INTELIGÊNCIA ARTIFICIAL E ALGORITMOS

ADRIANA GONÇALVES
LUÍSA TORRE
PAULO VICTOR MELO
[EDS.]

LABCOM
COMUNICAÇÃO
& ARTES

INTELIGÊNCIA ARTIFICIAL E ALGORITMOS

DESAFIOS E OPORTUNIDADES PARA OS MEDIA

ADRIANA GONÇALVES
LUÍSA TORRE
PAULO VICTOR MELO
[EDS]

LABCOM
COMUNICAÇÃO
& ARTES

Ficha Técnica

Título
Inteligência Artificial e Algoritmos:
Desafios e oportunidades para os media

Editores
Adriana Gonçalves, Luísa Torre & Paulo Victor Melo

Editora LabCom
www.labcom.ubi.pt

Coleção
Livros de Comunicação

Direção
Gisela Gonçalves

Design Gráfico
Cristina Lopes

[Imagens de capa e separadores criadas através
da ferramenta Midjourney AI]

ISBN
978-989-654-973-2 (papel)
978-989-654-975-6 (pdf)
978-989-654-974-9 (ePub)

Depósito Legal
527476/24

Tiragem
Print-on-demand

Universidade da Beira Interior
Rua Marquês D'Ávila e Bolama
6201-001 Covilhã
Portugal
www.ubi.pt

Covilhã, 2024

Índice

Introdução

Quais os papéis desempenhados pelas tecnologias de Inteligência Artificial (IA) e pelos algoritmos nas relações sociais, na arena política, nos aspetos económicos e, sobretudo, na conformação de sociedades democráticas? Que questões éticas devem ser consideradas na adoção dessas tecnologias por instituições públicas e privadas? Como garantir a aplicação de tecnologias digitais sem que discriminações e vieses preconceituosos sejam reforçados? De que modo o jornalismo é impactado pela ampliação de conteúdos produzidos de forma automatizada? Quais as implicações do uso de algoritmos nas rotinas produtivas do jornalismo e nos profissionais? Como a IA contribui para a disseminação e, por outro lado, pode ajudar a fazer frente à desinformação?

Estas foram algumas das inquietações partilhadas entre investigadores que desenvolviam pesquisas de doutoramento e pós-doutoramento no LabCom - Comunicação e Artes, que motivaram o debate. Assim surgiu o **Ciclo de Conversas "Inteligência Artificial, Algoritmos e Media"**, realizado ao longo do mês de janeiro de 2023, evento que reuniu investigadores/as, professores/as e especialistas oriundos de três países: Portugal, Brasil e Espanha. Em termos de participação, o Ciclo de Conversas juntou 217 inscritos, de universidades e institutos de pesquisa de oito países (Portugal, Brasil, Espanha, Cabo Verde, Reino Unido, Itália, Chile, Holanda e Alemanha), das áreas de Ciências da Comunicação, Direito, Sociologia, Ciências da Computação, entre outras.

Num contexto de cada vez maior relevância dos algoritmos e da IA nas nossas vidas individuais e coletivas, estes debates estimularam a reflexão sobre o impacto das tecnologias emergentes na vida contemporânea.

E, visando contribuir com esta necessidade democrática, convidámos os/as oradores/as a partilharem as suas ideias neste livro.

Inteligência Artificial e Algoritmos: Desafios e Oportunidades para os Media assume o papel de continuidade da discussão em torno da IA, dos algoritmos e das transformações incitadas por estas tecnologias na sociedade e nos media. Como estratégia de preservar a diversidade dos textos, mantivemos cada capítulo no idioma em que cada autor/a escreveu – português, espanhol e inglês. A partir dos contributos dos/as autores/as, o livro estrutura-se em sete capítulos.

O primeiro é da autoria de Joaquim Paulo Serra, da Universidade da Beira Interior. **Tecnologia, sociedade e democracia. A questão da regulação** recupera o conceito de técnica para delinear o contexto atual de procura pela automação para substituir algumas das funções anteriormente desempenhadas pelo homem. Partindo dessa contextualização, o autor aprofunda o debate acerca das tentativas de regulação dos usos da Internet, deixando em aberto inúmeros desafios que se colocam na investigação, na sociedade, na educação, na economia ou na política.

O segundo capítulo – **Inteligência artificial, algoritmos e media: diálogos de pesquisa** – é uma contribuição dos editores do livro, que propõe um olhar interdisciplinar para as problemáticas que envolvem o uso de IA e algoritmos nos media e na sociedade. Desta forma, assume-se que as tecnologias emergentes influenciam as relações sociais, os media e o jornalismo, ameaçando as democracias. Por isso, é fundamental discutir possíveis saídas para os desafios que são colocados hoje, como a desinformação e os vieses discriminatórios.

Intitulado **Como (e por que) os jornalistas devem investigar algoritmos**, o terceiro capítulo é da autora Krishma Carreira, da Universidade Metodista de São Paulo, e fala sobre os Sistemas de Decisões Automatizadas que são hoje utilizados em diversos setores. Neste texto, a autora discute uma estratégia de combater os erros e vieses discriminatórios presentes nestes sistemas, a partir de um olhar jornalístico. Com esse objetivo, a autora de-

senvolve um método denominado Reportagem Investigativa sobre Sistemas de Decisões Automatizadas (RISDA), a partir do qual lança as suas reflexões sobre as problemáticas que envolvem os sistemas de IA.

André Lemos, da Universidade Federal da Bahia, explora em **Fake News as Digital Disruption: Unravelling Algorithmic Logic in the Spread of Disinformation** a relação entre fake news e algoritmos, examinando os mecanismos através dos quais os algoritmos facilitam a propagação da desinformação, e introduzindo a ideia de que as fake news e a desinformação não são erros na cultura digital, e sim, uma disrupção. Na visão do autor, compreender estas dinâmicas será essencial para desenvolver estratégias de combate aos efeitos das fake news e preservar a integridade dos espaços digitais.

No quinto capítulo, **Inteligencia Artificial y Deepfakes: sesgos de género y agresión contra las mujeres,** Rosa Franquet**,** da Universitat Autònoma de Barcelona, discute os riscos para as mulheres de uma interseção entre Inteligência Artificial, desinformação e media, lançando luz sobre a disseminação de pornografia *deepfake* online e sobre o tratamento dispensado pelos meios de comunicação social ao tema, muitas vezes de forma descontextualizada, sem abordar causas profundas. Como resultado, este tipo de violência contra a mulher não encontra o tratamento adequado na esfera pública, impossibilitando uma ação preventiva destes abusos.

Em **Por que falar de raça quando falamos de dados pessoais, inteligência artificial e algoritmos?**, Johanna Monagreda, da Universidade Federal de Minas Gerais, discute sete riscos de ampliação de vulnerabilidades da população negra quando o assunto é o uso de dados pessoais: perda do direito à privacidade; dataficação da pobreza; reprodução e automatização do racismo; perfilamento racial e discriminação; hipervigilância e criminalização; impacto nos processos de subjetivação; e apagamento do caráter político das problemáticas sociais.

Por fim, no último capítulo, **Fighting Algorithmic Racism: reactions, remediations and re-appropriations**, Tarcízio Silva, da Universidade Federal

do ABC, demonstra que existe um conjunto de iniciativas a ser construídas por ativistas, desenvolvedores/as, cientistas e tecnólogos/as de diferentes áreas para o enfrentamento ao racismo algorítmico, como resposta aos riscos apontados no texto anterior. A partir de reflexões de intelectuais negros e negras, e de entrevistas com ativistas que discutem tecnologia e raça, o autor enfatiza que as modalidades de resistência, reações e remediação contra a transformação algorítmica do racismo estrutural envolvem a lembrança das diversas frentes dos movimentos negros nas batalhas sociais e na solidariedade diaspórica.

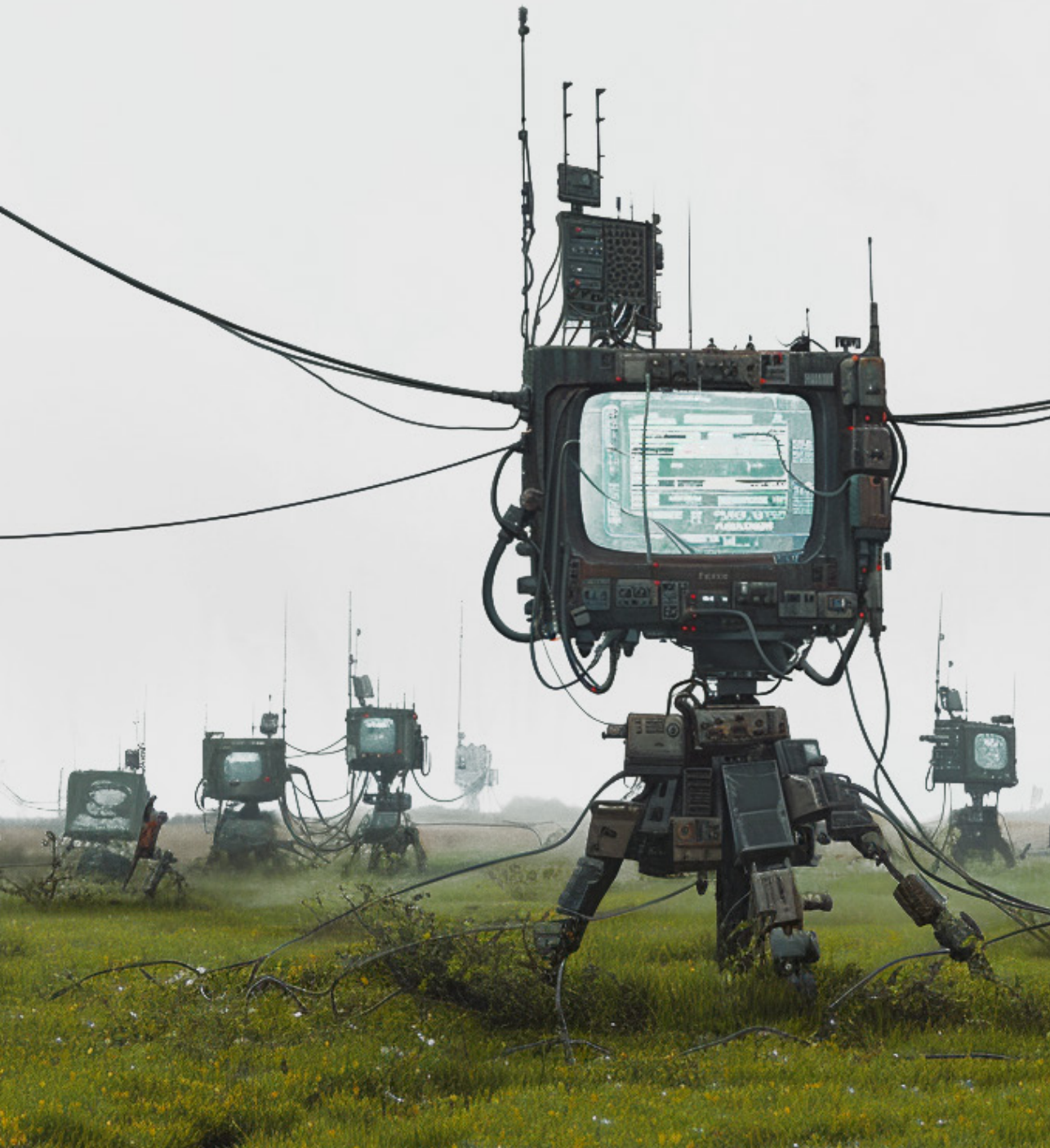
Como será possível perceber nas próximas páginas, mais do que respostas prontas e estanques às questões que envolvem as tecnologias de IA e os algoritmos, este livro apresenta novas preocupações, coloca outras perguntas e pretende alertar e inquietar os leitores. Mas há também, em todos os textos, algumas pistas e sugestões de possíveis caminhos para resolver alguns dos dilemas e desafios contemporâneos.

Boa leitura!

Adriana Gonçalves, Luísa Torre e Paulo Victor Melo

Parte 1

TECNOLOGIAS E OS MEDIA



TECNOLOGIA, SOCIEDADE E DEMOCRACIA: A QUESTÃO DA REGULAÇÃO

J. Paulo Serra
/ Universidade da Beira Interior - LabCom

Introdução

O texto que se segue resulta da minha intervenção na primeira das conversas do Ciclo de Conversas Inteligência Artificial, Algoritmos e Media organizado pelo LabCom – Comunicação e Artes. [1]

Nessa conversa, cada um dos intervenientes deveria responder a duas questões, uma de caráter geral, dirigida a todos os conferencistas, e outra de caráter mais específico, diferente para cada um deles.

A questão dirigida aos três conferencistas era a seguinte: "Concentração da propriedade em poucos grupos econômicos, utilização para projetos políticos autoritários, disseminação massiva de desinformação e discursos de ódio. Essas são algumas características globais quando falamos em internet e tecnologias digitais, inclusive num sentido oposto às promessas democratizadoras aquando do surgimento da internet. O que levou a esta situação e como sair dessa trama?"

A questão que me era especialmente dirigida era a seguinte: "Quais chaves teóricas e caminhos metodológicos você entende relevantes para a investigação sobre tecnologias digitais e internet atualmente?"

1. A conversa teve lugar em 10 de janeiro de 2023, na modalidade online. O programa do Ciclo pode ser consultado aqui: https://labcomca.ubi.pt/ciclo-de-conversas-inteligencia-artificial-algoritmos-e-media/

Compreende-se, assim, que o texto seguinte apresente duas seções referentes a cada umas das questões. A anteceder essas seções insiro uma outra secção sobre a questão da técnica, terminando o texto com algumas considerações finais acerca das três secções referidas.

1. A questão da técnica[2]

A técnica é vista, geralmente, de forma dicotómica: ora em termos de utopia, ora em termos de distopia; ora de forma eufórica, ora de forma disfórica. No entanto, qualquer consideração da técnica que não leve em conta a sua dimensão antropológica – a técnica é do homem, tanto quanto o homem é da técnica – carece de fundamento. De entre os autores que têm acentuado essa dimensão antropológica da técnica, Arnold Gehlen ocupa, ainda hoje, um lugar proeminente. Assim, na sua pequena obra *A alma na era da técnica*, publicada pela primeira vez em 1949 e reeditada em 1957, Gehlen afirma:

> A técnica é tão velha como o homem, pois é pelos vestígios da utilização de instrumentos de trabalho que podemos concluir seguramente que certos achados arqueológicos se relacionam com o homem. E já o mais rude "*coup de poing*" de sílex possuía a mesma ambiguidade que hoje caracteriza a energia atómica: era um instrumento útil e simultaneamente uma arma mortífera. (Gehlen, s.d., pp. 15-16)

De acordo com o autor, a técnica tem origem nas deficiências orgânicas do homem, e rege-se pelos princípios da substituição dos órgãos ("técnicas de compensação", como as armas ou o fogo), de fortalecimento dos órgãos ("técnicas de reforço", como a pedra na mão cerrada ou o martelo) e de redução do esforço dos órgãos ("técnicas de alívio", como o carro). (Gehlen, s.d., p. 16).[3]

2. Adotamos, neste texto, a distinção entre "técnica" e "tecnologia" feita por Mario Bunge (2012): enquanto a técnica é "todo o conjunto coerente de práticas ou regras de procedimento conducentes a um fim predeterminado" (p. 50), a tecnologia é "todo o sistema de técnicas práticas fundadas [numa disciplina científica], ou ao estudo das mesmas", pelo que se distingue da técnica pré-científica (p. 51).
3. Note-se que esta visão de Gehlen pode ser vista como uma revisitação do mito de Prometeu, que vê no fogo – dádiva divina -, o elemento que pode compensar as deficiências do homem relativamente aos outros animais (com penas, garras, etc.). Ora, o fogo é aqui a sinédoque da técnica – desde logo porque, para os gregos, o deus do fogo era Hefesto, "o deus-artífice" (Sottomayor, 2001, p. 134).

Estes princípios não são excludentes; assim, o avião, por exemplo, reúne as três técnicas referidas: "Quem viaja de avião encontra reunidos num os três princípios: substituição das asas que nos faltam, superação em grande escala de todas as possíveis capacidades orgânicas de vôo e economia de esforço próprio na deslocação a grandes distâncias." (Gehlen, s.d., p. 16).

A intelectualidade do homem, a sua capacidade de criar e utilizar técnicas que lhe permitem superar as suas deficiências orgânicas não é apenas uma característica acidental, que poderia não existir no homem e ele continuar a ser homem – já que sem ela não haveria homem. Por outra palavras, a técnica faz parte da própria essência do homem:

> Se entendermos por técnica os meios e as capacidades pelas quais o homem põe a natureza ao seu serviço, descobrindo as suas qualidades e leis que aproveita e põe em jogo umas contra as outras, temos de admitir que, neste sentido, a técnica pertence à essência do homem. (Gehlen, s.d., p. 17)

No seu conjunto, a história pode ser vista como uma substituição progressiva do orgânico pelo inorgânico: a matéria orgânica substituída por matérias artificiais (os metais, por exemplo), as forças orgânicas substituídas por forças anorgânicas (o vapor, a eletricidade...). (Gehlen, s.d., pp. 17-18)

A diferença essencial entre a técnica antiga e a moderna reside no facto de que, nesta última, se dá a conjugação de três elementos cruciais: as ciências da natureza, a técnica, e o sistema industrial, que dão origem a um "conjunto funcional" (Gehlen, s.d., p. 21).

A "fascinação do automatismo", já presente na magia, é o verdadeiro impulso da técnica:

> A fascinação do automatismo constitui o impulso pré-racional e extraprático da técnica, que primeiro se fez sentir na magia, técnica do supra-sensível, durante milénios, até encontrar nos tempos mais recentes a sua perfeita realização nos relógios, motores e máquinas rotativas de toda a ordem. (Gehlen, s.d., p. 21)

De acordo com H. Schmidt, citado por Gehlen (s.d., p. 28), podemos distinguir três "graus" (ou estádios) na evolução da técnica: o da ferramenta (assente na força física do sujeito), o da máquina de trabalho e de energia (assente na força de meios técnicos), e o do autómato (substituição do próprio esforço intelectual do sujeito por meios técnicos). Seguindo essa distinção, diz Gehlen:

> Da lei dos três estádios de H. Schmidt ressalta que a objetivação exterior de ações e capacidades humanas desloca-se, com referência ao homem, do exterior para o interior. Primeiro apenas se reforça, intensifica, aperfeiçoa e alivia o esforço orgânico. Depois, encarrega-se a natureza inanimada do dispêndio de energia orgânica, humana ou animal. No terceiro estádio, de que estamos tratando, objetiva-se o próprio ciclo de acção incluindo as funções intermediárias conscientes, de controle e direção. (Gehlen, s.d., p. 29)

É precisamente neste contexto (de procura) do automatismo, da substituição das funções conscientes de controlo e direção do homem pelas máquinas, que se coloca a questão da chamada "inteligência artificial".

Como se sabe, essa inteligência é vista, por Alan Turing (1950), em termos daquilo a que chama o "jogo da imitação",[4] o qual pode ser descrito, resumidamente do seguinte modo: se, dados um ser humano e uma máquina, ocultos ao observador, a máquina conseguir imitar o ser humano de forma tão perfeita que o observador em diálogo com ambos não consegue distinguir qual é o humano e qual é a máquina, então a máquina pode ser considerada como "inteligente".

Note-se que, no caso de Turing, estamos perante a "inteligência" de uma máquina ou, no limite, de várias máquinas que conjugam as suas capacidades de memória e de controlo. Ora, uma das características da atual inteligência artificial é o ser não de uma máquina – "inteligente" – mas de uma rede de máquinas utilizadas por milhões de seres humanos; ou seja, a inteligência

4. Este foi, precisamente, o título escolhido por Morten Tyldum para o seu filme sobre a vida de Alan Turing (*The Imitation Game*, Morten Tyldum, Reino Unido-EUA, 114m).

está na rede e nas suas interconexões, não propriamente nas máquinas. Deste modo, a inteligência da rede tem na sua base a inteligência dos seres humanos que alimentam a rede, aquilo a que Pierre Lévy designava como "inteligência coletiva" – ou seja, a inteligência dos múltiplos seres humanos conjugada por meios artificiais ou técnicos.

Descrevendo a situação da inteligência artificial na atualidade, Yuval Harari diz o seguinte:

> A revolução da automação está a surgir da confluência de duas marés científicas. Os cientistas da computação estão a desenvolver algoritmos de inteligência artificial (IA) que podem aprender, analisar grandes quantidades de dados e reconhecer padrões com eficiência sobre-humana. Ao mesmo tempo, biólogos e cientistas sociais estão a decifrar emoções, desejos e intuições humanas. A fusão da tecnologia da informação e da biotecnologia está a dar origem a algoritmos que podem analisar-nos e comunicar connosco de forma bem-sucedida, e que podem em breve superar médicos, motoristas, soldados e banqueiros humanos (...). (Harari, 2017, pp. 324-5)

2. A Internet e as tecnologias digitais

2.1. A Internet da utopia à distopia

Quando falamos da Internet e da sua história no âmbito da comunicação e dos media falamos, essencialmente, da sua versão como World Wide Web.

A World Wide Web, com origem na proposta apresentada por Tim Berners-Lee em 1989, foi vista inicialmente como um espaço de utopia, liberto dos condicionamentos económicos, políticos e militares de governos e corporações (Barlow e a sua "Declaração de Independência do Ciberespaço", 1996), assente na partilha do conhecimento (Berners-Lee et al., 1994; Berners-Lee, 2000), de construção da "comunidade virtual" (Rheingold, 1993), de confluência da "inteligência coletiva" (Lévy, 1997), configurando uma espécie de *agora* virtual.

Essa utopia – e mesmo euforia – teve um novo incremento quando, nos inícios dos anos 2000, emergiu a chamada "Web 2.0", que veio permitir que cada um dos utilizadores/consumidores se tornasse, também, um utilizador/produtor, inaugurando aquilo a que Castells (2009) chamou a "auto comunicação de massa".

No entanto, aquilo que parecia ser a grande viragem democrática da Web – de uma Web dos especialistas a uma web dos cidadãos –, foi dando cada vez mais lugar à distopia – e disforia – da comunicação populista das redes sociais e dos sites partilhados, que tem sido objeto de crítica acesa de, entre muitos outros, autores como Umberto Eco ou Zygmunt Baumann.

Assim, numa palestra realizada em 2014, no *Festival della comunicazione di Camogli,* Eco assaca vários problemas à Internet e às redes sociais. No que se refere à Internet em geral, Eco elenca os seguintes problemas (o resumo é meu): dificuldade ou mesmo ausência de validação das fontes das informações disponíveis; possibilidade de auto publicação de meros "aspirantes a escritores", que pode levar alguns a confundir o trigo com o joio; alegada democratização do gosto e do juízo do utilizador, que acaba por confundir qualidade com mediocridade; ultrapassagem dos *gatekeepers*, das instâncias de seleção e validação da informação. Relativamente ao Facebook (às redes sociais), Eco aponta os seguintes problemas: impossibilidade de verificar os produtores da mensagem (um pedófilo disfarçado de criança, por exemplo); redes como instrumento de vigilância e de controlo promovido pelos próprios utilizadores a ser vigiados e controlados;[5] predomínio da função fática, do contacto sobre o conteúdo, visível também na utilização dos "*like*"; transformação da conversa em rumor partilhado, que não se confunde com o consenso democrático (Eco, 2021, p. 189).

Tal não obsta a que Eco reconheça – tal como, noutro contexto, também Habermas o fez – que "uma das grandes revoluções da comunicação online" consistiu em permitir a cada um tomar a palavra e dirigir-se ao mundo de

5. Como sublinha Eco, "é a primeira vez na história da humanidade que os espiados colaboram com os que espiam, facilitando o seu trabalho e retirando dessa colaboração sentimentos de satisfação pois, ao serem vistos, existem." (Eco, 2021, p. 188).

forma livre, ultrapassando a censura dos regimes ditatoriais (Eco, 2021, p. 188). As redes permitiram também, a todos em geral e aos líderes políticos e religiosos em particular, estabelecer uma comunicação direta, livre de mediação, e dirigida a mais destinatários – fazendo com que as mensagens de redes como o Twitter sejam retomadas pelos outros media, nomeadamente a televisão, configurando o fenómeno a que Eco chama de "comunicação da comunicação" (o exemplo de Eco, Beppe Grillo, foi mais recentemente replicado por Donald Trump) (Eco, 2021, p. 189).

Numa outra intervenção oral, feita ao receber o título de doutor honoris causa em Comunicação e Cultura, na Universidade de Turim, na Itália, em 10 de junho de 2015, Umberto Eco afirmou que "o drama da internet é que ela promoveu o idiota da aldeia a portador da verdade" (Eco, citado em Silva, 2015, 23 de junho).[6] Acrescentou ainda Eco que "normalmente, eles [os imbecis] eram imediatamente calados, mas agora eles têm o mesmo direito à palavra de um Prêmio Nobel", e que "redes sociais deram voz à legião de imbecis" (Eco, citado em Silva, 2015, 23 de junho).

Quanto a Bauman, as palavras que são citadas no título de uma sua entrevista publicada no *EL País (Brasil)* em 8 de janeiro de 2016 exprimem bem a posição do sociólogo da "modernidade líquida" sobre as redes sociais: "As redes sociais são uma armadilha" (Bauman, citado em Querol, 2018). Bauman avança pelo menos três razões para esta sua posição: i) longe de promoverem a comunidade, as redes sociais centram-se no indivíduo que pretende construir a sua própria "comunidade" de amigos com os mesmos "gostos" – pelo que a suposta "comunidade" não passa, assim, de uma bolha; ii) as redes sociais não exigem as habilidades sociais necessárias para os indivíduos se relacionarem com os outros, as quais (só) se desenvolvem na interação presencial; iii) as redes sociais não envolvem um diálogo com o diferente e o controverso; o suposto "diálogo" em rede não passa de um monólogo em eco. Assim, mesmo se admite que "as redes são muito úteis,

6. Apesar de utilizarmos apenas este autor como fonte das palavras de Eco, tivemos ocasião de verificar que essas mesmas palavras foram citadas por diversas outras fontes. Abstenho-me de discutir a posição específica do autor sobre a intervenção de Eco, já que esse não é aqui o meu propósito.

oferecem serviços muito prazerosos", Bauman conclui que elas acabam por ser uma armadilha – uma armadilha do ponto de vista social e comunitário, fazendo exatamente o contrário daquilo que (os seus criadores) prometem.

As caraterísticas da Internet e das redes sociais identificadas (e criticadas) por Eco e Bauman decorrem do facto de que a Internet é, hoje, uma outra Internet: uma "pós-Internet", assente nos princípios da Cloud, dos Big Data e da Internet das Coisas (Mosco, 2016; 2017), submetida às exigências económicas e políticas do "capitalismo de vigilância" (Zuboff, 2019), com evidentes repercussões em áreas como a economia, a (re)distribuição do poder dos estados e das agências de informações militares, o meio ambiente, a vida privada e a segurança pessoal, o trabalho humano, etc.

Não admira assim que, em particular na sequência da venda e/ou utilização dos dados dos utilizadores pelas redes sociais envolvidos acontecimentos políticos como o referendo do Brexit (2016) ou a eleição de Donald Trump (2016),[7] "a crença ciberlibertária da década de 1990 de que a internet representa a última fronteira da liberdade em relação ao estado cedeu ainda mais terreno a uma aceitação generalizada de que é necessária uma maior regulação estatal da internet" (Rone, 2021, p. 175).

2.2. Governança e regulação da Internet

A importância económica, política, social e cultural da Internet fez emergir, sobretudo a partir do final dos anos 90, a questão crucial da sua governação.

Até finais dos anos 1990 a governança da Internet assentava numa espécie de acordo de cavalheiros entre cientistas, engenheiros e técnicos informáticos como Vint Cerf, Robert Kahn, John Postel, ou Tim Berners-Lee, que, na base da boa vontade e da partilha, iam gerindo a rede.

7. Referimo-nos, em particular, ao escândalo da Cambridge Analytica e da sua relação com o Facebook, que rebentou em 2018. Ver, por exemplo, https://www.theguardian.com/news/series/cambridge-analytica-files. O filme *The Great Hack* (Karim Amer, Jehane Noujaim, EUA, 113m) refere-se, também, a esse mesmo escândalo.

Com a criação da Internet Corporation for Assigned Names and Numbers (ICANN), em 1998, pela administração Clinton, assume-se a Internet como coisa estadunidense, gerida e controlada pelos EUA. Tornava-se visível, deste modo, a contradição entre o caráter global da Internet e o seu controlo efetivo pelos EUA (Castells, 2001, pp. 29-33).

Sobretudo com a Web 2.0, no início dos anos 2000, a ideia de governança da Internet, incluindo empresas, governos, associações da sociedade civil, movimentos sociais, cidadãos, etc., dá lugar ao governo de facto das chamadas "GAFAM" (Google, Apple, Facebook, Amazon e Microsoft). Deste modo, na "sociedade da plataforma" (van Dijck, Poell & De Waal, 2018), o processo de governança, entendido como um processo teleológico e deliberado, foi sendo substituído pela opacidade fática dos "sistemas sociotécnicos" (p. 139), colocados ao serviço da acumulação capitalista.

Em outubro de 2013, e ainda no rescaldo do caso Snowden, a presidente do Brasil Dilma Rousseff e o presidente do Internet Corporation for Assigned Names and Numbers (ICANN), Fadi Chehadi, deram início ao processo da NetMundial, envolvendo não apenas governos, mas muitos outros parceiros, trazendo definitivamente a questão da governança da Internet para o primeiro plano da política internacional e colocando mais uma vez em questão o controlo prático da Internet pelos EUA.

Nesse mesmo espírito – de defesa da democracia e do comum – a Unesco tem vindo a defender a "universalidade da Internet" como princípio fundamental, apontando para a necessidade de uma Internet fundada nos direitos do homem, aberta, acessível a todos e envolvendo a participação de múltiplos atores.

É certo que o caminho para a "universalidade da Internet" não é óbvio e está longe de ser fácil. Mas parece indubitável que a solução da questão da governança da Internet só pode ser política, por parte dos poderes públicos e dos movimentos sociais mundiais.

Estamos, assim, colocados perante um dilema: ou uma sociedade democrática, de informação acessível a todos, e em que a governança é partilhada por entidades centrais e descentralizadas, locais, regionais e internacionais; ou uma sociedade controlada pelas corporações transnacionais e pelos serviços de informação dos diversos países, sendo o "governo" deixado ao livre-arbítrio dos mercados (Mosco, 2016, p. 262).

Países como o Brasil ou organizações internacionais como a União Europeia têm dito uma importante palavra sobre este (e neste) futuro incerto. O Marco Civil do Brasil ou as recentes medidas de regulação (legal) da União Europeia – sobre a neutralidade da rede, os dados pessoais, ou o discurso do ódio – são já um importante passo nesse sentido.

As atuais tentativas de regulação por parte da União Europeia em relação às GAFAM, configurando aquilo que Santaniello (2021) chama "a viragem para a regulação da Internet pelo estado" (p. 29), assenta na oposição à tese delas de que serão apenas "plataformas", irresponsáveis relativamente aos conteúdos por si veiculados – discurso de ódio, racismo, xenofobia, etc. – para considerar que elas são "meios" (de comunicação) e, como tal, responsáveis pelos conteúdos que veiculam. No entanto, a regulação democrática da Internet e das redes sociais deve recusar, desde logo, quer o "modelo californiano", quer o "modelo chinês", ou seja, quer o *laissez faire, laissez passer* do liberalismo selvagem, quer o controlo autoritário do estado (Haggart, Tusikov & Scholte, 2021).

