e-ISSN nº 2447-4266
Vol. 6, n. 4, Julho-Setembro. 2020

CONTROVÉRSIAS SOBRE DANOS ALGORÍTMICOS: discursos corporativos sobre discriminação codificada

CONTROVERSIES ON ALGORITHMIC HARMS: corporate discourses on coded discrimination
CONTROVERSIAS SOBRE DAÑO ALGORÍTMICO: discursos corporativos sobre discriminación codificada

Sergio Amadeu da Silveira
Doutor e Mestre em Ciência Política pela Universidade de São Paulo (USP). Professor da Universidade Federal do ABC (UFABC). samadeu@gmail.com.

0000-0003-1029-9133

Tarcizio Roberto da Silva
Doutorando em Ciências Humanas e Sociais na Universidade Federal do ABC e Mestre em Comunicação pela Universidade Federal da Bahia (UFBA). eu@tarciziosilva.com.br.

0000-0002-7094-8708

Correspondência: Universidade Federal do ABC (UFABC). Avenida dos Estados, 5001, 09210-580 - Bangú, Santo André, SP – Brasil.

Recebido em: 03.04.2020.
Aceito em: 27.05.2020.
Publicado em: 01.07.2020.

RESUMO:
Impactos discriminatórios e danos de sistemas algorítmicos têm gerado discussões sobre o escopo da responsabilidade de empresas de tecnologia da comunicação e inteligência artificial. O artigo apresenta controvérsias públicas engatilhadas por 8 casos públicos de danos e discriminação algorítmica que geraram respostas públicas de empresas de tecnologia, abordando o esforço realizado pelas empresas de tecnologia em enquadrar o debate sobre responsabilidades no fluxo de planejamento, treinamento e implementação dos sistemas. Em seguida, discute como a opacidade dos sistemas é defendida pelas empresas comerciais que os desenvolvem, alegando prerrogativas como "segredo de negócio" e inescrutabilidade algorítmica.

PALAVRAS-CHAVES: Algoritmos; Auditoria algorítmica; Explicabilidade; Jornalismo de Tecnologia; Plataformas.

Introdução

Sistemas algorítmicos, explicabilidade e responsabilidade

Algoritmos nunca agem isoladamente (SEAVER, 2019; SILVEIRA, 2019). Definidos, em geral, como um conjunto de instruções ou regras para solucionar um problema ou para realizar uma tarefa, precisam estar em contato com uma estrutura de dados para agirem. Algoritmos integram uma rede de actantes (LATOUR, 2005). Suas conexões com dados de entrada, com o feedback, com os efeitos de suas próprias decisões e com os demais componentes dos sistemas que os implementam precisam ser considerados. Assim, utilizamos a expressão sistemas algorítmicos neste texto.

Esses sistemas podem ser confeccionados para seguirem regras de como executar suas ações a partir das informações que recebem. Podem ser criados para aprenderem com os dados que recebem em função dos objetivos prescritos. Também podem ter

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

como finalidade encontrar correlações fortes nos dados que recebem. Enfim, podem criar suas operações com base nos dados e não em regras fixadas.

Os chamados sistemas de aprendizado de máquina utilizam inúmeros modelos computacionais, entre eles, o modelo de redes neurais que têm obtido um grande sucesso em diversas áreas, tais como a robótica, os diagnósticos de medicina, o processamento de voz, a biometria, a mineração de dados, o reconhecimento automático de alvos, entre diversas aplicações. Como outros sistemas que aprendem a partir de dados, as redes neurais artificiais atuam onde a programação baseada em regras não tem obtido uma boa performance. Elas se inspiram no sistema nervoso central e buscam simular a ação dos neurônios.

O sucesso da chamada Inteligência Artificial (IA) que envolve modelos de redes neurais artificiais, aprendizagem profunda, modelos gráficos probabilísticos, entre outros, se deve ao seu alto desempenho (GUNNING; AHA, 2019). Esse desempenho em tratar dados, detectar padrões e realizar previsões tem sido útil ao avanço da competitividade em um cenário dominado pela doutrina econômica neoliberal. Assim, os modelos de negócios baseados em coleta, armazenamento e análise de dados dos consumidores com finalidades preditivas tem incentivado e ampliado a utilização de sistemas algorítmicos de aprendizado profundo.

Frank Pasquale (2015) demonstrou que esse processo informacional se realiza de modo opaco. Como nos alertou no livro *The Black Box Society*, a opacidade dos sistemas algorítmicos é defendida como indispensável para proteger os segredos de negócios, a propriedade intelectual dos códigos e evitar que os usuários possam anular a finalidade dos sistemas. Assim, a transparência é vista como um entrave para as grandes corporações, entretanto, as empresas, os consultores e as plataformas tecnológicas a consideram fundamental para "melhorar a experiência" dos usuários, clientes e consumidores. Desse modo, as pessoas são convencidas de que seus dados pessoais estarão em boas mãos se forem entregues às empresas privadas.

A vida de cada uma e cada um vai sendo convertida em um imenso fluxo de dados, uma vez que os modelos estatísticos de predição e seus algoritmos exigem uma grande quantidade e variedade de dados para extrair padrões e realizar previsões. A lógica da concorrência é um energizador que faz com que o mercado de dados seja um ecossistema em expansão agregando novos dispositivos geradores de dados às suas redes de actantes. Jose Van Dijck alertou para uma dupla alienação que a sociedade vai engendrando nesse processo. Primeiro, a crença que os dados são naturais e expressam

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

e-ISSN nº 2447-4266
Vol. 6, n. 4, Julho-Setembro. 2020

a realidade. Segundo, a confiança de que as plataformas de dados são tais como eles, neutras (VAN DIJCK, 2014).

