Inteligência artificial e internet: um olhar sobre o conteúdo de usuários e a sua moderação

Francisco Brito Cruz

Stockgiu/Freepik

resumo

O artigo organiza o papel que a IA tem e pode ter sobre a circulação de conteúdos na internet. À medida que a produção de conteúdo gerado por usuários atinge volumes sem precedentes, a IA emerge como uma ferramenta indispensável para criar, filtrar, organizar e avaliar a vasta quantidade de informações. De sua parte, a emergência contínua dessa tecnologia traz consigo desafios significativos, incluindo o risco de perpetuar vieses e injustiças sistêmicas. O trabalho explica os diferentes eixos de impacto da IA e mergulha no seu uso para a moderação de conteúdos na internet, crucial para a segurança de usuários e proteção a direitos. Abordando riscos e formas de proteção e mitigação, expõe um panorama dos debates sobre o tema no momento em que os grandes modelos de linguagem passam a ser considerados como possíveis aliados de plataformas de internet em seu trabalho de aplicar regras para conteúdos gerados por usuários.

Palavras-chave: IA e internet; plataformas; moderação de conteúdo; curadoria nas plataformas; riscos da IA na internet.

abstract

The article outlines the role that AI currently plays in the dynamics of production, distribution and consumption of content on the internet. As user-generated content reaches unprecedented volume, AI emerges as an indispensable tool for creating, filtering, organizing, and evaluating vast amounts of information. However, the continuous emergence of this technology brings significant challenges, including the risk of perpetuating biases and systemic injustices. The paper explains the different impact axes of AI and delves into its use for content moderation on the internet, which is crucial for user safety and protection of human rights. Addressing risks and their mitigation, it presents an overview of the ongoing conversations on the subject, especially as large language models are considered potential allies for internet platforms in their efforts to enforce rules for user-generated content.

Keywords: *AI and internet*; *platforms*; *content moderation*; *curation on platforms*; *AI and risks*.

Sketchepedia / Freepik

INTRODUÇÃO AO ESCOPO: INTERNET, PLATAFORMAS E CONTEÚDO GERADO POR USUÁRIOS

A inteligência artificial e internet são temas vastos por si sós. Ao apresentarem-se como tecnologias revolucionárias e repletas de ramificações e usos, escrever sobre o tema demanda escolhas e recortes. Na mesma interseção, por exemplo, é possível pesquisar o uso de tecnologias de aprendizagem de máquina para o oferecimento de uma miríade de serviços on-line pela internet, para a análise de dados pessoais e produção de perfis comportamentais para o direcionamento de propaganda ou, ainda, para vulnerabilizar a segurança da informação de uma série de atividades que fazemos pela rede.

Neste texto o recorte é o uso de IA no contexto da produção, curadoria e, mais especialmente, moderação de conteúdos gerados por usuários da internet, processos que em geral têm ocorrido a partir de plataformas que possuam componentes "sociais" ou decorrentes da geração de conteúdo por terceiros na internet. A escolha por esse recorte não é por acaso: apesar de uma série de usos de IA desconectados das chamadas "plataformas" apresentarem questões sociais e econômicas críticas (como o impacto no mundo do trabalho, as questões socioambientais

FRANCISCO BRITO CRUZ é diretor executivo do InternetLab e autor de *Novo jogo, velhas regras* (Letramento).

e de matriz energética ou de segurança cibernética), a emergência e permanência da discussão sobre "desinformação"[1] no debate público desde meados da década de 2010 impõe especial atenção a de que forma tais serviços se propõem a serem espaços para produção de conteúdo gerado por usuários, como por vezes realizam a curadoria automatizada de enorme volume deste conteúdo, e, finalmente, como criam e aplicam regras sobre o que pode ou não ser postado ou circular – e como.

Assim, se considerarmos que a desordem informacional[2] é um problema relevante a ser discutido e atravessado pelos avanços nas tecnologias de IA, é fundamental compreendermos como no coração desse fenômeno estão presentes serviços de compartilhamento de conteúdos na internet que possuem como base o recebimento e organização de informações geradas não por seus funcionários ou por pessoal contratado, mas sim por usuários alheios às estruturas dessas empresas.

Por mais que cada um desses serviços – ou "plataformas", para facilitar – escolha estratégias diferentes de geração de receitas (frequentemente ligadas à venda de publicidade direcionada por dados pessoais extraídos dos mesmos usuários), o que os une é que se construíram atraindo usuários que neles podem ao mesmo tempo colocar sua atenção para consumir conteúdo alheio e produzir conteúdo próprio em diversas atividades de acesso à informação, comunicação e entretenimento. No jargão, são plataformas que acondicionam, curam e apresentam "conteúdo gerado por usuários" (*user-generated content*, ou UGC) em diferentes formatos multimídia.