Essa regulação enfrenta, hoje, um novo desafio: o da inteligência artificial. Sobre esse mesmo desafio, Yuval Harari alertava, em conferência recente em Lisboa,[8] para a necessidade de os governos procederem à regulação da implementação da inteligência artificial nas sociedades, dada a impossibilidade de travar ou controlar o seu desenvolvimento. Note-se que esta mesma ideia já fora defendida por Harari noutros textos, por exemplo em artigo de 2017 na *Nature*, quando, ao analisar os possíveis efeitos da inteligência

8. https://www.publico.pt/2023/05/19/ciencia/noticia/yuval-harari-lisboa-inteligencia-artificial-trex-destruira-democracia-2050375

artificial no mercado de trabalho, na economia e na sociedade em geral, afirma que "os governos podem decidir desacelerar, deliberadamente, o ritmo da automação, para diminuir os choques dela resultantes e permitir tempo para reajustamentos" (Harari, 2017, p. 325).

3. Caminhos teóricos e metodológicos

Abordo, agora, a segunda questão que me foi colocada relativa às "chaves teóricas e caminhos metodológicos" a adotar para a investigação sobre tecnologias digitais e internet.

No que se refere às "chaves teóricas", penso que a perspetiva teórica geral a adotar na investigação sobre as tecnologias digitais e a internet implica, desde logo, recusar o determinismo tecnológico – e tecnofílico –, tendo sempre presente a ideia de que a tecnologia, em particular a de comunicação, não é tecnológica, mas política, social e cultural.

Quanto aos "caminhos metodológicos", vejo a necessidade de adoção de uma abordagem multidisciplinar e, *eo ipso*, metodologicamente pluralista (métodos mistos) das tecnologias digitais e a internet, que faça confluir nos estudos de comunicação disciplinas como a filosofia, a sociologia, a antropologia, a história, a economia, a ciência política e outras.

Essa abordagem multidisciplinar e este pluralismo metodológico decorrem, desde logo, da multiplicidade dos tópicos a investigar neste domínio. Destaco aqui alguns deles, que me parecem mais relevantes:

a. Investigação: o que significa passar de um modelo de trabalho empírico centrado na recolha de dados para outro centrado na extração de dados?

b. Educação: que novas competências devem os estudantes adquirir? E como? Quais os problemas éticos e políticos envolvidos? Como incorporar os novos programas de inteligência artificial na educação?

c. Profissões: que impactos têm a inteligência artificial e os algoritmos nas profissões em geral e, em particular, nas profissões da comunicação e

do jornalismo, no que se refere a emprego, rotinas de produção, ética, empreendedorismo, etc.?

d. Sociedade e cultura: quais as consequências culturais, éticas e políticas da *Next Internet* em termos de democracia, diversidade cultural e linguística, privacidade, neutralidade da rede, acesso, etc.?

e. Economia: que repercussões terá, nas economias nacionais e na economia mundial, a viragem para uma economia assente nos dados e na informação, em vez das tradicionais mercadorias e força de trabalho? Quem serão os novos ricos e os novos pobres? Como se fará a distribuição da riqueza em cada sociedade e a nível mundial? Que novas formas de desigualdade irão emergir?

f. Política: em que medida é que a inteligência artificial e os algoritmos alteram a política democrática, assente na existência de partidos, na luta de ideias não violenta, no governo da maioria com respeito pelas minorias? Quais os efeitos de uma democracia mais populista do que popular, assente nas redes sociais, no discurso de ódio e no ataque *ad hominem*?

g. Comunicação: como se faz e fará a articulação entre o artístico-cultural e o informativo, as narrativas intermediáticas, os projetos interdisciplinares, em suma, a criação com o cálculo?

Considerações finais

A enumeração destas questões, a que poderiam (poderão) juntar-se muitas outras, mostra a necessidade de continuar a investigar e a pensar a realidade multidimensional da internet e das tecnologias digitais – e, desde logo, as suas tendências emergentes, ligadas à inteligência artificial.

A história dos humanos tem mostrado que é impossível travar ou controlar o desenvolvimento da tecnologia, por mais arriscados que sejam os seus efeitos reais ou potenciais – a bomba atómica é um bom exemplo.[9] De facto,

9. Veja-se, a propósito, o conhecido texto de Günther Anders (1962).

a tecnologia, apesar de ser uma criação dos humanos, apenas em parte é controlável pelos humanos. Não só ela tem uma espécie de lógica autotélica, desenvolvendo-se por si própria, de acordo com as suas próprias exigências, como envolve impactos e consequências sociais que só muito tempo depois da invenção das várias tecnologias se tornam visíveis. A inteligência artificial é, também, um bom exemplo disso: preparada, pelo menos, após a II Guerra Mundial, apenas hoje ela se tornou um desafio crucial para todos nós.

No entanto, impossibilidade de travar ou controlar o desenvolvimento da tecnologia não significa que tal desenvolvimento não possa ser objeto de uma regulação política e jurídica, que procure fazer respeitar os direitos fundamentais de todos os cidadãos. Assim, contra o determinismo tecnológico, seja na sua versão utópica, seja na sua versão distópica, resta-nos, como hipótese, a regulação. O caminho que nesta matéria tem vindo a ser seguido, entre outros atores, pela União Europeia, dá-nos alguma esperança de que pelo menos alguns dos impactos mais perigosos das atuais tecnologias possam ser evitados ou, pelo menos, minorados.

Assim, se é verdade que não podemos passar sem a tecnologia, que ela é indissociável da história dos humanos, não é menos verdade que não podemos passar sem a política; e, ao mesmo tempo, passar sem (a tentativa de) regulação política da tecnologia.

Referências

Anders, G. (1962). Theses for the atomic age. *The Massachusetts Review*, *3*(3), 493-505.

Barlow, J. P. (1996). *A Declaration of Independence of Cyberspace*. http://www. eff.org/pub/Publications/John_Perry_Barlow/barlow_0296.declaration

Berners-Lee, T., Cailliau, R., Luotonen, A., Frystyk Nielsen, H. & Secret, A. (1994). The World-Wide Web. *Communications of the ACM*, 39(8), 76-82.

Berners-Lee, T., & Fischetti, M. (2000). *Weaving the Web: The original design of the World Wide Web by its inventor*. Harper Collins.

Bunge, M. (2012). *Filosofía de la tecnología y otros ensayos*. Fondo Editorial de la Universidad Inca Garcllaso de la Vega.

Castells, M. (2001). *The internet galaxy*. Oxford University Press.

Castells, M. (2009). *Communication power*. Oxford University Press.

Eco, U. (2021). Comunicação: Hard e soft. Umberto Eco Palestra realizada em 2014, Festival della comunicazione di Camogli1 (Curadoria, adaptação e edição por Gustavo Cardoso e Caterina Foá, a partir de vídeo e texto). *Observatorio (OBS*) Journal*, 15 (2), 185-193.

Gehlen, A. (s.d.). *A alma na era da técnica*. Livros do Brasil (original de 1949).

Haggart, B., Tusikov, N., & Scholte, J. A. (Eds.) (2021). *Power and authority in internet governance: Return of the state?* Routledge.

Harari, Y. N. (2017). Reboot for the AI revolution. *Nature*, 550, 324-327. https://doi.org/10.1038/550324a

Lévy, P. (1997). *L'intelligence collective : Pour une anthropologie du cyberespace*. La Decouverte.

Mosco, V. (2016). Après l'internet : Le cloud, les big data et l'internet des objets. *Les Enjeux de l'Information et de la Communication*, 18, 253-264.

Mosco, V. (2017). *Becoming digital: Toward a post-internet society*. Emerald Publishing Limited**.**

Querol, R. de (2016, 8 janeiro). Zygmunt Bauman: "As redes sociais são uma armadilha". *El País (Brasil)*. https://brasil.elpais.com/brasil/2015/12/30/cultura/1451504427_675885.html

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier*. Addison-Wesley.

Rone, J. (2021). The return of the state? Power and legitimacy challenges to the EU's regulation of online disinformation. In B. Haggart, N. Tusikov & J. A. Scholte (Eds.), *Power and authority in internet governance: Return of the state?* (pp. 171 - 194). Routledge.

Santaniello, M. (2021). From governance denial to state regulation: A controversy-based typology of internet governance models. In B. Haggart, N. Tusikov & J. A. Scholte (Eds.), *Power and authority in internet governance: Return of the state?* (pp. 15 – 36). Routledge.

Silva, M. F. L. (2015, 23 de junho). O idiota da aldeia e o portador da Verdade. *Observatório da Imprensa, Jornal de Debates*, 856. http://www.observatoriodaimprensa.com.br/jornal-de-debates/o-idiota-da-aldeia-e-o-portador-da-verdade/

Sottomayor, A. P. Q. (2001). O fogo de Prometeu. *Hvmanitas*, LIII, 133-140.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, New Series, *59*(236), 433-460.

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

INTELIGÊNCIA ARTIFICIAL, ALGORITMOS E MEDIA: DIÁLOGOS DE PESQUISA

Adriana Gonçalves
/ Universidade da Beira Interior - LabCom

Luísa Torre
/ Universidade da Beira Interior - LabCom

Paulo Victor Melo
/ Universidade Nova de Lisboa - ICNOVA

Introdução

As primeiras discussões sobre a Inteligência Artificial (IA[1]) levam-nos a recuar quase 100 anos, até ao ano de 1940, momento a partir do qual se começou a questionar se os computadores poderiam reagir de forma idêntica a um humano perante a mesma situação (Oliveira, 2019). Essa dúvida permaneceu por muitos anos uma problemática de investigação perseguida por investigadores de diversas áreas. Entre eles, destacou-se o matemático Alan Turing que inventou o Jogo da Imitação para responder à questão "as máquinas conseguem pensar?". Turing propunha aos observadores que se não conseguissem distinguir a ação da máquina do comportamento dos seres humanos, então a máquina seria

1. Neste trabalho, seguimos a definição de IA atribuída pela Comissão Europeia: "A IA refere-se a sistemas projetados por humanos que, dado um objetivo complexo, atuam no mundo físico ou digital, percebendo o seu ambiente, interpretando os dados coletados estruturados ou não estruturados, raciocinando sobre o conhecimento derivado desses dados e decidindo a(as) melhor(es) ação(ões) a realizar (de acordo com parâmetros predefinidos) para atingir o objetivo determinado. Os sistemas de IA também podem ser projetados para aprender a adaptar seu comportamento, analisando como o ambiente é afetado pelas suas ações anteriores (Comissão Europeia, 2018, p.7).

considerada inteligente (Turing, 1950). Este jogo ficou conhecido como o Teste de Turing, que continua a ser uma referência nos dias de hoje para descrever tecnologias 'inteligentes'.

Na verdade, a inspiração para a construção dos computadores modernos foi a arquitetura do cérebro humano: "o computador foi desde o início pensado teoricamente como um modelo do cérebro" e ainda "a ligação original entre cérebro e computador teve como consequência lógica a constituição da disciplina Inteligência Artificial" (Rosa, 2007, p. 44). Porém, a IA enquanto disciplina só surgiu em 1956 e, desde essa altura, tem atravessado períodos de rápida evolução e outros de estagnação, estes últimos conhecidos como invernos da IA.

O passado recente é marcado por um novo impulso de desenvolvimento tecnológico, conhecido por Quarta Revolução Industrial ou Indústria 4.0 (Vicente & Dias-Trindade, 2021). Esta revolução é marcada por processos de convergência de inovações digitais, biológicas e físicas (Schwab, 2018), nos quais os algoritmos, a IA e o Big Data participam ativamente, acelerando a revolução digital em todos os setores da sociedade (Kusters et al., 2020). A indústria 4.0 assenta na abundância informativa e no volume de dados estruturados que cresce de dia para dia. Organizar e fazer sentido desta vastidão de informação é um desafio que tem sido entregue às máquinas, mais especificamente aos algoritmos. Um algoritmo é uma sequência de comandos que instrui uma máquina, expresso em linguagens de programação computacional, que informa um aplicativo ou software que ações tomar (Bunz, 2017). Assim, os algoritmos organizam a informação em dados quantificáveis, um fenômeno conhecido por dataficação (*datification*) (Lemos, 2021; Gillespie, 2014). Neste sentido, os dados estruturados constituem a matéria-prima dos sistemas algorítmicos e, particularmente, da IA.

Em certa medida, estas tecnologias tornam as nossas vidas mais aceleradas e ajudam-nos a resolver problemas complexos em todos os setores, desde a justiça, a educação, a saúde, a segurança, a economia e ainda o jornalismo (Biswal & Kulkarni, 2024). Porém, também incitam novos problemas

na sociedade. Vicente (2023) refere que "a proliferação de sistemas inteligentes de classificação, seleção, recomendação e de apoio à tomada de decisão adquiriram uma relevância sem precedentes ao serem integrados no quotidiano das instituições" (p.10) e no quotidiano dos cidadãos, tornando urgente o debate sobre o seu funcionamento e ações.

Um dos momentos que contribuiu para a popularidade do uso da IA foi o lançamento do ChatGPT a 30 de novembro de 2022 (Sanin, 2023). A partir daí, a IA entrou na agenda mediática, nas publicações das redes sociais e no quotidiano de milhares de utilizadores em todo o mundo, originando uma multiplicidade de usos e interpretações, e potenciando o agravamento de fluxos de desinformação. O interesse pelo assunto é tanto que o número de pesquisas no Google pelo termo "AI" em todo o mundo registrou um crescimento exponencial durante os seis meses seguintes ao lançamento do ChatGPT, continuando a aumentar de dia para dia (Google Trends, 2024).

Este cenário provocou um aumento do leque de dúvidas e de problemáticas para a humanidade. Uma das principais preocupações deve-se à opacidade e à falta de explicabilidade associadas aos algoritmos e à IA, duas características que impedem a comunidade de compreender estas tecnologias, desenvolver métodos para o seu estudo e regras para a sua utilização (Wang, 2019; Kusters et al., 2020; Vicente, 2023). Este e outros problemas cruzam-se com os nossos interesses de investigação, nomeadamente: a utilização de algoritmos e IA no jornalismo; a problemática da desinformação; e as relações entre IA e o racismo algorítmico. Partindo destes três tópicos, neste capítulo debatemos alguns dos desafios do uso de algoritmos e de IA, trazendo exemplos que conectam as diversas questões com as nossas linhas de pesquisa.

Inteligência Artificial no jornalismo

A viragem algorítmica (Napoli, 2014) tem incitado transformações profundas na atividade jornalística, levantando questões em torno da objetividade, da autonomia e do serviço público (Milosavljević & Vobič, 2019), que

culminam em debates mais profundos sobre a conceptualização e essência do jornalismo (García-Orosa et al., 2023).

A chegada de bancos de dados, algoritmos e sistemas inteligentes afeta toda a cadeia informativa, onde se incluem as fases de produção, distribuição e recepção de conteúdos noticiosos (Diakopoulos, 2019). Os estudos sobre este tema revelam "o papel transformador das máquinas, especialmente nas fases de coleta e distribuição de notícias, e cada vez mais na fase de redação, especialmente em especialidades onde os dados são abundantes, como o desporto e a economia" (García-Orosa et al., 2023, p.16). Em vários países do mundo, multiplicam-se os usos de IA e automação, seja em tarefas rotineiras dos jornalistas, como transcrever entrevistas e traduzir conteúdos, ou noutro tipo de funções, como identificar tendências nas redes sociais, rastrear preferências do público em tempo real e analisar bancos de dados em busca de padrões (Beckett & Yaseen, 2023; Newman et al., 2023). A escrita de conteúdos por parte de softwares é uma das áreas em expansão, utilizada especialmente em notícias sobre desporto, oscilações da bolsa económica, alertas meteorológicos, mercado imobiliário ou crimes (Diakopoulos, 2019; Canavilhas, 2023).

Dados recentes revelam que 75% dos *media* (numa amostra de 105 organizações de 46 países), utiliza IA em pelo menos uma das áreas – recolha, produção e distribuição, mas apenas 1/3 tem uma estratégia institucional para o uso de IA implementada ou em desenvolvimento (Beckett & Yaseen, 2023). Estes números ajudam a enquadrar o uso crescente de IA nas redações, que carece de regras para uma utilização responsável (Helberger et al., 2022).

A estratégia de delegar algumas funções jornalísticas para ferramentas computacionais mostra uma tentativa por parte dos media de acompanhar a velocidade de circulação dos conteúdos na web e de superar a crise de financiamento, alicerçada na decadência da receita publicitária e das vendas em banca. Com estes novos recursos, os processos de produção informativa tornam-se mais céleres, surgem novas abordagens e espaço para conteúdos diferenciados (Newman, 2023). Dentro desses conteúdos diferenciados,

incluímos grandes reportagens, como por exemplo, as investigações em torno do *Wikileaks* ou dos *Panama Papers*, que constituíram marcos no jornalismo de dados e demonstraram o potencial do uso de automação no jornalismo para processar e analisar enormes quantidades de informações (Baack, 2016).

Embora Milosavljević e Vobič (2019) defendam que os jornalistas continuam a controlar todas as etapas da cadeia de notícias, Diakopoulos (2019) destaca que "já não existe nenhuma etapa do processo jornalístico que não seja tocada por algoritmos, desde coleta de informações, tomada de decisão, narrativa e distribuição de conteúdo" (p.3), o que torna pertinente a questão: qual o papel desempenhado pelo jornalista neste novo ecossistema informativo?

O papel do jornalista sempre foi o de informar com rigor e objetividade, servindo o interesse público e a democracia, mas a atuação dos algoritmos tem-se estendido a funções como a recolha de informações, a verificação, a escrita de determinados conteúdos e a distribuição nas redes, ocupando, de algum modo, o espaço do jornalista. Esta mudança tem impacto direto nas rotinas do jornalista que, por um lado, se liberta de tarefas rotineiras e repetitivas, mas por outro, teme que a sua profissão possa ser inteiramente ocupada pelas máquinas.

Os avanços tecnológicos são evidentes no quotidiano dos jornalistas, que há bem pouco tempo perdiam horas a transcrever entrevistas, hoje transcritas automaticamente por *softwares*. No entanto, algo que estas tecnologias não possuem é a capacidade de raciocínio, de interpretação e criatividade, muito menos reconhecem as normas éticas ou códigos deontológicos da profissão de jornalista. Nesse sentido, a sua inteligência pode ser questionada, pois na verdade, as máquinas cumprem uma série de regras matemáticas lógicas (Carreira, 2017), pré-estabelecidas pelos programadores. E o conjunto dessas limitações traduz-se na incapacidade de fornecer contexto, aprofundar os assuntos e na falta de faro jornalístico (Thurman et al., 2017).

Por essa razão, alguns autores ressaltam que a IA não substitui o jornalista nas suas funções primordiais, apenas o desloca para as funções de verificar,

interpretar e dar sentido à informação (Anderson et al., 2013; Quandt et al., 2021), reforçando o seu papel de contextualizar e explicar a informação ao público. Até porque, "a informação e os dados não têm profundidade. O pensamento humano é mais do que cálculo e resolução de problemas. Clarifica e ilumina o mundo. Faz surgir um mundo completamente diferente", aponta o filósofo Han (2021, p.50), reforçando a ideia de que os dados não são nada sem a interpretação humana.

Mesmo com estas diferenças claras entre os humanos e as máquinas, o cenário de aceleração tecnológica contribui para um clima de insegurança e incerteza na profissão do jornalista, que teme ser substituído por máquinas, receia os despedimentos e a consequente fragilização da sua atividade profissional (Wölker & Powell, 2021; Beckett, 2019). Dada esta soberania dos sistemas computacionais, adensam-se os debates sobre a ética (Pocino, 2022), redefinição de práticas e valores (Wu et al., 2019), regulação, entre outros.

Fruto dessas preocupações, em 2022, o Conselho de Informação da Catalunha (CIC) lançou um relatório pioneiro sobre ética, Inteligência Artificial e jornalismo, como forma de refletir sobre os desafios atuais e fazer recomendações para a integração ética da IA no jornalismo. Baseado nas reflexões propostas pelos principais *media* espanhóis, é proposto um conjunto de ações guiadas pelos valores éticos do jornalismo, tais como: salvaguardar as fontes e diversidade dos dados; dar a conhecer aos leitores a existência de algoritmos e como eles operam; ter o cuidado de manter os dados pessoais anonimizados para respeitar a privacidade dos cidadãos; e evitar o uso de mecanismos de personalização que prejudique a exposição a conteúdos diversificados (Pocino, 2022). No fundo, o relatório sugere que a IA deve ser utilizada para fomentar os valores do jornalismo e da democracia. Este documento é um exemplo a seguir, porque promove uma reflexão crítica e traça uma estratégia para o uso democrático dos sistemas inteligentes no jornalismo, sem que isso destrua os pressupostos base desta profissão centenária.

Enquanto a regulação não se torna efetiva, a experiência dos grandes grupos mediáticos, como a *BBC*, *Reuters, Los Angeles Times, The Guardian* e outros, mostra melhorias na integração da IA com o trabalho dos jornalistas humanos. Do outro lado do espectro, as redações mais pequenas ficam cada vez mais defasadas deste progresso tecnológico, uma vez que experimentar novas tecnologias requer esforços financeiros, tempo e disponibilidade de profissionais, que são cada vez em menor número (Rinehart & Rung, 2022). Além disso, a automação e a integração de algoritmos no processo de disseminação de informação traz outros desafios para a prática do jornalismo, nomeadamente, riscos associados à excessiva personalização dos conteúdos e os filtros-bolha (Pariser, 2012). Aqui entra também um outro problema relacionado com os contributos do público no ambiente noticioso, e mais especificamente, de contribuições antidemocráticas tais como conteúdos de desinformação e de fake news, como discutimos no próximo tópico.

Os algoritmos e a desinformação

As lógicas algorítmicas permeiam a internet, desde motores de busca até os feeds de notícias em redes sociais, onde organizam, relacionam e hierarquizam informação. Seja uma pessoa comum compartilhando uma foto ou uma organização de media divulgando notícias, todos os conteúdos estão sujeitos à ação desses algoritmos. Essa intervenção algorítmica não é neutra, e estes dispositivos realizam leituras interessadas da realidade (Rieder, 2018) a partir de fórmulas que são segredos industriais, como se fossem "caixas pretas" (Pasquale, 2015), eles controlam o que se vê e o que não se vê e transformam o mundo em que vivemos (Pariser, 2012).

Os algoritmos vão desempenhar um papel fundamental na forma como as plataformas funcionam: a recolha massiva de dados, resultado de um processo de vigilância constante e extensivo na internet. A vigilância realizada pelos algoritmos tornou-se a base deste novo paradigma, o capitalismo de vigilância, em que a experiência humana nos territórios digitais é transformada em dados comportamentais que, por sua vez, vão alimentar processos de treinamento de máquinas que vão antecipar o que o usuário vai fazer e

como ele irá agir no futuro (Serra, 1998; Wolton, 2010; Musso, 2004; Zuboff; 2019). Um aspecto crucial desse sistema é o imperativo de compartilhar, que gera dados sobre preferências, hábitos e opiniões dos usuários, fundamentais para a publicidade direcionada e a construção de sujeitos consumidores (Lyon, 2017).

Para as corporações, o tráfego de informações será a chave da geração de valor (Marshall, 2009). As cinco grandes corporações (Apple, Google, Microsoft, Amazon e Meta) "dominam não apenas a Internet, mas também o modo econômico de operação, que ultrapassou os modos de acumulação gerencial e financeira que caracterizaram o final do século XX e início do século XXI" (Lyon, 2017, p. 4). Apesar da origem da internet como um espaço livre de regulamentação, e de sua ideologia baseada em uma utopia de libertação das autoridades e de horizontalidade da comunicação que é própria da ideia de rede, a vigilância tornou-se naturalizada em suas estruturas (Wolton, 2010). Surpreendentemente, os indivíduos voluntariamente se submetem a dispositivos de vigilância, como smartphones, rastreando não apenas suas atividades online, mas também movimentos, horários e padrões de comportamento. Nas redes sociais, a vigilância é explorada por grandes empresas, como o Facebook, para coletar dados e direcionar publicidade personalizada, financiada por empresas diversas. Esses dados, entregues voluntariamente pelos usuários em troca de entretenimento e pertencimento ao grupo (Lyon, 2017), serão categorizados em perfis e classificados (Bruno, 2008) para tornarem-se alvos de anúncios e também de recomendações de novos conteúdos, em um cenário de total personalização (Pariser, 2012).

Plataformas de redes sociais funcionam com base em algoritmos de recomendação utilizados para criar personalização, reter os utilizadores e prever as suas preferências, bombardeando-os com informação que muito provavelmente irão gostar, num processo sem mediação humana, tudo feito por algoritmos que não são neutros (Risi & Pronzato, 2022; Heinderyckx & Vos, 2016; Zuboff, 2019; Rieder, 2018). Reter o usuário nas plataformas e fazê-lo se engajar nas publicações, comentando, compartilhando e interagindo, é a

base de sua relevância também no mercado publicitário. Enquanto as redes permitem a qualquer usuário monetizar conteúdos por meio de anúncios online e divulgação nas redes sociais, elas acabam por oferecer ferramentas para promover ativamente a desinformação e torná-la uma atividade rentável (Bakir & McStay, 2018; Lazer et al., 2018).

A disseminação de informações intencionalmente enganosas na internet e nas redes sociais pode ser vista como uma forma antidemocrática de intervenção no debate público ou uma "participação obscura" (Quandt, 2018), categoria que também inclui o discurso de ódio. Esta participação antidemocrática é promovida por fontes em que os jornalistas confiam, como agentes políticos, que empregam técnicas para manipular o debate e tiram partido das técnicas da cultura participativa e das possibilidades das redes sociais para manipular e influenciar a esfera pública (Marwick & Lewis, 2017).

Do ponto de vista do seu formato, o fenómeno da desinformação engloba uma variedade de conteúdos falsos, que incluem elementos deliberadamente enganadores no seu conteúdo ou no seu contexto, com a intenção de enganar (Bakir e McStay, 2018; Wardle e Derakhshan, 2018) e produzidos com um objetivo económico ou ideológico (Tandoc, Lim & Ling, 2018). O fenómeno também se refere a informações que visam poluir o ecossistema noticioso e dificultar que o público saiba em quem acreditar (Zuckerman, 2017), minando a confiança nas instituições públicas (Rivas-de-Roca, Morais & Jerónimo, 2022). Um desafio acrescido com alguns dos desenvolvimentos da IA, que vão resultar na disseminação de *deepfakes,* as imagens, vídeos ou áudios criados a partir de manipulação com altas doses de realismo que os fazem parecer autênticos (Souza & Santaella, 2021).

As informações falsas ou fraudulentas, que surgem em vários formatos, incluindo memes, são muitas vezes feitas à medida para se tornarem virais nas redes sociais, chegando, em regra, a muito mais pessoas do que as notícias publicadas pelos media – baseadas em polémicas, fatos curiosos ou absurdos, fake news têm 70% mais chances de viralizar que as notícias (Vosoughi, Roy & Aral, 2018; Batista & Gradim, 2020). Ao mesmo tempo, as redes sociais podem ser um espaço privilegiado para a disseminação da

desinformação devido ao seu formato que apresenta ao utilizador pequenas informações descontextualizadas num *newsfeed* contínuo, dificultando o julgamento da veracidade de um artigo; ao seu contexto e à mentalidade que produz, muito mais focada na dimensão social da partilha do que no discernimento da exatidão; e às suas "arquiteturas de escolha" que orientam e estimulam o comportamento do utilizador (Allcott & Gentzkow, 2017; Epstein et al., 2023; Kozyreva, Lewandowsky, & Hertwig, 2020).

Além disso, muitos dos filtros utilizados na espécie de curadoria do conteúdo através de algoritmos em plataformas como o Facebook estão baseados em processos de filtragem em que a própria rede dá visibilidade àquilo que considera importante, enquanto despreza aquilo que não é social (Recuero, Zago & Soares, 2017). A combinação entre essa filtragem social e a combinação com a ação dos algoritmos pode levar à personalização em excesso, processo que "é, em sua maior parte, invisível para os usuários" (Recuero, Zago & Soares, 2017, p. 7), e potencia a disseminação da desinformação por se basear em laços fortes ou fracos de interações em grupos sociais já estabelecidos fora do mundo online. Se a personalização for excessiva, poderá nos impedir de entrar em contato com experiências e ideias divergentes às nossas crenças, que poderão destruir preconceitos e mudar o modo como pensamos sobre o mundo (Pariser, 2012).

Outro aspecto que precisa ser levado em consideração é que a ecologia contemporânea das mídias digitais é "um espaço fértil para a ascensão do conteúdo de mídia direcionados e contextos de notícias (...) que provocam reações afetivas" (Bakir & McStay, 2018, p. 159). Neste contexto, *fake news* e desinformação funcionam como *clickbait*, ou seja, conteúdos criados para atrair cliques sem preocupação ética ou com a verdade, a partir de técnicas de sensacionalismo e imagens fortes. O atual fenómeno das *fake news*, aliás, permite a ascensão de "uma nova economia política assente no *clickbait*" (Monsees, 2023, p.154).

Estudos mostram que a exposição dos usuários a conteúdos emocionais, tanto positivos quanto negativos, os torna mais ativos e engajados na plataforma (Bruno, Bentes & Faltay, 2019). Como o objetivo das plataformas

de rede social é capturar e mobilizar a atenção dos usuários para que eles passem o máximo de tempo possível conectados, "as estratégias deste mercado se voltam para desenvolver mecanismos persuasivos de captura da atenção, nos quais o agenciamento algorítmico exerce um papel central" (Bentes, 2019, p. 222). Mais engajamento nas redes sociais significa mais tempo gasto nas plataformas, o que reverte em lucro para seus donos e para os disseminadores de desinformação. A lógica é, portanto, mais emoções, mais atenção, mais tráfego, mais tempo gasto nas redes sociais, mais dados coletados.

Se os algoritmos das redes sociais priorizam interações e engajamento, as *fake news* são, na verdade, um produto lucrativo deste modelo de negócio que sustenta as redes sociais, uma vez que recebem mais reações e despertam emoções. São o produto máximo do modelo de negócios da rede social e o resultado da engrenagem do capitalismo de vigilância. São também produto da personalização da informação e de critérios que não se baseiam na veracidade das informações para seleção e distribuição de informação.

Observa-se, portanto, que se desinformação e *fake news* circulam sem correção em comunidades fechadas, se as pessoas passam a desacreditar fatos verdadeiros divulgados pelos meios de comunicação social, e se são afetivas e inflamatórias, estamos indo na direção oposta da esfera pública de Habermas (1984), que pressupunha consenso através de um debate racional, após ouvir pontos de vista diferentes. Se o debate público é baseado em emoções, em separação entre ganhadores e perdedores, o resultado lógico é a polarização da sociedade, queda na legitimidade de governos e decisões democráticas equivocadas baseadas em desinformação (Bakir & McStay, 2018).

Se os processos algorítmicos que organizam e dão sentido à vastidão de informação da internet vão resultar em vigilância massiva, personalização extrema e favorecer a criação de bolhas onde informações enganosas também vão circular, a polarização e radicalização são os desfechos mais previsíveis para este cenário. É também difícil perceber, dentro dos "jardins

murados" da internet (Paterson, 2012), como será possível exercitar e desenvolver uma verdadeira participação democrática para construir uma sociedade mais justa e inclusiva.