O fato é que mesmo que existam *frameworks* de IA abertos, a maioria dos sistemas algoritmos de grande relevância pública (GILLESPIE, 2015) são fechados, opacos, sem nenhuma transparência. Basta lembrar do sistema de busca mais utilizado no planeta, o buscador do Google. Trata-se de um sistema algorítmico fechado. O mesmo ocorre com o sistema algorítmico do Facebook e das demais plataformas de intermediação digital.

Há uma relação entre falta de transparência dos sistemas algorítmicos e processos discriminatórios de pessoas e segmentos da população quando submetidos à governança praticada pelos algoritmos. Por isso, há movimentos pela transparência dos códigos e pelo reconhecimento que sistemas algorítmicos possuem viés, definições prévias embutidas em seus modelos (DIAKOPOULOS, 2014). É comum ouvir pessoas que afirmam que os desvios, os vieses não estão nos algoritmos, mas nos bancos de dados, ou melhor, nos dados coletados. Isso não parece consistente com informações encontradas sobre o desenvolvimento de vários sistemas algorítmicos que possuem objetivos de claramente de buscar diferenciações em traços físicos, em determinados comportamentos, em locais de moradia, em escolas cursadas etc.

Em 2019, o Conselho Municipal de San Francisco, na Califórnia, vetou o uso das tecnologias de reconhecimento facial pela polícia e outras agências públicas (FRANCE PRESS, 2019, texto eletrônico). O principal argumento é que os riscos para os direitos e as liberdades civis superam os possíveis benefícios. Além disso, a decisão dos conselheiros apontava que o reconhecimento facial poderia "exacerbar a injustiça racial e ameaçar nossa capacidade de viver sem a contínua vigilância do governo." (FRANCE PRESS, 2019, texto eletrônico). O debate se deu em torno dos riscos dos sistemas algorítmicos perseguir e discriminar minorias e grupos socialmente marginalizados.

A transparência dos sistemas algorítmicos pode não solucionar o problema de explicar como ele alcançou determinados resultados, muitos deles preconceituosos, racistas e discriminatórios. Em alguns modelos de IA, de aprendizado profundo, por exemplo, como as redes neurais artificiais, o modo como atua o algoritmo não permite a explicação dos seus procedimentos, dos seus passos que resultaram em uma dada decisão. São modelos algorítmicos considerados inescrutáveis, insondáveis ou incompreensíveis.

O Departamento de Defesa dos Estados Unidos se deparou com o problema da explicabilidade e da compreensão de como um sistema de aprendizado profundo oferece determinada proposta de ação para utilizar sistemas inteligentes nas ações de defesa

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

nacional. Esse foi a razão principal para a DARPA (Defense Advanced Research Projects Agency) ter criado o programa XAI, Explainable Artificial Intelligence, em português, Inteligência Artificial Explicável. O objetivo do XAI é criar um conjunto de técnicas de aprendizado de máquina que produzam modelos explicáveis, mantendo um alto desempenho de aprendizado, bem como, permita que as pessoas possam entender, confiar e gerenciar efetivamente esses sistemas algorítmicos (GUNNING, 2016).

A questão aqui tem uma grande dimensão sociotécnica ou tecnopolítica uma vez que se reconhece que existem modelos e sistemas algorítmicos que podem encontrar soluções ou propor decisões de grande relevância social sem que os seus gestores ou mesmo desenvolvedores saibam exatamente quais procedimentos ou cálculos foram realizados para tal. Mesmo que sejam soluções comerciais adquiridas por corporações privadas, em geral, os países de democracia liberal costumam possuir leis de defesa do consumidor que exigem explicações e responsabilidades pelas decisões adotadas pelas empresas.

O Regulamento Geral de Proteção de Dados europeu dá o direito a explicação e à revisão humana de decisões automatizadas, principalmente para evitar uma possível alegação empresarial ou governamental de que o seu sistema algorítmico não permite saber os motivos de certas ações. É provável que para se enfrentar o racismo e as discriminações socialmente e democraticamente inaceitáveis seja necessário que os sistemas algorítmicos sejam transparentes, explicáveis e que sejam supervisionados por responsáveis por reconfigurá-los com celeridade. Parece que com o avanço dos sistemas algorítmicos, os riscos de segregação, exclusão, marginalização possam aumentar sob o argumento de uma certa neutralidade e objetividade sistêmica que esconde decisões embutidas nos códigos ou vieses originados em bancos de dados.

O dano algorítmico como controvérsia: auditorias públicas e engajamentos civis

Parte da mobilização civil sobre possíveis danos algorítmicos têm sido realizada no campo público através de expedientes como auditorias públicas e matérias baseadas em jornalismo investigativo ou relatos espontâneos de usuários de sistemas. Como material empírico deste artigo citamos oito notas de casos que contaram com repercussão pública e declaração das organizações envolvidas através de recursos como notas à imprensa ou declarações públicas, listadas na Tabela 01. Antes de nos debruçarmos sobre as reações corporativas na seção a seguir, apresentaremos na presente seção o conceito de auditoria algorítmica e alguns dos casos de repercussão pública analisados mais adiante.