Compreendido este contexto, três atividades emergem como fundamentais na configuração de plataformas "de conteúdo gerado por usuários". Em primeiro lugar, a produção e postagem do conteúdo. Elas podem ser realizadas por indivíduos ou

1 Este é outro termo que merece comentário. Nos trabalhos desenvolvidos por este autor no âmbito do InternetLab, centro de pesquisa em direitos e tecnologias, *fake news* ou desinformação são alcunhas simplificadoras para um fenômeno que reflete a mudança de como as sociedades produzem, circulam e consomem informação política. Assim, o advento da internet como meio de comunicação política permitiu que qualquer indivíduo com conectividade se tornasse potencialmente um emissor de comunicação em massa, o que faz com que os novos conteúdos produzidos possivelmente se distanciem daqueles elaborados sob o imperativo do jornalismo profissional, de modo que as informações emitidas nem sempre têm como compromisso a busca por objetividade. Esse novo ambiente acaba por enfraquecer "ciclos de checagem da realidade", impulsionados pelo jornalismo profissional, em favor de ciclos de "retroalimentação de propaganda", nos quais a informação circula a partir de uma lógica político-partidária. A dinâmica da desinformação, assim, extrapola componentes de manipulação e de crise entre verdade e mentira, e envolve a própria forma como os indivíduos se relacionam com informação, em um processo de comunicação em rede no qual a autonomia desse indivíduo de produzir e compartilhar conteúdo ganha escala significativa (Brito Cruz, 2019).

2 Segundo Wardle e Derakshan (2017), o conceito de "desordem informacional" é o conjunto de três tipos de informação diferenciados de acordo com o seu grau de dano, sua inveracidade e a intenção presumida do seu remetente. São eles: (i) informação incorreta (*misinformation*); (ii) desinformação (*disinformation*); e (iii) conteúdo mal-informado (*malinformation*). Enquanto a informação incorreta corresponde ao fenômeno do compartilhamento de informações falsas sem a intenção de causar danos, a desinformação ocorre quando informações falsas são propagadas conscientemente para causar danos. Por sua vez, o conteúdo mal-informado contém informações genuínas, mas que são compartilhadas para causar danos, geralmente ao transferir para a esfera pública informações que deveriam permanecer privadas (por exemplo, vazamentos de e-mail, assédios on-line e discursos de ódio).

organizações de diversos tipos, de maneira mais ou menos organizada, e são sempre condicionadas pelos elementos arquitetônicos (ou *affordances*[3]) oferecidos pela plataforma. O exemplo mais clássico é o antigo Twitter, que apenas permitia postagens de texto que tivesse um número máximo de caracteres.

Em segundo, a atividade de curadoria e arquitetura de conteúdo gerado por usuários destina-se a organizar como tais peças multimídia irão viajar pela rede e serem entregues a diferentes audiências usuárias da plataforma, em diferentes contextos de uso. Na medida em que cada serviço possui uma estratégia de geração de receitas (que em geral passa por favorecer determinados usos e reter a atenção do público para revendê-la na forma de publicidade), cada um projeta a sua arquitetura a fim de proporcionar diferentes tipos de recomendação ou disposição gráfica de conteúdos gerados por terceiros. Essa distribuição pode ser simplesmente uma caixa de mensagens que atualiza conforme as mais recentes forem recebidas ou enviadas – como no WhatsApp – ou um construto complexo de ferramentas de busca e recomendação ativa (ou seja, aquela em que o usuário não precisa dizer o que procura) que dispõe o conteúdo em diversos formatos – como no Instagram e YouTube.

Por fim, a atividade de moderação de conteúdo caracteriza-se quando a plataforma formula, edita, publica e aplica normas sobre o que pode ou não ser realizado por seus usuários. Conforme já explicitado em outro trabalho no âmbito do centro de pesquisa InternetLab,

"[...] essa é uma atividade de elaboração e aplicação de regras [públicas]. Com a finalidade de construir diferentes ambientes digitais propícios a interações sociais e geração de conteúdos por seus usuários, cada plataforma decide que tipos de discurso (ou seja, de conteúdo em texto ou multimídia) serão permitidos, incentivados, desincentivados ou proibidos. [...] Internamente, por sua vez, detalham procedimentos de interpretação de tais regras privadas, organizam precedentes e constroem sistemas de aplicação para dar conta de todo o conteúdo que é gerado por seus usuários. Todas essas atividades, portanto, fazem parte do referido binômio de funções: elaborar e aplicar as normas que regem e gerenciam o comportamento de usuários de determinado espaço" (Brito Cruz et al., 2023).

A divisão proposta é esquemática e na realidade encontra evidentes sobreposições e afetações. A produção se sobrepõe à arquitetura na medida em

3 O conceito de *affordances* surgiu, primeiramente, com o psicólogo estadunidense James Gibson. De acordo com o estudioso, as *affordances* não se encontram no ambiente ou nos agentes, mas surgem a partir do encontro entre eles. Um exemplo dado por Gibson é de insetos que conseguem andar sobre a água. Um lago oferecerá *affordances* de andabilidade para indivíduos dessa espécie, mas não para os demais. No mesmo sentido, *affordances* em novas tecnologias não ditam o comportamento de seus usuários, mas configuram o ambiente de uma forma que molda o envolvimento destes. Segundo Boyd (2010), quatro tipos de *affordances* surgem na arquitetura de redes sociais: (i) persistência, pois as expressões on-line são automaticamente registradas e arquivadas; (ii) replicabilidade, pois o conteúdo produzido pode ser duplicado; (iii) escalabilidade, pois a visibilidade potencial do conteúdo é grande; e (iv) capacidade de pesquisa, pois o conteúdo produzido por outros usuários pode ser acessado por meio de pesquisa (Nascimento et al., 2022).

que só é possível produzir quando a plataforma é desenhada com essa possibilidade, por exemplo. A moderação também incide na produção quando uma proibição gera uma autocensura em um criador de conteúdo.