IA e racismo algorítmico

Uma outra questão que nos interessa nos debates sobre IA diz respeito à relação entre racialidades e tecnologias digitais. Diferentes investigações têm evidenciado que aplicações como reconhecimento facial, escores de crédito e filtros para selfies, dentre outras que usam sistemas de IA, apresentam um conjunto de implicações em grupos minorizados racialmente.

Para a compreensão dessa relação entre racialidades e tecnologias digitais, um conceito relevante é o de racismo algorítmico (Silva, 2023), sustentado numa dupla opacidade: por um lado, a ideia de neutralidade tecnológica; e, por outro, a invisibilidade ou negação do racismo enquanto uma categoria que estrutura as dinâmicas sociais.

> Uso o termo "racismo algorítmico" para explicar como tecnologias e imaginários sociotécnicos em um mundo moldado pelo privilégio branco fortalecem a ordenação racializada de conhecimentos, recursos, espaço e violência em detrimento de grupos não brancos. Então, muito além dos detalhes das linhas de programação, falamos aqui da promoção e implementação acríticas de tecnologias digitais que favorecem a reprodução dos desenhos de poder e opressão que já estão em vigor (Silva, 2023, s/p.).

Numa espécie de linha do tempo, que é constantemente atualizada, Silva (2023a) indica uma série manifestações do racismo algorítmico em sistemas de visão computacional, a exemplo de aplicações de análise de expressões faciais que associam emoções negativas a pessoas negras, ferramentas de branqueamento de pele que sugerem deixar as selfies "mais bonitas", softwares que confundem cabelos crespos com perucas.

Outros estudos têm apontado como modelos de IA generativa também expressam o racismo algorítmico. A título de exemplo, uma pesquisa realizada pela *Bloomberg Technology* (Nicoletti & Bass, 2023) sobre a geração de imagens de trabalhadores para diferentes tipos de empregos nos Estados Unidos da América, classificados por "alta remuneração" ou "baixa remuneração", produziu imagens de pessoas com tons de pele mais escuros 70% para a palavra-chave "trabalhador de fast-food", quando 70% dos trabalhadores de fast-food no país sejam brancos.

Um outro caso sobre vieses raciais e IA generativa aconteceu no Brasil. Ao utilizar a ferramenta de IA generativa da Microsoft, de *trend* da Disney Pixar, a deputada estadual do Rio de Janeiro, Renata Souza, teve como resultado uma ilustração de uma mulher negra segurando uma arma em uma favela (FIG 1). Vale ressaltar que, ao fazer a sua autodescrição para a geração da imagem, a deputada não fez qualquer menção a armas. Em uma postagem na internet, a parlamentar disse que a descrição pedida era de "uma mulher negra, de cabelos afro, com roupas de estampa africana num cenário de favela. E essa foi a imagem gerada. O que leva essa 'desinteligência artificial' a associar o meu corpo, a minha identidade, com uma arma?" (Motta, 2023, s/p.).

Figura 1. Imagem gerada por IA da Microsoft (esq.) e foto de Renata Souza (dir.)

Fonte: perfil da deputada no X.

Ainda que os exemplos citados até aqui se refiram a ferramentas utilizadas por empresas privadas, a questão do racismo algorítmico é ainda mais grave pelo fato de que, cada vez mais, governos e instituições públicas estão adotando tecnologias baseadas em IA.

Uma área em que é possível verificar a massificação deste uso é na segurança pública, através da adoção de sistemas de reconhecimento facial para vigilância nos espaços públicos. No Reino Unido, um relatório elaborado pela Universidade de Essex identificou uma taxa de 81% de erro no reconhecimento facial utilizado pela Polícia Metropolitana de Londres (Fussey & Murray, 2019), sendo a maior parte contra pessoas negras e migrantes. No Brasil, 80% das prisões errôneas por reconhecimento facial na cidade do Rio de Janeiro são de pessoas negras (Sampaio, 2022). O caso mais recente aconteceu em janeiro de 2024, quando uma mulher passou um dia presa após ter sido erroneamente identificada, pelo sistema de reconhecimento facial instalado na praia de Copacabana, como suspeita dos crimes de roubo e formação de quadrilha (Sousa, 2024).

Ainda sobre o Brasil, um edital da Prefeitura de São Paulo, que visava a contratação de empresa para instalação de 20 mil câmaras com reconhecimento facial na cidade, mencionava que a tecnologia deveria ser utilizada para rastrear "pessoas suspeitas, monitorando todos os movimentos e atividades", que a identificação deveria ser feita por "tipos de características como cor, face e outras" e que o monitoramento deveria apontar situações de "vadiagem" e "tempo de permanência" como comportamentos suspeitos.

Este exemplo da Prefeitura de São Paulo é emblemático do processo de naturalização do uso acrítico, pelo poder público no Brasil, das tecnologias digitais para vigilância e segregação (Melo & Serra, 2022). Os nomes de alguns programas para a segurança pública, como Muralha Digital (Curitiba), Cercamento eletrônico da cidade (Aracaju), City Câmeras (São Paulo), são indicadores desse processo.

As potenciais implicações da IA nas relações raciais têm sido objeto de preocupação de diferentes órgãos e instituições internacionais. Vale registar,

neste sentido, o relatório assinado por E. Tendaiy Achiume (2020), Relatora Especial da Organização das Nações Unidas sobre formas contemporâneas de racismo, discriminação racial, xenofobia e intolerância correlata. Ao afirmar que "o coração da questão é político, social e econômico, não apenas um problema tecnológico ou matemático", Achiume (2020, p. 5) indica que os Estados devem estabelecer comprometimento legal para realizar um amplo escrutínio dos potenciais discriminatórios contra minorias raciais ou étnicas.

Algumas das recomendações de caráter estrutural presentes no documento são: o banimento de tecnologias que tenham impacto racial discriminatório significativo, como o reconhecimento facial; a inclusão e representação das identidades raciais em todos os níveis do setor de tecnologia; e o protagonismo das pessoas diretamente afetadas na resolução dos problemas.

Considerações finais

A IA e os algoritmos são instrumentos poderosos que influenciam o quotidiano de milhões de utilizadores em todo o mundo e de todas as instituições. A ação destas tecnologias, a favor do capitalismo de vigilância, precisa ser melhor escrutinada por autoridades, investigadores, *media* e outros atores para que seja possível avançar em uma visão crítica sobre as questões que levanta e um debate profundo na esfera pública, especialmente no que concerne aos problemas éticos.

Frente à adoção acelerada de IA por governos, um desafio essencial relativamente ao racismo algorítmico é a regulação, que deve envolver a criação de legislações, políticas e diretrizes que estabeleçam padrões éticos, definam responsabilidades, promovam a transparência e protejam os direitos humanos. Nesta perspectiva, Silva e Silva (2023) propõem a formulação de "lentes antirracistas" sobre regulação de IA, considerando que "o combate ao racismo e ao aprofundamento de suas violências através das tecnologias digitais (...) ainda é abordado de forma secundária por formuladores e entidades ligadas a políticas públicas" (p. 11).

Dentre as possibilidades que contribuam para a superação do racismo algorítmico na IA, vale citar: o banimento de tecnologias com evidências de discriminação racializada, como a vigilância biométrica em massa; a previsão de análises de impacto racial no desenvolvimento de sistemas tecnológicos, com participação de avaliadores independentes e de representações das comunidades potencialmente impactadas; e políticas afirmativas de inclusão de grupos sub-representados na, idealização, desenvolvimento, análise, supervisão e implementação de tecnologias digitais emergentes.

Além disso, entendemos que os *media* devem assumir um papel ativo nesse debate, como forma de retomar a sua função de mediadores dos problemas da sociedade. Os *media*, também afetados pelos sistemas algorítmicos, têm a função primordial de desconstruir a realidade, explicar o funcionamento e ação destes atores "invisíveis" e alertar o público para as consequências sobre o seu uso antidemocrático. É preciso estar alerta, no entanto, para o risco de cair em uma espécie de "pânico moral" sobre questões tão caras à sociedade, como acaba por acontecer com o jornalismo e a desinformação (Carlson, 2020), podendo levar não só à descrença no que é falso mas também nas fontes fiáveis de informação como as jornalísticas (Van der Meer et al., 2023). Fortalecer os projetos de fact-checking, que não têm o efeito de reduzir as percepções da credibilidade das notícias reais, e promover a literacia mediática desde os bancos das escolas, mas também para grupos de adultos e idosos, parecem ser as soluções mais adequadas para engendrar a cidadania (Bulger & Davison, 2018).

O rápido avanço tecnológico não pode ser travado, por isso a implementação da IA não pode ser deixada de lado, é necessário delinear um caminho de adaptação a esta nova realidade, em que a tecnologia seja apropriada para servir os valores da democracia e do serviço público. Neste cenário, os cidadãos têm também de ser informados, e conhecer os meandros dos processos algorítmicos para resistir, ao invés de deixar-se manipular, cumprindo o papel de agente consciente das suas próprias decisões. Já as instituições e os governos devem encarregar-se de melhorar a transparência dos critérios algorítmicos e delinear estratégias de combate à ação nociva das

plataformas digitais, enquanto principais impulsionadoras de desinformação, *fake news* e *deep fakes*.

Assim, defendemos que a reflexão e debate público sobre estas questões deve unir especialistas de diferentes áreas do conhecimento, decisores políticos, os *media* e os cidadãos. Acreditamos que ao fomentar essa discussão, estamos a dar o primeiro passo para a concretização de uma estratégia democrática do uso de IA, assente em quatro eixos de atuação: a academia, os cidadãos, os *media* e os governos.

Referências bibliográficas

Referências bibliográficas

Achiume, E. T. (2020). *Racial Discrimination and Emerging Digital Techonologies: A Human Rights Analysys. United Nations General Asssembly.* https://www.ohchr.org/sites/default/files/HRBodies/HRC/RegularSessions/Session44/Documents/A_HRC_44_57_AdvanceEditedVersion.docx

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Anderson, C., Bell, E., & Shirky, C. (2013). Jornalismo pós-industrial: adaptação aos novos tempos. *Revista de Jornalismo ESPM*, *5*(3), 30-89. https://bit.ly/3NDHolM

Baack, S. (2016). What big data leaks tell us about the future of journalism--and its past. *Internet Policy Review.* https://policyreview.info/articles/news/what-big-data-leaks-tell-us-about-future-journalism-and-its-past/413

Bakir, V., & McStay, A. (2018). Fake News and The Economy of Emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

Baptista, J. P., Gradim, A. (2020). Understanding fake news consumption: A review. *Social Sciences, 9*(10), 185. https://doi.org/10.3390/socsci9100185

Beckett, C. (2019). *New Powers, New Responsibilities: A Global Survey of Journalism and Artificial Intelligence.* The London School of Economics and Political Science. http://blogs.lse.ac.uk/polis/2019/11/18/new-powers-new-responsibilities

Beckett, C. & Yaseen, M. (2023). *Generating Change. A global survey of what news organisations are doing with AI.* https://shre.ink/ruQw

Bentes, I. (2015) *Mídia-Multidão: estéticas da comunicação e biopolíticas.* Mauad X.

Biswal, S. & Kulkarni, A. (2024). *Exploring the Intersection of Artificial Intelligence and Journalism: The Emergence of a New Journalistic Paradigm*. Routledge.

Bruno, F. (2008) Monitoramento, classificação e controle nos dispositivos de vigilância digital. *Revista Famecos: mídia, cultura e tecnologia, 36*(2), 10-16. https://doi.org/10.15448/1980-3729.2008.36.4410

Bruno, F.; Bentes, A. C.; Faltay, P. (2019). Economia psíquica dos algoritmos e laboratório de plataforma: mercado, ciência e modulação do comportamento. *Revista FAMECOS, 26*(3), 33095. https://doi.org/10.15448/1980-3729.2019.3.33095

Bulger, M., & Davison, P. (2018). The Promises, Challenges, and Futures of Media Literacy. *Journal of Media Literacy Education*, *10*(1), 1–21. https://doi.org/10.23860/JMLE-2018-10-1-1

Canavilhas, J. (2023). Produção automática de texto jornalístico com IA: contributo para uma história. *Textual & Visual Media, 17*(2), 22-40. https://doi.org/10.56418/txt.17.1.2023.2

Carlson, M. (2020). Fake news as an informational moral panic: the symbolic deviancy of social media during the 2016 US presidential election. *Information Communication and Society, 23*(3), 374–388. https://doi.org/10.1080/1369118X.2018.1505934

Carreira, K. (2017). *Notícias automatizadas: A evolução que levou o jornalismo a ser feito por não humanos.* [Disseração de mestrado, Universidade Metodista de São Paulo]. http://tede.metodista.br/jspui/handle/tede/1671

Diakopoulos, N. (2019). *Automating the news: How algorithms are rewriting the media.* Harvard University Press.

Epstein, Z.; Sirlin, N.; Arechar, A.; Pennycook, G.; Rand, D. (2023). The social media context interferes with truth discernment. *Science Advances*, *9*, eabo6169(2023). DOI:10.1126/sciadv.abo6169

European Comission (2018). *A definition of AI: Main capabilities and scientific disciplines.* https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

Fussey, P.; Murray, D. (2019). *Independent report on the London Metropolitan Police Service's trial of live facial recognition technology.* Human Rights Centre – University of Essex. https://repository.essex.ac.uk/24946/1/London-Met-Police-Trial-of-Facial-Recognition-Tech-Report-2.pdf

García-Orosa, B., Canavilhas, J., & Herrero, J. V. (2023). Algoritmos y comunicación: Revisión sistematizada de la literatura. *Comunicar: Revista científica iberoamericana de comunicación y educación,* (74), 9-21. https://doi.org/10.3916/C74-2023-01

Gillespie, T. (2014). T*he relevance of algorithms.* In Gillespie, T., Boczkowski, P. J. & Foot, K. (Eds.) *Media technologies: Essays on communication, materiality, and society.* MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Google Trends (2024). *AI.* https://trends.google.com/trends/explore?date=all&q=AI&hl=pt-PT

Habermas, J. (1984) *Mudança estrutural da esfera pública: investigações quanto a uma categoria da sociedade burguesa.* Tempo Brasileiro.

Han, B. (2021). *Não-Coisas. Transformações no mundo em que vivemos.* Relógio D'Água.

Helberger, N., van Drunen, M., Moeller, J., Vrijenhoek, S., & Eskens, S. (2022). Towards a normative perspective on journalistic AI: Embracing the messy reality of normative ideals. *Digital Journalism, 10*(10), 1605-1626. https://doi.org/10.1080/21670811.2022.2152195

Heinderyckx, F., Vos, T. P. (2016) Reformed gatekeeping. *Communication and Media, XI*(36), 29–46, 2016. DOI: 10.5937/comman11-10306.

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools. *Psychological Science in the Public Interest, 21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Kusters, R., Misevic, D., Berry, H., Cully, A., Le Cunff, Y., Dandoy, L., Díaz-Rodríguez, N., Ficher, M., Grizou, J., Othmani, A., Palpanas, T., Komorowski, M., Loiseau, P., Frier, C., Nanini, S, Quercia, D., Sebag, M., Fogelman, F., Taleb, S., Tupikina, L, Sahu, V., Vie, J. & Wehbi, F. (2020). Interdisciplinary research in artificial intelligence: challenges and opportunities. *Frontiers in big data*, *3*, 577974. https://doi.org/10.3389/fdata.2020.577974

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science, 359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lemos, A. (2021). Dataficação da vida. *Civitas-Revista de Ciências Sociais, 21*, 193-202. https://doi.org/10.15448/1984-7289.2021.2.39638

Lyon, D. (2017) Surveillance Culture: Engagement, Exposure, and Ethics in Digital Modernity. *International Journal of Communication, 11*(2017), 1–18.

Marshall, D. P. (2009) New Media as transformed media industry. In Holt, J, & Perren, A. *Media industries: history, theory and method.* Wily-Blackwell.

Marwick, A., & Lewis, R. (2017). *Media Manipulation and Disinformation Online.* Data & Society.

Melo, P.; Serra, P. (2022). Tecnologia de Reconhecimento Facial e Segurança Pública nas capitais brasileiras: apontamentos e problematizações. *Comunicação e Sociedade*, *42*, 205–220. https://revistacomsoc.pt/index.php/revistacomsoc/article/download/3984/4789/19902

Milosavljević, M. & Vobič, I. (2019). Human Still in the Loop: Editors reconsider ideals of professional journalism through automation. *Digital Journalism, 7*(8), 1098-1116. Doi: 10.1080/21670811.2019.1601576

Monsees, L. (2023). Information disorder, fake news and the future of democracy. *Globalizations*, *20*(1), 153–168. https://doi.org/10.1080/14747731.2021.1927470

Motta, J. (2023, 26 de outubro). Deputada denuncia racismo em trend da Disney Pixar: desinteligência artificial. *Fórum*. https://revistaforum.com.br/politica/2023/10/26/deputada-denuncia-racismo-em-trend-da-disney-pixar-desinteligncia-artificial-146653.html

Musso, P. (2004). A filosofia da rede. In Parente, A. (Org.). *Tramas da Rede.* (pp. 17-38) Sulinas.

Napoli, P. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication theory 24*(3), 340-360. https://doi.org/10.1111/comt.12039

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., & Nielsen, R. K. (2023). *Digital News Report 2023*. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023

Nicoletti, L.; Bass, D. (2023). *Humans are biased. Generative IA is even worse*. Bloomberg Technology. https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Oliveira, A. (2019). *Inteligência Artificial.* Fundação Francisco Manuel dos Santos.

Pariser, E. (2012). *O filtro invisível: o que a internet está escondendo de você.* Zahar.

Pasquale, F. (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information.* Harvard University Press.

Paterson, N. (2012). Walled gardens: The new shape of the public internet. *ACM International Conference Proceeding Series*, 97–104. https://doi.org/10.1145/2132176.2132189

Pocino, P. (2022). *Algorithms in the newsrooms: Challenges and recommendations for artificial intelligence with the ethical values of journalism*. Catalan Press Council (CIC). https://t.ly/6XLa

Quandt, T. (2018). Dark participation. *Media and Communication, 6*(4), 36–48. https://doi.org/10.17645/mac.v6i4.1519

Quandt, N., Sant'Anna, R., Winques, K. & Máximo, M. (2021). Análise de apurações jornalísticas feitas com uso de Inteligência Artificial. *Redes-Revista Interdisciplinar do IELUSC*, (4), 39-52. https://bit.ly/3GnGdo7

Recuero, R., Zago, G., & Soares, F. B. (2017). Mídia social e filtro-bolha nas conversações políticas no twitter. *XXVI Encontro Anual da COMPÓS*, 1–27.

Rieder, B. (2018). Examinando uma técnica algorítmica: o classificador de bayes como uma leitura interessada da realidade. *Revista Parágrafo, 6* (1), 123-142. http://revistaseletronicas.fiamfaam.br/index.php/recicofi/article/view/726

Rinehart, A. & Kung, E. (2022). *Artificial Intelligence in Local News. A survey of US newsrooms' AI readiness.* The Associated Press. https://www.ap.org/assets/files/ap_local_news_ai_report_march_2022.pdf

Risi, E., & Pronzato, R. (2022) Algorithmic Prosumers. In Armano, E., Briziarelli, M., & Risi, E. (eds.), *Digital Platforms and Algorithmic Subjectivities* (pp. 149–165). University of Westminster Press. https://doi.org/10.16997/book54.l

Rivas-de-Roca, R., Morais, R., & Jerónimo, P. (2022). Comunicación y desinformación en elecciones: tendencias de investigación en España y Portugal. *Universitas*, *36*, 71–94. https://doi.org/10.17163/uni.n36.2022.03

Rosa, A. M. (2007). Nota sobre o processo de exteriorização da técnica: o lugar da interacção homem-computador. *Comunicação e Sociedade*, *12*, 39–49.

Sampaio, F. (2022, 12 de janeiro). 80% das prisões errôneas por reconhecimento facial no Rio de Janeiro são de negros. *Agência Brasil*. https://agenciabrasil.ebc.com.br/radioagencia-nacional/justica/audio/2022-01/80-das-prisoes-erroneas-por-reconhecimento-facial-no-rj-sao-de-negros

Sanin, C. (2023). Artificial Intelligence: Current Perspectives and Alternative Paths. *TecnoLógicas*, *26*(57), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Schwab, K. (2018). The Fourth Industrial Revolution. Encyclopædia Britannica. https://www.britannica.com/topic/The-Fourth-Industrial-Revolution-2119734

Serra, J. P. (1998). *A informação como utopia*. Universidade da Beira Interior.

Silva, T. (2023). *O racismo algorítmico é uma atualização do racismo estrutural*. Entrevista. Fiocruz. https://cee.fiocruz.br/?q=Tarcizio-Silva-O-racismo-algoritmico-e-uma-especie-de-atualizacao-do-racismo-estrutural

Silva, T. (2023a). *Linha do tempo do racismo algorítmico*: *casos, dados e reações.* https://tarciziosilva.com.br/blog/destaques/posts/racismo-algoritmico-linha-do-tempo/

Silva, T.; Silva, F. dos S. R. (orgs.). (2023). *Lentes Antirracistas sobre Regulação de Inteligência Artificial*. Desvelar. https://desvelar.org/2023/12/12/lentes-antirracistassobre-regulacao-de-inteligencia-artificial.

Sousa, A. (2024, 4 de janeiro). Mulher é solta após ser detida por erro no sistema de reconhecimento facial no Rio. *Folha de S. Paulo*. https://www1.folha.uol.com.br/cotidiano/2024/01/mulher-e-solta-apos-ser-detida-por-erro-no-sistema-de-reconhecimento-facial-no-rio.shtml

Souza, C., & Santaella, L. (2021). Deepfakes na perspectiva da semiótica. *TECCOGS: Revista Digital de Tecnologias Cognitivas,* (23). https://doi.org/10.23925/1984-3585.2021i23p26-44

Tandoc Jr., E. C., Lim, Z. W., & Ling, R. (2018) Defining "Fake News". *Digital Journalism, 6*(2), 137-153. DOI: 10.1080/21670811.2017.1360143.

Thurman, N., Dörr, K., & Kunert, J. (2017). When reporters get hands-on with robo-writing: Professionals consider automated journalism's capabilities and consequences. *Digital Journalism, 5(10), 1240-1259.* https://doi.org/10.1080/21670811.2017.1345318

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 441–460. https://academic.oup.com/mind/article/LIX/236/433/986238

van der Meer, T. G. L. A., Hameleers, M., & Ohme, J. (2023). Can Fighting Misinformation Have a Negative Spillover Effect? How Warnings for the Threat of Misinformation Can Decrease General News Credibility. *Journalism Studies*, *24*(6), 803–823. https://doi.org/10.1080/1461670X.2023.2187652

Vicente, P., & Dias-Trindade, S. (2021). Reframing sociotechnical imaginaries: The case of the Fourth Industrial Revolution. *Public Understanding of Science, 30*(6), 708-723. https://doi.org/10.1177/09636625211013513

Vicente, P. (2023). *Os Algoritmos e Nós*. Fundação Francisco Manuel dos Santos.

Vosoughi, S.; Roy, D.; Aral, S. (2018) The spread of true and false news online. *Science, 359*(6380), 1146-1151. Doi: 10.1126/science.aap9559.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence, 10*(2), 1-37. Doi: 10.2478/jagi-2019-0002

Wardle, C., & Derakhshan, H. (2018). Thinking about 'Information Disorder': formats of misinformation, disinformation and mal-information. In Ireton, C., & Posetti, J. (Eds.) *Journalism, fake news & disinformation: handbook for journalism education and training.* Unesco.

Wölker, A. & Powell, T. (2021), Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism, 22,* 86–103. Doi: 10.1177/1464884918757072

Wolton, D. (2010). *Informar não é comunicar.* Sulina.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: the fight for a human future at the new frontier of power.* PublicAffairs.

Zuckerman, E. (2017). Fake news is a red herring. *Deutsche Welle.* https://www.dw.com/en/fake-news-is-a-red-herring/a-37269377

COMO (E POR QUE) OS JORNALISTAS DEVEM INVESTIGAR ALGORITMOS QUE TOMAM DECISÕES AUTOMATIZADAS NOS SERVIÇOS PÚBLICOS?

Krishma Carreira
/ Universidade Metodista de São Paulo

Introdução

Todos os dias, bilhões de pessoas são afetadas no mundo todo em função de decisões tomadas automaticamente por algoritmos, muitas vezes sem o conhecimento de que eles foram os responsáveis. E mesmo quando sabem, quase sempre não fazem ideia de como foi o processo decisório, por causa da falta de transparência da operação algorítmica (Introna, 2016). Estas escolhas algorítmicas, onipresentes e pervasivas, que ocorrem em muitos casos sem supervisão humana, determinam os resultados de pesquisas nos sistemas de buscas, as notícias que recebemos, as vagas de emprego visualizadas, até a concessão (ou não) de um benefício social ou até mesmo de empréstimo.

Muitos Sistemas de Decisões Automatizadas (SDAs) são usados na administração pública, configurando uma categoria de serviços públicos algorítmicos (Kaun, 2020). Eles são justificados, assim como no mundo privado, pela busca de eficiência, agilidade, aumento de produtividade, otimização, controle e redução de custos e pela promessa de escolhas mais justas e com o potencial de liberarem os funcionários da execução de tarefas monótonas e repetitivas.

Os SDAs são sistemas complexos, que envolvem múltiplos atores (humanos ou não), diversas tecnologias (das mais simples a técnicas de inteligência artificial, como machine learning, deep learning, visão computacional, IA generativa[1]), e são fruto de escolhas humanas, corporativas e governamentais, que carregam subjetividades e ideologias (O'Oneil, 2020).

Suas decisões podem trazer inúmeros benefícios para a sociedade, mas são também capazes de gerar erros e impactos negativos ao exercício de direitos (privacidade, liberdade de expressão, acesso à justiça, etc.), de impor barreiras a serviços e direitos, de refletir relações sociais opressivas, racismo (Noble, 2018), sexismo (Carpenter, 2015) e outras formas de preconceito e discriminação.

Cabe destacar que existe uma grande assimetria de poder entre quem é impactado por uma decisão total ou parcialmente tomada por sistemas automatizados e quem os desenvolve ou os implementa, reforçada principalmente pela opacidade do processo.

Por tudo isso, os Sistemas de Decisões Automatizadas impactam os direitos humanos e configuram um grande tema de interesse público, que deve (e pode) ser objeto de investigação e monitoramento crítico por parte de jornalistas, desvendando interesses ocultos, erros, danos e vieses e possibilitando controle público e responsabilização (*accountability*). Não é uma tarefa fácil, devido principalmente à opacidade e complexidade dos sistemas, mas também pelas dificuldades enfrentadas pela mídia em geral. Mas é urgente!

Parte deste trabalho pode ser realizado com técnicas tradicionais do jornalismo investigativo, utilizando ou não processos, recursos e ferramentas do Jornalismo de Dados e do Jornalismo Computacional (também estruturado a partir de dados, mas inclui programação). Entretanto, devido à complexidade destes sistemas, a constituição de redes de profissionais com

1. O relatório do Capgemini Research Institute indica que, para 96% das 800 organizações ouvidas, a IA generativa é um tópico de discussão em reuniões, sendo que 59% de seus líderes são fortes defensores da tecnologia. Para 83% das organizações, os chatbots são a aplicação mais relevante. Outros 75% dos executivos citam que os aplicativos de dados podem ser usados de forma eficaz em suas organizações. Com base neste trabalho, é possível imaginar que o uso de IA generativa nos serviços públicos também deve crescer.

competências diversas (jornalistas, cientistas de dados, cientistas da computação, etc.) será cada vez mais necessária e pode ser, em determinados momentos, a única forma de concluir uma investigação.

Um caminho investigativo

Com o objetivo de auxiliar o trabalho de jornalistas interessados no tema, a autora desenvolveu o método RISDA (Reportagens Investigativas sobre Sistemas de Decisões Automatizadas), que será apresentado ao longo do artigo. Ele foi baseado em estudos sobre trabalhos como Diakopoulos (2014); Diakopoulos e Koliska (2016); Trielli e Diakopoulos (2020); Gray e Bounegru (2019); Marconi, Daldrup, Pant (2019) e produções da Transparência Algorítmica, iniciativa brasileira que faz parte da organização Transparência Brasil.

Com a RISDA, procura-se despertar nos jornalistas a possibilidade de investigar não só as ações humanas, mas os artefatos tecnológicos dos sistemas sociotécnicos que tomam decisões automatizadas que impactam vidas através de serviços públicos.

Neste trabalho entende-se jornalismo investigativo como aquele pautado na busca por revelar irregularidades, uma tarefa que resulta do trabalho original dos repórteres e não da investigação de autoridades. O jornalista investigativo revela algo desconhecido, que até então estava oculto, trabalhando com temas que interessam à opinião pública, produzindo uma matéria original, que não é típica do jornalismo diário (Nascimento, 2016).

Ao denunciar irregularidades, o jornalismo investigativo estimula um conjunto de repercussões que são essenciais para a democracia (Silvio Waisbord apud Nascimento, 2016): deliberativos (formação de comissões ou constituição de audiências); individualizados (punições individuais) e substanciais (geram novas leis, regulamentações e alterações administrativas), segundo a classificação de Protess (apud Nascimento, 2016).

Mas para investigar os algoritmos é preciso, antes de mais nada, entender seus componentes (dados de entrada, cálculos, modelos e dados processados), sabendo que cada um deles pode conter ou gerar erros, vieses e problemas diversos (Trielli e Diakopoulos, 2020).

- Dados de entrada: o jornalista investigativo deve procurar saber quais são os dados de entrada de um determinado sistema. Um SDA pode tomar decisões erradas com base em dados inadequados para uma determinada solução, ou porque não são representativos ou são tendenciosos. Veja os dados de códigos postais. Eles podem levar a cálculos capazes de causar prejuízos para determinadas populações.

- Cálculos (instruções): podem ocorrer desde erros nos códigos a problemas com as inferências que podem levar a falsos positivos (exemplo: um SDA pode entender que uma pessoa de baixo risco é perigosa) e falsos negativos (exemplo: uma pessoa de alto risco pode ser considerada inofensiva).

- Dados processados (*output*): nesta etapa são possíveis diversos tipos de problemas, entre eles falta de utilidade (exemplo: ativação de ações policiais desnecessárias, gerando gastos e, no mínimo, constrangimento para os envolvidos).

- Modelos: ainda que os dados usados por um SDA sejam corretos, que as instruções sejam adequadas e que as decisões sejam de fato úteis, os modelos são capazes de gerar erros, impactos negativos, preconceitos e discriminações (exemplo: *softwares* de reconhecimento facial modelados a partir de categorias problemáticas de raça e gênero, podendo levar, inclusive à detenção de pessoas inocentes, como ocorreu em inúmeros lugares no mundo (Silva, 2021) e ao desperdício de dinheiro público (Nunes, 2019)).

Todo modelo de decisão envolve um processo de filtragem (o que deve ser incluído ou excluído), priorização, classificação e associação (Diakopoulos, 2016), Uma escolha, com ingredientes subjetivos, que, ainda hoje, é tomada por profissionais com reduzido grau de diversidade.

Hammond (2016) aponta cinco fontes de vieses que podem ser identificadas nas aplicações de Inteligência Artificial (IA): dados tendenciosos; viés de interação (podem reproduzir os comportamentos identificados nos usuários); tendência emergente (decisões tomadas de acordo com o perfil do usuário, com possibilidade de gerar um ciclo vicioso); preconceito de similaridade (também seguem a lógica do filtro bolha) e preconceito de objetos conflitantes (sistemas baseados no comportamento de clique que querem maximizar resultados e, por isso, têm potencial de reforçar estereótipos).