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

e-ISSN nº 2447-4266
Vol. 6, n. 4, Julho-Setembro. 2020

Sandvig e colaboradores (2014) propõem metodologia inspirada nos Estudos de Auditoria para propor um conjunto de cinco abordagens possíveis para a auditoria de sistemas algorítmicos: auditoria de usuário não-invasiva; *sock-puppet audit*; auditoria colaborativa; auditoria de código; e auditoria de raspagem. A *Auditoria de Usuário Não-Invasiva* é, basicamente, a adaptação de métodos clássicos da ciência social como entrevistas em profundidade, *surveys* ou observação não-participante para investigar os modos, dinâmicas e percepções dos usuários quanto aos sistemas estudados. Ao ser uma "seleção não-invasiva de informação sobre interações normais de usuários em uma plataforma"[1] (SANDVIG *et al.*, 2014, p.11), relatos jornalísticos a partir de consulta a usuários se aproximam do modelo. É o caso dos recorrentes problemas dos algoritmos de recomendação do YouTube quando analisados em torno de vídeos relacionados a infância. Reportagens do New York Times[2] e Wired[3] descobriram, respectivamente em 2017 e 2019 (ver Notas 4 e 5), que vídeos perturbadores de animação com violência escatológica simulam conteúdo infantil para serem vistos por crianças, enganando os filtros automáticos da plataforma, e que uma rede de pedófilos usa as recomendações da plataforma para acessar vídeos de crianças semi-nuas dançando.

Bastante similar, uma segunda abordagem pode envolver a construção de sistemas *crowdsourced* ou *colaborativos* para avaliar alguns pontos do sistema através do uso, relato ou codificação distribuída. Tecnicamente e financeira mais complexo, um exemplo é o projeto *FeedVis* desenvolvido por Eslami e colaboradores (2015). Através do desenvolvimento de um aplicativo para Facebook que analisa, com consentimento, os dados da *timeline* dos participantes, as pesquisadoras puderam comparar as percepções daqueles quanto a ingerência algorítmica do Facebook nas interações interpessoais na plataforma. Descobriu-se que os participantes estavam "atribuindo as ações dos algoritmos à intencionalidade de seus próprios amigos e familiares. Os usuários concluíram incorretamente que estavam publicando opiniões impopulares ou sendo ignorados"[4] (ESLAMI *et al.*, 2015, p. 9) o que reforça a tese da influência da plataforma no distanciamento interpessoal.

[1] Tradução livre de: "noninvasive selection of information about users' normal interactions with a platform".
[2] Link: https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html.
[3] Link: https://www.wired.co.uk/article/youtube-pedophile-videos-advertising.
[4] Tradução livre de: "attributing the algorithm's actions to be the intent of their own friends and family. Users incorrectly concluded that they held unpopular views or were being given the cold shoulder".

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

e-ISSN nº 2447-4266
Vol. 6, n. 4, Julho-Setembro. 2020

Uma terceira abordagem proposta é chamada de *Sock-Puppet Audit* (Sandvig *et al.*, 2014) e envolve a simulação de usuários com variáveis controladas pelo desenho da pesquisa ou mesmo sistemas *bots*. Em um dos casos documentados publicamente a partir de denúncias de usuários sobre discriminação racial na plataforma de intermediação de hospedagem Airbnb, o Departamento de Justiça em Emprego e Habitação da California auditou a plataforma através da simulação de contas com as características demográficas variadas[5].

Quanto à análise dos aspectos vistos como estritamente técnicos dos sistemas, a *Auditoria de Raspagem (Scraping Audit)* engloba a coleta de dados nos sistemas, incluindo técnicas de raspagem de dados, acesso através de APIs, captura de tela e afins. Quando tratamos de sistemas focados em comunicação (como plataformas de mídias sociais e buscadores) ou com interfaces de autogestão do usuário (tais como formulários de seleção, ferramentas de escore de crédito e afins) esta abordagem é usada com frequência por permitir avaliar os resultados e requisições oferecidas aos usuários. A tática é coletar e analisar dados da plataforma através de simulações de uso ou interações em escala, de modo distinto aos anteriores por "acessar a plataforma diretamente através de uma API ou podem realizar requisições de busca que seriam improváveis de serem feitas por um usuário (ou ao menos em uma frequência improvável parra um usuário)"[6] (SANDVIG *et al.*, 2014, p. 12). Investigações recentes sobre os modos pelos quais características da interface e algoritmos do YouTube promovem canais extremistas, sobretudo da direita, seguiram por este caminho através da análise de redes de recomendações entre vídeos e canais (RIBEIRO *et al.*, 2019; RIEDER *et al.*, 2018).

A *Auditoria de Código*, através da qual efetivamente os códigos que incorporam cadeias de decisões, escolhas metodológicas, *datasets*, pacotes e módulos de programação costuma ser a mais recomenda. É a mais difícil de ser aplicada por questões institucionais (a maior parte das plataformas possui código fechado por questões comerciais e competitivas) e técnicas (a miríade de tecnologias empregadas ultrapassa em muito a capacidade de pesquisadores isolados). Assim, "mesmo se fornecidos os detalhes específicos de um algoritmo, ao nível normal de complexidade no qual estes sistemas operam um algoritmo não pode ser interpretado apenas pela sua leitura"[7]

[5] Link: https://www.usatoday.com/story/tech/news/2016/06/06/airbnb-openair-diversity-racism-airbnb-connect/85490536/.
[6] Tradução livre de: "accessing the platform directly via an API or they may be making queries that it is unlikely a user would ever make (or at least at a frequency a user is unlikely to ever make)".
[7] Tradução livre de: "even given the specific details of an algorithm, at the normal level of complexity at which these systems operate an algorithm cannot be interpreted just by reading it".