A sobreposição mais complexa de organizar é aquela entre curadoria e moderação, como classificar a atividade de diminuição na recomendação de um conteúdo extremista que continue permitido, por exemplo. O corte proposto neste trabalho baseia-se na publicidade prévia dessas regras e na sua possível aplicação. Assim, a moderação de conteúdo se diferencia da curadoria em dois aspectos. Primeiro, tais regras de comportamento e adequação (em geral derivadas de preocupações de segurança e integridade do serviço e dos demais usuários) são dispositivos públicos, não se confundindo com decisões comerciais de privilegiar recomendações disto ou daquilo, ou de fomentar algum estilo específico de curadoria. Segundo, se aplicam em situações as quais a arquitetura não conseguiu prever de antemão (quando a plataforma simplesmente não disponibiliza ferramentas para postagem de vídeos, eles não estariam proibidos por regra, mas impossibilitados arquitetonicamente).

Por fim, caracterizar tais atividades não pode prescindir de uma nota sobre a escala na qual se realizam. A expansão da conectividade e da capacidade de processamento em dispositivos acessíveis a importante fração da população global, nos anos 2000 e 2010, e as oportunidades e atrativos oferecidos por tais plataformas implicaram um crescimento vertiginoso que é fato notório, tornando-as suas controladoras empresas de valor trilionário. Com esse crescimento, veio um volume absolutamente sem precedentes de geração de conteúdo, que fez com que tais atividades (produção, curadoria/arquitetura e moderação) ocorressem em contexto, em escala industrial e massiva[4].

- No caso da *produção*, a escala multiplica exponencialmente as possibilidades do que será expresso por cada um dos usuários. Enorme diversidade interseccional abre imensas possibilidades de usos a partir de contextos sociais, econômicos, étnico-raciais, culturais e linguísticos – que passam a poder interagir entre si. Mesmo plataformas que oferecem serviços a determinados nichos não deixam de abarcar uma diversidade extraordinária na expressão.
- No eixo da *curadoria*, a escala é justamente o elemento que torna seus processos nevrálgicos na experiência de qualquer usuário. As enormes quantidades de spam presentes na internet são um dos elementos que podem impossibilitar que usuários encontrem aquilo que procuram, por exemplo. Essa organização ao mesmo tempo massiva e personalizada do que deve ser entregue/recomendado para cada audiência exigiu de muitos desses serviços o desenvolvimento de sistemas cada vez mais complexos de inferência sobre interesses e preferências baseados em dados pessoais.

4 Dados da edição mais recente da pesquisa Data Never Sleeps mostram, por exemplo, que a cada minuto 500 horas de vídeo são postadas no YouTube e mais de 347 mil tweets são postados. Disponível em: https://www.domo.com/data-never-sleeps#.

- Entendida sob este quadro, a *moderação* de conteúdo é um desafio ao mesmo tempo logístico e político/normativo. Político porque revela quais as escolhas que quem controla a plataforma faz sobre segurança e adequação, refletindo valores sociais e políticos e escolhas delicadas a respeito de questões de enorme tensão (como limites da expressão e discursos de ódio, por exemplo). Logístico porque requer que tais escolhas políticas sejam imbuídas em sistemas de detecção, avaliação e aplicação de regras que funcionem conforme quantidades colossais de denúncias, potenciais danos e fluxo de informações.

A IA E AS PLATAFORMAS: DA PRODUÇÃO À MODERAÇÃO DE CONTEÚDO

O atravessamento dos avanços tecnológicos no aprendizado de máquina ocorre em cada um dos tipos de atividades elencados acima – até porque tais processos já encontram usos correntes e atuais de ferramentas de IA. Tais usos se popularizaram conforme consolidados o desafio logístico e a dimensão industrial dessas atividades, em especial nos dois eixos que estão sob o controle direto das plataformas – o da curadoria e o da moderação.

Com efeito, sem tecnologias de aprendizagem de máquina e suas aplicações, plataformas de internet de conteúdo gerado por usuários não funcionariam da maneira como funcionam hoje. Não haveria feeds que recomendam conteúdo inferindo qual o nosso interesse ou o que cativa a nossa atenção, não existiriam boas ferramentas de busca para encontrar vídeos ou imagens que gostaríamos de lembrar, redes sociais estariam potencialmente repletas de *spam*[5] e conteúdo indesejado produzido por seus milhões de usuários. Não existiriam filtros ou bons mecanismos de edição de imagem e som para criadores de conteúdo, não seriam geradas legendas automáticas em vídeos e denúncias de conteúdo impróprio possivelmente demorariam muito mais para serem analisadas.

Para esquematizar o atravessamento mencionado, alguns dos usos de IA em plataformas de internet são elencados como exemplos na tabela 1.