Em um artigo sobre vieses de aprendizado de máquina e suas implicações, a partir de um estudo de caso de reconhecimento facial, que pode ser extrapolado para outros modelos, Ruback, Ávila e Contero (2021) apontam quatro vieses,: histórico (etapa anterior à coleta de dados); de representação ou de amostra (fase de coleta de dados); de avaliação (o tipo de métrica também tem capacidade também de introduzir vieses) e de interpretação humana (feito em uma possível fase de checagem da decisão).

Existem, portanto, dois grandes tipos de riscos a direitos (Transparência Brasil, 2020): por natureza da ferramenta, com base no que ela foi desenhada para entregar, e por discriminação algorítimica, falta de representatividade nos dados.

Para investigar os algoritmos, Trielli e Diakopoulos (2020) apontam quatro caminhos possíveis: 1) acesso ao código para checar se teve algum problema na fase de cálculos e instruções do código, utilizando técnicas do Jornalismo Computacional; 2) acesso aos dados de entrada (*input)* e de saída (*output)* para possibilitar comparações, utilizando a técnica de engenharia reversa; 3) acesso aos dados de entrada (*input)* ou de saída (*output*), sendo que os procedimentos variam e que nalguns casos só os *outputs* são suficientes para a análise e a comparação com outros conjunto de dados; 4) acesso a informações complementares e contextuais, usando técnicas tradicionais do jornalismo, como entrevistas com desenvolvedores, autoridades, funcionários públicos, especialistas, pessoas afetadas, etc., e buscando provas e documentos. Esta última fase é essencial com ou sem acesso ao código e

aos dados, pois ela permite preencher as lacunas e entender, mesmo que parcialmente, as escolhas, regras e procedimentos dos SDAs.

O roteiro de investigação da RISDA

A reportagem investigativa pautada nos Sistemas de Decisões Automatizadas usados por órgãos públicos deve se concentrar em quatro eixos centrais: informações gerais sobre os SDAs, dados, erros, danos e vieses, e informações complementares. A seguir, são elencadas uma série de questões para serem pesquisadas, que podem ser ajustadas conforme a investigação. Estas instruções são um ponto de partida, que pode (e deve) ser ampliado.

Informações gerais:
- Qual é o nome do sistema, da empresa ou órgão público que desenvolveu ou adquiriu e implantou o SDA?
- O código é aberto?
- É possível acessar o código?
- Qual é a categoria do SDA (chatbots, etc.)?
- Qual é o modelo estatístico envolvido (processamento de linguagem natural, etc,)?
- Qual é o objetivo do SDA?
- Qual é o grau de apoio gerado pelo sistema (toma decisões, suporta ou sugere)?
- Existem métricas de avaliação?
- Quais são os resultados?
- Tem supervisão humana? Se sim, em que fases?

Dados:
- Quais são os tipos de dados de entrada?
- Qual é a fonte de origem dos dados?
- Usa dados pessoais?
- Se sim, os usuários sabem disso?
- Como o uso é comunicado?
- Houve consentimento de fato do usuário?
- Os usuários podem checar os dados usados?

- Podem retirá-los quando quiserem?
- Os dados podem ser compartilhados com terceiros? Se sim, como?
- São usados dados sigilosos? Se sim, quais são as aplicações de segurança?
- Dados sensíveis são utilizados para treinamentos?
- Qual é o grau de privacidade?
- Os dados são anônimos ou não?
- Onde os dados ficam armazenados?
- Qual é o nível de proteção contra possíveis ataques?
- Quais são os protocolos de segurança?

Erros, danos e vieses

- O erro identificado é único? Qual é o histórico de erros?
- Quais são os tipos de erros encontrados?
- Os nomes das vítimas de erros continuam em alguma base de dados, perpetuando o problema? Se sim, por quanto tempo?
- Quem são os prejudicados e beneficiados por erros?
- O que causa erros? E em que parte do processo eles ocorrem?
- Quando um erro é corrigido, a informação é usada para aprimorar (ou não) o sistema?
- O SDA investigado impacta direitos humanos fundamentais? Se sim, quais são os grupos afetados? Eles foram representados nas amostras de dados?
- Grupos particulares levam vantagens ou desvantagens com o SDA? Se sim, os erros e prejuízos afetam diferentes grupos da mesma forma?
- Há vieses? O órgão público buscou entender se foram considerados ou não no desenvolvimento, aquisição ou implementação?
- Em caso de identificação de vieses, eles foram corrigidos? Reduzidos? Como ocorreram?
- Foram feitos testes antes e durante a implementação dos SDAs? Quais? As taxas de erros encontradas são as mesmas para todos os grupos?
- As amostras de dados de treinamento representam diferentes grupos?
- O SDA contempla as características populacionais de onde está sendo implementado? Caso não, como a taxa de acurácia difere para cada público?
- Quem é responsável pelo SDA quando alguém é prejudicado?

Informações complementares:

- Os cidadãos impactados pelos SDAs sabem que as decisões foram tomadas ou suportadas por eles?
- As equipes de desenvolvimento dos sistemas são diversas? Em que nível?
- Eles respeitam as legislações, regulamentações e princípios éticos?
- Os SDAs implementados são proibidos em outras cidades ou países? Se sim, por quê?
- Eles são usados com propósito autoritário? Ajudam a monitorar pessoas e grupos minoritários?

Conclusões

Investigar os Sistemas de Decisões Automatizadas não é uma tarefa fácil. Muito pelo contrário. Muitos SDAs são formados por diversos e complexos algoritmos. Alguns usam *deep learning* (aprendizagem profunda), uma técnica inspirada na rede neural do cérebro humano, que, ao contrário do *machine learning* (aprendizagem de máquina), que é projetado para resolver uma questão específica, pode realizar várias tarefas complexas.

Mas é possível avançar em algumas investigações, mesmo que até certo ponto, incorporando técnicas tradicionais do Jornalismo com outras vinculadas ao Jornalismo de Dados e ao Jornalismo Computacional. A formação de redes com repórteres, estatísticos, programadores, cientistas de dados, entre outros profissionais com competências complementares necessárias, também pode possibilitar a execução de investigações sobre os Sistemas de Decisões Automatizadas.

Foi estimulada a investigação sobre os SDAs desenvolvidos e implementados pela administração pública, uma vez que eles impactam cidadãos em todo mundo, e são usados por órgãos que deveriam se pautar pelo princípio da transparência. Portanto, atendem ao critério jornalístico de interesse público. Mas os repórteres também podem se debruçar sobre as decisões tomadas de forma automatizada que são empregadas em diversas organizações privadas, uma vez que elas também podem reproduzir vieses,

discriminações, afetar os direitos humanos e cometer erros. A RISDA foi elaborada para servir como base de apoio para as investigações, oferecendo um norte. Mas assim como os sistemas investigados avançam a cada dia, ela também precisa incorporar, com o tempo, novas questões e métodos.

Referências bibliográficas

Carpenter, J. (2015, julho) *Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you. The Washington Post*. Disponível em: https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/

Carreira, K. (2021). Reportagens Investigativas sobre os Sistemas de Decisões Automatizadas. Tese. Universidade Metodista de São Paulo. São Bernardo do Campo. http://tede.metodista.br/jspui/handle/tede/2190

Diakopoulos, N. (2013). Algorithmic accountability reporting: on the investigation of black boxes. *Town Center for Digital Journalism*. http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf

Diakopoulos, N. (2014). Algorithmic accountability. Journalistic investigation of computational power structures. *Digital Journalis.* 2014, vol. 3(3), 398-415. https://doi.org/10.1080/21670811.2014.976411

Diakopoulos, N. (2014). Algorithmic accountability: on the investigation of black boxes. *Tow Center for Digital Journalism*. Columbia Journal School. Disponível em: https://academiccommons.columbia.edu/doi/10.7916/D8ZK5TW2

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59(2), 56-62*. https://dl.acm.org/doi/10.1145/2844110 Communications of the ACM.2016, Vol. 59 No. 2, Pág. 56-62. Disponível em: https://cacm.acm.org/magazines/2016/2/197421-accountability-in-algorithmic-decision-making/fulltext. Acesso em: 26 jun. 2018.

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital journalism, 5(7), 809-828.* https://doi.org/10.1080/21670811.2016.1208053*Digital journalism,*2016.

Diakopoulos, N. (2021). 31. The Algorithms Beat: Angles and Methods for Investigation. In L., Bounegru & J. Gray (eds), The Data Journalism Handbook Towards a Critical Data Practice (pp.219-229). https://library.oapen.org/bitstream/handle/20.500.12657/47509/1/9789048542079.pdf#page=222*The Algorithms Beat.*2018. Disponível em:

Bounegru, L., Gray, J., & Chambers, L. (2014). Manual de Jornalismo de Dados. https://kclpure.kcl.ac.uk/portal/en/publications/manual-de-jornalismo-de-dados *Manual de Jornalismo de Dados,* 2014. Disponível em: http://datajournalismhandbook.org/pt/. Acesso em: 27 mai. 2018.

The algorithms beat: angles and methods for investigation. In:*The Data Journalism Handbook 2*: towards a critical data practice Disponível em: https://datajournalism.com/read/handbook/two/investigating-data-platforms-and-algorithms/the-algorithms-beat-angles-and-methods-for-investigation Acesso: 23 mai, 2019.

Hammond, K. (2016, dezembro 10). *5 unexpected sources of bias in Artificial Intelligence.* Blog *Narrative Science*. http://resources.narrativescience.com/h/i/312890970-5-unexpected-sources-of-bias-inartificial-intelligence.

Introna, L. (2016). Algorithms, governance, and governmentality: On governing academic writing. *Science, Technology, & Human Values, 41(1), 17-49*. https://doi.org/10.1177/0162243915587360 *Algorithms, governance, and governmentality: on governing academic writing.* 2016.*Science, Technology & Human Values.*Vol.41. Pág. 17-49. 2016. Disponível em: https://journals.sagepub.com/doi/10.1177/0162243915587360. Acesso: 13 jan. 2016.

Kaun, A. (2022). Suing the algorithm: The mundanization of automated decision-making in public services through litigation. *Information, Communication & Society, 25(14), 2046-2062*. https://doi.org/10.1080/1369118X.2021.1924827 *Suing the algorithm: the mundanization of automated decision-making in public services through liti-*

gation. 2021. *Information Communication & Society*Disponível em: https://www.tandfonline.com/doi/full/10.1080/1369118X.2021.1924827. Acesso: 20 out. 2021.

Marconi, F., Daldrup, T. & Pant, R. (2019, fevereiro 14). *Acing the algorithmic beat, journalism new frontier. Nieman Lab.* https://www.niemanlab.org/2019/02/acing-the-algorithmic-beat-journalisms-next-frontier/

Nascimento, S. (2016). *Os novos escribas: o fenômeno do jornalismo sobre investigações no Brasil.* Porto Alegre: Arquipélago Editorial.

Noble, S. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: New York University Press.

Nunes, P. (2019). Novas ferramentas, velhas práticas: reconhecimento facial e policiamento no Brasil Retratos da violência: cinco meses de monitoramento, análise e descobertas. *Rede de Observatório de Segurança.* https://drive.google.com/file/d/18CEwZynKosnnS7Nh6-YOpEN7msZN9f77/view

O'Neil, C. (2021). *Algoritmos de destruição em massa.* Editora Rua do Sabão. *Algoritmos de destruição em massa:* como o big data aumenta a desigualdade e ameaça a democracia. Tradução: Rafael Abraham. 1ed. Santo André: Rua do Sabão, 2020.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge. Harvard University Press.

Vieses no Aprendizado de Máquina e suas Implicações Sociais: Um Estudo de Caso no Reconhecimento Facial. In: *Anais do workshop sobre As Implicações da Computação na Sociedade.* 2021.

Ruback, L., Avila, S., & Cantero, L. (2021, julho). Vieses no aprendizado de máquina e suas implicaçoes sociais: Um estudo de caso no reconhecimento facial. In Anais do II Workshop sobre as Implicaçoes da Computaçao na Sociedade (pp. 90-101). SBC. https://doi.org/10.5753/wics.2021.15967

Silva, T. (2021, maio 16). *O reconhecimento facial deve ser banido: veja dez razões.* https://tarciziosilva.com.br/blog/reconhecimento-facial-deve-ser-banido-aqui-estao-dez-razoes/#:~:text=Entre%20exemplos%20recentes%2C%20est%C3%A3o%3A%20o,Facebook%20em%20todo%20o%20mundo%20

TRANSPARÊNCIA BRASIL. Estrutura de avaliação de riscos a direitos e de transparência. Fev. 2020b. https://www.transparencia.org.br/downloads/publicacoes/Estrutura_Avaliacao_Risco.pdf. Acesso: 30 nov. 2020.

Trielli, D. & Diakopoulos, N. (2017, maio 30). *How to report on algorithms even if you're not a data whiz. Town Center.* https://www.cjr.org/tow_center/algorithms-reporting-algorithmtips.php

Trielli, D., & Diakopoulos, N. (2020). How journalists can systematically critique algorithms. In *Proc. Computation+ Journalism Symposium (5), 1-30.* https://bpb-us-w2.wpmucdn.com/sites.northeastern.edu/dist/0/367/files/2020/02/CJ_2020_paper_14.pdf . *Journalism Symposium*. 2020. Disponível em: http://www.nickdiakopoulos.com/wp-content/uploads/2020/02/How-journalists-can-systematically-critique-algorithms.pdf. Acesso: 13 nov. 2020.

Parte 2

INTELIGÊNCIA ARTIFICIAL, ALGORITMOS E VIESES

FAKE NEWS AS DIGITAL DISRUPTION: UNRAVELLING ALGORITHMIC LOGIC IN THE SPREAD OF DISINFORMATION

André Lemos
/ Faculdade de Comunicação / Universidade Federal da Bahia

Introduction

This paper explores the intricate relationship between fake news (FN) and algorithms, shedding light on their profound impact on the modern information landscape. In an era of rapidly advancing technology, algorithms play a pivotal role in shaping our online experiences and influencing the dissemination of information. However, the dark side of algorithms emerges when they are exploited to propagate disinformation, leading to the erosion of trust, the polarisation of societies, and the undermining of democratic principles. By examining the mechanisms through which algorithms facilitate the spread of disinformation, this paper delves into the notion that disinformation is a disruption rather than an error within digital culture. FN is deeply intertwined with the underlying logic of algorithms designed to optimise user experience and engagement, contributing to the disinformation's amplification and propagation.

Understanding this digital disruption is essential for developing effective strategies to combat the harmful FN effects and preserve the integrity of digital spaces. The first part of this article defines what is meant by FN. Next, it discusses digital errors, placing the debate on three levels: errors, failures and disruptions, defining FN at this last level. The third part points out the prob-

lem of FN from Generative Artificial Intelligence (GAI), such as the ChatGPT. New developments in the field of AI have fostered intense debates about FN proliferation. By exploring the interplay between disinformation and algorithmic logic, we can gain insights into the mechanisms that amplify and perpetuate disruption in digital culture. In conclusion, the paper highlights the need to implement regulatory measures and algorithmic transparency to stop the amplification of FN and preserve the informational public space.

Fake News, Disinformation, and Algorithms

The digital age has revolutionised how information is created, distributed, and consumed. Algorithms designed to personalise and optimise user experiences have become integral components of online platforms and services (Amoore, 2020; Crawford, 2021; Gillespie, 2014; van Dijck et al., 2018). These complex mathematical formulae are responsible for curating and presenting content, making them powerful information gatekeepers. There is unanimous recognition that FN has become a significant concern in the digital era, with algorithms playing a crucial role in its dissemination.

Algorithms have the potential to amplify the spread of disinformation. The analysis of user preferences, behaviours, and engagement patterns can create filter bubbles (Pariser, 2011) and echo chambers (Nguyen, 2020), reinforcing pre-existing beliefs and preferences. This personalised delivery of content on digital platforms creates an environment where individuals are exposed to a limited range of viewpoints making them more susceptible to FN. Algorithms optimise engagement metrics, prioritising sensational or controversial content that often includes misinformation, propaganda or conspiracy theories, thus amplifying their reach and impact (Gillespie, 2010; Nieborg & Helmond, 2019). The algorithmic logic inherent biases and consequences contribute to the spread of false information.

Digital platforms operate through recommendation and boosting algorithms, data extraction, analysis, and profiling (datafication) (Lemos, 2021). They constitute one of the central infrastructures of society today, also

referred to as the "platform society" (van Dijck et al., 2018) or, in its economic version, data or surveillance capitalism (Srnicek & De Sutter, 2016; Zuboff, 2019). Today, a large part of the economy, work, culture, and education passes through digital platforms. FN exploits the platform ecosystem (e.g., disseminating disinformation on Telegram and linking to specific videos on YouTube) to gain traction and go viral.

We can define FN as the intentional actions of disseminating false information (retaining aspects of "misinformation" – false information spread as rumours without malicious intent –; or "disinformation" – false content, intentionally disseminated, ideologically biassed information that distorts facts or narratives as propaganda[1]) to harm groups or individuals using social media platforms, which may or may not simulate traditional media (Fetzer, 2004; Zimdars & McLeod, 2020)[2]. It is false information and ideological propaganda disseminated through social media and algorithmic logic. FN is not a strictly journalistic phenomenon but a political one. Professional journalism does not produce FN but what we could call "invisible news," omitting topics and agendas due to corporate bias.

The main characteristics of FN are two. First, its materiality, as it exploits the algorithmic logic of platforms (platform ecology) to achieve its objectives. Second, the spread of unquestioned disinformation in a "tribal", highly emotional and affective sharing culture that seeks to reinforce ideology and a sense of belonging to a group. FN's effectiveness lies in these two central dynamics: algorithmic circulation and emotional consumption of disinformation. This appeal to a "religious audience mode" involves the adherence of the same interested community to a single narrative, without contradiction, reinforcing their values. The rationale is: "*If this information is within my network of friends or relatives, who share the same belief/ideology, there is no need to fact-check because it is true. I will pass it on!*". Therefore, with

1. See Temple University Guide (https://guides.temple.edu/fakenews) and "Understanding Digital Disorder (https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf?x32994).
2. We can use Fake News (FN) or disinformation. In this paper, I'll use the term FN.

weak chains of reference, FN works because they are accepted in a logic that moves away from that of a rational subject who searches for verified sources of information.

This corresponds, therefore, to the difficulty of fact-checking tools and projects to effectively combat misinformation (Lemos & Oliveira, 2021). Fact-checking projects are necessary but are not very effective for two reasons. First, they do not go viral like FN. Once misinformation reaches groups, the damage is done, and it is challenging to reverse. It cannot go backwards. The second reason is the religious mode or "tribal" logic. When receiving information from someone who thinks like them, citizens/believers do not seek out fact-checking projects to verify it because they believe such projects (as well as the mainstream media) are created to manipulate them. Reversing this logic is challenging, even with technical or legal mechanisms.

FN is, consequently, "fake politics" (Lemos et al., 2021). Politics is exercised by circulating words for debate, seeking a minimum consensus for the common good. In modern democracies, this circulation takes place through the rule of law, freedom of expression, and the plurality of information, with a free and responsible press. The public sphere must be informed in order to support debate and foster actions in the public interest.

FN does not mobilise words in a republican sense for the public interest. Instead, they stimulate emotions for private and individual interests, producing a public sphere grounded in lies. By creating content that triggers emotional responses, exploiting cognitive biases, or leveraging viral sharing patterns, purveyors of disinformation use algorithmic logic to promote their false narratives. This strategic exploitation allows FN to flourish and spread rapidly, posing significant challenges to truth and reliable information.

FN and Digital Error

FN is a phenomenon directly linked to platforms and their algorithmic performance. However, they are not digital errors that can be attributed to an

author or a platform and thus corrected. It is a broader issue that I will call digital disruption. It is essential to recognise this in order to address the problem effectively.

Platforms position themselves as common carriers and try to absolve themselves of responsibilities by claiming that they cannot control content. With their expansion, they have become important agents in mobilising public opinion and have had to take responsibility (through the actions of states and global society). However, the circulation and virality of information are the platforms' central objectives (commercial and strategic). In this sense, FN is not a platform's digital error but disruptive effects caused by socially recognised harmful uses.

Digital errors and failures are generative, allowing controversies to emerge and ethical value issues to be located in the dimension of the commons (Barker & Korolkova, 2022; Korolkova & Bowes, 2020; Nunes, 2011). In current algorithmic systems, chains of disinformation should be seen as systemic disruptions. They are not digital errors or failures necessarily (although they may be due to logical errors or external failures, as infrastructure blackouts can produce disturbances). They operate according to the logic of the platforms, which monetise any circulation of information. They are disruptions caused by ethical and moral issues related to the use of artificial intelligence.

Digital disruption points to issues of interest and where attention needs to be directed. Therefore, understanding algorithmic errors, failures, and disruptions should be considered a methodological and epistemological issue. They highlight not what is obvious and works but what generates controversy, placing objects as the focus of ethical-political discussion and helping to identify what needs to be qualitatively analysed. Errors are problems of principle, logical, internal discrepancies, and events that deviate from the norm or principle and lead to incorrect results. Failures are problems created due to external causes, like accidents, infrastructure blackouts etc.

Disruptions are disruptive events within the logic of systems or platforms, which may or may not be caused by errors and failures. They arise from malicious uses that exploit the grammar of systems and platforms, generating disruptive effects concerning intended uses. In this sense, disruptions are neither errors (although they can also be when a platform tries to monitor FN and fails, for example), nor failures, but anomalous events (socially identified as such) that allow controversies to be located and qualitative analyses of the society to be expanded.

Digital disruptions can be seen as anomalies (Parikka & Sampson, 2009) that arise from ethical and moral positioning regarding the performance of devices and systems. In this sense, they are linked to value judgments that change according to the moods of the times. The ethical question arises from using algorithmic systems that transform the indiscernible world into a discernible one (Amoore, 2019, 2020). That is, what should be the object of intervention with AI, what should be platformed, and what should be automated must always be a matter of ethical-political choices.

Disruptions are caused by errors and failures or by the everyday use of information devices. Mobile phones, computers, and the internet have changed and continue to change society. New technologies, products and services can produce harmful effects on social achievements (gender and race bias, environmental problems, precarious work in the gig economy, and regimes of control and surveillance, among others) (Eubanks, 2017; Noble, 2018; van Dijck et al., 2018; Velkova, 2019; Zuboff, 2019).

As examples of digital disruption (which are neither errors nor failures), we can mention spam (unsolicited email), viruses (programs designed to cause damage), deep fakes (videos with fake images), algorithmic bias (gender, race or ethnicity), stalking and nudes (attacks on people, or sending unsolicited images through the use of social networks), among others. None of these cases are system errors or external failures but anomalies caused by the use of digital systems that are recognised in context as abusive.

The same applies to FN on social media networks. There is no logical or algorithmic error in the platforms for disseminating false information. FN are not errors to be corrected by identifying an author but disruptions to be addressed within a broad action network. As we said, it is a political, journalistic, but also an infrastructural, algorithmic/technical, and legal phenomenon. They are not errors from the companies' perspective because their dissemination follows the correct functioning of the platforms within their logical parameters, generating significant monetisation and profit for the companies.

Initially, social media platforms claimed to be "common carriers" and, as such, were not responsible for the content. However, their importance in shaping a new media public sphere has led society to demand legal and technical mechanisms. They are now under pressure to exercise regulatory power over what circulates through their algorithmic structures. Europe already has a law to regulate the actions of platforms, which will take effect in the coming years. Brazil is presenting a similar bill. Therefore, analysing FN as disruptions allows us to address significant ethical and moral questions of digital culture and to locate controversies that can instruct us about the organisation of our society: Can we eliminate FN by correcting the code? With algorithms to control, monitor, and filter? How can we regulate platforms to address FN without risking censorship? Through fact-checking projects? By teaching media literacy and debating FN from an early age? How can we escape the religious mode of the internet underground?

As the philosopher and biologist Donna Haraway (Haraway, 2016) states, it is vital that "*we stay with the trouble*" with the disruptions of digital culture, as they allow us to understand the present and confront the challenge of the common. Recognising disinformation as a digital disruption requires proactive measures to mitigate its harmful effects. Algorithmic transparency and accountability are paramount in combating the spread of false information. Platforms should adopt more stringent regulations, undergo independent audits, and disclose algorithmic processes to identify and rectify flaws that

enable the amplification of FN. Additionally, fostering media literacy and critical thinking skills among users is crucial to empowering individuals to navigate the digital landscape more discerningly.

Fake News, Hallucination and Artificial Intelligence

The AI algorithms of platforms and new artificial intelligence systems have driven the rapid spread of FN. Platformisation has been one of the differentiators in the definition of FN (Dan et al., 2021). Now, with Generative Artificial Intelligence (GAI), such as Chat GPT[3], an expansion of FN's circulation is already a social disruption. Recent texts are pointing to this dimension[4]. GAI can be defined as an AI system that interacts with humans and can produce high-quality texts, videos, and sounds. They can be used to identify false information or to generate texts with false information, mimicking human-like behaviour.

Disinformation and generative artificial intelligence pose significant challenges in today's information landscape. The proliferation of AI-powered technologies has made it unprecedentedly easier for malicious actors to create and spread disinformation[5]. Generative AI models, such as deep learning algorithms, can generate highly realistic and convincing content, including text, images, and videos, which can be exploited to manipulate public opinion, deceive individuals, and undermine trust in reliable sources of information. This emerging issue raises concerns about the potential misuse of AI and the need for robust solutions to detect, mitigate, and combat disinformation.

3. ChatGPT generative algorithm, released on November 30, 2022, is a natural language processing system (Large Language Model) that uses neural networks to string together words in a conversation with a human. He is trained through extensive information (texts, images, code) available on the internet (and until then collected up to September 2021).
4. See Santos (2023). See also "Disinformation generated by AI may be more convincing than disinformation written by humans" (https://www.technologyreview.com/2023/06/28/1075683/humans-may-be-more-likely-to-believe-disinformation-generated-by-ai/); and "Nearly 50 news websites are 'AI-generated'". (https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study).
5. When an FN is generated by fair use of AI, it would be an error, but when it is produced with the intention of generating misinformation, it would be a disruptive effect.

Several scientific papers highlight the urgency of addressing the issues surrounding disinformation and generative AI (Akhtar, 2023; Goldstein et al., [s.d.]). They provide valuable insights into the potential threats posed by AI-generated content and propose approaches to mitigate these challenges, contributing to the ongoing efforts in combating disinformation in the digital age. According to some research, the development of generative artificial intelligence will increase the risk of producing FN in an automated manner without direct human intervention, either intentionally, through errors, or through artificial hallucinations.

When a GAI makes a mistake, this error is called "algorithmic hallucination", generating erroneous information or surreal images. The concept is recent, arising in the field of AI computer vision. According to the company Open AI (cited by Alkaissi & McFarlane, 2023, p. 3):

> Artificial hallucination refers to the phenomenon of a machine, such as a chatbot, generating seemingly realistic sensory experiences that do not correspond to any real-world input. This can include visual, auditory, or other types of hallucinations. Artificial hallucination is not common in chatbots (...). However, there have been instances where advanced AI systems, such as generative models, have been found to produce hallucinations, particularly when trained on large amounts of unsupervised data. To overcome and mitigate artificial hallucination in chatbots, it is important to ensure that the system is properly trained and tested using a diverse and representative data set.

The term hallucination indicates very different actions, such as producing an output with misinformation about James Webb[6], claiming to love a human being[7], writing racist texts from scientific literature[8], or lying to get your way[9].

6. https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo
7. https://www.nytimes.com/2023/02/17/opinion/letters/bing-chatbot-kevin-roose.html
8. Edwards, Benj (18 November 2022). "New Meta AI demo writes racist and inaccurate scientific literature, gets pulled". *Ars Technica*. Retrieved 30 December 2022.
9. https://shorturl.at/uBFR8

These examples have generated failures in arguments, identification of historical events, and other disturbing effects that call into question the use of GAI.

The everyday use of GAI is already disturbing. According to some[10], it threatens jobs and creative making and could annihilate the human species. A recent letter asking for a moratorium on GAI's development, written by experts, business people and celebrities, clearly illustrates the disturbances caused by the very existence of GAI. This positioning shifts concerns about the urgent challenges of the current society of platforms that also use artificial intelligence systems (disinformation, data surveillance, privacy, data colonialism, the threat to sovereignty from global corporate control in the clouds) to an abstract future. Hallucinating or not, GAI will undoubtedly be an element that will aggravate the spread of FN, amplifying this digital disruption.

In the case of GAI, such as the GPT chat, the discussion about its errors, failures and disturbances brings the debate to the agenda, points out its benefits, or reveals its potential problems. They reveal the dimensions of this object (IA), the complexity of the problems and potentialities for its realisation, and the multiple arrangements that touch different domains (education, employment, science, politics, management...). Looking at errors, failures, and disturbances is a methodological and epistemological strategy to reveal issues of interest for qualitative research on digital culture in general and AI in particular. So, let us move on to what is controversial. As Ernst states, "*only in case of failure or error, media become apparent as technological beings (...).*" (Barker & Korolkova, 2022, pp. 83–84).

10. See "Pause Giant AI Experiments: An Open Letter." In https://futureoflife.org/open-letter/pause-giant-ai-experiments/ . "The open letter to stop dangerous AI race is a huge mess". In https://www.vice.com/en/article/qjvppm/the-open-letter-to-stop-dangerous-ai-race-is-a-huge-mess

Conclusion

The proliferation of false narratives and conspiracy theories can contribute to social polarisation, leading to divisions and the erosion of social cohesion. Disinformation also threatens public health, as seen during the COVID-19 pandemic, when false information about vaccines and treatments spread rapidly, endangering lives (Cotter et al., 2022).

Errors, failures, and disruptions highlight what generates controversy and guide research toward questions of interest. They place objects as the focus of ethical-political discussion and can help identify what we should qualitatively analyse in AI. There is an epistemological vector that helps to think about digital culture (Maalsen, 2023, p. 11).

Addressing the disinformation crisis requires a multifaceted approach, with AI and the platform's algorithmic regulation and accountability playing a crucial role. Stricter oversight and transparency measures should be implemented to ensure that algorithms are not weaponised to propagate disinformation. Increased scrutiny and auditing of algorithmic systems can help identify biases, vulnerabilities, and potential avenues for abuse. Platforms should give users greater control over the content they are exposed to, fostering a more diverse and balanced information environment (Hermann, 2021).

FN cannot be reduced to a mere digital error but must be acknowledged as a disruption within the digital landscape. Understanding the symbiotic relationship between disinformation, tribal groups and algorithmic logic is crucial to effectively combat the harmful effects of false information. By recognising the unintended consequences of algorithms and implementing robust regulatory measures and transparency, we can mitigate the amplification of disinformation and preserve the credibility of digital information spaces, fostering a more informed and resilient society.

References

Akhtar, Z. (2023). Deepfakes Generation and Detection: A Short Survey. *Journal of Imaging*, *9*(1), 18. https://doi.org/10.3390/jimaging9010018

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. https://doi.org/10.7759/cureus.35179

Amoore, L. (2019). Doubt and the Algorithm: On the Partial Accounts of Machine Learning. *Theory, Culture & Society*, *36*(6), 147–169. https://doi.org/10.1177/0263276419851846

Amoore, L. (2020). *Cloud ethics: Algorithms and the attributes of ourselves and others*. Duke University Press.