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

(SANDVIG *et al.*, 2014, p. 10), mas um profundo conhecimento dos processos técnicos envolvidos em um determinado sistema permitem pesquisadoras e pesquisadores atacarem as raízes de problemas através da mesma lógica computacional, mas com sensibilidade aos danos algorítmicos possíveis de acordo com as variáveis demográficas e diversidade de usos.

Entre os casos que mesclaram técnicas de auditoria de raspagem e auditoria de código com investigação de maior impacto, a série de estudos em torno do projeto *Gender Shades da Algorithmic Justice League* merece menção especial. As pesquisadoras analisaram a precisão de recursos de identificação de características de gênero e idade em reconhecimento facial em três das principais tecnologias no mercado, das empresas IBM, Microsoft e Face++. Foi descoberta uma desigualdade interseccional: os sistemas erram mais com pessoas negras e mais em mulheres, resultando em taxas de erros enormes em fotos mulheres negras, algo com impacto abrangendo de aplicativos de mídias sociais a vigilância policial. Além da identificação das raízes dos problemas - sobretudo uso acrítico de dados de treinamento enviesados -, as pesquisadoras identificam que a "análise interseccional de erro fenotípico e demográfico pode nos ajudar a informar métodos para melhorar composição de bases de dados, seleção de recursos e desenhos de redes neurais"[8] (BUOLAMWINI; GEBRU, 2018, p.12). Para além do mérito científico do texto acadêmico, essencial para o projeto foi a publicação de um site interativo[9] e apresentações públicas dos dados, gerando cobertura midiática e interesse público nas descobertas.

Tornando a questão de um "problema de fato" em um "problema de interesse" (LATOUR, 2004), o impacto público das descobertas forçou as empresas envolvidas a se pronunciarem publicamente através de notas públicas e compromisso de melhorias dos sistemas. Em trabalho subsequente (RAJI; BUOLAMWINI, 2019), as pesquisadoras do projeto *Gender Shades* revisaram as taxas de erros dos sistemas analisados, identificando melhorias efetivas, e compararam com mais dois fornecedores, Amazon e Kairos. Na trajetória do projeto descrito pelas autoras, depois da fase de identificação dos problemas, é ofertada às empresas envolvidas o conhecimento prévio sobre o estudo e um período para reação, antes da exposição pública dos resultados. Após este período, os resultados são divulgados em conferências científicas, imprensa e, no caso do Gender Shades, um website interativo - que posteriormente incluiu as respostas corporativas. Ao

[8] Tradução livre de: "intersectional phenotypic and demographic error analysis can help inform methods to improve dataset composition, feature selection, and neural network architectures".
[9] http://gendershades.org/.

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

voltar aos dados e identificar a diminuição nas taxas de erros, as autoras propõem o conceito de "auditoria pública acionável" como "um mecanismo para incentivar corporações lidarem com o viés algorítmico presente em tecnologias datacêntricas"[10] (RAJI; BUOLAMWINI, 2019, p.1).

Fatores como explicabilidade e responsabilização dos danos algorítmicos são ainda controversos e dependentes de consideráveis redes de fluxos de poder regulatório e legislativo, então propostas como a de Pasquale ao tratar de intermediários especializados e obrigatórios como uma possibilidade de agir em prol da governança algorítmica através de regulamentação prévia sobre os sistemas com a força de uma "quarta lei da robótica" (PASQUALE, 2017) parece algo ainda distante. Dados do NeurIPS, o maior evento de inteligência artificial e redes neurais do mundo, mostram que o número de *papers* com propostas de novos modelos supera em 10 vezes o número de *papers* analisando modelos existentes, configurando um *gap* de conhecimento sobre os sistemas algorítmicos (EPSTEIN *et al.*, 2018).

Relatos ou auditorias públicas mais ou menos sistemáticas sobre danos algorítmicos poderiam forçar corporações a se pronunciar sobre suas responsabilidades, através do poder de pressão pública e da imprensa. Bucher alerta que o caráter performativo dos algoritmos assim como os discursos sobre eles construídos nas interfaces entre seus usos, opiniões públicas, imprensa e engajamentos civis sobre estes constroem um "imaginário algorítmico", que seriam os "modos de pensar sobre o que algoritmos são, o que deveriam ser, como eles funciona e o quê esta imaginação torna possível"[11] (Bucher, 2016a, p. 39-40).

O espaço de comunicação pública e jornalística para além das trocas de *experts*, portanto, é uma fértil fonte de investigação sobre as estratégias discursivas de busca por enquadramento corporativo em casos de publicização de dados algorítmicos. Na seção a seguir percorreremos casos documentados de danos algorítmicos nos quais as corporações envolvidas reagiram publicamente aos erros encontrados.

Reações corporativas: evasão de responsabilidade

No panorama de relações decorrentes da disseminação de sistemas algorítmicos nas esferas sociais, o paradigma da invisibilidade do funcionamento dos sistemas é

[10] Tradução livre de: "one mechanism to incentivize corporations to address the algorithmic bias present in data-centric technologies".
[11] Tradução livre de: "ways of thinking about what algorithms are, what they should be, how they function and what these imaginations in turn make possible".

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

decorrente de sua integração ao cotidiano. Em momentos de publicização dos danos algorítmicos, os sistemas se tornam "assunto de interesse" (LATOUR, 2004) abrindo a controvérsia sobre a neutralidade ou objetividade da tecnologia já incorporada no cotidiano. As notas públicas e de imprensa que analisaremos a seguir fazem parte do esforço organizado, através de técnicas de relações públicas e gestão da comunicação organizacional, em enquadrar o caso de modo a minimizar danos à percepção dos valores positivos da organização ou de suas tecnologias.