Cada um dos eixos enseja discussões científicas relevantes no campo da regulação e de políticas públicas, em especial se consideradas as questões já mencionadas relacionadas ao tema da desordem informacional. O da produção de conteúdo, por exemplo, pode explorar tanto as possibilidades de pesquisa no âmbito da criatividade quanto a interseção com a tutela de direitos autorais, seja nas bases de treinamento ou nos resultados gerados pelas ferramentas de IA generativa. Esse eixo também pode abordar como tais conteúdos "sintéticos" serão recebidos por usuários de internet, anabolizando pro-

5 *Spam* é um termo em inglês utilizado para significar o envio em massa de mensagens não solicitadas pelo destinatário. Peças de spam não são, necessariamente, ilegais. Vistas individualmente, podem ser totalmente inofensivas e protegidas pela liberdade de expressão. Todavia, caso esse tipo de conteúdo passe a dominar a experiência de um usuário em uma rede social, esta poderá ter perdido sua utilidade (Monteiro et al., 2021).

TABELA 1

Usos de IA por tipo de atividade em plataformas de internet

	Produção	Curadoria/Arquitetura	Moderação
Usos de ferramentas de IA	• Geração de imagens e textos em IA generativa para postagem; • Edição de itens multimídia de maneira acelerada e facilitada, como em aplicativos que criam filtros de imagem, dublam vídeos ou geram legendas; • Controle de perfis, contas ou canais de forma automatizada, inclusive considerando interação com usuários humanos (como chatbots).	• Seleção e organização automatizada de peças de conteúdo a serem entregues em estrutura de feed ou similar em determinada ordem, a partir de inferências feitas sobre comportamentos de usuários; • Fornecimento de perfis ou canais sugeridos ao usuário a partir de seus interesses; • Seleçao e disposição de conteúdo relacionado a conteúdo consumido.	• Filtragem prévia de conteúdo a ser publicado a partir de regras específicas, como o treinamento de ferramentas que previnem o compartilhamento de conteúdo que contenha material protegido por direito autoral; • Detecção de postagens com conteúdo que potencialmente viola termos de uso e ranqueamento das possibilidades de violação a partir de precedentes; • Organização de ordem de análise de conteúdos por ação humana por inferência sobre sua urgência frente a critérios de risco.

cessos de desinformação e/ou polarização política. Na frente da arquitetura e curadoria, algoritmos de recomendação sendo capazes de fazer inferências mais certeiras ou relevantes sob determinados propósitos.

Este artigo vai explorar de forma mais detida as possibilidades e riscos em apenas um desses eixos, o da moderação de conteúdo. A escolha tem motivo na centralidade que o tema tem encontrado nas discussões regulatórias ao redor do globo, com especial eco na aprovação do Digital Services Act (DSA) pela União Europeia[6]. Normativas como o DSA exa-

6 O Digital Services Act (DSA) é um regulamento europeu cujo objetivo é regular os serviços digitais e criar um ambiente digital seguro, em que os direitos fundamentais dos usuários são respeitados. Embora não mencione explicitamente o termo "inteligência artificial", o DSA inclui várias referências ao uso de algoritmos e sistemas automatizados. No artigo 34, por exemplo, a legislação pontua que provedores de plataforma on-line e mecanismos de busca de grande porte devem identificar, analisar e avaliar diligentemente quaisquer riscos sistêmicos decorrentes da concepção, do funcionamento de seu serviço e de seus sistemas relacionados ou da utilização de seus serviços, incluindo sistemas algorítmicos. Por sua vez, os artigos 14 e 27 do DSA indicam que as plataformas são obrigadas a indicar nos termos e condições de seus serviços informações sobre moderação de conteúdo e sistemas de recomendação por algoritmos. Disponível em: https://eur-lex.europa.eu/eli/reg/2022/2065/oj. Acesso em: 25/jan./2024.

cerbam como a moderação de conteúdos é crescentemente percebida como uma tarefa crítica para mitigação dos danos que podem ser catalisados por ferramentas digitais.

A IA NA MODERAÇÃO DE CONTEÚDO: FUNCIONALIDADES E LIMITAÇÕES

O campo da IA soma-se nesse esforço na medida em que torna possível "ensinar" computadores a reconhecerem padrões e sinais típicos em conteúdos postados por usuários, entrando em cena quando falamos, por exemplo, das tarefas de detecção, priorização e análise de conteúdos que violam o conjunto de regras das plataformas.

Assim, é necessário destrinchar exemplificativamente a cadeia de processos que podem receber diferentes tipos de aplicações de IA. Na frente da detecção, por exemplo, é possível utilizar ferramentas de reconhecimento de padrões (em imagens ou outros conteúdos) para separá-los para uma análise. Como explica um glossário feito pela Digital Trust & Safety Partnership (DTSP), uma parceria de mais de 20 empresas donas de plataformas de internet para debaterem questões de moderação de conteúdo e segurança, a "automação pode ser usada para detectar potenciais abusos, por meio de métodos como filtragem de palavras-chave, correspondência de *hash* [que é a comparação entre o conteúdo analisado e conteúdos anteriores a partir de técnicas específicas], análise comportamental, aprendizado de máquina e inteligência artificial" (DTSP, 2023).

Em uma linha de priorização, por sua vez, é possível automatizar a análise de critérios de prioridade na "fila" de conteúdos a serem analisados por revisores humanos, a fim de tornar seu trabalho mais eficiente ou de treinar a partir de como essa priorização pode ser otimizada. Por mais que essa discussão sobre priorização seja um dos pontos mais obscuros ao debate público, casos recentes demonstram sua importância e sensibilidade (Brito Cruz [coord.]; Lana; Jost, 2023)[7].

