automação, a pré-moderação por IA só pode sinalizar candidatos potenciais a serem removidos. Além disso, a IA pode ser implementada para sintetizar os dados de treinamento para melhorar o desempenho da pré-moderação. Além disso, a IA pode auxiliar os moderadores humanos na pós-moderação, reduzindo o efeito de moderadores individuais sobre o resultado final. Analisando os dados das redes sociais YouTube, Facebook, e Instagram expressos nas Figuras 2 e 3, podemos notar que há um aumento na participação da Inteligência Artificial no processo de moderação. No caso do YouTube, podemos ver que a porcentagem de conteúdo removido automaticamente pela IA aumentou de 76% em 2018 para quase 95% em 2021. O mesmo fenômeno pode ser observado no caso do Facebook.

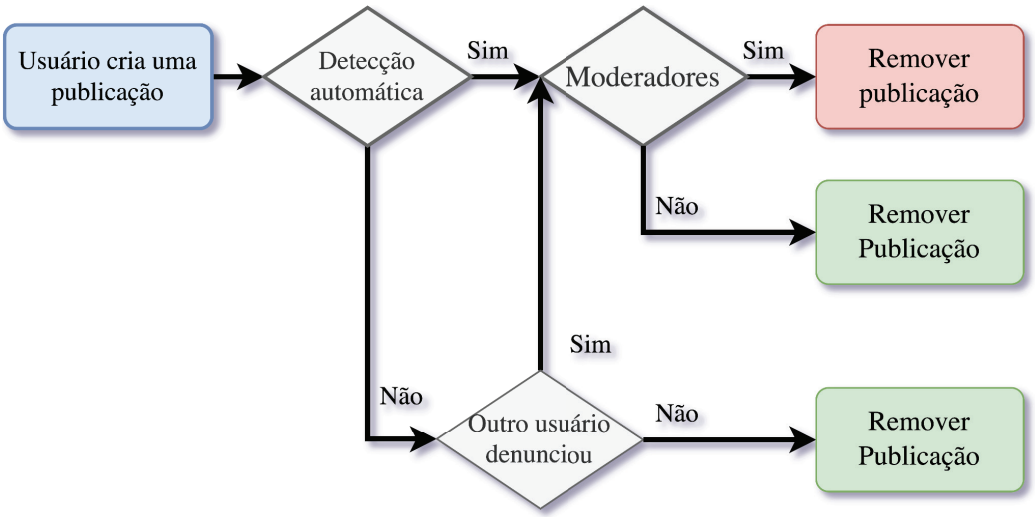


Facebook Instagram

Fonte: Elaboração própria baseada nos dados do Relatório de Transparência do Facebook (Meta, 2021).

Figura 3 – Quantidade de conteúdo removido no Facebook e Instagram ao longo dos anos, não considerando a remoção de spam.

Ultimamente, o Facebook tem implementado uma abordagem mista para moderar o conteúdo[1] apresentada na Figura 4. Ela utiliza moderadores humanos, algoritmos automatizados e denúncias feitas pelos usuários para analisar o conteúdo. Começa usando algoritmos automatizados para decidir se o material será analisado posteriormente – essa verificação acontece usando um moderador humano. O Facebook alega que a maioria dos conteúdos que geram maior preocupação, como terrorismo, exploração infantil ou autoflagelação, são classificados para serem moderados em primeiro lugar por seus analistas, enquanto conteúdos como spam são classificados em último lugar. Se o conteúdo não for escolhido por este moderador, então existe a opção de ser denunciado por um usuário, e então novamente um moderador humano irá investigá-lo.



Fonte: Elaboração própria baseada no Facebook Transparency Report (Meta, 2021).

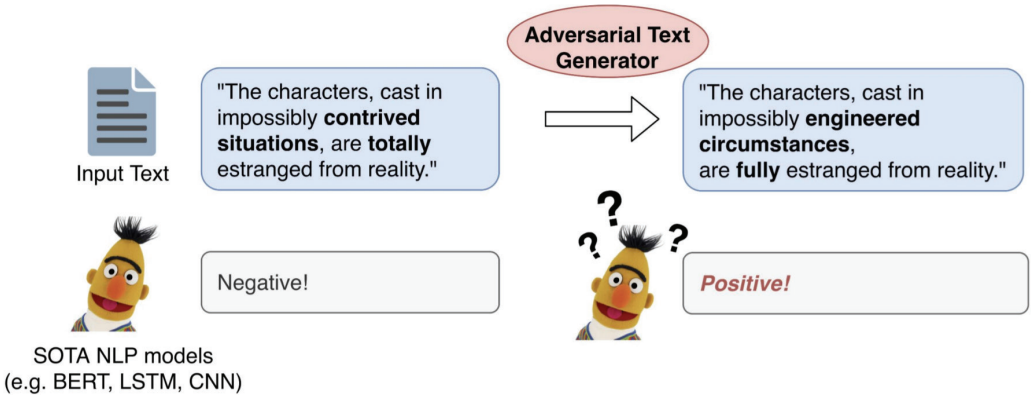
Figura 4 – Novo procedimento de moderação proposto pela Meta em 2021.