Barker, T., & Korolkova, M. (Orgs.). (2022). *Miscommunications: Errors, Mistakes, Media*. Bloomsbury Academic.

Cotter, K., DeCook, J. R., & Kanthawala, S. (2022). Fact-Checking the Crisis: COVID-19, Infodemics, and the Platformization of Truth. *Social Media + Society*, *8*(1). https://doi.org/10.1177/20563051211069048

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor* (First Edition). St. Martin's Press.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Orgs.), *Media Technologies* (p. 167–194). The MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. ([s.d.]). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*.

Haraway, D. J. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.

Hermann, E. (2021). Artificial intelligence and mass personalization of communication content—An ethical and literacy perspective. *New Media & Society*. https://doi.org/10.1177/14614448211022702

Korolkova, M., & Bowes, S. (2020). Mistake as method: Towards an epistemology of errors in creative practice and research. *European Journal of Media Studies*, *9*(2), 139–157. https://necsus-ejms.org/mistake-as-method-towards-an-epistemology-of-errors-in-creative-practice-and-research/

Lemos, A. (2021). Dataficação da vida. *Civitas - Revista de Ciências Sociais*, *21*(2), 193–202. https://doi.org/10.15448/1984-7289.2021.2.39638

Lemos, A. L. M., Bitencourt, E. C., & Dos Santos, J. G. B. (2021). Fake news as fake politics: The digital materialities of YouTube misinformation videos about Brazilian oil spill catastrophe. *Media, Culture & Society*, *43*(5), 886–905. https://doi.org/10.1177/0163443720977301

Lemos, A., & Oliveira, F. (2021). Fake news e cadeias de referência: A desinformação sobre Covid-19 e o projeto de verificação do Facebook. *Fronteiras - estudos midiáticos*, *23*(2), 73–88. https://doi.org/10.4013/fem.2021.232.06

Maalsen, S. (2023). Algorithmic epistemologies and methodologies: Algorithmic harm, algorithmic care and situated algorithmic knowledges. *Progress in Human Geography*, 030913252211494. https://doi.org/10.1177/03091325221149439

Nieborg, D. B., & Helmond, A. (2019). The political economy of Facebook's platformization in the mobile ecosystem: Facebook Messenger as a platform instance. *Media, Culture & Society*, *41*(2), 196–218. https://doi.org/10.1177/0163443718818384

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*.

Nunes, M. (Org.). (2011). *Error: Glitch, noise, and jam in new media cultures*. Continuum.

Parikka, J., & Sampson, T. D. (2009). On anomalous objects of digital culture. An Introduction. *The Spam book. On Viruses, Porn, and Other Anomalies from the Dark Side of Digital Culture*. Hampton Press.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Srnicek, N., & De Sutter, L. (2016). *Platform capitalism*. Polity Press.

van Dijck, J., Poell, T., & de Waal, M. (2018). *The Platform Society*. Oxford University Press.

Velkova, J. (2019). Data Centers as Impermanent Infrastructures. *Culture Machine*, 1–11.

Zimdars, M., & McLeod, K. (Orgs.). (2020). *Fake news: Understanding media and misinformation in the digital age*. The MIT Press.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power* (First edition). PublicAffairs.

INTELIGENCIA ARTIFICIAL Y *DEEPFAKES:* SESGOS DE GÉNERO *Y* AGRESIÓN CONTRA LAS MUJERES

Rosa Franquet
/ Universitat Autonòma de Barcelona

Introducción

La Inteligencia Artificial (IA) se presenta como uno de los grandes logros del siglo XXI, una especie de panacea que resolverá los acuciantes problemas de nuestra sociedad. Desde propiciar un crecimiento económico sostenible que mitigue el cambio climático a lidiar con nuevas pandemias o terminar con la desigualdad social. Sin embargo, su implementación resulta compleja para algunos sectores y conlleva enormes interrogantes acerca de si su impulso será capaz de mitigar los enormes desequilibrios entre los territorios y las personas de etnias, clases y géneros diferentes. La adopción generalizada de herramientas de IA generativa, como los programas que permiten la generación automática de textos, voces e imágenes, esboza diferentes problemáticas en una sociedad centrada en la lógica del capitalismo extractivo, y donde las personas encargadas de desarrollar la tecnología se sienten poco comprometidas con las consecuencias que se derivan de su proliferación y uso en los distintos sectores económicos y grupos sociales, sobre todo aquellos más vulnerables. Una de estas secuelas tiene que ver con la desigualdad entre hombres y mujeres y, en concreto, la lacra que supone la violencia que se ejerce contra ellas, en todo el mundo, que lejos de desvanecerse ha encontrado nuevos instrumentos virtuales para seguir ejerciéndola.

En el texto se revisan críticamente algunas de las prácticas sociales que utilizan la inteligencia artificial para perpetuar actitudes misóginas y de subyugación hacia las mujeres al utilizar los recursos de IA para ejercer nuevas violencias y continuar cosificándolas y sometiéndolas virtualmente. Finalmente, se revisará el papel de los medios de comunicación en este nuevo contexto de desarrollo de la IA y la tímida legislación que la EU está desarrollando para afrontar algunos de los retos que la IA plantea en la actualidad y, así, mitigar sus efectos adversos.

Omnipresencia de la IA

A mediados de la década de los 1950, la disciplina de la IA emerge dentro de la informática con el cometido de construir sistemas de computación capaces de emular la inteligencia humana. Una ocurrencia muy atractiva e inspiradora que congregaría a investigadores/as a lo largo de décadas con el propósito de alcanzar esa idea utópica y arriesgada de emular el pensamiento humano, desde que John McCarthy acuñó el término en 1956 (Kurzweil,1985).

Con los sucesivos avances en IA, en las últimas décadas del siglo XX, encontramos pensadores/as que nos han alertado sobre las consecuencias que conlleva su uso generalizado en todos los sectores económicos y sociales. La Inteligencia Artificial puede aplicarse a cualquier campo y afecta a todos los ámbitos de nuestras vidas como la educación, la sanidad, el mundo laboral o la información y el entretenimiento que consumimos a diario. Sabemos que no podemos categorizar los progresos de la tecnología y la IA como simples equipamientos y, en consecuencia, un análisis reflexivo y crítico es imprescindible para conocer si su adopción resulta inevitable y que problemas se presentan con su uso generalizado.

Durante la última década, es conocida la expectación que la IA ha generado en muchas personas que ven en sus desarrollos nuevas posibilidades para el desempeño de sus tareas cotidianas desde escribir, componer música, crear imágenes o generar voz humana. Unas facilidades que están afectando

profundamente a todos los sectores productivos y a la mayoría de las prácticas sociales que ejecutamos diariamente.

La ayuda que nos dispensa la IA en la toma de decisiones se ha incrementado, en la última década, y son muchas las áreas que se benefician de su uso, pero es necesario comprender y abordar su impacto social. Los progresos acelerados nos invitan a pensar sobre los beneficios de la IA, pero al mismo tiempo a evaluar los riesgos y a abordar los desafíos de todo tipo, como los éticos o legislativos, que se derivan de la subversión de las reglas sociales actuales. Pensemos en el control social que se ejerce mediante las tecnologías digitales, como el sistema de crédito social chino implementado en 2014, y que permite a los gobiernos locales clasificar a los y las ciudadanas en categorías de "buenos/as" y "malos/as"[1], según sus actitudes y comportamientos, o la identificación biométrica de individuos en espacios públicos y en tiempo real, donde las aplicaciones de IA permiten escanear el rostro e identificarlo automáticamente sin autorización alguna en la mayoría de los países del planeta. Como apunta Harari (2019), antes aceptábamos la autoridad divina y "la autoridad humana estaba justificada por el relato liberal, así la revolución tecnológica que se avecina podría establecer la autoridad de los algoritmos de macrodatos, al tiempo que socavaría la idea misma de libertad individual" (p. 68). La omnipresencia de los algoritmos altera nuestra anterior forma de tomar decisiones y la IA continúa ganando protagonismo al introducirse en muchas de nuestras prácticas cotidianas.

Por otra parte, entre los efectos indeseados del uso de los algoritmos encontramos las investigaciones que apuntan los sesgos que estos sistemas tienen y que ayudan a perpetuar los estereotipos, amplificándolos y, por tanto, justificando ciertos patrones de discriminación. Algunos modelos de aprendizaje automático tienen los llamados bucles de retroalimentación integrada y eso significa que el modelo produce una predicción que ayuda a tomar decisiones con los resultados obtenidos y esos resultados se agregan

1. Ver el artículo de Mo, Ch. y Grossklags, J. (2022). Social Control in the Digital Transformation of Society: A Case Study of the Chinese Social Credit System. Social Sciences 11: 229. https://doi.org/10.3390/socsci11060229

a los datos de entrenamiento para la próxima ronda de entrenamiento. En consecuencia, las predicciones hechas por el sistema influyen en los datos que se generan para futuras predicciones. Además, el hecho de que la IA y los sistemas algorítmicos a menudo carezcan de transparencia complica la detección de los sesgos y la posible discriminación.

La equidad en el tratamiento de género de los sistemas de IA es una quimera y el sesgo de género entendido como "la diferencia sistémica e injusta en la forma en que se trata a hombres y mujeres en un ámbito particular" (Masiero y Aaltonen, 2020, p. 1) ha sido evidenciado por numerosos estudios (Agarwal, 2020; Nadeem et al., 2020; Levy, 2018). El análisis de la UE, "*Bias in Algorithms. Artificial intelligence and discrimination"* tiene un apartado sobre sesgos étnicos y de género en los modelos de detección y predicción del habla y recomienda una evaluación para superar dichos sesgos en la implementación de algoritmos.

Geoffrey Hinton, uno de los pioneros de *Deep learning* señala en una entrevista a la MIT Technology Review[2] que los sistemas como el GPT-4 pueden aprender cosas nuevas muy rápidamente una vez que los y las investigadoras los entrenan adecuadamente. Nos propone unas reflexiones sobre "las consecuencias de un adiestramiento efectuado en un entorno muy desequilibrado con grandes sesgos y que a penas se tiene conciencia para corregirlos". En ese proceso de entrenamiento tienen claras desventajas las lenguas minorizadas, las mujeres, las personas racializadas, las ideas de-coloniales, etc. Para Hinton el siguiente paso de las máquinas inteligentes será "su capacidad para crear sus propios sub-objetivos, pasos intermedios necesarios para llevar a cabo una tarea". Y se pregunta "¿Qué sucede, cuando esa capacidad se aplica a algo intrínsecamente inmoral?"

2. "Geoffrey Hinton tells us why he's now scared of the tech he helped build" por Will Douglas Heaven. 2 de Mayo 2023. Consultado. 15 junio de 2023. https://www.technologyreview.com/2023/05/02/1072528/geoffrey-hinton-google-why-scared-ai/

Efectos colaterales de la IA: del reconocimiento facial al *deepfake*

En el siglo XXI, el software ha adquirido una enorme centralidad y su trascendencia es comparable a avances como la electricidad o el motor de combustión del siglo XX. Unos servicios de software que han incrementado sus prestaciones y que usan tecnologías de automatización "que a menudo presentan soluciones de 'caja negra', aunque el software consiga resultados deseados, no sabemos las reglas que sigue" (Manovich, 2017, p. 25). La opacidad y la falta de transparencia de las aplicaciones informáticas basadas en la IA es lo que hace vulnerables a los y las usuarias que los utilizan, al no conocer las consecuencias de su funcionamiento que son difíciles de prever y/o descubrir.

El primer nivel de IA engloba sistemas competentes para efectuar trabajos concretos o resolver problemas específicos, que tienen un rendimiento mayor que el de las personas. Los chatbots como ChatGPT, los asistentes de voz como Siri y Alexa, los sistemas de reconocimiento de imágenes y los algoritmos de recomendación pertenecen a esta categoría.

La información suministrada a buscadores, a aplicaciones de todo tipo a las que nos enganchamos, con interacciones constantes, delata nuestros comportamientos, deseos, gustos y opiniones y nos convierten en sujetos rastreados y mercantilizados. La lectura de las tediosas explicaciones sobre la política de privacidad de webs o aplicaciones nos daría una idea de lo vulnerables que podemos llegar a ser. Unos datos cedidos alegremente que pueden utilizarse en nuestra contra como ocurre con todo tipo de manipulaciones derivadas de las fotografías subidas a las diferentes webs o aplicaciones.

Las posibilidades de ejercer control o manipular datos e informaciones no son nuevas y, tampoco, las prácticas creativas destinadas a enganchar una cara en otro cuerpo, o exponer a personas en escenarios o actividades distintas e inventadas. La historia nos ha mostrado como las imágenes de famosos personajes se han incrustado en espacios imaginados o completamente ficticios, incluidos los escenarios pornográficos, así pues, la pornografía falsa

o las imágenes sexuales no consentidas llevan mucho tiempo mostrándose en la esfera pública (Burkell y Gosse, 2019). La novedad reside en la facilidad de creación y distribución de estas imágenes no consentidas con la permeabilización en la sociedad del código informático, desde mediados del siglo pasado, y la llegada de internet después. Cada vez resulta más viable que ciertos grupos adquieran las habilidades para hacer funcionar los sistemas de generación de imágenes inventadas. Las consecuencias pueden ser devastadoras para las víctimas de esas usurpaciones. Únicamente se necesitan habilidades de retoque fotográfico para modificar las fotografías y conseguir efectos no deseados relacionados con la cosificación sexual que, a menudo, se ejerce contra la víctima como venganza (Gander, 2016).

A partir de 2014, la posibilidad de generar contenidos simulados se beneficia de los avances de los *Generative Adversarial Networks* (GANs) que permiten la manipulación de imágenes de manera mucho más rápida y sencilla y con resultados de gran realismo y con buena calidad visual. En 2017, la aplicación FaceApp, publicitada como un servicio inofensivo para conocer nuestro aspecto, pasados los años, a partir de una *selfie* enviada por un o una usuaria alcanzó importantes cuotas de popularidad. Todas las fotografías enviadas se almacenarían en los servidores de la empresa que obtendría una enorme cantidad de todo tipo de rostros humanos. Un banco de datos susceptible de ser utilizado por la propia empresa o vendido a terceros. Las personas al subir su imagen a la aplicación y transferirla al servidor del desarrollador perdían todo el control sobre la fotografía enviada y sobre las consecuencias que dicha usurpación podría conllevar. El debate público sobre los avances tecnológicos debe cuestionar los principios éticos sobre los que se asienta, por ejemplo, esa cesión no consciente y de los peligros derivados del uso que pueda hacerse posteriormente con las imágenes.

Las herramientas que generan imágenes a partir de sistemas de entrenamiento que utilizan enormes bases de datos sin consentimiento ni compensación y que violan la privacidad, cada vez son más sofisticadas y necesitan menos habilidades informáticas o creativas para su utilización. Históricamente la creación de imágenes y videos falsos había estado

restringida a los y las especialistas de la industria del entretenimiento que disponían de la destreza suficiente para crearlos. Con los nuevos avances, la generación de textos, voces, imágenes o contenidos videográficos son más fáciles de crear por personas aficionadas.

Estas alertas y preocupaciones se han disparado desde el lanzamiento, en noviembre de 2022, del chatbot de base generativa ChatGPT, de Open AI, que alcanzó enorme popularidad en un tiempo récord y vino a competir con los desarrollos de Google con BERT y los avances de Meta. progresos informáticos creados por gigantes tecnológicos que utilizan enormes recursos financieros y energéticos y se entrenan con ingentes cantidades de datos procedentes de textos, músicas, voces, imágenes y códigos generados por instituciones y/o personas anónimas que no tendrán ninguna compensación por ello.

La concentración de poder por parte de las élites extractivas no tiene precedentes y los GAFAM (Google, Amazon, Facebook, Apple y Microsoft), o las BATX (Baidu, Alibaba, Tencent y Xiaomi) no sólo acumulan bienes materiales, sino que quieren los beneficios de los bienes intangibles generados por otros actores sin respetar la propiedad de autoría intelectual. Una concentración de poder en manos de oligopolios que controlan los sistemas de IA, las redes sociales y otras muchas divisiones de bienes y servicios. La impunidad para operar en los distintos mercados sin a penas control hace preguntarse al pensador García Canclini (2020, p.84): "¿Cuál es la autoridad de los Estados que ni siquiera se proponen tener algún tipo de política para controlar a las corporaciones dedicadas a comercializar la vida privada de sus ciudadanos?"

La IA permite generar voz, textos, imágenes y videos manipulados digitalmente que pueden adquirir una gran viralidad en las redes sociales independientemente de su veracidad. Estos contenidos falsos facilitan la manipulación de la opinión pública y pueden interferir en los procesos electorales. Y, sin ánimo de exhaustividad, entre los múltiples efectos dañinos podemos reseñar las prácticas extendidas de ciberacoso o la destrucción de la reputación personal mediante la generación de voces o imágenes

manipuladas sexuales no consentidas, que se ceban especialmente contra las mujeres. Prácticas que violan la privacidad e intimidad de millones de personas y que exigen una regulación centrada en los ciber-acosadores y en los sitios que las alojan.

El término *deepfakes* apareció en 2017, cuando un moderador en creó un subreddit con ese nombre donde publicaba y compartía vídeos pornográficos falsos de mujeres famosas sin su consentimiento, como los de Scarlett Johansson o Taylor Swift, utilizando un algoritmo de IA. Una práctica que rápidamente se extendió y se hizo popular entre los grupos masculinos de distintas redes. En enero de 2023, los *deepfakes* de porno se difundieron en medios de comunicación y redes cuando el popular *streamer* de Twitch, Brandon Ewing, apodado "Atrioc", tenía un "sitio web de porno falso abierto en su navegador durante una transmisión en vivo en la plataforma. El sitio presentaba imágenes manipuladas de otros streamers y amigos de Twitch"[3]. Como apunta Sophie Compton, cofundadora de la coalición *My Image My Choice* (Mi imagen, mi elección), la mirada debe focalizarse, también, sobre las personas cómplices que son compañeros de clase, amigos, hermanos y novios. "Puede que no se den cuenta de que su participación también forma parte del problema. Están ayudando a validar algo que es extremadamente perjudicial para las mujeres"[4]. En un primer momento, las manipulaciones de videos no consentidos se desplegaron contra políticos o figuras públicas, por ejemplo, recordemos el famoso video de Obama producido por Jordan Peele. En este caso, la producción se concibió para el entrenamiento del sistema de IA y para demostrar el poder de la herramienta tecnológica *deepfake* que con esa sub-plantación se conseguiría una gran popularidad.

En el inicio, la popularidad sirvió como enganche y las víctimas de las falsificaciones empezaron siendo políticos o mujeres famosas. Luego emergió una plaga de vídeos sexuales no consentidos misóginos, degradantes y

3. Ver https://www.euronews.com/next/2023/04/22/a-lifelong-sentence-the-women-trapped-in-a-deepfake-porn-hell.Consultadp. Consultado 6 Mayo de 2023.
4. Ver el artículo de Imane El Atillah. "El infierno de las mujeres atrapadas por el porno "deepfake", las falsificaciones pornográficas". Euronews 24/4/2023. https://es.euronews.com/next/2023/04/24/el-infierno-de-las-mujeres-atrapadas-por-el-porno-deepfake-las-falsificaciones-pornografic. Consultado 7 junio de 2023.

humillantes contra multitud de mujeres. Una explicación del aumento exponencial de los *deepfake*, basados en las imágenes pornográficas, puede hallarse en la comercialización de herramientas y servicios que facilitan la creación de falsificaciones por parte de personas no expertas ya que con ciertas habilidades informáticas y con una tarjeta de gráfica adecuada se puede empezar a manipular imágenes para generar todo tipo de vídeos no consentidos. A partir de un *selfie* o un vídeo de una persona es posible crear unas imágenes pornográficas no consentidas.

Los amateurs que quieren crear *deepfakes*, vídeos o imágenes alterados digitalmente para situar a alguien en un escenario falso, encuentran facilidades al contar con el acompañamiento de comunidades o fórums de creadores experimentados, que les inician. "Encontramos que la mayoría de comunidades de creación y foros estaban alojadas en webs *deepfakes* pornográficas, sitios webs basados en foros incluidos Reddit, 4chan, 8chan y Voat. Algunas de estas webs como 4chan and 8chanson conocidas por albergar actividades ilegales y poco éticas" (Ajder, et. alt. 2019, p.4).

En poco tiempo, estos vídeos sexuales no consentidos alcanzarán una mayor calidad y serán más realistas y convincentes. Circunstancias que imposibilita distinguirlos de los contenidos reales y etiquetarlos como falsos, dejando a la persona falsificada en una posición de total indefensión. Si bien con la tecnología *deepfake* pueden generarse contenidos inofensivos, satíricos, de humor, etc. El Deeptrace Labs reveló que el 96 % del contenido *deepfake* en Internet era pornografía no autorizada. Existen millones de imágenes y vídeos pululando en internet que pueden utilizarse para dañar a las mujeres (Ajder, et. alt. 2019).

Las imágenes obscenas y degradantes ejemplifican otra forma de posesión y propiedad de la persona conocida sobre la que se ejerce esa manipulación y que se intensifica si la persona es famosa. El placer y la gratificación se consigue al explotar sin su consentimiento las identidades sexuales inventadas de las personas agredidas. Como van der Nagel (2020) argumenta, que los rostros de las mujeres, al ser tratados como un recurso digital que

se puede editar en cuerpos sexuales gracias a la IA, "se refuerza la idea de que las mujeres existen como objetos sexuales." (p. 3). Objetos sexualizados que consiguen formar parte del imaginario colectivo al viralizarse en redes sociales y al convertirse en la puerta de entrada de la iniciación sexual de adolescentes que construyen su identidad y precisan descubrir nuevas experiencias.

Los d*eepfakes* sexuales no consentidos presentan a las mujeres en roles vergonzosos, degradados, cosificados y humillantes; experimentando la víctima una sensación de que su cuerpo ya no les pertenece al haberse codificado y cosificado para el placer de los demás. Las víctimas, además de todas las secuelas sicológicas, tienen dificultades para continuar utilizando los servicios online, conseguir o mantener un trabajo o sentirse seguras.

La pornografía *deepfake* es un fenómeno que se dirige y daña casi exclusivamente a las mujeres y que encuentra un altavoz en los medios de comunicación y en las redes sociales que amplifican la angustia y desesperación de la persona agredida que no dispone de medidas legales adecuadas para luchar contra las falsificaciones creadas con imágenes sexuales no consentidas, encontrándose en una total indefensión al ser desatendidas por las instituciones y las leyes.

Los medios de comunicación ampliadores del ciberacoso

Los modelos pre-entrenados de IA generativa (GPT) han sido adoptados por millones de personas en el planeta en un tiempo récord, situación que está propiciando una tensión y un desconcierto entre sectores importantes de la población. La preocupación reside, entre otros efectos, en que los *deepfakes* aceleran la desconfianza ya erosionada en las cosas que leemos, escuchamos y vemos, ya que se desdibuja la frontera entre la realidad y la ficción.

En este contexto, los medios de comunicación atrapados en la lógica de la rentabilidad a corto plazo y la dinámica de espectacularización de la información no aportan a la opinión pública los elementos clarificadores de las controversias en curso relativas a los efectos de la IA. Las rutinas

productivas de la inmediatez y la dinámica del caza click no propician opiniones reflexivas con argumentarios sosegados que descubra los pros y los contras de la adopción de los sistemas de IA generativos. En el caso de los *deepfakes* sexuales no consentido, si bien los medios de comunicación dispensan atención a este fenómeno, el tratamiento que efectúan descontextualizado sin incidir en las causas profundas que los generan no ayudan a entender los desafíos en curso y a encontrar las posibles vías de solución.

Los medios informativos sí han presentado como un problema serio las falsificaciones y manipulaciones destinadas a influir en las elecciones, el fraude en los negocios o los intentos para alterar la seguridad o soberanía nacional. El análisis realizado por Gosse y Burkell (2020) encontró que las informaciones de los medios se centraron de manera preferente en el uso negativo de *deepfakes* políticos, económicos o de seguridad y confirió una escasa atención al daño causado por la creación de *deepfakes* sexuales, a pesar de la abundancia apabullante de este tipo de contenidos en las redes.

Esta lacra social de violencia contra las mujeres no encuentra en el espacio público una actuación adecuada que las proteja contra los abusos. Además, el tratamiento dispensado por los medios de comunicación contribuye al efecto cascada, al propagar las falsificaciones sin atender a las características del delito y a las necesidades de las víctimas. La estrategia mediática acrecienta el dolor de las mujeres difamadas con la difusión repetida de imágenes y videos no consentidos de carácter sexual. Unos daños asociados a la publicitación de las informaciones de *deepfakes* sexuales que son difíciles de reparar, aunque se retiren dichos contenidos de la esfera mediática a posteriori.

La víctima, que puede ser conocida o no, debe enfrentarse a un juicio auspiciado por los medios con un alto impacto, sometiéndola al escrutinio de la opinión pública, donde los medios online actúan como si fueran los garantes del proceso y a menudo se atribuyen un rol acusatorio. Un proceso donde se etiqueta a la víctima en una situación vergonzante y de total indefensión. Eso se produce debido a que la tecnología digital y los medios online

promueven "la participación del público en el control social y, por lo tanto, puede ayudar a alterar las relaciones de poder tradicionales en los procesos de avergonzar y etiquetar" (Mo y Grossklags, 2022, p.5).

Los y las periodistas deben recordar que la tarea de los medios es autentificar los contenidos y detectar e identificar los *deepfakes*, sabiendo de la dificultad que dicha tarea conlleva al desarrollarse nuevas técnicas para eludir los controles. Y, también, ser conscientes que al escribir sobre *deepfake* sexuales se puede amplificar el daño causado por la propia falsificación. Una situación que se agrava si no se pone el énfasis en la responsabilidad de las personas que utilizan las herramientas técnicas para perjudicar a terceros. Las informaciones no pueden presentarse como si fuera un daño colateral que debemos aceptar inevitablemente si queremos sumergirnos en el ecosistema digital. Ni se puede exculpar al perpetrador de la agresión, ni justificar a los difusores y consumidores que contribuyen al daño causado y fomentan la culpabilidad de las víctimas al ser presentadas como ilusas que han permitido esa usurpación de sus imágenes procedentes de los múltiples espacios digitales o de sus propios perfiles de las redes sociales.

Los *deepfakes* son una forma relativamente nueva de desplegar la violencia de género, aprovechando la inteligencia artificial para explotar, humillar y acosar a través de la táctica milenaria de despojar a las mujeres de su autonomía sexual[5]". El informe "*The State of Deepfakes 2019 Landscape, Threats, and Impact"* identificó la prominencia de la pornografía *deepfake* no consentida, que representó el 96 % del total de videos *deepfake* en línea y detectó "los cuatro principales sitios web dedicados a la pornografía falsa recibieron más de 134 millones de visitas de vídeos de cientos de celebridades femeninas de todo el mundo. Esta audiencia significativa demuestra un mercado para los sitios web que crean y alojan pornografía falsa, una tendencia que seguirá creciendo a menos que se tomen medidas decisivas" (Adjet et. alt. 2019, p. 5).

5. Ver el artículo de Suzie Dunn de 2021. Women, Not Politicians, Are Targeted Most Often by Deepfake Videos. https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/. Consultado: 10 de Junio de 2022.

Todo y ser el ciberacoso y los *deepfake* un problema detectado de dimensiones colosales no se dispone de protocolos efectivos para su tratamiento en los medios y tampoco una reglamentación efectiva que proteja a las víctimas y les ayude a mitigar los efectos de las agresiones en el entorno digital. El fenómeno de agredir mediante las plataformas digitales es antiguo y la encuesta sobre la violencia contra las mujeres de 2012, ya detectó que el ciberacoso estaba muy extendido entre las mujeres de la Unión Europea, ya que 1 de cada 20 había experimentado este tipo de prácticas contra ellas[6].

No se encontrarán vías para su erradicación sino se establecen mecanismos de protección de las víctimas contra los *deepfake* sexuales no consentidos, ya que una ausencia de ayuda constituye una injusticia y una nueva agresión que deja la víctima en una situación de total indefensión.

El tímido intento de regulación: arbitrar en los conflictos con perspectiva de género

En España, existe la Carta de Derechos Digitales que no posee carácter normativo como se anuncia en el texto. En concreto, nos fijaremos en dos preceptos referidos al entorno digital: el primero se refiere a los derechos de libertad, que incluye: derechos y libertades en el entorno digital, el derecho a la identidad en el entorno digital, derecho a la protección de datos, derecho al pseudo anonimato, derecho a la persona a no ser localizada y perfilada, derecho a la ciberseguridad y derecho a la herencia digital. Unos derechos que en ningún caso se cumplen en la actualidad.

El segundo contempla los derechos entorno a la igualdad en el entorno digital: el derecho a la igualdad y a la no discriminación, el derecho de acceso a internet, la protección de las personas menores de edad, la accesibilidad universal, las brechas de acceso al entorno digital. Unos principios que tampoco se cumplen y que al no ser normativos quedan en papel mojado, ya que no protegen a millones de mujeres afectadas por el ciberacoso y los *deepfakes* sexuales. Tampoco se contemplan "los derechos de nueva generación" que

6. Ver. EU. (2015). Violence against women: an EU-wide survey.

permiten "existir o no existir digitalmente, el derecho a la identidad y reputación digital, el derecho al olvido en internet y en redes sociales, el derecho a la desconexión digital, el derecho a establecer el propio legado digital, el derecho a ser protegido en la integridad personal frente al a tecnología, el derecho a disponer de una última instancia humana en las decisiones de sistemas de inteligencia artificial, el derecho a la actualización de las informaciones. Unos derechos sobre el papel que precisan incluirse en las normativas correspondientes y en las declaraciones de derechos humanos.

Por otra parte, una normativa a nivel nacional tiene poca efectividad sino se despliega una de ámbito internacional, ya que muchos de los sitios web eluden a los países con normativas más restrictivas y se alojan en "paraísos cibernéticos" donde no necesitan rendir cuentas y eluden cualquier responsabilidad. En esta dimensión trabaja la Unión Europea (UE) desde 2018 para construir un modelo de gobernanza de la IA que sea un marco de referencia común y "que proteja los derechos y libertades que caracterizan a las democracias europeas" (Ortíz, 2023, p.8). A su vez, el Grupo de Expertos de Alto Nivel de la UE establece cuatro principios éticos relativos a los algoritmos, que son: el respeto por la autonomía humana, la prevención de daños, la equidad y el entendimiento (High-Level Expert Group on AI, 2019).

Sin embargo, a pesar de estar en la agenda europea el despliegue de normativas sobre IA por su propia naturaleza, y debido a los constantes desarrollos, es difícil y complejo establecer una legislación consensuada. Todo y así, la Unión Europea preparó un proyecto de ley sobre la IA que se aprobó en junio de 2023 por el Parlamento Europeo. El trámite siguiente es pasar por el Consejo Europeo para atender las distintas consideraciones de cada país miembro. El documento incluye a los sistemas de IA generativa como ChatGPT y se clasifica a la IA en distintos niveles de peligrosidad según su uso. A su vez, los y las europarlamentarias "han impulsado una prohibición adicional sobre el uso de la IA para sistemas de vigilancia biométrica."[7] Las previsiones son que la ley se apruebe a finales de 2023 o principios de 2024

7. https://cincodias.elpais.com/economia/2023-06-14/la-primera-ley-europea-sobre-inteligencia-artificial-se-amplia-para-incluir-a-chatgpt-y-entra-en-su-recta-final.html. Consultado el 6 de Junio de 2023.

con un periodo de gracia de dos años para que las empresas puedan adaptarse a la nueva normativa. Un laxo calendario que contrasta con la rápida evolución de los sistemas de IA.