Por fim, na frente da análise, máquinas podem aprender a identificar sinais de possíveis problemas a partir de bancos de dados de violações pregressas, fornecendo índices ou inferências sobre um conteúdo detectado ainda não analisado. Se a probabilidade de violação for muito alta, plataformas poderão programar seus sistemas para agir inclusive sem revisão humana prévia. Nesses casos, as definições do que é "muito alto" e os casos apropriados para esse tipo de decisão podem gerar polêmicas importantes[8].

Por mais que sejam tarefas diferentes, tais casos se conectam: um conteúdo pode ser detectado já com um índice alto de possível violação aos termos de uso – e, consequentemente, priorizado nas "filas" de análise. Como explicou Emma Llansó (2020), determinadas plataformas inclusive utilizam tais índices para realizar "filtragem" e aplicação de regras no momento do upload (ou seja, no tempo entre o ato da postagem pelo usuário e a publicação

7 Como já descrevemos em trabalho sobre o caso do sistema XCheck, operado pela Meta.

8 Em caso recente, a detecção de conteúdo de nudez infantil e a tomada de decisão rápida por seu alto risco tiveram consequências drásticas em uma família nos EUA após uma criança cometer um erro (Hill, 2024).

do conteúdo na internet). Tais práticas surgiram com força para detectar conteúdo protegido por propriedade intelectual (Hartmann; Silva, 2020) e abuso sexual infantil, mas há pressão para o seu uso em outros contextos.

Se considerada a escala necessária para dar conta do fluxo de informações geradas nos ambientes das plataformas digitais, seria impossível contar com processos "artesanais", ou seja, sem qualquer tipo de aprendizado de máquina. Os serviços e as qualidades por eles oferecidos estariam possivelmente inviabilizados e os sistemas de segurança seriam significativamente mais custosos. Em outro trabalho desenvolvido no âmbito do InternetLab, explicamos como "entre as boas práticas do setor está a composição equilibrada e complementar entre sistemas automatizados (e 'inteligentes') e a participação de colaboradores humanos que supervisionam, revisam e direcionam o trabalho de máquinas" (Brito Cruz et al., 2023).

RISCOS E SUAS ESTRATÉGIAS DE MITIGAÇÃO

Abuso, erros e problemas ocorrem na moderação de conteúdo – e nem sempre eles estão ligados ao uso da IA[9]. Mesmo que uma miríade de casos ocorra por má prática corporativa e ética, falta de investimento e erro humano, é certo que também toda ferramenta de IA poderá vir a cometer erros e apresentar vieses indesejados[10]. Pesquisas continuamente têm demonstrado como tais tecnologias podem reproduzir desigualdades estruturantes e potencializar violências preexistentes. Sua natureza é de serem produzidas e terem seu treinamento desenhado por pessoas que carregam em si mesmas experiências e perspectivas.

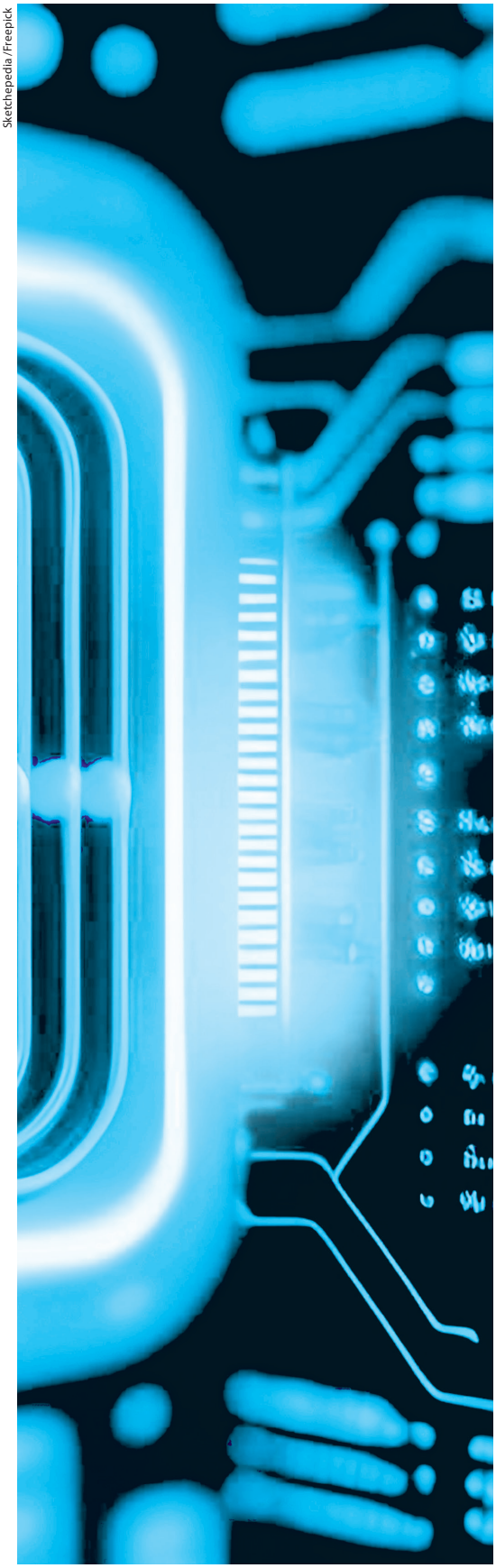
Com efeito, um dos problemas está relacionado às bases de dados dedicadas aos treinamentos de aprendizado de máquina. Tais bases podem possuir uma série de limitações ou mais dados sobre um fenômeno e não outro, bem como podem não estar atualizadas. Se uma máquina aprende padrões a partir do comportamento de usuários de um tipo de serviço, por exemplo, isso pode fazer com que não esteja preparada para lidar com uma demografia de usuários que usam tal serviço de maneira diferente, por qualquer motivo que seja.