Bucher critica o conceito de "caixa preta" algorítmica quando é enunciado apenas como uma questão de investigação sobre as entradas (*inputs*) e saídas (*outputs*) de um sistema, uma vez que os sistemas crescentemente construídos para adaptar cálculos e procedimentos através de aprendizado de máquina reconfigurariam os status de entrada e saída (2016b). É possível expandir o escopo de observação sobre os algoritmos para além de suas imediações técnicas evidentes, em busca das redes sociais e relações de poder materializadas e/ou intermediadas nos fluxos de performatividade (INTRONA, 2015).

Podemos aproximar o conceito de "tecnografia" de Bucher, elaborado para abordar os modos pelos quais software se intersecta com socialidades, "as normas e valores que foram delegados e materializados em tecnologia"[12] (2016b, p.86) com a abordagem proposta por Brock de *Análise Tecnocultural Crítica do Discurso*. Para Brock, parte dos princípios de análise do discurso: que este exibe padrões recorrentes; que envolvem escolhas do emissor; e que o discurso mediado por computadores, assim como em outros meios e formatos, pode ser moldado e adaptado a características do ambiente (BROCK, 2016).

Para compreender os esforços corporativos no enquadramento de casos de danos algorítmicos, selecionamos 8 notas e declarações de corporações e seus representantes, como pode ser visto na Tabela 1.

Tabela 1 Notas de Imprensa / Notas Públicas Analisadas

Número	Nota Pública / Reportagem	Empresa Envolvida	Ano
1	MICROSOFT JANUARY 2018 STATEMENT to lead author of "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification"[13]	Microsoft	2018

[12] Tradução livre de: "the norms and values that have been delegated to and materialized in technology".
[13] http://gendershades.org/docs/ibm.pdf.

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

2	IBM Response to "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification"[14]	IBM	2018
3	FaceApp apologises for 'racist' filter that lightens users' skintone[15]	Faceapp	2017
4	On YouTube, a network of paedophiles is hiding in plain sight[16]	Youtube	2019
5	On YouTube Kids, Startling Videos Slip Past Filter[17]	Youtube	2017
6	YouTube won't stop recommending videos with children, despite pedophilia problem[18]	Youtube	2019
7	Google tweaks algorithm to show less porn when searching for 'lesbian' content[19]	Google	2019
8	Pesquise 'tranças bonitas' e 'tranças feias' no Google: um caso de racismo algorítmico[20]	Google	2019

Os textos mencionados na Tabela 1 foram analisados abaixo em seu contexto sociotécnico "como um processo comunicativo, ao desvelar sobre o que um artefato específico de TIC é baseado e para o quê é desenhado a "fazer" e, criticamente, como os usuários se articulam em relação ao[21] (BROCK, 2016, p. 15). Sob a luz dos conceitos de explicabilidade e inescrutabilidade, exploramos esse grupo de declarações para propor as categorias abaixo.

Processo Contínuo de Otimização

A ideia de *beta perpétuo* ganhou corpo nas tecnologias da comunicação como plataformas e softwares na medida em que os dispositivos de acesso e largura de banda avançaram em qualidade e eficiência técnicas, permitindo que o paradigma do "software como serviço" ("SaaS") dominasse a oferta de produtos nos últimos anos (ROMANI & KUKLINSKI, 2007), culminando na emergência dos aplicativos mobile e plataformização (SRNICEK, 2017). Além de ser um modo de abordagem de desenvolvimento de software, também se tornou em tática comercial: ao enquadrar produtos informacionais como *beta*, as corporações ao mesmo tempo buscavam uma aura de inovação criativa e também gerenciavam expectativas quanto a *bugs* ou falhas nos sistemas.

[14]http://gendershades.org/docs/ibm.pdf.
[15]https://www.mirror.co.uk/tech/faceapp-apologises-hot-selfie-filter-10293590.
[16]https://www.wired.co.uk/article/youtube-pedophile-videos-advertising.
[17]https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html.
[18]https://www.theverge.com/2019/6/3/18650318/youtube-child-predator-pedophilia-family-vlogging-comments-recommendation-algorithm.
[19]https://thenextweb.com/tech/2019/08/07/google-tweaks-algorithm-to-show-less-porn-when-searching-for-lesbian-content/.
[20]https://blogs.oglobo.globo.com/ancelmo/post/pesquise-trancas-bonitas-e-trancas-feias-no-google-um-caso-de-racismo-algoritmico.html.
[21] Tradução livre de: "as a communicative process, by unpacking what a specific ICT artifact is based upon what it is designed to "do" and critically, how users articulate themselves in and about the artifact".

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

e-ISSN nº 2447-4266
Vol. 6, n. 4, Julho-Setembro. 2020

Atualmente os principais fornecedores de tecnologia abandonaram o termo *beta* como qualificador de seus produtos monetizáveis, mas resgatam o princípio de acordo com a necessidade. Dois casos relacionados no YouTube relacionados a recomendação de conteúdo problemático foram abordados pela empresa através da tática de evocação da melhoria contínua ou beta perpétuo. Como resposta ao caso de recomendação de vídeos de crianças a pedófilos, citado na seção anterior, YouTube declarou em nota oficial em seu blog que "ao longo dos últimos 2 anos ou mais, nós temos implementado melhorias regulares ao classificador de aprendizado de máquina que nos ajuda a proteger menores e famílias. Lançamos a melhoria mais recente no início deste mês"[22] (Nota 6). A otimização contínua como um jogo de gato e rato resultante da excessiva complexidade da produção de conteúdo web também foi mencionada por Malik Ducard, responsável por supervisionar conteúdo familiar e educacional na plataforma alegou. Em caso de 2017, Ducard direciona o aprendizado de máquina como solução, uma vez que o processo de monitoramento contínuo seria "multidimensional e dependente de muito aprendizado de máquina"[23] (Nota 5), reforçando a tecnicidade como foco para as soluções.