En febrero de 2023, Sam Altman, director ejecutivo de Open AI, el desarrollador de ChatGPT, dijo que "los beneficios de las herramientas que hemos desarrollado hasta ahora superan ampliamente los riesgos"[8], y que la regulación sería fundamental. Entre las vías de control propuso la formación de una agencia estadounidense o global que otorgaría licencias a los sistemas de inteligencia artificial más potentes y que debería tener la autoridad para retirar esa licencia y garantizar el cumplimiento de las normas de seguridad. Aún así, para el Congreso estadounidense, elaborar reglas claras sobre la IA no es una prioridad, al contrario de lo que está haciendo el Parlamento Europeo.

La preocupación en la ciudadanía sigue y un grupo de expertos en IA, periodistas, legisladores e individuos de la sociedad civil han promovido la "*Statement on AI Risk AI experts and public figures express their concern about AI risk*" para potenciar el debate y crear conocimiento común. La declaración apunta los peligros de la IA y marca como prioridad mundial "mitigar el riesgo de extinción a causa de la IA" junto con otros riesgos a "escala social, como las pandemias y la guerra nuclear".[9]

A título de conclusión

El diálogo entre especialistas de distintas disciplinas de ámbitos científicos, del campo de las humanidades y las ciencias sociales pueden dar sus frutos al presentar los desafíos que afronta la especie humana derivados de la proliferación de herramientas de IA disponibles en todos los sectores y sentar las bases para desarrollar una legislación que atienda principios éticos y garantice una justicia a la ciudadanía. Unos marcos normativos, regulatorios y

8. El creador de ChatGPT pide al Congreso de EE.UU. que regule el desarrollo de la IA. https://efe.com/ciencia-y-tecnologia/2023-05-16/sam-altman-pide-al-congreso-de-ee-uu-que-regule-el-desarrollo-ia/ Consultado el 23 de Mayo de 2023.
9. Ver: https://www.safe.ai/statement-on-ai-risk#open-letter. Consultado el 20 de Junio de 2023.

éticos efectivos que deben considerar las fronteras que modelen los desarrollos de la IA y donde se atiendan las necesidades integrales de la ciudadanía (Watkins y Human, 2022).

Es el momento de actuar con arrojo, sabiduría y valor, buscando la cooperación multidisciplinar y considerando las aportaciones de los enfoques feministas sobre IA con el propósito de estimular un diálogo social que nos ilustre acerca de cómo aprovechar las ventajas de los nuevos desarrollos de la IA, sin pagar un alto precio que signifique un retroceso en los derechos humanos a nivel individual y colectivo.

La confianza en la IA se asienta en tres componentes como se recoge en el sumario ejecutivo de la guía de la Comisión Europea "*Ethics Guidelines for Trustworthy AI*": la legalidad, la ética y la robustez. Los y las promotoras, desarrolladoras y usuarios/as deben cumplir las leyes y reglamentos para asegurar que no se subvierten los principios y valores éticos. Finalmente, el sistema de IA debe ser robusto tanto técnicamente como socialmente ya que los sistemas de IA, "incluso si las intenciones son buenas, pueden provocar daños accidentales" (Comisión Europea, 2019, p.5).

Sabemos que la misoginia no está "incorporada" a la tecnología y que la decisión de usar la tecnología para crear falsificaciones sexuales profundas recae en los usuarios que sí reflejan la cultura misógina dentro de la cual se implementa la tecnología (Burkell y Gosse, 2019). Y que "las raíces del sesgo de género en los sistemas de toma de decisiones basados en IA no son tecnológicos y, por lo tanto, las soluciones técnicas pueden no ser suficientes" (...) "Los científicos computacionales han desarrollado técnicas matemáticas para detectar y mitigar los sesgos en los algoritmos" (Nadeem et al., 2022, p.13). En consecuencia, Nadeem et al. (2022) explica que:

> La conceptualización del sesgo de género en los sistemas de toma de decisiones basados en IA debe ser "multicapa, multidimensional y socio-técnico, y el desarrollo y la implementación de los sistemas requieren una combinación e integración de enfoques técnicos, organizacionales y sociales. (p.13)

Los sistemas generativos, como GPT, no pueden ser opacos y deben basar su actividad en la transparencia que permita detectar los contenidos falsos y evitar el contenido ilegal como los *deepfake* sexuales no consentidos.

Es esencial encontrar un equilibrio regulatorio entre la libertad de expresión y la creatividad de todo tipo de contenidos, como textos, audios, imágenes, vídeos, etc., sin correspondencia con el mundo real que pueden generar los sistemas de IA generativa. Estamos por fomentar la creatividad sin perjudicar a terceros que no han consentido que su imagen o su trabajo se utilice por personas que buscan un enriquecimiento a costa de los demás o infringen un daño irreparable al manipular la imagen personal de alguien con objetivos espurios y que son a ojos del público creíbles.

Finalmente, la academia por su prestigio y neutralidad debería encargarse de orientar y articular el debate social en su dimensión política y jurídica, junto a otras instituciones públicas y privadas, sobre las ventajas e inconvenientes de el IA a nivel planetario, generando investigaciones independientes que analicen en profundidad este fenómeno con el propósito de aportar elementos científicos e históricos para promover el diálogo, la comprensión y el conocimiento necesario para evaluar los futuros desarrollos, incentivar los beneficios de la IA y encontrar los remedios a los posibles perjuicios causados a la ciudadanía.

Bibliografía citada

Ajder, H.; Patrini, G., Cavalli, F. y Cullen, L. (2019). *The State of Deepfakes: Landscape, Threats, and Impact, Deeptrace.*

Agarwal, P. (2020). Gender bias in STEM: Women in Tech still facing discrimination. *Forbes*. Retrieved from https://www.forbes.com/sites/pragyaagarwaleurope/2020/03/04/gender-bias-in-stem-women-in-tech-report-facing-discrimination/?sh=72c9e78670fb

Burkell, J., y Gosse, C. (2019). Nothing new here: Emphasizing the social and cultural context of deepfakes. *First Monday*, 24(12). https://doi.org/10.5210/fm.v24i12.10287

EU. (2015). *Violence against women: an EU-wide survey.*

EU (2022). *Bias in Algorithms. Artificial Intelligence and Discrimination.*

García Canclini, N. (2020). *Ciudadanos reemplazados por algoritmos*. Bielefeld University Press.

Gander, K. (2016). The people who photoshop friends and family onto porn. *Independent* (13 October), at https://www.independent.co.uk/life-style/love-sex/porn-photoshopping-4chan-family-friends-superimposed-into-sex-scenes-world-a7358706.html.

Gosse, Ch. y Burkell, J. (2020). Politics and porn: how news media characterizes problems presented by deepfakes**.** *Critical Studies in Media Communication,* 37(5), 497-511. https://doi.org/10.1080/15295036.2020.1832697

Harari, Y. N. (2019). *21 lecciones para el siglo XXI*. Penguin Random House Grupo Editorial. 4ª edición.

High-Level Expert Group on AI (2019) *Ethics guidelines for trustworthy AI*. https ://ec.europ a.eu/futur ium/en/ai-allia nce-consu ltati on/guide lines #Top

Kurzweil, R. (1985). What Is Artificial Intelligence Anyway? As the techniques of computing grow more sophisticated, machines are beginning to appear intelligent—but can they actually think? *American Scientist*, *73*(3), 258–264. http://www.jstor.org/stable/27853237

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in Machine learning. *2018 IEEE/ACM First international workshop on gender equality in software engineering,* Gothenburg, Sweden. doi: https://ieeexplore.ieee.org/document/8452744

Maddocks, S. (2020). 'A deepfake porn plot intended to silence me': Exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, 1–9. https://doi.org/10.1080/23268743.2020.1757499

Manovich, L. (2017). Los algoritmos de nuestras vidas, en CIC. *Cuadernos de Información y Comunicación,* 22, 19-25. http://dx.doi.org/10.5209/CIYC.55960

Masiero, S., Aaltonen, A. (2020). Gender bias in IS research: A literature review. In *Proceedings of the 41st International Conference on Information Systems*, Hyderabad, India. doi: http://dx.doi.org/10.2139/ssrn.3751440

Mo, Ch. y Grossklags, J. (2022). Social Control in the Digital Transformation of Society: A Case Study of the Chinese Social Credit System. *Social Sciences,* 11: 229. https://doi.org/10.3390/socsci11060229

Nadeem, A., Abedin, B., Marjanovic, O. (2020). Gender bias in AI: A review of contributing factors and mitigating strategies. In *Proceedings of the Australasian Conference on Information Systems*, Wellington, New Zealand. https://aisel.aisnet.org/acis2020/27

Nadeem,A., Marjanovic, O., y Abedin, B. (2022). Gender bias in AI-based decision-making systems: a systematic literature review. Australasian Journal of Information Systems, 26, 1-34. oi: https://doi.org/10.3127/ajis.v26i0.3835

Ortíz de Zárate Alcarazo, L. 2023. Sesgos de género en la inteligencia artificial. *Revista de Occidente,* 502, 5-20.

Rayón, M.C. (2022). La regulación de derechos humanos en el entorno digital: ¿es necesaria la actualización de las declaraciones de derechos para crear un nuevo marco de referencia para la humanidad?. In D'Ávila Lopes, A.M.; Paredes, F.; Pereira. A.J. y Passos, A. (Org.). *Neurodireito, Neurotecnologia e Direitos Humanos*. Livraria do Advogado Editora.

Rivadeneira, C. C., Rivera Cuellar, D., & Riomaña, K. (2023). ¿Por qué nuestros datos importan? Conceptos claves sobre los impactos de la inteligencia artificial en la protección de los datos personales y sus marcos de regulación. *Revista Lecciones Vitales*, (I), lv0104. https://doi.org/10.18046/rlv.2023.6122

van der Nagel, E. (2020). Verifying images: Deepfakes, control, and consent. Porn Studies, 1–6. https://doi.org/10.1080/23268743.2020.1741434

Watkins, R. y Human, S. (2022). Needs-aware artificial intelligence: AI that 'serves [human] needs'. *AI Ethics,* 3, 49–52 (2023) https://doi.org/10.1007/s43681-022-00181-5

POR QUE FALAR DE RAÇA QUANDO FALAMOS DE DADOS PESSOAIS, INTELIGÊNCIA ARTIFICIAL E ALGORITMOS?

Johanna K Monagreda
/ Universidade Federal de Minas Gerais

> *O racismo continua sendo a base para o colonialismo, inclusive o colonialismo digital.*
> Deivison Faustino e Walter Lippold

Introdução

Da mesma forma que as opressões estruturam a vida em sociedade, estruturam também o mundo digital. As decisões sobre tratamento de dados pessoais e uso de tecnologia são tomadas por agentes sociais com interesses, valores, visões, posições de poder, etc. Nestas condições, as representações, perspectivas, prioridades, oportunidades e premissas dos grupos oprimidos ou em situação de desvantagem acabam sendo definidas pelo discurso, interesses e decisões dos grupos sociais dominantes, tomadas, artificialmente, como neutras e universais.

É sabido que, como fenômeno biológico, a raça não existe, mas, assim como o gênero, são construções sociais com implicações simbólicas e materiais na vida das pessoas. Tanto na sua dimensão histórico-material, quanto na sua dimensão simbólica, o machismo e o racismo se combinam para produzir e perpetuar a exclusão social, econômica e política de sujeitos racializados na subalternidade. Daí que Lélia González insista em tratar o racismo e o sexismo como "um duplo fenômeno" ou

triplo fenômeno, trazendo os conceitos de racismo-patriarcal e capitalismo-racista-patriarcal para salientar o caráter imbricado que adquirem as opressões, e o caráter racial e de gênero na organização da exploração capitalista (Gonzalez, 1984).

Pelo seu caráter estruturante, a categoria social raça nem sempre precisa ser explicitamente capturada para contribuir com a reprodução ou produção de injustiça. Uma vez que a própria categoria se relaciona com muitas outras condições materiais e simbólicas que permitem tanto inferir raça quanto direcionar a discriminação, nem sempre é preciso explicitar a raça para que esta influencie o resultado da aplicação de determinada tecnologia, ou inclusive justifique a invenção de uma tecnologia.

São diversos os autores que têm escrito sobre a relação intrínseca entre mundo digital, tecnologia e a racialização da sociedade. Através de diversos conceitos como colonialismo digital, colonialismo de dados, racismo codificado, racialização digital (Lippold & Faustino, 2022); algoritmização do racismo, microagressões algorítmicas (T. Silva, 2022); algoritmos de opressão (Noble, 2018); racismo algorítmico (Noble, 2018; Silva, 2022); Black Opticon (Allen, 2022); racismo plataformizado (Matamoros-Fernández, 2017), diversos autores têm se proposto a desvendar as formas racializadas de controle, vigilância, discriminação, exclusão, exploração e violência mediada pelo uso de tecnologia orientada por dados.

Centralizar esse tipo de análise crítica sobre a sociedade no debate sobre internet, plataformas digitais, redes sociais, tecnologia, inteligência artificial, uso de dados pessoais, etc., é relevante porque instiga a tirar o véu de neutralidade, objetividade e infalibilidade nas discussões sobre internet e uso de tecnologia. De fato, tanto na sua dimensão simbólica, quanto na sua dimensão material, os problemas estruturais da sociedade se fazem presentes e atravessam o mundo digital.

Em muitos casos, a opacidade caracteriza a captura de dados pessoais, especialmente quando processados por grandes corporações de tecnologia, as

chamadas *Big Techs*. Ainda, um dos maiores desafios está em entender o protagonismo da decisão humana em situações que envolvem uso de tecnologias de inteligência artificial e que os atores sociais imprimem seus valores nos funcionamento deste tipo de tecnologia, nas diferentes etapas (Noble, 2018; Prince & Schwarcz, 2020; T. Silva, 2022).

Levando em consideração todo esse contexto, este ensaio apresenta uma proposta de organização de diversos argumentos que salientam os possíveis efeitos nocivos da utilização de tecnologia de inteligência artificial para a população racializada negra, a partir das decisões sobre uso e tratamento de dados pessoais. Ao mesmo tempo, traz uma proposta de organização do campo de discussão sobre uso de tecnologia movida por dados, e manutenção da dominação racista-capitalista-patriarcal e atualização do racismo.

A discussão se organiza a partir de sete riscos com efeitos acentuados sobre a população negra, envolvendo uso de dados pessoais ou tecnologia orientada por dados: (1) perda de privacidade e apropriação indevida de dados pessoais sensíveis, (2) mercantilização da vida cotidiana e dataficação da pobreza, (3) reprodução e automatização do racismo, (4) perfilamento racial e discriminação, (5) vigilância excessiva e criminalização, (6) impacto nos processos de subjetivação e (7) apagamento do caráter político das problemáticas sociais.

Argumenta-se que a reprodução do racismo a partir de algoritmos que se alimentam de informações sobre a sociedade é inevitável. Algoritmos são códigos baseados em padrões socialmente determinados (Lippold & Faustino, 2022). Portanto, as respostas lógicas, a partir da informação subministrada, tenderão a indicar o padrão histórico de marginalização e exclusão da sociedade. Defende-se que a possibilidade de corrigir a reprodução de discriminações pode ser encontrada na adoção de uma linguagem moral de reparação nas decisões envolvendo uso de dados pessoais e tecnologia de inteligência artificial.

1. Perda de privacidade e apropriação indevida de dados sensíveis

A maior exposição das pessoas, suas vidas, conversas privadas, preferências, deslocamento, etc. tem produzido um aumento da capacidade de monitoramento, por parte do mercado, e de vigilância e controle por parte do Estado, e aumenta os riscos de apropriação indevida de dados pessoais.

Esses dados pessoais que alimentam o *Big Data* podem ser capturados por grandes corporações de tecnologia sem que sequer se tenha conhecimento, mas também por câmeras de reconhecimento facial em espaços públicos, no trânsito pela internet, pela burocracia estatal para garantir previdência social ou outros direitos básicos, nas farmácias, nas delegacias, etc. Independentemente da forma de captura e armazenamento dos dados, estes podem acabar sendo tratados com fins de mercado e governamentais que colocam em risco o direito à privacidade.

> Representando o impulso para coletar grandes quantidades de dados analisáveis com a finalidade de discernir informações valiosas, o *Big Data* apresenta grandes desafios ao ideal de privacidade pessoal, que inclui direitos de acesso limitado a informações pessoais e controle sobre informações pessoais (Allen, 2016, s/p).

A garantia do direito à privacidade tem uma dimensão racial, de gênero e classe que precisa ser centralizada nas análises sobre riscos do uso abusivo de dados pessoais. Tensionar esse direito fundamental, junto com uma compreensão do caráter histórico, estrutural e interseccional das desigualdades na sociedade contemporânea, exige observar os distintos aspectos em que pessoas negras encontrarão barreiras, materiais e simbólicas, para exercer o direito à privacidade e ter real controle sobre as suas informações pessoais.

Historicamente raça é um marcador de discriminação, é uma categoria socio-política que surge para a invenção, classificação e hierarquização de grupos sociais, que organiza e legitima a exploração, a exclusão, a marginalização, a violência. Pelo caráter histórico-estrutural e pelo potencial de

colocar as pessoas pertencentes a determinados grupos sociais em uma situação de vulnerabilidade maior, os dados sobre raça são dados pessoais sensíveis, quer dizer, são dados que requerem um cuidado especial para seu tratamento, tornando relevante debater sobre quando e como a datificação e maior automatização da sociedade pode produzir um acirramento da capacidade de vulneração do mercado e do Estado sobre a população negra.

O termo *Black Opticon*, cunhado por Anita Allen (2022), justamente faz referência aos custos sociais que recaem de forma desproporcional sobre a população negra (e outros grupos marcados pela história de colonização e escravidão), pelo uso de tecnologia de inteligência artificial e plataformas *online*, a partir de uma leitura das falhas à privacidade e à proteção de dados como vulnerações aos direitos civis.

Ao mesmo tempo que a interseção raça-classe-gênero aumenta as vulnerabilidades às garantias do direito à privacidade e os riscos de apropriação indevida de dados pessoais, especialmente dados sensíveis, a vedação da coleta de dados sobre raça, como medida antidiscriminatória, parece pouco eficaz como se verá em alguns exemplos ao longo do texto. Pelo contrário, argumenta-se aqui que, em muitos casos, o tratamento de *proxies* raciais pode ser indispensável para atingir resultados em prol de uma maior igualdade racial.

2. Mercantilização da vida cotidiana e datificação da pobreza

A crescente automatização da sociedade elevou a capacidade de capturar, em dados, as interações, sentimentos, comportamentos e a intimidade das relações humanas com potencial mercadológico (Zuboff, 1994; 2019). Cada vez mais os dados pessoais podem ser armazenados, compartilhados, vendidos e/ou utilizados para a tomada de decisões dos mais distintos âmbitos da vida em sociedade e, inclusive, para a produção e aprimoramento de tecnologias de inteligência artificial.

As tecnologias de inteligência artificial, especialmente as tecnologias de controle e preditiva, são produtos a serem vendidos e que se tornam mais

rentáveis com o aproveitamento dos dados pessoais, o que produz tensão com outros direitos fundamentais, que dependem tanto da privacidade quanto da garantia de uso não abusivo ou discriminatório dos dados pessoais.

Na mineração de dados, nas buscas por legalização da comercialização de dados pessoais e na competição pelo tratamento de dados sensíveis e sigilosos como dados de saúde, educação, justiça por parte de grandes corporações de tecnologia se evidencia a importância dos dados pessoais na produção de capital. Para alguns autores a dataficação da sociedade é também a mercantilização das relações humanas (Silveira, 2017; Zuboff, 2019; Lippold & Faustino, 2022).

> Essas operações de aprovisionamento visam nossa personalidade, nosso humor, nossas emoções, nossas mentiras e nossas fragilidades. Todos os níveis de nossa vida pessoal seriam automaticamente capturados e compactados em um fluxo de dados destinado às linhas de montagem que produzem a certeza. Realizado sob o disfarce da "personalização", grande parte desse trabalho consiste em uma extração intrusiva dos aspectos mais íntimos de nosso cotidiano (Zuboff, 2019, s/p).

Alguns autores salientam como as condições socioeconômicas são determinantes nos processos de mercantilização da cotidianidade e dataficação da pobreza (Taylor, 2017; Dencik et al., 2019; Masiero & Das, 2019; Martin & Taylor, 2021). As práticas históricas de monitoramento, vigilância e classificação de pessoas negras tornam esse grupo ainda mais vulnerável a mercantilização e dataficação das informações pessoais.

Pessoas negras, especialmente mulheres, estão mais expostas a terem suas informações pessoais administradas pela burocracia estatal, tanto como requisito para o acesso a políticas sociais ou de distribuição de renda, quanto como parte da prática de vigilância excessiva que recai sobre pessoas negras e marginalizadas. Por outro lado, abdicar do direito à proteção de dados pode parecer bem atraente frente às carências materiais quando implica acessos, descontos ou outras formas de monetização dos dados pessoais. Ainda mais em uma dimensão material concreta, ter acesso a ferramentas

de segurança de dados ou a dispositivos eletrônicos mais seguros implica um custo econômico que pessoas de escassos recursos podem ter dificuldades em assumir.

Algumas pesquisas sobre dataficação e acesso a políticas públicas permitem entrever que estereótipos e preconceitos incidem negativamente sobre a compreensão de pessoas negras e marginalizadas como sujeitos de direito, pessoas que têm direito a serem informadas sobre a finalidade, necessidade, possível transferência e compartilhamento de seus dados pessoais com terceiros, riscos de vazamento de dados, etc.

Argumenta-se aqui que a noção de corpos impolíticos (*body impolitic*), que recai sobre corpos negros, se apresenta como mais uma limitante ao conceito do consentimento como uma base legal suficiente para a proteção do tratamento dos dados pessoais. A noção de corpos impolíticos foi apresentada por Charles Mills (1997) para explicar como a racialização dos corpos se estabelece através de uma partição ontológica entre pessoas (brancas) e subpessoas (não-brancas) que coloca sujeitos negros como "[...] corpos impolíticos cujas donas e donos são julgados incapazes de formar ou entrar totalmente em um corpo político" (Mills, 1997, p. 53). Isso significa dizer que o *status* moral de pessoa e de pessoa de direito não está dado para pessoas negras e precisa ser constantemente reivindicado.

Isto tem implicações no entendimento da responsabilidade do Estado, e também do mercado, para com as pessoas, as comunidades e os territórios negros (e indígenas), ao tempo que impacta nas dinâmicas das interações públicas. Corpos impolíticos são corpos aos quais não correspondem a administração do poder, portanto a própria condição de titular dos seus dados pessoais pode se ver enfraquecida frente a essa construção ontológica. Em dois sentidos, tanto no desinteresse por explicar o tratamento dos dados pessoais, quanto na internalização da ausência de direito pelas pessoas negras.

A pesquisa da InternetLab (Fragoso et al., 2021; Valente et al., 2021), ajuda a ilustrar esse ponto. Um dos múltiplos achados dessa pesquisa foi que servidores públicos tenderiam a negligenciar a explicação sobre uso dos dados

pessoais para as possíveis beneficiárias do programa Bolsa Família. A pesquisa mostra que, em se tratando da coleta de dados pessoais, as ideias de mulheres negras e carentes como sujeitos que pretendem fraudar o sistema prevaleceu na dinâmica de tratamento dos dados por sobre o direito do titular à proteção do dado pessoal.

Por fim, a disposição de câmeras de reconhecimento facial em espaços públicos, especialmente em espaços de serviços, como transporte, hospitais, ilustra muito bem o impacto diferenciado que adquire a coleta expansiva de dados pessoais, para alguns sujeitos, quando vinculado à prestação de um serviço público. Ainda, o exemplo permite tecer considerações sobre as formas em que o monitoramento ou a vigilância excessiva reflete práticas históricas de criminalização da população negra e marginalizada.

3. Reprodução e automatização dos vieses existentes na sociedade

> *Today, we pose this question to new powers*
> *Making bets on artificial intelligence, hope towers*
> *The Amazonians peek through*
> *Windows blocking Deep Blues*
> *As Faces increment scars*
> *Old burns, new urns*
> *Collecting data chronicling our past*
> *Often forgetting to deal with*
> *Gender race and class, again I ask*
> *"Ain't I a Woman?"*
> AI, Ain't I A Woman? - Joy Buolamwini

Diversas pesquisas salientam a reprodução dos vieses existentes na sociedade como uma das falhas do uso de tecnologia de inteligência artificial. Os vieses raciais expressam uma tendência recorrente à produção de erros e à emissão de juízos distorcidos sobre a população não-branca. Os vieses raciais acontecem porque a categoria raça remete a estereótipos e preconceitos que conduzem a pensamentos, interpretações e decisões errôneas.

Identificar a causa do viés na utilização de tecnologia de inteligência artificial, quando involuntário, representa um desafio por diversos fatores: pela complexidade do funcionamento dos algoritmos, pelo opacidade da tecnologia, pelo ocultamento como segredo comercial, etc., mas quando o resultado do uso de determinada tecnologia expressa um dano, intencional ou não, sobre a população negra, sobre mulheres, pessoas transexuais, ou sobre pessoas com deficiência, entre outros grupos minoritários, estamos em evidência da reprodução de vieses.

Argumenta-se aqui que a automatização dos vieses existentes na sociedade pode acontecer de duas formas: como uma falha da tecnologia, por exemplo quando o conjunto de dados pessoais utilizados para treinamento da máquina reproduz o "padrão de normalidade" da sociedade; ou quando a máquina aprende a reproduzir a lógica discriminatória de funcionamento da sociedade. Enquanto o primeiro tipo de viés pode ser considerado uma falha de falta de diversidade ou de insuficiência de dados de treinamento, o segundo expressa o resultado de uma decisão lógica embasada na injustiça social. Em ambos os casos, há responsabilidade humana e relações de poder informando a tecnologia.

Veja-se:

O conjunto de dados de treinamento da máquina reproduz o "padrão de normalidade" pré-existente na sociedade

A raça, ao tempo que inventa o corpo/território negro, inventa também o corpo/território branco. A brancura, no corpo e no espaço, é definida como padrão de normalidade a partir do qual se produz a hipervisibilidade negativa e a invisibilização do corpo/território negro.

São diversas as evidências de que raça e gênero atravessam o funcionamento da tecnologia de inteligência artificial. "Black Desi" e "White Wanda" em 2009 publicaram um vídeo noYouTube[1] denunciando falhas no funcionamento do *software* de rastreamento facial de HP incapaz de identificar e seguir

1. HP computers are racist

o rosto negro. Em 2017, Joy Boulamwini constatou que um determinado sistema de inteligência artificial não conseguia detectar seu rosto de mulher negra, mas funcionava perfeitamente quando ela utilizava uma máscara branca[2]. Em 2018, Buolamwini & Gebru auditaram a precisão de três programas comerciais de classificação por gênero, sendo esses Microsoft, IBM, Face++. Nos três aplicativos, a identificação foi mais precisa quando referida a homens brancos e menos precisa quando referida a mulheres negras.

Em alguns casos, os erros associados ao uso de algoritmos de visão computacional ou câmeras de reconhecimento facial podem entrar nesta categoria, como problemas de precisão no desempenho da tecnologia, que podem acontecer em diferentes etapas, produzindo falhas na detecção do rosto, falha na autenticação ou problemas de identificação como falsos positivos ou falsos negativos.

No entanto, ao ser um problema de precisão com desempenho diferenciado por condição demográfica (raça-gênero), deve conduzir a reflexões sobre como o uso de determinada tecnologia pode reforçar a lógica de opressão presente na sociedade. As tecnologias de análise facial apresentam altas taxas de erro para pessoas com deficiências, pessoas negras, e mulheres, especialmente mulheres negras. As taxas de erros são maiores em pessoas não-brancas e não-masculinas.

Um dos aspectos a partir do qual tem se tentado explicar esse tipo de falhas nas tecnologias de inteligência artificial, especialmente na tecnologia de visão computacional, quando se trata de raça e gênero, é a configuração do conjunto de dados a partir do qual a inteligência artificial é treinada para identificar, classificar e tomar decisões. Segundo isto, é possível explicar as maiores taxas de erros em pessoas não-brancas e não-masculinas, ou com deficiências, por causa dos vieses no treinamento da máquina (Buolamwini & Gebru, 2018; Noble, 2018; T. Silva, 2022) associados à falta de diversidade, o mesmo padrão de exclusão que acontece no mundo *offline* é reproduzido no conjunto de dados de treinamento da máquina.

2. Joy Buolamwini: How I'm fighting bias in algorithms | TED Talk

Muitas dessas pesquisas trabalham com a hipótese explicativa de que a base de dados a partir da qual é treinado o algoritmo pode não ser representativa da heterogeneidade da sociedade. Isso porque o banco de imagens ou o conjunto de imagens (*dataset)* que a máquina usa como referência (*benchmark*) para aprender a identificar rostos humanos seria composto principalmente por homens brancos e em muitos casos de regiões ou estados dos Estados Unidos muito específicas (Center for democracy & tecnology, 2022).

Informações sobre codificação ou funcionamento do algoritmo usualmente não estão disponíveis ou se tornam indecifráveis pela própria complexidade da tecnologia, portanto não é possível afirmar com certeza a origem da falha. Contudo, a partir do resultado, é preciso compreender as tecnologias de inteligência artificial "em vista do contexto em que operam, refletindo não apenas vieses contidos nos *datasets*, mas, também, formas de opressão e exclusão que condicionaram a constituição tanto dos *datasets* quanto do *software,* incluindo ai a pouca diversidade das equipes de desenvolvimento" (T. Silva, 2022, p. 90).

É possível apontar aqui para uma problemática que mulheres, pessoas negras, indígenas, pessoas com deficiência sinalizam como um problema da sociedade, que é a criação de um padrão de normalidade que se argumenta neutral, mas que na verdade revela um lugar social definido, porém propositalmente invisibilizado, que é a localização social ou a perspectiva social do homem-branco-heterosexual-de classe social privilegiada.

Da mesma forma em que o capitalismo racista patriarcal privilegia a presença de homens brancos nos lugares de poder e destaque da sociedade, haveria uma sobre-representação desses homens brancos do norte global nos conjuntos de dados a partir dos quais as máquinas aprendem a ler e interpretar rostos humanos, produzindo algoritmos incapazes de ler com precisão a diversidade da humanidade, incapazes de ler as pessoas que não se encaixam no padrão de normalidade que são a maioria (mulheres, pessoas pretas, pessoas não-binárias, pessoas com deficiência, trabalhadores braçais), contribuindo para invisibilizações ou representações limitadas da diversidade

racial, étnica, de gênero, social, cultural, etc. Erros desse tipo expressam os valores, interesses, as assimetrias de poder dos grupos dominantes.

> Assim como outras modalidades de inteligência artificial, os sistemas algorítmicos com recursos de visão computacional trazem em si valores políticos e estéticos racializados, que se manifestam em invisibilização, hipervisibilização, estereotipização ou mesmo em embranquecimento literal dos indivíduos (T. Silva, 2022, p. 75)

Até aqui argumentou-se que a reprodução de vieses, voluntários ou involuntários, no uso de tecnologia de inteligência artificial poderia ter sua origem em que o banco de dados de referência pode não ser heterogêneo o suficiente para abarcar a heterogeneidade da sociedade produzindo erros de desempenho no sistema. Nesse sentido, algumas soluções caminham para alargar o *dataset* com miras a corrigir as falhas de leitura ou precisão do algoritmo.