Impactos éticos e sociais do uso da Inteligência Artificial

Andrew Ng (2016) define que provavelmente podemos automatizar usando IA, seja agora, seja no futuro próximo, qualquer tarefa mental que um humano normal levaria menos de um segundo de reflexão. Entretanto, assim como os humanos, as máquinas são suscetíveis a erros. Quando falamos de moderação de conteúdo na Internet, estamos falando de um problema de classes inerentemente desbalanceadas, onde certas categorias estão naturalmente menos presentes no mundo real. Este é um problema clássico no aprendizado de máquinas, que tem sido discutido por vários autores, incluindo Chawla, Japkowicz e Kotcz (2004), Fernández et al. (2018), Ferraz et al. (2021), Krawczyk (2016), e He e Garcia (2009). No caso das mídias sociais, em termos gerais, há mais conteúdo a ser mantido do que a ser removido. Para que os modelos possam aprender nestes cenários, são feitas escolhas de compromisso que podem levar à perda de desempenho. Por exemplo, maximizar o *recall*[2] (revocação ou sensibilidade do algoritmo) pode garantir que um modelo de detecção de *fake news* classifique todas as notícias falsas corretamente, mas também pode levá-lo a classificar muitas notícias verdadeiras como falsas (erro tipo I, ou falso positivo). A isto se somam os desafios já destacados ao lidar com diferentes tipos de mídia (áudio, vídeo, imagem, texto), em diferentes cenários. Dessa forma, é praticamente impossível ter um modelo com desempenho perfeito, mesmo nos casos em que foram treinados para acertá-lo (que constam nos seus dados de treinamento) (Duarte; Llanso; Loup, 2018). Em muitos casos, os modelos podem superar os humanos em velocidade e até mesmo em padronização de julgamento, mas nem sempre eles conseguirão acertar. E o que fazer quando eles erram? Qual é o custo de um erro destes sistemas? Isto, evidentemente, sempre dependerá da aplicação, ou seja, do seu propósito e quão sensível ele pode ser.

Como demonstrado por diversos estudos, incluindo o mais recente trazido por Mehrabi et al. (2021), os modelos baseados em aprendizado são sempre suscetíveis a diferentes tipos de vieses. O estudo relata que os dados enviesados geram modelos enviesados (*Bias from Data to Algorithm*). Esse viés pode ocorrer de várias maneiras, inclusive: Viés de Medição; Viés de Variável Omitida; Viés de Representação; Viés de Agregação - quando são tiradas conclusões falsas sobre indivíduos a partir da observação de toda a população; Viés de Ligação – quando atributos de rede obtidos a partir de interações de usuários deturpam o verdadeiro comportamento dos usuários; entre outros. Ainda existem vieses que são adicionados ao usuário pelo algoritmo (*Bias from Algorithm to User*), que são vieses resultantes de resultados algorítmicos e, como consequência modulam, o comportamento do usuário. Isso pode ocorrer a partir de coisas simples como escolhas de *design* da interface, como a informação é apresentada, os resultados que aparecem primeiro em uma busca, até mesmo viés de popularidade (comentários falsos ou uso de *bots* tornam um item popular e mais exposto) (Ciampaglia et al., 2018), e viés decorrentes da escolha do algoritmo e do método de otimização (Danks; London, 2017). E finalmente, existem vieses adicionados pelo usuário aos dados (*Bias from User to Data*) quando o comportamento do usuário é afetado por um algoritmo, quaisquer vieses presentes nesses algoritmos podem introduzir vieses no processo de geração de dados.

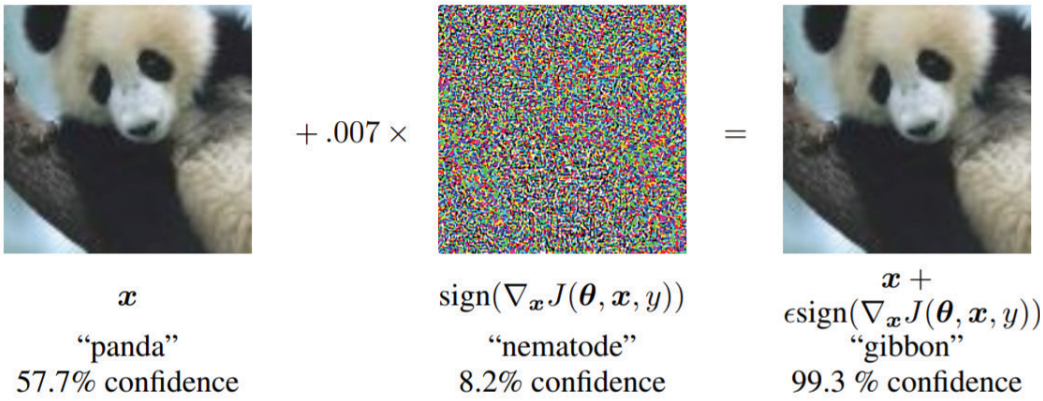
Uma consequência muito clara dos vieses em modelos de moderação de conteúdo na Internet é a possibilidade de estes serem seletivos para grupos sub-representados nos dados ou mesmo em contextos sub-representados. Como exemplo, Oliva, Antonialli e Gomes (2021) trazem evidências de que os sistemas de IA podem não interpretar corretamente o contexto social do discurso, deixando de reconhecer casos em que palavras, que convencionalmente poderiam ser vistas como ofensivas, carregam significados diferentes e positivos no discurso LGBTQIA+. Nesse sentido, Harrison et al. (2020) conduziram um estudo sobre a percepção humana de justiça (*fairness*) e ausência de preconceito nos sistemas automáticos de estipulação de valor de fiança a presos, e concluíram que modelos realísticos são, consequentemente, necessariamente imperfeitos em relação às diferentes definições de *fairness* na IA. A verdade é que a natureza caixa-preta dos modelos tradicionais de IA torna difícil entender por que uma IA falha em certos casos, e quais são esses casos, tornando difícil prever seu comportamento.



Fonte: D. Jin et al. (2020).

Figura 5 – Demonstração (em inglês) da suscetibilidade a "Adversarial Attacks" (ataques adversariais) de um sistema de análise de sentimentos nas críticas de filmes. Observa-se que ao modificar as palavras por sinônimos menos utilizados leva o classificador a uma decisão errada.

E como os modelos de IA se comportam nos casos em que ele não foi treinado para lidar (casos que não constam dos dados de treinamento)? A resposta é que os modelos baseados em aprendizado podem produzir saídas inesperadas a entradas inesperadas. Sua vulnerabilidade, acima de tudo, pode ser avaliada através de *Adversarial Attacks*, ou ataques adversariais (Goodfellow; Shlens; Szegedy, 2015), onde os dados são modificados de modo a não se apresentarem da maneira convencional, confundindo o modelo. Os métodos de ataques adversariais buscam descobrir quais são os casos em que o modelo vai falhar e geram uma entrada que provoque este efeito. Um exemplo claro é quando uma palavra é escrita com erros ortográficos ou usando *leetspeak*, como relatado na seção anterior. Nesse caso, o modelo nunca viu esta palavra e não pode reconhecê-la. Entretanto, isso pode acontecer de uma forma mais sutil, como no exemplo da Figura 5, a troca de palavras por seus sinônimos levou a uma previsão diferente e incorreta do classificador. Ou mesmo em imagens, como mostra a Figura 6 onde a introdução de ruído na imagem engana o classificador na previsão errada da classe. A noção de *Adversarial Robustness* (robustez adversarial) dos sistemas de IA é, então, a capacidade do modelo de considerar iguais duas coisas que são iguais aos olhos humanos, como os exemplos nas Figuras 5 e 6.