Reprodução da Sociedade

Como corporação especialmente posicionada estrategicamente em torno do acesso e organização da informação, o Google possui interesse de manter a percepção de neutralidade. Como a missão declarada de "organizar as informações do mundo para que sejam universalmente acessíveis e úteis para todos"[24], a corporação engloba diversos produtos e serviços informacionais através do guarda-chuva da Alphabet, mas possui o buscador como um dos seus principais feitos indispensáveis como base para outros serviços, como fornecimento de computação em nuvem.

Dois casos recentes de cobertura da imprensa no Brasil sobre vieses ofensivos na apresentação de resultados receberam comentários da corporação. Em julho de 2019 viralizou nas mídias sociais a comparação dos resultados de imagens resultantes para as buscas "tranças feias" e "tranças bonitas", fato coberto por veículos como *O Globo*, através da coluna *O Blog do Ancelmo*. Neste espaço, a Google se pronunciou alegando que os resultados apenas reproduzem os "estereótipos" existentes: "Como nossos sistemas

[22] Tradução livre de: "over the last 2+ years, we've been making regular improvements to the machine learning classifier that helps us protect minors and families. We rolled out our most recent improvement earlier this month".
[23] Tradução livre de: ""multilayered and uses a lot of machine learning".
[24] https://about.google/.

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

encontram e organizam informações disponíveis na web, eventualmente, a busca pode espelhar estereótipos existentes na internet e no mundo real em função da maneira como alguns autores criam e rotulam seu conteúdo" (Nota 8).

Em novembro de 2019 outro caso repercutiu: usuárias do buscador identificaram que consultas a termos como "mulher negra dando aula" traziam basicamente resultados pornográficos. Consultada pela reportagem publicada no site *Universa da UOL*, a empresa reconhece que "o conjunto de resultados para o termo mencionado não está à altura desse princípio" (Nota 9) e alegam que irão "buscar uma solução para aprimorar os resultados não somente para este termo, como também para outras pesquisas que possam apresentar desafios semelhantes" (Nota 9) propõe também aos usuários que realizem uma ação extra: adicionar o recurso de SafeSearch, originalmente criado para esconder conteúdo pornográfico a menores.

Assim como nos casos anteriores, o enquadramento da questão como surpresa pelas corporações vai de encontro à bibliografia acadêmica e especializada que tem abordado nos últimos 10 anos (NOBLE, 2011; EDELMAN, 2011; SWEENEY, 2013) os potenciais danos algorítmicos de buscadores. Em *Algorithms of Oppression*, Safiya Noble desvela os modos pelos quais os buscadores performam representações "descontextualizadas em um tipo específico de processo de resgate de informações, particularmente para grupos sobre os quais imagens, identidade e histórias sociais são enquadradas através de formas de dominação sistêmica"[25] (NOBLE, 2018, pos.2467).

Reações Diferenciais

A postura da Google nos dois casos citados acima contrasta diretamente com o caso da mobilização da organização francesa SEOLesbienne. O projeto, ligado à organização de combate à violência sexual e de gênero Nous Totes, pressionou o buscador a mudar seu algoritmo contra a hiper-sexualização dos resultados a buscas como "lésbica" e "lesbianidade" na plataforma, para priorizar resultados informativos, noticiosos e culturais sobre as identidades lésbicas, em contraposição ao conteúdo misógino frequente em sites pornográficos.

Em reportagem da The Next Web a partir de dados do portal francês Numerama, o Vice-Presidente de Qualidade do buscador, Pandu Nayak, reconhece que há problemas como este em várias línguas e buscas e explica a decisão de tomar medidas "em casos

[25] Tradução livre de: "decontextualized in one specific type of information-retrieval process, particularly for groups whose images, identities, and social histories are framed through forms of systemic domination".

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

e-ISSN nº 2447-4266
Vol. 6, n. 4, Julho-Setembro. 2020

onde e quando há uma razão para a palavra ser interpretada de um modo não-pornográfico, que esta interpretação seja priorizada"[26] (Nota 7).

O contraste pode ser observado na atribuição de julgamento aos resultados. Enquanto Nayak expressa claramente que "Eu acredito que estes resultados [de busca] são horríveis, não há dúvida sobre isto"[27] (Nota 7), as declarações brasileiras pedem desculpas "àqueles que se sentiram impactados ou ofendidos" (Nota 8), deslocando a percepção da ofensa ao público afetado ao mesmo tempo que minimiza as relações poder sobre grupos minorizados ao dizer que "pessoas de todos as raças, gêneros e grupos podem ser afetadas" (Nota 8) e resgata a ideologia de cegueira racial ao enfatizar que supostamente fará o mesmo "também para outras pesquisas que possam apresentar desafios semelhantes" (Nota 9).