Algumas evidências da detecção de vieses em sistemas de moderação de conteúdo que usam IA são relevantes para servir de alerta e projetar formas de mitigação a serem adotadas.

- O primeiro foi detectado por pesquisa realizada pelo InternetLab (Dias Oliva; Antonialli; Gomes, 2021; 2019). Neste caso, pesquisadores testaram uma ferramenta criada por uma subsidiária do Google para análise do "nível de toxicidade" de discursos em textos. O teste comparou como a tecnologia classificava escritos em inglês postados por *drag*

9 Como organizado em bom mapeamento em: Silva; Cesar (2022).

10 Para mais informações sobre viés algorítmico, ver: Noble (2018).



Sketchepedia/Freepick

queens, de um lado, e por reconhecidos nacionalistas brancos dos EUA, de outro. O achado é de que esse tipo de tecnologia pode não ser capaz de distinguir nuances de contextos, como, por exemplo, a diferença entre discursos de ódio contra pessoas LGBTQIA+ e conteúdo publicado por pessoas LGBTQIA+, que frequentemente ressignificam termos compreendidos como ofensivos, positivando-os para comunicação entre pares. Os perfis de *drag queens* foram considerados potencialmente mais tóxicos do que os perfis de supremacistas brancos, pois termos comumente usados por pessoas LGBTQIA+, como *sissy*, *gay*, *lesbian* e *queer* eram considerados potencialmente tóxicos pela ferramenta de IA, independentemente do contexto. O caso da pesquisa sobre *drag queens* demonstra distorções e vieses da IA, que não foi capaz de identificar as matizes específicas da comunicação de pessoas *queers*, em que o uso de uma linguagem "pseudo-ofensiva" é uma forma comum de comunicação entre membros da própria comunidade. Pesquisadores como Udupa, Maronikolakis e Wisiorek (2023) argumentam sobre a dificuldade de classificar e treinar sistemas para reconhecer "discurso extremo" mesmo dentro de um mesmo idioma.

- Um segundo exemplo se refere a detecção e moderação da nudez em plataformas de internet. Algumas plataformas proíbem, de forma legítima, a disseminação de imagens contendo partes de um corpo desnudo, como, por exemplo, a Meta. Isso ocorre não só por decisões de negócio sobre qual o tipo de conteúdo erótico é ou não aceitável, mas também porque uma

escolha como essa pode reduzir riscos de circulação de conteúdos de exploração sexual de crianças e adolescentes pela dificuldade de aferição de idade precisa em imagens e vídeos[11]. Ao mesmo tempo, quando há previsão do banimento desse tipo de conteúdo, empresas podem elencar exceções a essas políticas de imagens com nudez que façam referência a (i) momentos de parto e pós-parto, (ii) procedimentos cirúrgicos e médicos de confirmação de gênero, e (iii) autoexames de câncer ou publicações sobre prevenção e avaliação de doenças em genitálias e partes íntimas[12]. No entanto, existem casos reincidentes de remoção de conteúdo, especialmente de pessoas trans e não binárias, com publicações sobre procedimentos de afirmação de gênero, com base nas políticas de nudez e de proposta de cunho sexual. Em resposta às remoções, pessoas trans e não binárias engajaram-se no movimento #DeserveToBeHere (em tradução livre, #MerecemosEstarAqui), em que contestavam o banimento de fotos com nudez sobre transição de gênero[13].

- Um terceiro exemplo é dado pelo desafio colocado a partir da diversidade de idiomas a serem analisados na moderação de conteúdo. Recentes relatórios (Duarte; Llansó; Loup, 2017; Nicholas; Bhatia, 2023) do *think tank* Center for Democracy and Technology demonstram como tecnologias de reconhecimento e análise de linguagem (como grandes modelos de linguagem que chamam a atenção a partir de 2022) são desproporcionalmente testados em inglês, o que limita seu uso no processamento de conteúdos – e possivelmente ajuda na moderação.

Riscos como esse aquecem o debate sobre quais mecanismos de mitigação – ou segurança – devem ser cobrados como medidas protetivas razoáveis para os usuários e seus direitos. Vistas sob diversos prismas, tais questões levantam tanto propostas de banimento de determinadas possibilidades (inclusive no âmbito da própria arquitetura dos serviços das plataformas, não só de sua moderação) como visões mais *laissez-faire*, que confiam no aperfeiçoamento contínuo das máquinas.

Três vertentes mais pragmáticas se destacam. Não esgotam as possibilidades e podem se combinar, organizando caminhos não excludentes para enfrentar o problema.