Fonte: Goodfellow, Shlens e Szegedy (2015).

Figura 6 – Uma demonstração do método de geração rápida de exemplos adversariais (Goodfellow; Shlens; Szegedy, 2015) aplicado à rede GoogLeNet (Szegedy et al., 2015) no conjunto de dados da ImageNet (Krizhevsky; Sutskever; Hinton, 2012). Ao adicionar um vetor imperceptivelmente pequeno cujos elementos são iguais ao sinal dos elementos do gradiente da função de custo com relação à entrada, a classificação da imagem da GoogLeNet pode ser alterada, sem que a mudança seja perceptível ao humano.

Embora haja forte pesquisa sobre como mitigar esses efeitos na Inteligência Artificial, as falhas potenciais apresentadas têm levantado várias preocupações, questões e discussões sobre a evolução da IA (Sichman, 2021), especialmente no que diz respeito à aplicabilidade das ferramentas habilitadas. A interação humano-computador de agentes baseados em aprendizado é um aspecto importante e muito discutido, mas o papel desses agentes como moderadores, julgando efetivamente o conteúdo dos usuários, é uma nova característica que acrescenta complexidade ao tema. Além disso, a falta de transparência em como se dá o processo de moderação bem como o acesso aos dados dificulta que pesquisadores entendam melhor o processo, em busca de aperfeiçoá-lo. A natureza opaca desse processo pode semear teorias conspiratórias como a investigada e refutada por Jiang, Robertson e Wilson (2020), de que a moderação das mídias sociais tem um viés a favor da esquerda. Além disso, as alegações trazidas são difíceis de validar, pois nem os pesquisadores nem os críticos podem acessar dados referentes a conteúdos removidos por decisões de moderação. Seria, portanto, de grande importância que o conteúdo moderado/removido fosse preservado e protegido por grandes plataformas, permitindo pesquisas que busquem melhorar o processo de decisão em relação à moderação.

Nesse contexto, é importante observar que abordagens que não utilizam IA podem mitigar o efeito do conteúdo nocivo nas plataformas. As técnicas já utilizadas nas redes sociais variam desde algoritmos automáticos, baseados em conjuntos de regras especificadas pelo ser humano, até o reforço ativo das

políticas de autenticação, visando reduzir o risco de contas falsas e forçando os usuários a saírem do anonimato, expondo-os a responsabilização legal pelo compartilhamento de conteúdo nocivo. Além disso, políticas de monetização que observem a natureza do conteúdo postado também podem igualmente desencorajar o compartilhamento desse tipo de conteúdo. Winchcomb (2019) detalha essas técnicas, citando também a censura dos usuários infratores, que procura impor restrições sociais (como a limitação do número de interações que o usuário pode ter), e, finalmente, a curadoria de conteúdo realizada por algoritmos. Essa curadoria, relacionada à indicação de conteúdo aos usuários, pode priorizar a indicação de conteúdo nocivo ou incentivar a produção de conteúdo nocivo, dependendo dos seus critérios. Por exemplo, priorizar conteúdo com muitos acessos ou interações é potencialmente perigoso, pois o conteúdo nocivo tem como pressuposto gerar engajamento e compartilhamento. Versões modernas de algoritmos de curadoria tratam deste problema e não necessariamente empregam técnicas de IA para fazê-lo. Por exemplo, plataformas como Instagram e YouTube minimizam o engajamento e até desmonetizam conteúdo que contém termos listados como nocivos.

Dilemas legais do atual paradigma de moderação

Nesta seção, pretende-se fornecer um breve panorama sobre como a moderação de conteúdo é debatida em todo o mundo, para que se possa compreender como modelos de *Explainable AI* podem influenciar – ajudando ou dificultando – a proteção dos direitos na Internet. A moderação de conteúdo online tem suscitado importantes discussões que relacionam aspectos teóricos e proposições com limitações práticas tanto dos Estados quanto das plataformas digitais para implementar regulamentações na esfera virtual.

Pode-se considerar que o ponto de partida da discussão é a questão de saber se os direitos humanos no domínio digital foram suficientemente conceituados e discutidos para que, com uma estrutura teórica robusta, possamos avaliar problemas concretos e planejar a política em conformidade. Uma política concebida para proteger os direitos relacionados às novas tecnologias pode ser vista como necessitando de mais evidências factuais (Waldron, 2003) ou como equivalente aos direitos *offline*, o que é seguido como premissa normativa nos instrumentos jurídicos internacionais (Tuori, 2019).

De qualquer forma, o atual paradigma de balanceamento dos direitos humanos a fim de protegê-los é o que guia as interações e o planejamento de políticas a partir de hoje. Os paradigmas podem mudar se novos direitos surgirem com o desenvolvimento da tecnologia, mas se não for este o caso, os paradigmas atuais ainda permanecerão (Jóri, 2016). Teorias como a visão de Robert Alexy (2014) sobre proporcionalidade são utilizadas para equilibrar direitos em estudos para reunir mais evidências factuais ou aplicadas aos direitos on-line da mesma forma que seriam aplicadas aos direitos *offline*. Em vez de mudar o paradigma, vale a pena observar os padrões pelos quais uma análise de propor-

cionalidade se orientaria para equilibrar os direitos, para que seja possível avaliar melhor como a *Explainable AI* pode melhorar a moderação do conteúdo.

Idealmente, além da legalidade, a moderação de conteúdo online deveria seguir quatro ideias principais a fim de proteger um ambiente de debate verdadeiramente democrático e o compartilhamento de ideias: *legitimidade, transparência, controle* e *capacidade de execução* (Sander, 2020).