Pretende-se avançar agora, para uma segunda explicação mais complexa. A heterogeneidade da sociedade pode estar sendo embutida na tecnologia na forma de estereótipos e preconceitos, de modo que a reprodução do racismo e do sexismo pode não ser uma falha do sistema, mas parte do funcionamento correto de uma tecnologia que opera sob a lógica discriminatória que estrutura a sociedade.

Quando a IA reproduz a lógica racista, sexista, o capacitismo e a xenofobia existente na sociedade

Se é fato que dados de treinamento enviesados produzem erros de precisão do sistema, leituras erradas ou incompletas, também existe um segundo fenômeno na reprodução do racismo, que é quando a tecnologia produz interpretações da realidade que expressam os valores discriminatórios embutidos na sociedade. Ou seja: as respostas da máquina são expressões lógicas, ainda que injustas, do aprendizado classista, racista e sexista da máquina.

Se no primeiro caso a reprodução do racismo pode ser considerado um erro de falta de diversidade ou de dados incompletos; no segundo, a reprodução

de vieses é uma resposta lógica em concordância com a forma em que o racismo (e a interseção de outras opressões) estrutura a sociedade e é codificado, pelos agentes, na tecnologia.

Para a compreensão deste fenômeno reforçamos a compreensão sobre algoritmos como códigos baseados em padrões socialmente determinados (Lippold & Faustino, 2022). Aqui entenderemos os padrões socialmente determinados, tanto na sua dimensão intersubjetiva – implicando a reprodução de valores e premissas que se correspondem com as sociedades racistas, sexistas, heteronormativas, etc., – quanto na sua dimensão estruturante das desigualdades materiais.

Os estudos de Noble (2018) sobre as formas em que plataformas de busca orientadas por algoritmos, como Google, representam mulheres negras, não expressam apenas erros de funcionamento do sistema, ou falhas na leitura da diversidade da sociedade, mas mostram como a tecnologia reproduz representações sobre mulheres negras ancoradas nos vieses racistas e sexistas que recaem sobre esse grupo social e representam os valores que são priorizados nos sistemas automatizados de tomada de decisões, assim como as assimetrias de poder existentes na sociedade.

Gênero e raça são categorias que revelam o posicionamento dentro de uma hierarquia social de poder. Sendo assim, diversos elementos indiretos associados às condições estruturais e simbólicas vivenciadas por mulheres e pessoas negras permitem identificar gênero e raça, mesmo quando a categoria não está sendo nomeada como um dispositivo explícito de discriminação. Além de embutir os valores humanos nas máquinas, por exemplo, através da codificação, desconsiderar os problemas estruturais da sociedade também produz decisões automatizadas enviesadas.

Um estudo publicado pela revista Science denunciou o enviesamento racial de um algoritmo utilizado por diversos provedores de saúde nos Estados Unidos (Lee, 2019). O algoritmo que deveria prever quais pacientes iriam precisar de maiores cuidados médicos acabou beneficiando pessoas brancas em detrimento de pessoas negras em condições de saúde semelhantes.

O critério para a tomada de decisão não era a raça, nem o registro médico, mas o gasto prévio em saúde. Por causa da acumulação histórica de desvantagens, pessoas negras teriam menos condições de investir grandes quantias na sua saúde, mesmo em condições graves. A partir desse critério, aparentemente neutro, e pela interseccionalidade de raça e classe, o algoritmo acabaria reproduzindo as práticas históricas de privatização da saúde e de negligenciamento de atendimento médico da população afro-americana.

A utilização de *softwares* de condenações no sistema de justiça também é um exemplo interessante para se pensar que nem sempre o erro é falta de diversidade nos dados de treinamento da máquina, mas que a heterogeneidade de formas de estar no mundo está sendo incorporada a partir de premissas racistas e classistas. Em 2016, a ProPublica, um meio independente de investigação jornalística, publicou uma pesquisa que evidenciou que um *software* de condenações utilizado nos Estados Unidos estaria enviesado contra réus negros, produzindo condenações maiores por crimes semelhantes quando comparado com as condenações dos réus brancos (Angwin et al., 2016). Diversas pesquisas, algumas delas sintetizadas no livro de O´Neil (2020), mostram que a raça tem sido um fator historicamente decisivo nas condenações da população negra nos Estados Unidos. Portanto, mal poderia se esperar que a utilização de modelos de riscos computadorizados contribuísse a reduzir os vieses humanos nas sentenças. De fato, o que a ProPublica encontrou na sua pesquisa é a reprodução da mesma tendência de criminalização da população negra e pobre, com o agravante de que "o funcionamento de um modelo de reincidência está escondido em algoritmos, compreensíveis somente para uma pequena elite" (O'Neil, 2020, p. 50).

O´Neil (2020) analisou o questionário-base do modelo de reincidência LSI-R, encontrando que o tipo de pergunta para determinar o risco de reincidência e a periculosidade do réu acaba capturando informações sobre a vida da pessoa, ficando evidente que, a partir desses critérios, pessoas privilegiadas serão consideradas menos perigosas por causa do contexto socioeconômico onde estão inseridas, enquanto pessoas negras, latinas e pobres serão

punidas mais pelas condições socioeconômicas adversas, pelo bairro onde moram, pelos lugares que frequentam, pela trajetória de amigos e familiares, do que pelo próprio crime. Dessa forma, estrutura de desigualdade e o sistema punitivo estariam sendo codificadas na tecnologia.

As condenações maiores para réus negros não representam um erro de precisão da tecnologia, mas o funcionamento correto da máquina dentro da lógica racista da sociedade. Os modelos de reincidência se alimentam de dados que refletem a lógica racista e classista da sociedade sobre a criminalidade. Portanto, é esperado que resultem na reprodução da prática histórica de criminalização de pessoas negras.

A partição ontológica do corpo/território que explica a criminalização da pobreza e da negritude também nos permite explicar como a lógica racista da sociedade funciona através da tecnologia. Em 2022, uma ilustração da imagem do Fórmula 1 em que mostra Lewis Hamilton numa favela brasileira foi lida e interpretada pelo algoritmo de moderação de conteúdo do Instagram/Facebook como incitação à violência e exposição de armas. Na análise de T. Silva (2022) sobre a decisão enviesada do algoritmo:

> Na imagem, [...] é possível reconhecer os elementos visuais de favela ou bairro popular, mas nada que indique violência. Porém, no Brasil, a cultura imagética hegemônica privilegia a violência na cobertura jornalística ou nas narrativas ficcionais ambientadas em tais espaços (T. Silva, 2022, p. 121).

A construção histórica do corpo/território negro como espaço de violência, assim sendo insumo para a inteligência artificial, produz uma leitura dentro da lógica racista, onde uma imagem remetendo a favela, remeteria também, dentro da lógica racista, à violência.

Em alguns casos, aquilo que convencionou se chamar “falhas do sistema” ou “uso abusivo dos dados” pode ser melhor entendido como sistemas automatizados que aprenderam a funcionar perfeitamente dentro da lógica racista e sexista da sociedade (Noble, 2018).

Ao se definir vieses racistas como tendências é possível abrir margem para a compreensão de que a fantasia de igualdade em ambiente digital e a pretensão de neutralidade e imparcialidade na automatização de decisões, enquanto as desigualdades de fato persistem, apenas contribuem para a reprodução de opressões. Tecnologias que funcionam a partir de algoritmos enviesados ou que reproduzem os vieses existentes na sociedade resultam em discriminação algorítmica. Portanto, a correção dos vieses racistas é intrínseca às decisões humanas e ações concretas de reparação.

4. Perfilamento racial e discriminação algorítmica

A forma de coleta, uso, armazenamento e tratamento de dados tem efeitos coletivos, no sentido de categorizar e hierarquizar populações com o potencial de discriminar, disciplinar e controlar, de formas complexas, especialmente os grupos mais vulneráveis. Isso porque sistemas de informação são também sistemas disciplinares (Johnson, 2014; Zuboff, 1994; 2019).

O perfilamento racial, junto com a reprodução de vieses raciais, é um dos mecanismos de utilização de dados pessoais para a produção de discriminação. O perfilamento, em si, remete a coleta de dados pessoais para produzir um perfil e criar um alvo. Entende-se perfilamento racial (*Racial profiling*) como o ato de classificar determinado corpo/território a partir de estereótipos racistas sobre o grupo/espaço, tomar decisões sobre ele, sobre consumo, acesso, abordagem policial, informação, oportunidades, etc; e criar um alvo. Alguns autores têm proposto a categoria *redlining tecnológico* (Noble, 2018) para salientar o papel dos algoritmos e da tecnologia de inteligência artificial nos atuais processos de perfilamento racial.

A discriminação algorítima pode ser entendida como a automatização de processos de exclusão racializada. Pode acontecer tanto como consequência da reprodução e automatização dos vieses existentes na sociedade, quanto como consequência da prática de perfilamento racial, e funciona das mais diversas formas e nos mais diversos âmbitos da vida em sociedade, limitando o exercício das capacidades e o acesso a oportunidades para membros de diferentes grupos sociais.

A discriminação algorítmica, como está sendo aqui definida, guarda relação direta com as três formas de discriminação descritas por Anita Allen (2022), como elementos constitutivos do *Black Opticon*. O *Panopticon* ou *Discriminatory Oversurveillance,* que consiste em marcar a população negra e marginalizada como alvo preferencial da vigilância excessiva; o *Ban-opticon* ou *Discriminatory Exclusion,* que consiste em marcar a população negra como alvo de exclusão de oportunidades benéficas com base na raça; e o *Con-panopticon* ou *Discriminatory Predation* consiste em marcar a população negra e marginalizada como alvo de golpes de consumo e fraude.

Discriminação e exclusão

A prática de perfilamento racial, especialmente *redlining*, tem sido usada no sistema financeiro e bancário nos Estados Unidos, criando e aprofundando desigualdades de raça ao taxar determinados territórios pela pobreza (Noble, 2018). Mas o perfilamento para a exclusão pode acontecer nas mais diversas áreas, análise de crédito, acesso à saúde, na classificação por risco, em processos de contratação, nas decisões de policiamento, etc.

Talvez um dos casos de perfilamento racial (e de gênero) mais discutidos seja a utilização de algoritmo de direcionamento de anúncios pelo Facebook/Metha ou Google. A discriminação na exibição de diferentes tipos de anúncios para pessoas negras e para mulheres mostra que a máquina tem apreendido o padrão patriarcal de divisão sexual e racial do trabalho, e o padrão de segregação racial histórico na ocupação do território. Assim, a máquina, ao criar um alvo preferencial para cada tipo de anúncio, reproduz padrões de exclusão e seleciona mostras de anúncios de emprego em ocupações consideradas femininas ou pauperizadas para mulheres e pessoas negras; e mostras de opções de moradia discriminadas por premissas racializadas de acesso econômico para pessoas negras e brancas.

No Brasil, um dos casos mais notórios de exclusão discriminatória por perfilamento racial aconteceu em 2018. Foi identificado que o site Decolar.com[3] estabelecia perfis territoriais e étnicos a partir da localização geográfica,

3. Em ACP, Ministério Público acusa Decolar.com de geo-blocking e geo-pricing | Jusbrasil

capturada pelo endereço IP. A depender do lugar geográfico a partir do qual fosse feita a busca, a empresa poderia oferecer vagas com preços diferentes (*Geo-pricing*), ou poderia bloquear ou negar a disponibilidade de determinada vaga (*Geo-blocking*). Isso para estimular o consumo de pessoas de determinados países e de determinadas raças e etnias; e desencorajar a presença de consumidores de nacionalidades não desejadas.

Target para fraudes

A predação discriminatória (con-ópticon), outro tipo de exclusão orientada por dados, acontece mediante o uso dos dados pessoais para identificar pessoas negras vulneráveis (ou outras comunidades racializadas pela marginalização) para explorar essa vulnerabilidade e as induzir a golpes de consumo, fraude e acordos danosos, como trabalho forçado e prostituição, etc. (Allen, 2022).

Um exemplo de discriminação predatória para o contexto latino-americano é encontrado entre os beneficiários de programas sociais, a exemplo do Bolsa Família, um programa de transferência de renda do governo brasileiro, cujos beneficiários, principalmente mulheres, já foram marcados como alvo de propaganda política dirigida, golpes, desinformação e instalação de vírus em dispositivos móveis (Valente & Fragoso, 2020).

Como mencionado, nem sempre é preciso explicitar raça para que essa funcione como uma categoria de exclusão. E a raça, na interseção de outras opressões como classe, gênero, sexualidade, e inclusive etarismo, agrava o cenário de vulnerabilidade. *Proxies*, como endereço, nome, tipo de material que se consome na internet, podem contribuir para a identificação racial e criação de alvos.

Todos esses aspectos podem implicar em vulneração de direitos, quando sistemas automatizados baseados em dados são usados para substituir os humanos na tomada de decisões em aspectos fundamentais para a vida em sociedade, como quem pode ser contratado, quem pode ter acesso a um crédito, quem é suspeito, etc. (O'Neil, 2020).

5. Vigilância excessiva e criminalização

Diversos autores explicitam a relação entre o discurso político racial e as práticas históricas de vigilância excessiva. A partir do conceito de necropolítica, pode se entender a invenção da raça como o elemento central que organiza e legitima a ação do Estado de deixar morrer, fazer morrer e implantar zonas de morte para a dominação (Mbembe, 2017). Assim, a partição ontológica do ser legitima as diversas violências e formas de genocídio sobre o corpo/território negro e as formas de cuidado e proteção sobre o corpo/território branco (Mills, 1997; 2013).

Dentro da política do medo, criminalização e encarceramento da população negra, os discursos de segurança pública fazem da vigilância, do controle e da punição (Borges, 2019) elementos centrais no tratamento de dados pessoais e no uso de tecnologia movida a dados, revelando uma relação intrínseca entre perfilamento racial, vigilância e criminalização.

O risco de vigilância excessiva é, também, uma das três dimensões do *Black Opticon* proposto por Anita Allen sob o termo *Discriminatory Oversurveillance*. A hipervigilância discriminatória expressa continuidade com práticas históricas de vigilância panóptica sobre a população negra e compreende o uso da tecnologia tanto para a sobre-exposição de informação sensível, íntima, incriminatória, quanto o aumento da capacidade de vigilância governamental e de controle do comportamento humano com impacto acentuado sobre a população afroamericana.

A hiper-vigilância excessiva, na sua relação com a raça, ficou muito evidente para o caso norte-americano no uso do aplicativo Geofeedia para perseguir ativistas do movimento *Black Lives Matters* com fins de conter e desencorajar a manifestação política (Allen, 2022).

No Brasil, talvez seja o uso expansivo de câmeras de reconhecimento facial no estado da Bahia, o estado com maior população negra no Brasil, a prática que melhor ilustra o uso discriminatório da vigilância panóptica com impacto acentuado sobre a população negra (T. Silva, 2021).

O uso de câmeras de reconhecimento facial se junta a, e atualiza, práticas históricas de vigilância biométrica sobre o corpo racializado negro, que prometem fazer predições sobre valor moral e propensão ao crime a partir da medição das características do rosto humano[4].

Para alguns autores, o uso de câmeras de reconhecimento facial na segurança pública deve ser entendido como uma tecnologia carcerária algorítimica que avança em prol do encarceramento em massa e do genocídio negro (T. Silva, 2022). O Brasil tem desenvolvido algumas pesquisas que mostram que pessoas negras são alvos preferenciais do encarceramento por monitoramento facial (90%) e que são as principais vítimas de reconhecimento errôneo pelo reconhecimento fotográfico[5]. A condição de suspeito, assim, se refere menos a uma atitude e mais à negritude.

O caráter racial da vigilância excessiva também se expressa na resistência em produzir dados que permitiriam maior visibilidade, representação ou melhor acesso a políticas públicas para pessoas negras. Bem como é possível evidenciar uma sobre-exposição de pessoas negras na produção de dados com fins de criminalização e discriminatórios, há também uma invisibilidade em aspectos que poderiam representar benefícios para o grupo social[6] (Richardson, 2020).

Sub-representados no conjunto de dados que treinam as máquinas, mas sobre-representados nas bases de dados de suspeitos, pelos estereótipos que existem sobre pessoas negras, esse grupo social se torna alvo preferencial do perfilamento racial com fins de vigilância policial.

A vigilância excessiva não acontece só sobre indivíduos, acontece também sobre territórios negros ou marginalizados. A vigilância excessiva sobre esses territórios aumenta o assédio policial, as detenções falsas, a persecução

4. SP lança edital para sistema de câmeras que identifica cor e 'vadiagem'
5. Álbuns de suspeitos é mais uma forma de perfilamento racial que está sendo automatizada com o uso de tecnologia de reconhecimento facial na segurança pública.
6. No contexto brasileiro, a atual discussão sobre o uso de tecnologia na segurança pública ilustra bem o caráter racial da vigilância excessiva. Enquanto há uma defesa na utilização de tecnologia de reconhecimento facial, há uma negativa na utilização de câmeras corporais em agentes policiais.

por situações irrelevantes ou crimes sem vítimas. Por sua vez, os próprios dados gerados dessas ações retroalimentam a base de dados de territórios alvos de vigilância. Assim, se estabelece um ciclo de criminalização sobre corpos/territórios negros e empobrecidos.

Não há evidências de que as tecnologias preditivas ajudem a diminuir a criminalidade. Pelo contrário, se argumenta que o policiamento orientado a dados ajuda a automatizar a lógica policial já existente dando aparência de ciência, precisão e neutralidade. Dados sobre policiamento preditivos não preveem criminalidade, apenas revelam a tendência de vigilância excessiva sobre determinado território. Sendo assim, com base nos dados disponíveis, algoritmos sobre policiamento preditivo são treinados para identificar apenas as preferências históricas em prisões e batidas policiais e replicá-las como uma profecia que se auto-cumpre.

Os poucos casos de sucesso são produzidos a um alto custo em termos econômicos, em criminalização da pobreza, perda da privacidade, assédio, abordagens invasivas e vulnerabilização de direitos – como proteção de dados pessoais, direito de ir e vir, presunção de inocência, entre muitos outros.

A vigilância excessiva também contribui para manter a segregação racializada do espaço. Inclusive em países onde não houve práticas de segregação racializada é possível perceber um padrão racializado de utilização do espaço, seja para moradia, lazer ou inclusive trabalho. Simone Browne, com o termo vigilância racializante (*racializing surveillance*) salienta como as tecnologias de vigilância funcionam para garantir que os sujeitos se mantenham "no seu lugar".

O risco de vigilância excessiva não acontece apenas com fins de segurança pública, mas se dá também no âmbito das relações laborais. O relatório de Matescuu sobre o funcionamento do aplicativo de verificação de visitas (*Electronic Visit Verification* - EVV) utilizado nos Estados Unidos, com o fim de, através da vigilância, aumentar a produtividade no trabalho, reduzir fraudes, modernizar e melhorar o serviço prestado por servidoras e servidores públicos que trabalham com beneficiários de assistência social, na prática,

contribuiu para criminalizar trabalhadores de baixa renda e beneficiários da política de assistência social com efeitos negativos sobre o trabalho, uma vez que um tempo substancial é utilizado para preencher os requisitos do aplicativo. A tecnologia, nesse caso, foi explicitamente configurada a partir de estereótipos sobre servidores públicos do setor de assistência social e beneficiários desse sistema, sobre propensão a cometer fraude. Entre outros impactos negativos, além de afetar a privacidade e execução do trabalho, as exigências de geolocalização para verificação acabaram afetando o direito de ir e vir das pessoas beneficiárias, ao produzir uma espécie de "arresto domiciliar digital" (Mateescu, 2021).

6. Impacto nos processos de subjetivação e socialização

> *O corpo negro, portanto, vê continuada sua posição de colonizado, como um repositório de representações fixas em prol da construção de soberania branca.*
>
> Achille Mbembe

Processos de construção de subjetividade são complexos, multidirecionados e multifacetados. São múltiplas as variantes que intervêm nas formas de produção do "eu" e do "outro". Mais do que empreender uma discussão teoricamente embasada sobre modos de construção de sujeitos e subjetividades, interessa aqui apontar para preocupações iniciais, mas que podem definir agendas de pesquisas futuras.

A subjetividade, em nosso entendimento, é socialmente construída na interação com o outro, mas mediada pelas estruturas e instituições sociais, os processos históricos, o território, a natureza, os dispositivos tecnológicos, as criações humanas, as rupturas, as lutas políticas... enfim, por tudo aquilo que impacta nosso ser, estar, sentir, pensar, desejar no mundo. Os modos de subjetivação se referem aos modos de construção do si, os modos em que pode se ser/existir como sujeitos, como as pessoas são lidas, interpretadas e reconhecidas.

Diversos autores têm refletido sobre como a colonização, a escravidão e as formas de atualização do racismo definem modos de subjetivação coletiva para o grupo social negro e como o racismo e as representações sobre raça e gênero impactam a vida das pessoas negras. A raça é tanto uma categoria material quanto uma categoria subjetiva, o que torna relevante se perguntar sobre os riscos do uso de tecnologia movida a dados sobre os processos subjetivos de construção e desconstrução da raça; e sobre as formas em que a tecnologia movida a dados constrói, na atualidade, corpos negros.

Para alguns autores, sistemas de dados têm o potencial de questionar ou restituir regimes de invisibilidade de sujeitos marginalizados e suas vidas (Dencik et al., 2019; Martin & Taylor, 2021; Richardson, 2020) e de influir na produção de entendimentos sobre o tipo de sociedade em que vivemos e sobre como se organiza a sociedade como um todo.

O argumento de autoridade científica – a pretendida neutralidade e infalibilidade – que acompanha as tecnologias orientadas por dados tem o potencial de produzir uma espécie de validação da construção ontológica de sujeitos negros como suspeitos, perigosos, menos humanos, etc., e legitimar a necropolítica, a vigilância, a criminalização, o encarceramento, o extermínio. Além de considerar a capacidade de decisões algorítmicas de mediar regimes de existência, Introna (2016) desafia entender os algoritmos, em si, como atores capazes de criar mundos.

Num momento em que raça, assim como gênero, está sendo cada vez mais compreendida como produto de processos históricos de dominação, se torna relevante questionar como grupos sociais estão sendo produzidos por sistemas biométricos (Kloppenburg & Van Der Ploeg, 2020) , sobre os riscos de "auto-essencialização automatizada" (Scheuerman et al., 2021), de que forma o uso da tecnologia automatizada ancorada nas métricas faciais poderia estar reinscrevendo noções de diferença que reforçam a ideia de biologização da raça (e do gênero) ocultando o fato de que são categorias sociais de poder (M. R. Silva, 2020), assim como permitindo o ressurgimento

de noções de racismo científico já rejeitadas, como as que estabeleciam relações entre características físicas, especialmente, padrões faciais, e moralidade ou criminalidade (T. Silva, 2022).

Além disso, ainda é pertinente se perguntar quais são os riscos de reificação do pensamento patriarcal de divisão sexual e racial do trabalho, no direcionamento enviesado de anúncios de empregos em tecnologias que aprenderam a discriminar por sexo e raça. Entre muitíssimas outras questões pertinentes de serem formuladas como parte dos riscos de uso de tecnologia e o tipo de valores que interessa descontruir para a construção de sociedades mais justas.

7. Apagamento do caráter político das problemáticas sociais

> *A diluição de responsabilidade que se verifica na atribuição à tecnologia de agência sobre decisões relacionadas a abordagem, identificação, tipificação ou condenação, por meio de dispositivos como reconhecimento facial, policiamento preditivo e escores de risco, é um dos maiores perigos do racismo algorítmico.*
>
> Tarcizio Silva

O último risco que se pretende discutir aqui é o apagamento do caráter político das problemáticas sociais, que pode acontecer, como muito bem apontado por T. Silva (2022), pela diluição da responsabilidade humana sobre as decisões algorítmicas e os resultados do uso de tecnologia de inteligência artificial, e pelo uso da tecnologia como principal instrumento para resolver problemas sociais complexos.

O uso de tecnologias de inteligência artificial ao ser incorporado nos sistemas estatais – seja de trabalho, assistência social ou segurança pública –, pode acabar reduzindo questões sociais que são complexas, deslocando o foco de questões que na verdade são políticas, que dependem de processos de tomada de decisões e que envolvem diversos atores com interesses particulares que requerem grande investimento público, etc., e reduzindo uma

problemática complexa à implementação de uma tecnologia, cuja aplicação usualmente é inclusive terceirizada.

Considerações finais: Uma linguagem moral de reparação

> *I believe that artificial intelligence will become a major human rights issue in the twenty-first century.*
> Safiya Umoja Noble

Neste ensaio, discutiu-se sobre os diversos riscos presentes na utilização de dados pessoais e tecnologia movida a dados para a tomada de decisões. A substituição dos humanos por processos automatizados não necessariamente representa um avanço na busca por sociedades mais igualitárias. Ao contrário, a opressão e as hierarquias de poder podem estar sendo embutidas nas tecnologias de inteligência artificial, através da reprodução e automatização de vieses e da utilização dos dados para perfilamento, com o agravante de que a máquina torna as práticas discriminatórias ainda mais sombrias e torna difusa a responsabilidade humana.

A reprodução e atualização do racismo através de ferramentas tecnológicas de inteligência artificial causa danos simbólicos e materiais concretos na vida das pessoas; têm consequências na distribuição do poder político, na produção e valorização de saberes, na distribuição e exploração econômica, no acesso a políticas sociais, etc.

Usualmente, as pessoas mais afetadas têm menos conhecimento sobre como a discriminação automatizada atua e também têm menos recursos para se opor ou procurar a restauração dos seus direitos. Aspectos-chave sobre decisões automatizadas que representam vulnerabilização de direitos parecem inescrutáveis para a sociedade. A monopolização da informação em grandes empresas de tecnologia e mídia digital (Noble, 2018) e a dupla opacidade, tanto do funcionamento do racismo quanto da linguagem e funcionamento da tecnologia (T. Silva, 2022), agravam ainda mais os problemas.

Programas automatizados de inteligência artificial reproduzem padrões estatísticos sem possuir meios para aprender com seus erros (O'Neil, 2020). Erros, estes, que são lógicos, porque usualmente estão em concordância com as informações subministradas e com os objetivos de aplicação da tecnologia, mas que também são injustos.

Neste caso, como em muitos outros, a pretensão de neutralidade só contribui para manter a reprodução da hierarquia racial. Isso porque os problemas da tecnologia se combinam muito bem com os problemas históricos estruturais da nossa sociedade.

Corrigir a reprodução e manutenção do racismo implica, na nossa proposta, assumir, nas discussões sobre usos de dados pessoais, inteligência artificial e ambiente digital, a lógica da *reparação*, uma linguagem moral presente no sistema internacional de direitos humanos, segundo a qual se faz necessário a adoção de medidas concretas, efetivas e adequadas para deter e reverter as consequências duradouras do racismo.

Se os algoritmos, a partir dos quais a tecnologia aprende a reproduzir o pensamento humano, são desenhados por humanos – a partir de valores sociais que garantem a manutenção do capitalismo-racista-patriarcal; a partir de interesses e relações de poder que privilegiam uns sobre outros por atores sociais privilegiados e poderosos; a partir de práticas históricas de produção de desigualdade, etc. –, não é de se esperar respostas automatizadas que possam conduzir a um mundo mais justo.

O caminho que aqui propomos não é o da neutralidade, mas o de centralizar raça, entender que as questões envolvendo sistemas de dados se relacionam com os sistemas de opressão que estruturam a sociedade, e pautar esse debate a partir de uma compreensão de que existem assimetrias de poder (como o racismo) que tornam vulneráveis determinados grupos sociais em benefício de outros. Assim, a raça deve ser incorporada na captura, processamento, armazenamento de dados, nos objetivos da produção e utilização da tecnologia, na codificação do problema e da solução

esperada, para propositalmente produzir decisões favoráveis aos grupos historicamente oprimidos.

Evitar a discriminação negativa, ilícita ou abusiva, e produzir ações afirmativas devem ser objetivos explícitos nas decisões automatizadas que envolvem garantia de direitos. Isso porque, caso contrário, a tendência será a reprodução das discriminações, involuntária ou voluntariamente.

O combate ao racismo e/ou a produção de sociedades igualitárias deveriam ser incorporados na definição do problema que requer solução algorítmica. Assim, algoritmos de avaliação de créditos poderiam ser codificados para considerar que é uma questão de justiça social e reparação garantir o acesso a créditos para pessoas negras e não apenas levar em consideração o lucro; algoritmos de atendimento à saúde poderiam ser codificados para considerar que o gasto em saúde não pode ser um critério para priorização de atendimento. Priorizar a justiça social também poderia levar a concluir que alguns usos da tecnologia, como câmeras de reconhecimento facial na segurança pública, deveriam ser simplesmente banidos devido ao seu potencial nocivo para a sociedade como um todo.

Nossa proposta implica interrogar o projeto político de sociedade por trás das decisões sobre o que pode ser contado, coletado, e como serão tratados os dados; sobre o projeto político que sustenta o uso da tecnologia; sobre a utilização de dados e tecnologia para a construção de outro horizonte de possibilidades para apoiar projetos emancipatórios.

O objetivo aqui foi apontar para os riscos. No entanto, um aspecto importante de se tratar é como pessoas negras, indígenas e grupos minorizados, em geral, disputam os domínios que sustentam a matriz de dominação (Hill Collins, 2014) capitalista-racista-patriarcal (Gonzalez, 2018) e produzem oportunidades de agência política, resistência e de aproveitamento da tecnologia para projetos de justiça social e para avançar numa agenda de luta contra o racismo.

Referências

Allen, A. L. (2016). Protecting one's own privacy in a big data economy. *Harv. L. Rev. F.*, *130*, 71.

Allen, A. L. (2022). Dismantling the "Black Opticon": Privacy, Race Equity, and Online Data-Protection Reform. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4022653

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, maio 23). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Borges, J. (2019). *Encarceramento em massa* (D. Ribeiro, Org.). Pólen.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Em S. A. Friedler & C. Wilson (Orgs.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, p. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Center for democracy & tecnology. (2022). *Request for Information (RFI) on Public and Private Sector Uses of Biometric Technologies: Responses* (Federal Register Notice 86 FR 56300). CENTER FOR DEMOCRACY & TECHNOLOGY (CDT).

Dencik, L., Hintz, A., Redden, J., & Treré, E. (2019). Exploring Data Justice: Conceptions, Applications and Directions. *Information, Communication & Society*, *22*(7), 873–881. https://doi.org/10.1080/1369118X.2019.1606268

Fragoso, N., Valente, M., LANGENEGGEr, N., & RUIZ, J. P. (2021). *Proteção de dados em Políticas de Proteção Social: Contribuições a partir do Programa Bolsa Família* (Diagnósticos e Recomendações n. 6). InternetLab. https://internetlab.org.br/wp-content/uploads/2021/10/Protecao-de-Dados-Pessoais-em-Politicas-de-Protecao-Social.pdf

G. Valente, M., Neris, N., & Fragoso, N. (2021). Presa na rede de proteção social: Privacidade, gênero e justiça de dados no Programa Bolsa Família. *Novos Estudos - CEBRAP*, *40*(1), 11–31. https://doi.org/10.25091/s01013300202100010001

Gonzalez, L. (2018). *Primavera para as rosas negras*. Diáspora Africana.

Gonzalez, L. (1984). Racismo e sexismo na sociedade brasileira. *Revista Ciências Sociais Hoje, Anpocs*, 223-244.