Em primeiro lugar há a defesa por maior transparência na moderação de conteúdo, com ampliação de formas de prestação de contas e da consolidação de direitos de usuários por um "devido processo" na análise de suas atividades. Buscando aumentar o acesso à informação e produzir uma moderação com maior legitimidade e justificação, essa vertente contempla proposições de normas e ter-

11 Em resposta ao Comitê de Supervisão do Facebook, a Meta afirmou, por exemplo, que na elaboração dos princípios gerais de suas políticas sobre nudez considerou: "(1) a natureza privada ou sensível das imagens; (2) se foi dado consentimento para a obtenção e compartilhamento de imagens de nudez; (3) o risco de exploração sexual; e (4) se a divulgação das imagens pode levar a assédio fora da plataforma, particularmente em países onde elas podem ser culturalmente ofensivas". Disponível em: https://oversightboard.com/decision/BUN-IH313ZHJ/.

12 Política de nudez adulta e atividades sexuais da Meta, disponível em: https://transparency.fb.com/en-gb/policies/community-standards/adult-nudity-sexual activity/.

13 Disponível em: https://www.thepinknews.com/2021/04/22/instagram-trans-bodies-censorhip we-deserve-to-be-here/.

mos de uso sobre o que é proibido e sobre como será a análise de eventual violação, escritos em linguagem clara e acessível, traduzidos para os idiomas de todas as regiões onde são navegados e dispostos de maneira que qualquer pessoa possa encontrá-los, por exemplo. Propõe também que existam métodos de prestação de contas nos quais pessoas que acessam as redes sociais possam dar opiniões e feedbacks sobre o exercício da moderação. Contempla também a ideia de que as plataformas possuam instâncias recursais, de maneira a possibilitar que as pessoas possam, por meio de procedimentos claros e formais, conhecer a razão das decisões que influenciam sua expressão nas redes, e contestá-las, caso julguem justo e necessário. Algumas dessas propostas inclusive visam submeter tais recursos à revisão humana.

Na esteira desse último componente, a segunda vertente é sobre a incorporação de etapas de supervisão realizadas por pessoas qualificadas que revisem e direcionem o trabalho das máquinas nas mais variadas etapas. A inserção desse tipo de camada pressupõe que esses processos precisam ser complementados com um elemento humano que os aprimore e agregue nuances, freios e contrapesos às suas análises para abrandar eventuais vieses. Tal revisão adicionaria matizes às decisões maquinadas e somaria especificidades e contextos que podem ser peças-chave para uma gestão saudável do discurso que circula nas plataformas ou para a proteção de populações historicamente marginalizadas.

Uma terceira vertente pragmática é a que contempla propostas para o constante investimento, reavaliação e inovação tecnológica. Com efeito, sistemas automatizados para moderação de conteúdo não só precisam de constante mirada por serem parte do produto (e valor) das plataformas, mas também porque as ameaças não são estáticas. Frente à enorme quantidade de conteúdo, de sua velocidade de circulação e das frequentes mudanças de contextos que envolvem a expressão, é necessário que as plataformas dediquem um olhar cuidadoso e continuado para essa questão. Isso deve envolver, por exemplo, diálogo com especialistas locais, equipes interdisciplinares com pontos focais regionais atentos e especializados, capazes de abordar desafios atrelados a idiomas e conjunturas específicos, como um processo eleitoral, a promulgação de determinada legislação ou o início de um conflito. Neste viés, percebe-se especial atenção à existência de métricas específicas e transparentes de análise, e que todos esses mecanismos combinados sejam capazes de fomentar melhorias nas ferramentas tecnológicas e, por consequência, na gestão da moderação de conteúdo.

Um dos exemplos marcantes nessa última vertente vem da sugestão dada por pesquisadores da OpenAI sobre o uso do seu GPT-4 na moderação de conteúdo (Weng; Goel; Vallone, 2023). Na apresentação dessa alternativa, os seus defensores explicam que conseguiram reduzir significativamente o tempo de "curva de aprendizado" dos sistemas de moderação na implementação de novas regras por conta do uso do modelo, ainda que ele conseguia detectar pontos de aperfeiçoamento nas regras para sua aplicação consistente. Em uma coleta de opiniões de *experts* sobre o experimento, o colunista Casey

Newton (2023) apontou que a promessa é bem vista, mas gera uma série de questionamentos: se, por um lado, recursos humanos podem ser mais bem aplicados em tarefas mais complexas e específicas, por outro, o uso de IA generativa para moderação de conteúdo pode inclusive criar problemas de "explicabilidade" nas decisões por remoção, o que pode gerar problemas legais para a plataforma perante a nova lei europeia, o DSA.

CONSIDERAÇÕES FINAIS

A promissora evolução das tecnologias de IA deve acentuar ainda mais as tendências mapeadas neste modesto artigo em todos os eixos aqui elencados. As possibilidades para produção de conteúdos por ferramentas de IA generativa, por exemplo, já são um importante ponto de atenção para reguladores ao redor do mundo, inclusive na seara eleitoral. Inovação atrás da outra, crescem os riscos de conteúdos "sintéticos" se passarem por "orgânicos" (ou seja, não tratados ou gerados por mecanismos de IA) e engatilharem diferentes crises. Esse tema está no centro de incipientes – porém aceleradas – discussões regulatórias (em especial de autoridades eleitorais, como o Tribunal Superior Eleitoral brasileiro), com alternativas que vão do banimento à rotulagem obrigatória de conteúdos.