No caso da legitimidade, a restrição da liberdade de expressão dos usuários deve ser prevista por lei (Comitê de Direitos Humanos das Nações Unidas, 2011), conforme o Comentário Geral n.34 do Comitê de Direitos Humanos das Nações Unidas. Com isso em mente, vemos que os termos de serviço e os padrões comunitários de empresas como as plataformas de mídia social se adaptaram para cumprir várias regulamentações diferentes ao mesmo tempo, criando um diálogo entre as legislações de diferentes países que regulamentam o mesmo objeto, por exemplo, a interação dos usuários e a moderação dos comentários. Isso não quer dizer que os Estados abdicaram de sua soberania, mas sim que, ao cumprir com várias regulamentações de direitos humanos, a abordagem necessária para as empresas é extensa.

Isso nos leva aos aspectos de controle e capacidade de execução (*enforcement*), que estão intrinsecamente ligados. Embora os Estados tenham legitimidade para criar a regulamentação sobre moderação de comentários, eles não necessariamente possuem a capacidade de agir e monitorar os comentários, mesmo se levarmos em conta as agências especializadas.

A questão do tempo é central neste caso, pois os direitos em jogo podem ser perdidos pela capacidade do ambiente virtual de disseminar e perpetuar informações que possam violar a privacidade de um indivíduo. A tendência geral para garantir controle é a criação de normas e diretrizes legais pelos Estados sobre como as empresas – as entidades que têm mais controle sobre tais situações – podem agir sobre tais situações.

Tais caminhos podem variar, mas ainda assim se enquadram no mesmo guarda-chuva de criar padrões para que as empresas exerçam sua capacidade de executar com moderação. O caso Delphi mostrou que a Corte Europeia de Direitos Humanos decidiu aplicar *online* os mesmos padrões usados para a mídia tradicional:

> Discursos difamatórios e outros tipos de discurso claramente ilegais, incluindo discursos de ódio e incitação à violência, podem ser divulgados como nunca antes, em todo o mundo, numa questão de segundos, e por vezes permanecem persistentemente disponíveis online. Estas duas realidades conflituosas estão no cerne deste caso. Tendo em conta a necessidade de proteger os valores subjacentes à Convenção, e considerando que os direitos ao abrigo dos artigos 10 e 8 da Convenção merecem o mesmo respeito, deve ser encontrado um equilíbrio que mantenha a essência de ambos os direitos. (ECtHR, 10 de Outubro de 2013, § 110)

Isso nos mostra que, para conseguir transparência, é preciso normas legais claras para regular a moderação e a tecnologia associada a ela. EsSa tem sido a razão de ser tanto do Regulamento Geral de Proteção de Dados da União EuropEia (GDPR) quanto da Carta de Direitos da Internet do Brasil (*Marco Civil da Internet*). Limitações claras e princípios ainda mais claros a serem aplicados nas normas de moderação de conteúdo são usados para proporcionar controle e permitir a capacidade de executar e regular comentários, o primeiro incluindo a obrigação das empresas de equilibrar direitos conflitantes *online*.[3]

Como o paradigma do equilíbrio de direitos se afirma, a criação de IA transparente que possa ter seus padrões em conformidade com tais regulamentações só pode ser alcançada através da tecnologia de *Explainable AI*.

Explainable AI e Oportunidades na Moderação de Conteúdo

Um dos maiores desafios na IA é lidar com sua complexidade e opacidade. Entender como a tecnologia funciona é um passo crítico para a realização de outros princípios, tais como responsabilidade, controle humano da tecnologia, segurança e proteção, e justiça e não discriminação (Kiritchenko; Nejadgholi; Fraser, 2021). A IA Transparente pretende lançar luz sobre o processo de criação de um sistema automático e torná-lo compreensível para as diferentes partes interessadas (Arbix, 2020). A transparência pode se referir a várias práticas, dependendo de quem é o público e dos beneficiários das explicações (Weller, 2019). Para os usuários, alude à *explicabilidade* (ou interpretabilidade) associada à boa documentação que lhes permite utilizá-la. Neste sentido, a explicabilidade pode ser vista como a capacidade dos modelos de fornecer explicações para suas decisões.

Quando se trata de explicabilidade, modelos mais complexos (como as redes neurais profundas) tendem a ter suas decisões mais difíceis de serem interpretadas. Um sistema de *Explainable AI* pode consistir em um modelo que tem menos complexidade e, portanto, é intrinsecamente capaz de fornecer explicações para suas decisões (*Model-Specific XAI*), ou pode confiar em outro modelo que é responsável por auditar suas decisões e fornecer um conjunto de explicações acrescentadas *post-hoc* (*Model-Agnostic XAI*). Estas explicações podem ser geradas para cada decisão (Explicação Local), identificando as razões para esta decisão específica, ou podem ser trazidas globalmente (Explicação Global), quando o que importa é a compreensão de toda a lógica de um modelo, e seguindo todo o raciocínio que leva a todos os diferentes resultados possíveis (Adadi; Berrada, 2018). EsSas explicações são geralmente fornecidas de forma mais técnica, fazendo referências às variáveis, e cabendo ao sistema gerar visualizações para o usuário. Entretanto, a demanda pelo desenvolvimento de sistemas capazes de produzir Explicações da Linguagem Natural (XAI Natural) (Camburu et al., 2018) também está crescendo.

A promoção da explicabilidade nas decisões tomadas pelos algoritmos tem uma série de vantagens. Na subseção seguinte, são apresentados os beneficiários e é discutido casos concretos de aplicação.

Aplicação concreta da Explainable AI

Shin (2021) traz para a discussão o conceito de causabilidade, que precede a explicabilidade. A causabilidade justifica o que deve ser explicado, e por quê. Isto é feito determinando a importância relativa da explicabilidade de cada *feature*, um passo subsequente para promover a transparência dos algoritmos. No estudo de caso apresentado, é evidente que dar aos usuários explicações sobre por que certas notícias são recomendadas gera confiança, ao mesmo tempo em que oferecer informações causais a respeito da explicação gerada promove segurança emocional para os usuários. Os resultados destacam as oportunidades que se discute neste trabalho, mostrando a importância de incluir a causabilidade e a explicabilidade nos sistemas de IA.