Negação de Escopo da Responsabilidade

A negação de responsabilidade para além do objetivo explícito dos aplicativos e sistemas algorítmicos é realizada através do apelo à complexidade dos produtos informacionais em questão. Um dos casos mais emblemáticos deste tipo de estratégia são as repetidas controvérsias do aplicativo de registro e manipulação de selfies *FaceApp*. Em abril de 2017, durante uma das primeiras rodadas de popularização do aplicativo, usuários perceberam que um dos filtros "embelezadores" embranquecia consistentemente usuários de grupos com pele escura, como afro-americanos e indianos. Questionado pela *Techcrunch*, o CEO do aplicativo Yaroslav Goncharov alegou que o problema era "um infeliz efeito colateral da rede neural subjacente causado pelo viés da base de dados para treinamento, não comportamento intencional"[28] (Nota 3) e que iriam trabalhar em um "ajuste completo" para breve. Entretanto, em agosto do mesmo ano o aplicativo lançou um recurso problemático de simulação racial e em 2019 percebeu-se que o novo filtro de simulação de envelhecimento também embranquecia os usuários em termos de cor e traços faciais.

Apesar de tentar evadir-se da responsabilidade alegando o viés na base de dados de treinamento, Goncharov admitiu que a empresa usou uma base própria de dados de criação própria. A contradição é apontada na reportagem do portal de tecnologia (Nota

[26] Tradução livre de: "in cases where, when there is a reason for the word to be interpreted in a non-pornographic way, that interpretation is put forward".
[27] Tradução livre de: "I find that these [search] results are terrible, there is no doubt about it".
[28] Tradução livre de: "an unfortunate side-effect of the underlying neural network caused by the training set bias, not intended behaviour".

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

3) e trata-se de paralelo relevante com o caso das notas (Notas 1 e 2) em reação ao projeto *Gender Shades* citados anteriormente.

Através de extensa resposta, a IBM aproveitou o caso de auditoria algorítmica para realizar experimento replicando parte da metodologia do trabalho inicial, alegando taxas de erros menores que a dos concorrentes (Nota 2), dados contraditados pelo estudo posterior das pesquisadoras (RAJI & BUOLAMWINI, 2019). A corporação alegou sem mais detalhes que "agora usa dados de treinamento e recursos de reconhecimento diferentes do momento avaliado por este estudo"[29] (Nota 2) e que busca apoiar "projetos contínuos para abordar vieses nos dados de treinamento"[30] (Nota 2). Entretanto, vale notar a ausência de menção a uma das principais descobertas do trabalho original, o fato de que as duas bases abertas de dados visuais para treinamento mais usadas pelo segmento são extremamente enviesadas.

Considerações finais

Como vimos acima, podemos identificar que as corporações buscam simplificar o debate sobre danos algorítmicos no discurso público através de variados expedientes. O combate aos danos algorítmicos nas sociedades contemporâneas caracterizadas por oligopólios das plataformas digitais passa necessariamente por problematizar a noção que algoritmos são caixas pretas inescrutáveis, pois isto os garantiria "um lugar especial no mundo das coisas desconhecidas que talvez não seja totalmente merecido"[31] (BUCHER, 2016b, p. 85-86).

Pelo contrário, o escopo de responsabilidade da implementação de sistemas algorítmicos em sistemas comerciais ou públicos envolve abordar as controvérsias sobre os seus limites e os modos pelos quais a evasão de responsabilidade e agência (RUBEL *et al.*, 2019) é posta em prática através de declarações públicas na imprensa ou canais corporativos de comunicação.

Identificamos em casos de danos algorítmicos com grande repercussão três expedientes pelos quais as empresas ou corporações envolvidas reagem às críticas e auditorias algorítmica: evocação de um processo contínuo de otimização como característico das tecnologias digitais contemporâneas; alegação que os sistemas apenas reproduzem as desigualdades e problemáticas já presentes na sociedade, portanto ações

[29] Tradução livre de: "now uses different training data and different recognition capabilities than the service evaluated in this study".
[30] Tradução livre de: "projects to address dataset bias".
[31] Tradução livre de: "a special place in the world of unknowns that perhaps is not fully deserved".

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

restaurativas seriam opcionais ou até mesmo injustas; e enquadramento do escopo de responsabilidade nas minúcias supostamente estritamente técnicas, deixando os procedimentos prévios de treinamento dos sistemas e impactos na sociedade em externalidades. Entretanto, a comparação entre declarações sobre casos nos centros de poder em relação ao Sul Global demonstram reações diferenciais a depender de nacionalidade, raça e classe, minando os argumentos comuns de neutralidade da tecnologia. No atual cenário de confusão midiática e crise de autoridade dos meios jornalísticos tradicionais, o engajamento crítico do público sobre controvérsias algorítmicas que modulam o acesso ou restrição ao uso igualitário e seguro dos meios de comunicação mostra-se essencial.

Referências

BROCK, Andre. Análise Crítica Tecnocultural do Discurso. In: SILVA, T. Comunidades, Algoritmos e Ativismos Digitais: olhares afrodiaspóricos. São Paulo, LiteraRUA, 2020.

BUCHER, Taina. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. Information, Communication & Society, v. 20, n. 1, p. 30-44, 2016b.

BUCHER, Taina. Neither black nor box: ways of knowing algorithms. In: KUBITSCHKO, S. & KAUN, A. (orgs.) Innovative methods in media and communication research. Palgrave Macmillan, Cham, 2016b. p. 81-98.

BUOLAMWINI, Joy; GEBRU, Timnit. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceeedings of Conference on fairness, accountability and transparency, 2018. pp. 77-91.