No plano da curadoria/arquitetura, tais grandes modelos de linguagem cada vez mais estão embarcados na experiência dos usuários, potencialmente reestruturando sua dieta de informações. Na moderação, como vimos, o uso de ferramentas de ponta promete, mas requer supervisão, transparência e investimento para ser mais do que uma maneira de diminuir custos de uma atividade bastante sensível a direitos humanos.

A incerteza gerada nesse campo não é necessariamente mais impressionante do que aquela produzida no mundo do trabalho, dos negócios ou das mudanças climáticas, mas tem uma característica especial. Considerando a forma como a população mundial hoje conversa, se informa e se polariza, ela se esparrama sobre todas as outras. As capacidades de criarmos ambientes para expressão e troca, para informar e ser informado, é que estão em jogo.

REFERÊNCIAS

BRITO CRUZ, F. (coord.), LANA, A. de P.; JOST, I. *Iguais perante as plataformas? Equidade e transparência na moderação de conteúdo em plataformas digitais*. São Paulo, InternetLab, 2023. Disponível em: https://internetlab.org.br/wp-content/uploads/2023/08/relatorio_internetlab_crosscheck_PORTUGUES_ok2.pdf.

BRITO CRUZ, F. et al. Contribuição do InternetLab ao Expediente T-8.764.298 Acción de Tutela instaurada por Esperanza Gómez Silva contra Facebook Colombia S.A.S, Instagram Colombia y Meta Platforms, Inc., Ministerio de las Tecnologías de la Información y las Comunicaciones (MinTIC) y Superintendencia de Industria y Comercio. São Paulo, InternetLab, 2023.

BRITO CRUZ, F. et al. *Internet e eleições no Brasil: diagnósticos e recomendações*. São Paulo, InternetLab, 2019. Disponível em: http://www.internetlab.org.br/wp-content/uploads/2019/09/policy-infopol-26919_4.pdf.

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

DIAS OLIVA, T.; ANTONIALLI, D. M.; GOMES, A. "Drag queens e inteligência artificial: computadores devem decidir o que é 'tóxico' na internet?". InternetLab, 28/jun./2019. Disponível em: https://internetlab.org.br/pt/noticias/drag-queens-e-inteligencia-artificial-computadores-devem-decidir-o-que-e-toxico-na-internet/

DTSP – Digital Trust & Safety Partnership. *Trust and safety glossary of terms* Jul./2023. Disponível em: https://dtspartnership.org/wp-content/uploads/2023/07/DTSP_Trust-Safety-Glossary_July-2023.pdf.

DUARTE, N.; LLANSÓ, E.; LOUP, A. "Mixed messages? The limits of automated social media content analysis". Center for Democracy & Technology, 2017. Disponível em: https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf.

HARTMANN, I. A.; SILVA, L. A. da. "Inteligência artificial e moderação de conteúdo: o sistema CONTENT ID e a proteção dos direitos autorais na plataforma YouTube". *IUS Gentium*, v. 10, n. 3, 2020, pp. 145-65.

HILL, K. "Como um erro on-line de seu filho pode arruinar a vida digital dos pais". *O Estado de S. Paulo*, 2024. Disponível em: https://www.estadao.com.br/lifestyle/como-um-erro-on-line-de-seu-filho-pode-arruinar-a-vida-digital-dos-pais.

LLANSÓ, E. "No amount of 'AI' in content moderation will solve filtering's prior-restraint problem". *Big Data & Society*, v. 7, n. 1, 2020.

MONTEIRO, A. et al. *Armadilhas e caminhos na regulação da moderação de conteúdo. Diagnósticos & recomendações*. São Paulo, InternetLab, 2021. Disponível em: https://internetlab.org.br/wp-content/uploads/2021/09/internetlab_armadilhas-caminho-moderacao.pdf.

NASCIMENTO, L. et al. "Públicos refratados: a atuação de grupos de extrema direita brasileiros na plataforma Telegram". *Internet & Sociedade*, v. 3, n. 1, 2022.

NICHOLAS, G.; BHATIA, A. "Lost in translation: large language models in non-english content analysis". *Center for Democracy & Technology*, 2023. Disponível em: https://cdt.org/insights/lostin-translation-large-language-models-in-non-english-content-analysis/.

NEWTON, C. "OpenAI wants to moderate your content". *Platformer*, 2023. Disponível em: https://platformer.substack.com/p/openai-wants-to-moderate-your-content.

NOBLE, S. U. *Algorithms of oppression: How search engines reinforce racism*. Nova York, New York University Press, 2018.

SILVA, S. P. da; CESAR, D. J. T. "Inteligência artificial, moderação de conteúdos no YouTube e a proteção de direitos: características, problemas e impactos políticos". *Liinc em Revista*, v. 18, n. 2, nov./2022, pp. 1-21.

UDUPA, S.; MARONIKOLAKIS, A.; WISIOREK, A. "Ethical scaling for content moderation: extreme speech and the (in)significance of artificial intelligence". *Big Data & Society*, v. 10, n. 1, 2023.

WARDLE, C.; DERAKSHAN, H. "Information disorder: toward an interdisciplinary framework for research and policy making". *Council of Europe*, 27/set./2017. Disponível em: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c.

WENG, L.; GOEL, V.; VALLONE, A. "Using GPT-4 for content moderation". *OpenAI Blog*, 2023. Disponível em: https://openai.com/blog/using-gpt-4-for-content-moderation.