Esta importância tem múltiplos aspectos. Por um lado, é comum associar a XAI com a possibilidade de auditar tecnicamente as decisões de modelos, ajudando engenheiros a interpretar o que produziram. No entanto, XAI pode e deve ser utilizada para gerar explicações para os usuários. Apesar disso, Bhatt et al. (2020) mostram como há uma distância entre a transparência desejada e XAI na prática, pois a adoção da explicabilidade ultimamente ainda tem servido mais às partes interessadas internas do processo do que aos usuários finais.

Outro aspecto importante é a falta de unidade no que concerne aos objetivos/oportunidades do XAI: grupos de diferentes áreas, ao estudar a adoção do XAI, buscam objetivos diferentes. Portanto, Mohseni et al. (2021) estudam esse fenômeno numa tentativa de sistematizar os métodos de avaliação e objetivos do projeto do XAI. Quatro objetivos/oportunidades principais são definidos, relacionados aos usuários leigos em IA (novatos em IA) a serem buscados quando se oferecem explicações:

- **Transparência algorítmica:** explicar como funciona o sistema inteligente.
- **Confiança dos usuários:** melhorar a confiança dos usuários finais no sistema inteligente.
- **Mitigação de vieses:** ajudar os usuários humanos a inspecionar se os sistemas inteligentes são tendenciosos.
- **Conscientização sobre privacidade:** fornecer um meio para que os usuários finais avaliem sua privacidade de dados.

O trabalho realizado por Mohseni et al. (2021) também analisa as oportunidades geradas para outros tipos de usuários, definindo especialistas em dados e especialistas em IA. Assim, ele também define os objetivos:

- **Visualização e Inspeção de Modelos**: similar aos novatos da IA, *os usuários especializados* também podem se beneficiar da capacidade de interpretação da aprendizagem da máquina. Isto lhes permite inspecionar a incerteza e confiabilidade dos modelos.
- **Seleção e Ajuste de Modelos:** abordagens analíticas visuais, por exemplo, podem ajudar *os especialistas* a fazer um ajuste fino dos parâmetros de apren-

dizagem de máquinas para seus domínios específicos. Isto facilita a comparação de vários modelos e a seleção do modelo certo para seus dados.

• **Interpretabilidade de Modelos:** permite obter novos conhecimentos sobre os padrões de aprendizagem de modelos profundos.

• **Depuração de Modelos:** aumenta a capacidade dos *pesquisadores* de usar técnicas de interpretabilidade com o objetivo de melhorar a arquitetura de modelos e o processo de treinamento.

Nesse sentido, a XAI cria benefícios para os usuários finais e também permite a responsabilização dos desenvolvedores e mantenedores. A promoção da explicabilidade para usuários leigos pode gerar impactos sociais relevantes, uma vez que a esmagadora maioria dos usuários está incluída nesta categoria. No entanto, no contexto de engenharia de algoritmos e modelos, a sociedade é representada pelas outras duas categorias (especialistas em dados e especialistas em IA). Os efeitos da promoção da explicabilidade visando melhorar a visualização e inspeção dos modelos ou sua seleção e treinamento, por exemplo, podem melhorar *a confiança da sociedade* nas decisões desses modelos. Além disso, promover a interpretabilidade e viabilizar a depuração dos modelos permite que possíveis erros aos quais toda IA está sujeita sejam mais facilmente abstraídos pela sociedade, *reduzindo a desconfiança no processo*. A seguir, discutimos alguns casos em que a IA explicável foi aplicada no contexto da moderação de conteúdo.

Mohseni et al. (2021) foram capazes de produzir evidências de que critérios transparentes e decisões explicadas ao usuário têm o potencial de reduzir a recorrência do compartilhamento de notícias falsas. Além disso, eles mostraram que a transparência na moderação de conteúdo trazida pela XAI ajuda a construir uma confiança apropriada nos modelos de IA, que entende suas limitações, mas também seu potencial. Este resultado aumenta o potencial pedagógico que o uso de *Explainable AI* tem, o que pode ajudar o usuário a compreender os riscos de compartilhar um certo tipo de conteúdo.

Por outro lado, Kou e Gui (2020), ao analisar uma grande quantidade de comentários e interações em relação ao sistema de IA de punição a jogadores por quebra de regras do jogo *League of Legends (LoL)*, notaram a necessidade de uma explicação para as decisões dos critérios de IA, especialmente em funções de seus critérios em razão dos seus valores e normas sociais, de clarificação quanto às especificidades de seu funcionamento e de maneiras de como se evitar penalidades no futuro. O estudo conclui que no caso ideal a *Explainable AI* não conseguiria ser uma resposta universal satisfatória para todo e qualquer caso, uma vez que a comunidade tem um papel importante em ajudar a resolver essas dúvidas. Entretanto, a *Explainable AI* poderia desempenhar um papel essencial para aproximar a comunidade dos desenvolvedores, contribuindo principalmente em três pontos: tornar acessível explicações a nível técnico mais complexo; permitir que os próprios usuários tenham mais ferramentas para um melhor

diálogo e entendimento sobre como funciona a moderação; e contextualizar as decisões da IA explicando quais regras ela levou em conta em cada decisão.

Um projeto que pode ser inspirador e relevante neste sentido é o *Twitter Birdwatch* (Coleman, 2021), que permite que as pessoas identifiquem informações em Tweets que elas acreditam serem enganosas e que escrevam notas que forneçam um contexto informativo a estas. Isto permite uma resposta rápida quando informações enganosas se espalham, acrescentando contexto que as pessoas confiam e acham valioso. As notas são tornadas visíveis diretamente no Tweets para o público global do Twitter quando há consenso de um conjunto amplo e diversificado de colaboradores. Em vez de usar uma IA para gerar as explicações, este sistema usa as pessoas da comunidade para gerar explicações que podem ser usadas na tomada de decisões do algoritmo de moderação e cumprir um papel pedagógico entre os usuários.