DIAKOPOULOS, Nicholas. Accountability in algorithmic decision making. Communications of the ACM, v. 59, n. 2, p. 56-62, 2016.

EPSTEIN, Ziv *et al*. Closing the AI Knowledge Gap. arXiv preprint arXiv:1803.07233, 2018.

ESLAMI, Motahhare *et al*. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems. ACM, 2015. p. 153-162

FRANCE PRESS. (2019). San Francisco proíbe a polícia de usar reconhecimento facial Oito dos nove conselheiros municipais são contrários à tecnologia. G1, 16/05/2019, online. Disponível em: https://g1.globo.com/pop-arte/noticia/2019/05/16/san-francisco-proibe-a-policia-de-usar-reconhecimento-facial.ghtml Acesso em 22/04/2020.

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

GILLESPIE, Tarleton. A relevância dos algoritmos. Parágrafo, 6(1), 2018, pp. 95-121.

GUNNING, D. Broad Agency Announcement Explainable Artificial Intelligence (XAI). Technical report, 2016.

GUNNING, David. Explainable artificial intelligence (xai) Program. AI Magazine, v. 40, n. 2, 2019. pp.44-58.

LATOUR, Bruno. Why has critique run out of steam? From matters of fact to matters of concern. Critical inquiry, v. 30, n. 2, p. 225-248, 2004.

LATOUR, Bruno. Reassembling the Social: An Introduction to Actor-Network-Theory. New York: Oxford University Press, 2005..

NOBLE, Safiya Umoja. Searching for Black Girls: Ranking Race and Gender in Commercial Search Engines. Tese de Doutorado defenfida na Urbana-Champaign: University of Illinois at Urbana-Champaign, 2011.

NOBLE, Safiya Umoja. Algorithms of oppression: How search engines reinforce racism. New York: NYU Press, 2018.

PASQUALE, Frank. The black box society. Harvard University Press, 2015.

PASQUALE, Frank. Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society. Ohio St. LJ, v. 78, p. 1243, 2017.

RAJI, Inioluwa Deborah; BUOLAMWINI, Joy. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In: AAAI/ACM Conf. on AI Ethics and Society, 2019.

RIBEIRO, Manoel Horta *et al.* Auditing radicalization pathways on youtube. arXiv preprint arXiv:1908.08313, 2019.

RIEDER, Bernhard; MATAMOROS-FERNÁNDEZ, Ariadna; COROMINA, Òscar. From ranking algorithms to 'ranking cultures' Investigating the modulation of visibility in YouTube search results. Convergence, v. 24, n. 1, p. 50-68, 2018.

ROMANI, Cristóbal C.; KUKLINSKI, Hugo P. Planeta Web 2.0: Inteligencia colectiva o medios fast food. Barcelona: Grup de Recerca d'Interaccions Digitals, Universitat de Vic. Flacso, 2007.

RUBEL, Alan; PHAM, Adam; CASTRO, Clinton. Agency Laundering and Algorithmic Decision Systems. In: International Conference on Information. Springer, Cham, 2019. p. 590-598.

SANDVIG, Christian *et al.* Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry, v. 22, 2014.

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt

SEAVER, N. Knowing Algorithms. In: VERTESI, J.; RIBES, D. (orgs.) digitalSTS: A Field Guide for Science & Technology Studies. Princeton Univrsity Press, 2019. pp.412-422.

SILVEIRA, S. A. Democracia e os códigos invisíveis: como os algoritmos estão modulando comportamentos e escolhas políticas. São Paulo: Edições SESC-SP, 2019.

SRNICEK, Nick. Platform capitalism. John Wiley & Sons, 2017.

SWEENEY, Latanya. Discrimination in online ad delivery. arXiv preprint arXiv:1301.6822, 2013.

VAN DIJCK, José. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. Surveillance & Society, 12(2), 2014. pp. 197-208.

ABSTRACT:
Discriminatory impacts and the damages due to algorithmic systems have opened discussions regarding the scope of responsibility of communication technology and artificial intelligence companies. The article presents public controversies triggered by eight public cases of harm and algorithmic discrimination that generated public responses from technology companies, addressing the efforts made by them in framing the debate about responsibility in the course of planning, training and implementation of systems. Following that, it discusses how the opacity of systems is defended by the commercial companies that develop them, alleging prerogatives such as "industry secrets" and algorithmic inscrutability.

KEYWORDS: Algorithms; Algorithmic Auditing; Explainability; Technology Journalism; Platforms.

RESUMEN:
Los impactos y daños discriminatorios por sistemas algorítmicos han abierto discusiones sobre el alcance de responsabilidad de las empresas de tecnología de la comunicación e inteligencia artificial. El artículo presenta controversias públicas desencadenadas por ocho casos públicos de daño y discriminación algorítmica que generaron respuestas públicas por parte de las empresas, abordando los esfuerzos realizados por ellas en enmarcar el debate sobre la responsabilidad en el transcurso de la planeamento, alimentación con dadtos e implementación de sistemas. A continuación, se analiza cómo la opacidad de los sistemas es defendida por las empresas comerciales que los desarrollan, alegando prerrogativas como los "secretos de la industria" y la inescrutabilidad algorítmica.

PALABRAS-CLAVES: Algoritmos; Auditoría algorítmica; Explicabilidad; Periodismo Tecnológico; Plataformas.

Revista Observatório, Palmas, v. 6, n. 4, p. 1-17, jul-set. 2020 DOI: http://dx.doi.org/10.20873/uft.2447-4266.2020v6n4a1pt