Caminhos para um novo paradigma de moderação de conteúdo

Apesar de serem entidades privadas, as plataformas digitais se tornaram o espaço público de discussão. Elas são tão grandes que não podem estar alheias à sociedade. Isso levanta questões importantes quanto a legitimidade e a transparência com que desempenham a moderação e definem suas leis internas (normas da comunidade), já que suas decisões passam a afetar uma parcela enorme da população. Neste trabalho, argumenta-se que *Explainable AI* tem um potencial de atenuar algumas destas deficiências apontadas na moderação de conteúdo.

Com relação à legitimidade da moderação, é possível que a sociedade, através do Estado e das vias democráticas, crie padrões de moderação universal que garantam o respeito aos direitos fundamentais, inclusive o da liberdade de expressão, mas também garantindo que esta liberdade não ultrapasse os limites da lei. O uso das ferramentas de *accountability* apresentadas no arcabouço da *Explainable AI* pode permitir que estas regras definidas em sociedade possam ser implementadas pelas plataformas, e deste modo possam ser acompanhadas, auditadas e aperfeiçoadas.

Permitir que a sociedade possa auditar os modelos de IA em implementação e opinar nas políticas de moderação torna o processo mais legítimo. Há estudos, como o realizado por Vaccaro et al. (2021), que apontam a necessidade de uma moderação representativa, onde se garanta que as pessoas possam ter representantes dentro dos órgãos moderadores das plataformas digitais. Nesse sentido, é importante que o máximo possível de dados produzidos no processo de moderação seja de domínio público: conteúdos removidos, políticas de moderação, critérios de decisão, métricas e probabilidade/confiança dos modelos, entre outros. Isso é essencial para viabilizar que pesquisadores investiguem e colaborem nestes processos.

Com relação à transparência para o usuário final, a produção de explicações locais em linguagem natural (XAI Natural) sobre a razão pela qual um determinado conteúdo foi removido pode desempenhar um papel ainda mais

importante do que a própria remoção por si para a segurança do debate público. Em vez de excluir esse cidadão do debate, isso inclui e pedagogicamente dá aos usuários a oportunidade de aprender a serem mais críticos sobre o conteúdo que compartilham nas redes. Existe um potencial, a ser explorado em futuras pesquisas sobre a interação humana-computador em *Explainable AI*, de que o uso do XAI possa ajudar a reduzir a reincidência do compartilhamento de conteúdo nocivo. Ainda abre-se a oportunidade de usar outras abordagens ao invés da remoção direta do conteúdo, como sinalizar (*flagging*), limitar o número de interações que o post pode ter, e desmonetizar, dependendo do escopo que a publicação tenha tido.

Mais pesquisas interdisciplinares, envolvendo especialistas de todas as áreas (direito, sociologia, ciência da computação, ética, política), se fazem necessárias para que possamos construir um modelo mais concreto para moderação de conteúdo que abarque os conceitos aqui apresentados. Este estudo se limitou a uma pesquisa exploratória interdisciplinar que traz uma proposta preliminar de como deveria ser a moderação de conteúdo na internet. No entanto, foi possível demonstrar que a introdução da *Explainable AI* nesse debate tem um potencial muito grande de colaborar para que a internet seja um espaço de debate mais justo e saudável.

Conclusão

Neste trabalho, proporcionamos um panorama abrangente sobre a aplicação de Inteligência Artificial na Moderação de Conteúdo da Internet. Foi adotada uma abordagem interdisciplinar, com o intuito de oferecer uma perspectiva abrangente do processo de moderação empregado em grandes plataformas. Entre os resultados obtidos, destaca-se o crescimento exponencial da porcentagem de conteúdo removido automaticamente por meio de IA. Ademais, exploramos as implicações éticas e sociais associadas à utilização de sistemas de IA, delineando suas limitações e os desafios legais que surgem em decorrência de sua implementação. Introduzimos, também, a *Explainable AI,* destacando as oportunidades que o conceito proporciona, bem como ilustrando casos práticos nos quais essa abordagem tem sido aplicada com sucesso na moderação de conteúdo, potencialmente servindo como fonte de inspiração.

A principal contribuição deste estudo foi apresentar uma alternativa para tornar o uso da inteligência artificial na moderação de conteúdo mais transparente para o usuário final e legítimo para a sociedade, e isto envolve a adoção de *Explainable AI* associada a critérios de moderação definidos em conjunto pela sociedade. Esta configuração poderia constituir um novo paradigma de moderação de conteúdo, mais justo e mais ético, no qual o Estado, as Plataformas Digitais e os próprios cidadãos exercem papéis relevantes e definidos.

A adoção de um processo mais legítimo garante o respeito às liberdades individuais sem descuidar dos limites legais. Por sua vez, adotar uma abordagem transparente de moderação traz benefícios que se estendem além da proteção

dos direitos das minorias. A transparência no centro do processo de moderação tem o potencial de torná-lo mais eficaz, inclusive reduzindo a recorrência de abusos virtuais, compartilhando notícias falsas, etc., contribuindo para um ambiente virtual mais saudável.

Notas

1 Disponível em: <https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/>.

2 A métrica *recall* em estatística é a razão entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas. É uma métrica relevante porém limitada, já que não mede quantos falsos positivos o sistema produziu. Conceitos relacionados que podem ser úteis para a compreensão são: precisão, acurácia e matriz de confusão.

3 Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho de 27 de abril de 2016 relativo à proteção das pessoas físicas no que diz respeito ao processamento de dados pessoais e à livre circulação desses dados, e que revoga a Diretiva 95/46/CE (Regulamento Geral de Proteção de Dados).

Referências

ADADI, A.; BERRADA, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, IEEE, v.6, p.52138-60, 2018.

ALCOFORADO, A. et al. Zeroberto: Leveraging zero-shot text classification by topic modeling. In: PINHEIRO, V. et al. (Ed.) *Computational Processing of the Portuguese Language*. Cham: Springer International Publishing, 2022. p.125-36. ISBN 978-3-030-98305-5.

ALEXY, R. Constitutional rights and proportionality. *Revus. Journal for Constitutional Theory and Philosophy of Law/Revija za ustavno teorijo in filozofijo prava*, Klub Revus–Center za raziskovanje evropske ustavnosti in demokracije, n.22, p.51-65, 2014.

ARBIX, G. A transparência no centro da construção de uma IA ética. *Novos estudos Cebrap*, SciELO Brasil, v.39, p.395-413, 2020.

ARRIETA, A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, Elsevier, v.58, p.82-115, 2020.

BHATT, U. et al. Explainable machine learning in deployment. In: PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY. [S.l.: s.n.], 2020. p.648-57.

BROWN, T. B. et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*, v.33, p.1877-901, 2020.

CAMBURU, O.-M. et al. E-SNLI: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, v.31, p.9539-49, 2018.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, ACM New York, NY, USA, v.6, n.1, p.1-6, 2004.

CIAMPAGLIA, G. L. et al. How algorithmic popularity bias hinders or promotes quality. *Scientific reports*, Nature Publishing Group, v.8, n.1, p.1-7, 2018.

COLEMAN, K. *Introducing Birdwatch, a Community-Based Approach to Misinformation*. [S.l.]: Twitter, 2021.

DANKS, D.; LONDON, A. J. Algorithmic bias in autonomous systems. In: PROCEEDINGS OF THE 26TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. [S.l.: s.n.], 2017. p.4691-7.

DEVLIN, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: PROCEEDINGS OF THE 2019 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ACL. [S.l.: s.n.], 2019. p.4171-86.

DUARTE, N.; LLANSO, E.; LOUP, A. Mixed Messages? The Limits of Automated Social Media Content Analysis. In: PMLR. *Conference on Fairness, Accountability and Transparency*. [S.l.], 2018. p.106.

ECTHR. Delfi v. Estonia 64569/09. 2013. Delfi (n 9) para 110, Ibid para 59 (10 october 2013).

ESTARQUE, M.; ACHERGAS, J. V. *Redes Sociais e Moderação de Conteúdo*: criando regras para o debate público a partir da esfera privada. [S.l.], 2021. Disponível em: <https://itsrio.org/wp-content/uploads/2021/04/Relatorio_RedesSociaisModeracaoDeConteudo.pdf>.

FELZMANN, H. et al. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, Springer, v.26, n.6, p.3333-61, 2020.

FERNÁNDEZ, A. et al. *Learning from imbalanced data sets*. [S.l.]: Springer, 2018. v.10.

FERRAZ, T. P. et al. DEBACER: a method for slicing moderated debates. In: SBC. *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2021. p.667-8.

GOODFELLOW, I.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS. [S.l.: s.n.], 2015.

GOODMAN, B.; FLAXMAN, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, v.38, n.3, p.50-7, 2017.

GOOGLE. *Google Transparency Report*. 2021. URL: <https://transparencyreport.google.com≥ [Online;accessed 15-Out-2021]. Disponível em: <https://transparencyreport.google.com≥.

HARRISON, G. et al. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In: PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY. [S.l.: s.n.], 2020. p.392-402.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, IEEE, v.21, n.9, p.1263-84, 2009.

JIANG, S.; ROBERTSON, R. E.; WILSON, C. Reasoning about political bias in content moderation. In: PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE. [s.l.: s.n.], v.34, n.9, p.13669-72, 2020.

JIN, D. et al. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In: PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE. [S.l.: s.n.], v.34, n.5, p.8018-25, 2020.

JÓRI, A. Protection of fundamental rights and the internet: a comparative appraisal of german and central european constitutional case law. *The Internet and Constitutional Law: The protection of fundamental rights and constitutional adjudication in Europe*. London; New York: Routledge Taylor and Francis Group, 2016.

KIRITCHENKO, S.; NEJADGHOLI, I.; FRASER, K. C. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, v.71, p.431-78, 2021.

KLONICK, K. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, v.131, p.1598, 2017.

_____. The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression. *Yale Law Journal*, HeinOnline, v.129, p.2418, 2019.

KOU, Y.; GUI, X. Mediating community-ai interaction through situated explanation: The case of ai-led moderation. In: PROCEEDINGS OF THE ACM ON HUMAN--COMPUTER INTERACTION, ACM New York, NY, USA, v.4, n. CSCW2, p.1-27, 2020.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, Springer, v.5, n.4, p.221-32, 2016.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v.25, p.1097-105, 2012.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Publishing Group, v.521, n.7553, p.436-44, 2015.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v.54, n.6, p.1-35, 2021.

META. *Facebook Transparency Report*. 2021. Disponível em: <https://transparency.fb.com≥. Acesso em: 10 nov.2021

MOHSENI, S. et al. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. In: PROCEEDINGS OF THE INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA. [S.l.: s.n.], v.15, p.421-31, 2021.

MOHSENI, S.; ZAREI, N.; RAGAN, E. D. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, ACM New York, NY, v.11, n.3-4, p.1-45, 2021.

NAHMIAS, Y.; PEREL, M. The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations. *Harvard Journal on Legislation, Forthcoming*, 2020.

NG, A. What artificial intelligence can and can't do right now. *Harvard Business Review*, v.9, n.11, 2016.

OLIVA, T. D.; ANTONIALLI, D. M.; GOMES, A. Fighting hate speech, silencing drag queens? Artificial Intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, Springer, v.25, n.2, p.700-32, 2021.

REIS, J. C. et al. Explainable Machine Learning for Fake News detection. In: PROCEEDINGS OF THE 10TH ACM CONFERENCE ON WEB SCIENCE. [S.l.: s.n.], p.17-26, 2019.

SANDER, B. Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation. *Fordham International Law Journal*, v.43, n.4, 2020.

SHIN, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, Elsevier, v.146, p.102551, 2021.

SICHMAN, J. S. Inteligência artificial e sociedade: avanços e riscos. *Estudos Avançados*, v.35, p.37-50, 2021.

SUSI, M. The Internet Balancing Formula. *European Law Journal*, v.25, n.2, p.198-212, 2019.

SZEGEDY, C. et al. Going deeper with convolutions. In: PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. [S.l.: s.n.], p.1-9, 2015.

TAN, F. et al. TNT: Text Normalization based Pre-training of Transformers for Content Moderation. In: PROCEEDINGS OF THE 2020 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP). [S.l.: s.n.], p.4735-41, 2020.

TUORI, K. Principles and policies: once more. In:_____. *The Quest for Rights*. [S.l.]: Edward Elgar Publishing, 2019.

UN Human Rights Committee. General comment no. 34: Article 19: Freedom of opinion and expression, u.n. doc. ccpr/c/gc/34. 2011. 12 Sept. 2011 [hereinafter General Comment No. 34], para. 25.

VACCARO, K. et al. Contestability for content moderation. *Proceedings of the ACM on Human-Computer Interaction*, ACM New York, NY, USA, v.5, n.CSCW2, p.1-28, 2021.

WALDRON, J. Security and liberty: The image of balance. *The Journal of Political Philosophy*, v.11, n.2, p.191-210, 2003.

WELLER, A. Transparency: motivations and challenges. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. [S.l.]: Springer, 2019. p.23-40.

WINCHCOMB, T. *Use of AI in Online Content Moderation*. [S.l.], 2019. Disponível em: <https://www.cambridgeconsultants.com/insights/whitepaper/ofcom-use-ai--online-content-moderation>.

RESUMO – O uso massivo de Inteligência Artificial na moderação de conteúdo na internet é uma realidade dos tempos atuais. No entanto, isso levanta uma série de questionamentos, seja sobre a pertinência do uso de sistemas automáticos opacos, seja se as plataformas podem sozinhas tomar decisões que antes cabiam ao Estado. Nesse contexto, o uso de IA "caixa-preta" passa a ser considerado uma ameaça à liberdade de expressão. Por outro lado, manter conteúdos que promovam abuso virtual é igualmente danoso a este direito fundamental. Nesse cenário, este estudo sumariza os principais problemas apontados pela literatura quanto ao paradigma atual, avalia as respostas que as novas tecnologias trazem, e propõe um caminho para um novo paradigma de moderação que seja justo e ético, no qual Estado e plataformas de mídias sociais têm papel relevante. Esse passa pela adoção de IA explicável associada a critérios transparentes e legítimos definidos pela sociedade.

PALAVRAS-CHAVE: Humanidades digitais, Moderação automática de conteúdo, IA explicável, Liberdade de expressão na internet, Ética na Inteligência Artificial.

ABSTRACT – The massive use of Artificial Intelligence in Content Moderation on the internet is a reality of our times. However, this raises a number of questions, such as whether the use of opaque automatic systems is pertinent, or even whether platforms alone can make decisions that used to be made by the State. In this context, the use of *black box* AI comes to be considered a threat to freedom of expression. On the other hand, keeping content that promotes virtual abuse is equally harmful to this fundamental right. In this scenario, this study summarizes the main problems pointed out by the literature regarding the current paradigm, evaluates the responses that new technologies bring, and proposes a path for a new moderation paradigm that is fair and ethical in which the State and social media platforms play a relevant role. That involves the adoption of *Explainable AI* associated with transparent and legitimate criteria defined by society.

KEYWORDS: Digital humanities, Automatic content moderation, Explainable AI, Freedom of expression on the internet, Ethics in Artificial Intelligence.

Thomas Palmeira Ferraz é doutorando em Ciência da Computação na Télécom Paris e École Polytechnique, Institut Polytechnique de Paris. É engenheiro pela Escola Politécnica da Universidade de São Paulo (USP) e mestre em Matemática Aplicada e Inteligência Artificial pela École Normale Supérieure Paris-Saclay (ENS).
@ – thomas.palmeira@telecom-paris.fr / http://orcid.org/0000-0002-5385-9164.

Caio Henrique Dias Duarte é mestrando no Departamento de Direito Internacional e Comparado da Faculdade de Direito da Universidade de São Paulo (USP), e bacharel em Direito pela mesma universidade. @ – caio.henrique.duarte@usp.br / https://orcid.org/0000-0002-1720-7249.

Maria Fernanda Ribeiro é pesquisadora colaboradora da Universidade de São Paulo (USP) e é engenheira graduada pela mesma instituição.
@ – maria.fernanda.ribeiro@alumni.usp.br / http://orcid.org/0000-0002-4340-9901.

Gabriel Goes Braga Takayanagi é estudante do Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo.
@ – gabriel.takayanagi@usp.br / https://orcid.org/0000-0002-6498-6716.

Alexandre Alcoforado é mestrando em Engenharia de Computação na Escola Politécnica da Universidade de São Paulo, e engenheiro elétrico pela mesma instituição. @ – alexandre.alcoforado@usp.br / http://orcid.org/0000-0003-3184-1534.

Roseli de Deus Lopes é professora titular da Escola Politécnica da Universidade de São Paulo (USP), vice-coordenadora do Centro Interdisciplinar de Tecnologias Interativas (Citi-USP) e pesquisadora do Laboratório de Sistemas Integrados (LSI-USP), onde coordena projetos de pesquisa em mídia eletrônica interativa, com ênfase em aplicações na educação, inclusão e saúde. Atualmente, é diretora no Instituto de Estudos Avançados da USP. @ – roseli.lopes@usp.br / https://orcid.org/0000-0001-8556-6473.

Mart Susi é professor de Direito de Direitos Humanos e chefe de Estudos Jurídicos na Tallinn University, Estônia. @ – martsusi@tlu.ee / https://orcid.org/0000-0002-2624-4797.

Recebido em 27.2.2022 e aceito em 22.1.2024.

[I] Institut Polytechnique de Paris, École Polytechnique, Paris, França.

[II] Universidade de São Paulo, Faculdade de Direito, São Paulo, Brasil.

[III] Universidade de São Paulo, Faculdade de Engenharia, São Paulo, Brasil.

[I,IV,V] Universidade de São Paulo, Escola Politécnica, São Paulo, Brasil.

[VI] Universidade de São Paulo, Instituto de Estudos Avançados, São Paulo, Brasil.

[VII]Law School, Tallinn University, Tallinn, Estônia.