Hill Collins, P. (2014). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*.

Introna, L. D. (2016). Algorithms, Governance, and Governmentality: On Governing Academic Writing. *Science, Technology, & Human Values*, *41*(1), 17–49. https://doi.org/10.1177/0162243915587360

Johnson, J. A. (2014). From open data to information justice. *Ethics and Information Technology*, *16*(4), 263–274. https://doi.org/10.1007/s10676-014-9351-8

Kloppenburg, S., & Van Der Ploeg, I. (2020). Securing Identities: Biometric Technologies and the Enactment of Human Bodily Differences. *Science as Culture*, *29*(1), 57–76. https://doi.org/10.1080/09505431.2018.1519534

Lee, C. (2019, outubro 25). *A biased medical algorithm favored white people for health-care programs*. MIT Technology Review. https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/

Lippold, W., & Faustino, D. (2022). Colonialismo digital, racismo e acumulação primitiva de dados. *Germinal: Marxismo E educação Em Debate*, 56–78.

Martin, A., & Taylor, L. (2021). Exclusion and inclusion in identification: Regulation, displacement and data justice. *Information Technology for Development*, *27*(1), 50–66. https://doi.org/10.1080/02681102.2020.1811943

Masiero, S., & Das, S. (2019). Datafying anti-poverty programmes: Implications for data justice. *Information, Communication & Society*, *22*(7), 916–933. https://doi.org/10.1080/1369118X.2019.1575448

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Mateescu, A. (2021). *Electronic Visit Verification: The weight of surveillance and the fracturing of care*. Data & Society Research Institute. https://datasociety.net/wp-content/uploads/2021/11/EVV_REPORT_11162021.pdf

Mbembe, A. (2017). Necropolítica. *Arte & Ensaios; n. 32 (2016): Eclipse*. https://revistas.ufrj.br/index.php/ae/article/view/8993/7169

Mills, C. W. (1997). *The racial contract*. Cornell University Press.

Mills, C. W. (2013). *1 O contrato de dominação*. *8*(2), 56.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

O'Neil, C. (2020). *Algoritmos de Destruição em Massa: Como o big data aumenta a desigualdade e ameaça a democracia*. Apple Books.

Prince, A., & Schwarcz, D. (2020, março). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review*, *105*(3).

Richardson, R. (2020, outubro 8). Government Data Practices as Necropolitics and Racial Arithmetic. *Global Data Justice, Essay #1 in the Data and Pandemic Politics series on data justice and COVID-19*. https://globaldatajustice.org/gdj/1977/

Scheuerman, M. K., Pape, M., & Hanna, A. (2021). Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society*, *8*(2), 205395172110537. https://doi.org/10.1177/20539517211053712

Silva, M. R. (2020). Código da ameça: Trans; classe de risco: Preta. *São Paulo: N-1 Edições*.

Silva, T. (2022). *Racismo algorítmico: Inteligência artificial e discriminação nas redes digitais*. Edições Sesc SP.

Silva, T. (2021, setembro 20). *O horror do reconhecimento facial na Bahia, onde poderia ser diferente*. https://tarciziosilva.com.br/blog/o-horror-do-reconhecimento-facial-na-bahia-onde-poderia-ser-diferente/

Silveira, S. A. (2017). GOVERNO DOS ALGORITMOS. *Revista de Políticas Públicas*, *21*(1), 267. https://doi.org/10.18764/2178-2865.v21n1p267-281

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, *4*(2), 205395171773633. https://doi.org/10.1177/2053951717736335

Valente, M., & Fragoso, N. (2020). *Data Rights and Collective Needs: A New Framework for Social Protection in a Digitized World*. https://itforchange.net/digital-new-deal/2020/10/29/data-rights-collective-needs-framework-social-protection-digitized-world/

Zuboff, S. (1994). Automatizar/informatizar: As duas faces da tecnologia inteligente. *Revista de Administração de Empresas*, *34*(6), 80–91. https://doi.org/10.1590/S0034-75901994000600009

Zuboff, S. (2019, janeiro 3). Tua escova de dentes te espiona: Um capitalismo de vigilância. *Le monde diplomatique Brasil*.

FIGHTING ALGORITHMIC RACISM: REACTIONS, REMEDIATIONS AND RE-APPROPRIATIONS

Tarcizio Silva
/ Mozilla Foundation

There are many reactions and paths being charted by activists, developers, scientists and technologists of many fields and disciplines. Algorithmic racism is not a phenomenon that can be addressed in a simple way. More than handling "racist algorithms", the matter is the algorithmization of racism through some key points: reproduction and machine reinforcement of social, political and cultural inequality; growth of opacity regarding racial relationships and resulting oppressions; reproduction of necropolitics biopower as the foundation of the contemporary technocrat neoliberalism; and deepening of the colonial and racialized extraction of data and labour in the Global South.

Therefore, modalities of resistance, reactions, and remediation against the algorithmic transformation of structural racism involves remembering of the diverse fronts of Black movements in social battles and in diasporic solidarity. As well as in the refuse to disaggregate the identities and to not adhere the upkeep of the *status quo*, as said by Jurema Werneck (2010), reminding us that our steps come from afar.

Public Audits and Awareness

In the studies about algorithm racism, we come across the exposing of manifestations of algorithmic racism through public audits, journalism, research and

activism. Gathering evidence about the fragility of algorithmic systems is a multidisciplinary task. The evidence are not only computational or a result of code audits, but also from data gathering, ethnographies and journalistic investigations – all can have distinct impacts depending on the power relations, professional authority and social-political contexts.

The project *Gender Shades* brought light about the intersectional disparities that computational vision systems performed by making unacceptable mistakes against black women. Beyond pointing out the problems, the researchers created a benchmark of photographs to test the systems, in a reproducible format. Through careful curatorship, the *Pilot Parliaments Benchmark* is an instrument that allows any developer or company to analyze the precision of their system regarding gender and skin tones (Buolamwini & Gebru 2018).

That way, besides the audit of the three first systems – from IBM, Microsoft and Face++, the project *Gender Shades* suited the community as an instrument of analysis that has a potential for large scale replication. Two years after, Joy Buolamwini and Deborah Raji (2019) conducted a new audit about the analyzed systems in the first phase of the study and included comparisons between other suppliers, from Amazon and Kairos. The researchers found out that the systems approached previously improved the error rates, but the ones which were analyzed for the first time on this round followed the same tendencies of inter-sectional mistakes, with bigger imprecision regarding black women.

But beyond the metrics of the softwares from the targeted companies in the study and their practices, the project can be understood as an "actionable public audit". As far as its repercussion was based on principles of scientific disclosure built on the connection between academic spaces of power and visibility, it also generated debates in the public sphere, often mentioned in mobilized groups and even in regulatory propositions.

To address the racializing politics in the construction and functioning of algorithmic systems, however, should not be seen as something pertain-

ing solely to computer science and affiliated areas, limited by disciplinary borders. Understanding algorithmic systems do not only involve the path of analyzing code lines, but moving through its networks of delegation, which behaviors are normalized, which data is accepted, what types of mistakes are or are not considered between inputs and outputs of the system, its potential of transparency or opacity and which presences of absences are implemented – finally, the networks of racial-political relations in the changing materializations within technology.

The experiential knowledge about algorithmic systems transformed not only through scientific approaches but also through vernacular approaches reminds us that some dynamics of discrimination present themselves on the surface, even if they require enunciation. An example was the campaign *#BuscaPorIgualdade*, carried by "Desabafo Social". In short videos, the organization exhibits queries for key words of popular categories, such as "people", "family" or "skin", in stock photo websites like *Shutterstock* or *Deposit Photos*, resulting in pages filled with almost exclusively white people. The contrapostion with the searches supplemented with the qualifier "black" in terms like "family" is the hook to remind the stock photos banks that "Black family is also family" (Ferreira, 2017). The success of the campaign, along with other types of pressure and market reactions, which included vertical stock photos banks focused on black people, was one of the factors that came to advance the filter options in providers like Shutterstock (2017).

In an interesting typology of algorithm audits proposed by Sandvig and associates (2014), the project *#BuscaPorIgualdade* would be an example of what is called *Sock Puppet Audit*, because it simulates the behavior of users and reflects critically about the patterns of the results. Another accessible approach to non-technical investigations is the *Non-Invasive User Audit*, through of which traditional research methods from social sciences are applied, such as *surveys*, interviews or systematic observation. In a way, the initiatives coming from journalists and the activists' campaigns fit into these categories, in carrying out "non-invasive selection of information about normal interactions from users on a platform" (Sandvig et al., 2014, p.11).

To emphasize the legitimacy of those approaches is especially relevant in order to give the appropriate importance to the role of the possible harmful impacts, regardless of explicit intentionality by the developers of the systems or of the technically mensurable fragility of the codes. Thus, combating harmful impacts from algorithmic systems do not only involve the aspects seen strictly as technical or deriving from computing – neither it involves solely programmers and engineers. The audit concept, therefore, can be expanded so we can also think how the visibility of the incorporated dynamics in these systems generate arguments in order to "promote campaigns and activities which incorporate an accessible and objective language, but with impact and relevance for the citizens in general" (Nunes, 2020).

Ziv Epstein's and associates research alerts about the "knowledge gap about artificial intelligence " as far as the "number of unique AI systems grows faster than the number of studies that characterize these systems' behavior" (2018, p.1). Analyzing more than 7 thousand documents published in the major conference *Neural Information Processing Systems* from 1987 to 2017, the authors discovered that there are approximately 10 times more propositions of new computing models than studies of existing models, in a growing gap.

The tendency is also identified by Pablo Nunes, coordinator of Rede Observatório da Segurança, calling out attention for the mismatch between the "reflection of the effects and the efficiency of certain implementations of algorithms with the number of projects and applications of these same technologies that are already in development" (Nunes, 2020) when we address the implementation of facial recognition for public security.

The interdisciplinary dialogue for the development of algorithmic systems and its positive applications of artificial intelligence can engender ways in which the computing approaches can be part of the solution. Rediet Abebe points some directions in how computing can be used as diagnostic and rebuttal. Abebe defends that the studies about risk algorithmic scores and criminal and behavior prediction, for example, show insurmountable limitations for the use of AI in these cases, therefore

> no matter what algorithm is employed, any way of assigning risk estimates to two groups of differing base rates will necessarily produce a specific kind of disparity in outcomes between the groups; we cannot eliminate the problem through a better choice of algorithm. Formal results of this type can expose the limitations of an entire category of approaches — in this case, the assignment of numerical risks to individuals. (Abebe, 2019, p. 190).

It could also be possible to, paradoxically, formalize the problems within the algorithmic systems to the point of exposing the networks of delegations incorporated in the technologies. The computer, or the algorithmic system, becomes *synecdoche*, a discourse and critique tool about the society representing it as a part of a whole and "can offer us a tractable focus through which to notice anew, and bring renewed attention to, old problems" (Abebe, p.192).

Abebe's proposition goes in the direction of what Charô Nunes highlighted previously in writing about the computing algorithms in overlap to the algorithms of the society understood as a "group of social, economical, ideological and even semiotic rules that are a result of the disputes and all sort of interaction between many segments of the population" (Nunes, 2018, para. 7). Paradoxically, the audits of algorithmic systems from racial critiques, even though are few in relation to numbers of disastrous implementations, are transformed in possibilities of "auditing" the same dynamics proposed by the hegemonic society elects as the central to gain the status of automatized and opaque reproduction.

"Fuck the algorithm": Youth protests

The COVID-19 pandemic became an unexpected dream for the corporations and *startups* committed with collecting data for more layers of life on behalf of offering systems of algorithmic management. The *contact tracing* as a tool to control the transmission was rapidly co-opted by agents representing state violence, that began to use the same lexicon in order to monitor activists' actions (Castro, 2020).

But the suspension of essential activities like basic education generated new challenges in a society that promotes competition for public resources – to manage the challenges, the technocentrism was the mistaken solution chosen around the world. In the United Kingdom an algorithmic system was implemented to attribute scores to the students about to try openings at universities, since the normality of the school year was compromised. The data that fed the system was not only based on previous performance of the students, as well as on a ranking established by the teachers about which grade they thought the students could reach at the end of the period, and also the historical performance of the school. As it was expected, the private schools were benefited – in those, the number of maximum grades went up about 5%, the double of the average of the historic difference (Yasin, 2020).

The system, therefore, favored the elites as well as the class determinism and geography in linking the school to the score – in addition to adding a discriminatory variable in the arbitrary attribution of grades by the teachers. The students protested with explicit mottoes, literally "*Fuck the algorithm!*" in front of the Department of Education, pressing to suspend the method. Despite the suspension, however, the perceived damages of the selection processes already in motion should have been adjusted through new appeals by the schools, generating barriers to the harmed students (Katwala, 2020).

The juvenile sincerity of the screams of protest rang globally and one of the posters showed the refusal to forfeit self-determination, saying: "the algorithm doesn't know who I am". One of the first organized mobilizations organized in a public space against the imprecision of an algorithm system add up to others aimed to pressure public institutions and corporations to not serve certain ends.

Mobilizations and leaks of internal information by employees possess a long history of impact, including globally. One of the historical peaks was the mobilization of American employees against the colluding of Polaroid with the apartheid regime in South Africa. The company secretly sold supplies for photographies used on the abject "passes" used to identify South

Africans in terms of race, ethnicity and places where they were able to circulate. From the end of the 1960's to the following years, employees of the company got organized as the "*Polaroid Revolutionary Workers Movement*", and they advocated in favor of the termination of the contracts of the company with the racist regimen, for the public and global announcement of this refusal to make business in the country as long as apartheid was happening and for the company to contribute with the efforts in favor of the African liberation (Morgan, 2006).

Initially, Polaroid tried to hide ties with apartheid and persecuted employees who denounced the problems. Only after many years of fight and resulting coverage from the press, the company ceased the sales. However, this effort inspired the debate, close to other similar groups from the same time, which established a base for the proposition, in 1977, of the "Sullivan Principles", a set of principles for social responsibility from companies (Alexis, 2010), referenced to this day.

During the 2019 annual Amazon Web Services conference, in New York, hundreds of citizens protested requesting that the corporation stop offering services to harmful organizations from the American government like ICE (US Immigration and Custom Enforcement), responsible for persecuting immigrants and for the implementation of concentration camps (Paul, 2019). The protest adds to a long history of mobilizations against the e-commerce and AI giant, related to its impact of gentrification (Bradshaw, 2020), local business disruption and *dumping*, besides the growing exploitation of employees and the precarious work conditions (Golledge, 2019).

During the intensification of the waves of protests against racist police violence on the United States around 2020, part of the employees from the major technology and artificial intelligence companies signed petitions in order to pressure their employers. At Google, the internal petition "*No Police Contracts*" argues that " saying Black Lives Matter is not enough, we need to show it in our thinking, in our words and in our actions" (Elias, 2020, para. 3). Google, IBM, Amazon and others reacted to the demonstrations

and regulatory pressures (Wiewiórowski, 2020) implementing temporary moratoria to sales of facial recognition to certain uses, but through elusive lexicons like the IBM delimitation on terms for applications for "mass surveillance, racial profiling, violations of basic human rights and freedoms" (Krishna, 2020, para. 9) or the short moratorium proposed by Amazon until new regulatory ethical standards are implemented by the Congress, while it lobbies to influence such rules (Matsakis, 2020).

However, if the public protests made by citizens and consumers generate financial damage to companies with valuable brands which are partially dependent on the final consumer, like Amazon and Google, the same does not occur with gigantic corporations totally directed to contracts with governments and the financial market. It's the case of Palantir, that offers services of biometric management to persecute migrants and technologies to police forces and it's the target of public mobilizations (Chan, 2019).

The technological elite from Silicon Valley "suppress the questionings about racism and discrimination, even when the products of the digital elites are infused with markers of race, class and gender" (Noble & Roberts, p.45). Neoliberalism and post racial technocentric myths intentionally make it harder, however, to achieve honest comprehension of the racialization of algorithmic technologies because such comprehension would be diametrically opposed to their business models, which are based in carceral imaginaries and economical disparity.

For example, Amazon developed patents of "intelligent bracelets" to track employees and the movement of their hands under auspices of efficiency, to improve alongside *big data* the work logistics and the movement of assets, reifying the employees for increased levels of productivity (Yeginsu, 2018). Subsequently, the infrastructure has been shamelessly applied to analyze, through heat maps of movement and "emotional analysis" of the employees, which stores from Whole Foods – retail company bought by Amazon in 2017 – would be under risk of unionization (Peter, 2020). Specific job positions for intelligence analysts are being molded specifically to fight worker

organization with the support of data, algorithms and new mechanisms of biometric surveillance (Holmes, 2020).

In this way, the same algorithmic management of strife inside the corporations is a tendency applied by capitalist groups that move forward with the implementation of new mechanisms of control and surveillance of their teams – reminding them of the inherent limits to mobilization directed to values of corporation responsibility.

Resistance through re-inventions

The record or erasure of inventions and technologies is a sociopolitical and historical process employed to privilege Eurocentric conceptions of scientific progress for centuries. Reactions are undertaken from the revolutionary independent research about the kemetic civilization by Diop (Njeri & Ribeiro, 2020) to contemporary redemptions of the history of African technologies (Machado & Loras, 2017), and valuable studies about the violent appropriation of technologies (Cunha Júnior, 2010) during the slavery at the Black Atlantic. The technologist Ramon Vilarino alerts us that, despite this rich history being hidden, the Brazilian elites do not encourage the "creation of truly local and contextualized technology, which ends by frequently producing caricatures or bad thought adaptations" (Vilarino, 2020) of what is made in the technology centers of the global North.

In contemporary times, Afrodiasporic populations in countries like Brazil fight against the cumulative disadvantages in hostile environments molded by and for white supremacy, but, notwithstanding, developed many strategies of technological innovation. A typology offered by Rayvon Fouché offers a special lens to think what he calls "black vernacular technological creativity" (Fouché, 2006). The term aims to conceptualize the way in which African American inventors adapted, reinvented, or created technologies to their specific realities – despite of the constant erasure of their authorships or even the underestimating when it comes from peripheral regions.

The three categories proposed by Fouché are: a) redeploying, the process in which the material and symbolic power of technology is reinterpreted, but maintains its use and the traditional physical shape; b) reconceiving, the active redefinition of a technology in a way that it transgress the function and/or main significance; e c) re-creation, like the redesign and production of a new material artifact after an existing shape and/or function was previously rejected (Fouché, 2006, p.658).

The non hegemonical history of the internet has many examples (McIlwain, 2019) of technological creativity initiatives, created by minority groups, like the Black populations. A software created that was especially interesting was the *Blackbird*, launched in 2008 as an experimental browser, stemming from *Mozilla Firefox*, an open code browser. Directed to African Americans, there were some specifications that rejected interstitial whiteness present on the browsers of the time, even if they presented themselves as neutral. Through its characteristics, two could be highlighted (Brock, 2020): areas of content recommendation produced by African American people, selected through curatorship; and deliberate spotlighting of fund raising for social purposes, like the initiative *Give Back*. Its creators and the development community already rejected the idea of a browser that would only function as a window to the cyberspace, not forgetting of interacting with the physical and social world beyond the online and not necessarily for the market.

Rejecting false neutralities of technologies, databases and representations is a continuous effort against the oppressions and limitations imposed to minority groups. Regarding image banks, specifically, initiatives like *Nappy*[1] or *Young, Gifted & Black*[2] aim to fight negative representations and fill positive gaps in providing photography banks focused in representing Black people. *Nappy's* website explains how the modality of license and distribution works, *Creative Commons*, allowing the free use, but going beyond and saying that "we encourage it. The more you use it, the more we'll help to better the representation of Black and brown people in the media".

1. https://nappy.co
2. https://ygb.black.

If we take a look in how generalist search engines or professional stock photo banks produce visibility, invisibility and stereotypes that penalize mainly Black women, initiatives like these respond to what Patricia Hill Collins evokes in calling for the centrality of self-determination of the image representation as well as the epistemic ones by black women in order to fight the controlling images (Collins, 2002).

Addressing the issues with the transformation of *selfies* made with social media image filters that whiten faces and promote Eurocentric standards of beauty, the Brazilian designer Joyce Gomes created the project *Black Beauty Filters*. After gathering information through qualitative and active listening of lived experiences, the designer produced filters with the augmented reality framework *Spark AR*. Beyond the filters she created for herself, she points out the importance of the project in developing a racial aesthetic literacy, in "instructing the content creator, Black or not, into having a more dedicated view of issues that concern blackness and the filter universe", encouraging the decolonization of knowledge and the interdisciplinary creation by many groups" (Gomes, 2020).

Inverting the usual gaze of crime maps, the project *White Collar Crime Zones* (Lavigne, Clifton & Tseng, 2017) is a poignant critical parody. The developers produced a crime map[3] and a risk prediction system concerned to "white collar crimes". The project differs from the most famous spatial models of crime prediction because it does not focus in "street crimes" like drug trafficking, theft and vandalism, but in financial crimes of big sums and impact, which usually result in few penalties. The developers used data from financial regulatory institutions and crossed the information with legal drugs data, such as alcohol, besides the density of organizations that evade taxes.

Beyond that, the *White Collar Crime Zones* also constructs the average face of the criminal on its base, from the gathering and computing processing of the similarities between the face photographs of 7.000 executives from financial corporations, extracted from LinkedIn. The pictures of this

3. https://whitecollar.thenewinquiry.com/

"average" criminal compose, in navigating the map, variations of a prototypical face of a young white male. Ruha Benjamin argues that, in deliberately and creatively questioning the *status quo* of predictive technology, "analysts can better understand and expose the many forms of discrimination embedded in and enabled by technology" (Benjamin, 2019, p.197).

Training new imaginaries

Maybe one of the most controversial paths to overcome the damage of algorithm discrimination is the promotion of demographical diversity of the ones who develop the technologies, like computing scientists, engineers and developers. The majority of the professional positions of big apparent impact are employed by the oligopoly of "big tech", which molds global technology.

Shortly after the *BlackLivesMatter* protests in May and June of 2020 which gained global scale after the murder of George Floyd, big USA technology corporations were pressured to act on structural racism on the society and on the industry. The main ones promised, with a lot of fuss, dozens of millions of dollars to black community initiatives and to promote diversity. From 61 thousand of dollars from Dell to 209 million from Microsoft, the numbers caused an impression due to the difficulty of glimpsing its small significance in the billion dollar scale of the big American technology corporations.

However, a journalist made the interesting comparison of the donations, stemming from the question: "If the companies were people, how much money did they donate?" (Peters, 2020). Comparing with the average annual income of 63 thousand dollars in the USA, it shows that the equivalent of the Microsoft donation would be 99 dollars while Dell's only 4 cents.

If the technochauvinism and its resulting damages favor and are boosted by the platforms and AI oligopolies, initiatives that generate new outlooks to the learning of programing, technology, and digital safety deserve doubled praise. In Brazil, the connections between Black women around initiatives that opened paths (Barros, 2019) such as the Geledés Instituto da Mulher Negra website and the group Blogueiras Negras promote epistemic

constructions (Barros, 2020) about digital care which, in the words of Larissa Santiago, generate "philosophical and practical changes in relation to the use of technologies and tools of information and communication" (Santiago, 2020), incorporating other ways of doing and constructing.

Director of Olabi and founder of PretaLab, Silvana Bahia led an innovative data gathering project about the presence of Black and indigenous women in privileged fields of innovation and technology, defending that "the lack of representation is a problem not only for the ecosystem of technology and innovation, but also for the human rights and freedom of expression" (Pretalab, 2020). The data shows how the beginning of the contact with the area was predominantly by informal means, and that the relation with activism was the third type of motivation for their insertion in the practices of technological development.

Education initiatives like PretaLab aim to reinterpret the technologies from teaching of programing languages and *maker* cultures, as well as reflecting about the learning of the bases of social pyramids. Many innovations rise from the concreteness of the everyday life and Silvana Bahia reminds us that technology is a

> big umbrella, and for us it was always important mixing analog technology, *low tech*, with the *high tech*, because we tend to think that this is an effective way of making people understand the importance of it and being able to look in a more critical way

Initiatives focused on horizontal education, sharing of knowledges and mutual emerging support about technology such as Minas Programam, Conexão Malunga, Kilombotech, Perifacode, AqualtuneLab, Tecnogueto, Afropython, Afrotech, Quebradev, InspirAda and others, arise from groups that look into producing not only formal knowledge, but also alternative narratives.

Decenter commercial technology of the discourse about programming skills is something that Bárbara Paes, co-founder of *Minas Programam*, points out when talking about how the imagination of the students about

the *hackatons*, learning experiences and collective projects promoted by the group are crossed by a duality. Some students, while they train for technology-focused industries, from startups to Silicon Valley, also collectively de-construct the usage of the learning, emerging transforming knowledges. Between the collaborative ideas during the learning process in programing, Bárbara Paes points out the inclusion of conversations about

> what you can do with it, such as the learning can be one more tool for you, not only in your professional and work life, but also thinking on your community, thinking about the people that are around you, how this can be useful for *you*

In the co-organization of Perifacode, the software engineer Carla Vieira approaches this issue in a similar way, in sharing the polysemy of combating the slants in technology, adapting the term to a positive and transformative aspect in technology

> when there are many different people, with their different points of view, like the world is, it makes more sense [...]. What is created will be inclusive, will represent the world as it is, with diversity: they are not only not represented in technology, as well as in other areas.

To the developer Roselma Mendes, this interdisciplinary connection should be promoted, because she believes that the importance of the digital technologies in contemporary society makes that the arising problems do not talk only about technology. They are also meaningful on "how we approach our work [...] I believe that the multidisciplinarity and the inclusion/visibility are complementary" in companies that develop software. In some way, we can connect such perceptions and initiatives to a promotion of racial literacy in technologies, that aims to exhibit the false neutrality that only replicates oppressions and erases alternative imaginaries.

Jessie Daniels, Mutale Nkonde and Darakhshan Mir propose the advance of racial literacy in technology in a multisectoral way, pointing out the limitations of the propositions of big corporations. Three pillars are proposed

by the authors to leave innocuous patterns of action, including "an intellectual understanding of how structural racism operates in algorithms, social media platforms, and technologies not yet developed" (Nkonde & Mir, 2019, p.4).

More than promoting diversity as a token or in an isolated way, a transformative commitment is needed. The emphasis in fighting the damages due to racism goes through uncentering purely technical aspects from discourse (Gangadharan & Niklas, 2019) opens up possibilities for addressing different modalities of discrimination and algorithmic harm.

Regulation beyond the ethical principles

The abbreviation *FAT* or *FAccT* to account for the effort to avoid algorithm harms stems from the triad *Fairness, Accountability and Transparency*, and became a global synonym of the debate about ethics in algorithmic systems going on in the computation industry and in communities around machine learning and neural networks. Critiques about the *FAccT* approach involve specially the tendency to delimit the problem of algorithmic harm as a matter of coding or management. A common critique question the necessity of developing new concepts for prerogatives of respect to human rights, considering that framing discriminatory impacts as new can erase already known political and racial aspects of technologies.

An exploratory survey (Floridi & Cowls, 2019) studied the consensus between propositions by international initiatives about principles for artificial intelligence. It organized them into five principles: beneficence, non-maleficence, autonomy, justice and explicability. The first four are already discussed and applied frequently in bioethics, reminding how interdisciplinarity can fight tendency of perceiving problems in digital artifacts as something completely new. Regarding autonomy, Floridi and Cowls defend that humans should retain the power of deciding which types of decisions are made, having the possibility of intervention when needed, and, at last, collectively deciding in which cases the loss of control of the decision process is worth it in terms

of the benefits compared to the costs or possible damages (Floridi & Cowls, 2019).

It's more about who has the power of "classify, to determine the repercussions / policies associated thereof and their relation to historical and accumulated injustice" (Abdurahman, 2019, para. 8) in the words of J. Khadijah Abdurahman criticizing the *FaccT* approach. Catherine D'Ignazio e Lauren F. Klein agrees, in offering an alternative set of guiding concepts to the field, advocating for a transition of "data ethics" to "data justice". The goal is to dislocate the source from the individual problems and technical systems to the comprehension of the power relations as well as fighting them (D'Ignazio & Klein, 2020), like we can see on the following table:

Table 1: From Data Ethics to Data Justice

Concepts that secure power	**Concepts that challenge power**
Ethics	Justice
Bias	Oppression
Fairness	Co-liberation
Transparency	Reflexivity
Understanding algorithms	Understanding history, culture, and context

Source: D'Ignazio & Klein, 2020, p.60

The proposition of this shift means to embrace concepts that dialogue with the legacy of collective organization, inters-sectional feminist activism and critical thinking, rejecting the idea that radically new "ethical" principles would be necessary to grasp problems based on *big data* and artificial intelligence.

In this way, the idea of algorithmic inexplicability should not be acceptable in systems that have relevant harmful potential to individuals or groups. Moving forward with the implementation of an algorithm system with inexplicable decisions means making the possible harm acceptable – which

implements computationally the hierarchies of society based on race, gender, class and others. For Abeba Birhane and Jelle van Dijk, who analyze how the debate about "robots rights" have been used as a diversion tactic about the impacts of artificial intelligence,

> One of the pressing issues in this day and age is that 'intelligent' machines are increasingly used in sustaining forms of oppression. We do not 'blame' the machines (they can take no blame), nor do we say machines must bear 'responsibility' [15], precisely because thiswould relieve those actually responsible from their duties. (Birhane & van Dijk, 2020, p.6)

In this sense, the principle of explicability can be seen as an essential prerogative in fighting algorithmic racism if it's seen as pertinent not only to code lines, but also to the processes of planning, implementing and to whom do the systems benefit or exclude.

In many cases, like in the implementation of biometric surveillance for public security, the predominance of social injustice is evident in the production of imageries – carceral and racialized – in use of the artifact and notions of how to explain the functioning of a system. The search for fairness should take into consideration the patterns of action and conceptualization around the problem that supposedly should be solved by the algorithmic system (Hanna *et al.*, 2020). Sérgio Amadeu da Silveira highlights the contradictions between opaqueness and the implementation by the State in relevant areas, questioning the "convenience and legitimacy of the use by the State of algorithmic systems that not even its employers could explain all its operations" (Silveira, 2019, p.13).

In a report to the United Nations, E. Tendayi Achiume proposes an analysis based on human rights about racial discrimination in digital emerging technologies. For Achiume, "the heart of the issue is a political, social and economic one, not solely a technological or mathematical problem" (Achiume, 2020, p.15) and, therefore, the States should stablish legal commitment to perform ample scrutiny of the discriminatory possibilities against minorities.

Some of recommendations to the States to fight racial discrimination in designing and using emerging digital technologies are:

- States should adopt immediate and effective measures, particularly in the fields of teaching, education, culture and information, with a view to combating prejudices which lead to racial discrimination;
- Prevent and eliminate racial discrimination in the design and use of emerging digital technologies require addressing the "diversity crisis";
- Make racial equality and non-discrimination human rights impact assessments a prerequisite for the adoption of systems based on such technologies by public authorities;
- States should ensure transparency and accountability for public sector use of emerging digital technologies, and enable independent analysis and oversight, including by only using systems that are auditable;
- Frameworks and guidelines developed to provide flexible, practical and effective regulation and governance of emerging digital technologies are grounded in legally binding international human rights principles. (Achiume, 2020, p.16-17)

The recommendations also re-affirm that the scope of obligations should involve a perspective based in intersectional analysis that can be applied to multiple and overlapping forms of discrimination. The relevance of the debate and the dispute in international organizations around the obligations may impact positively as much in the practices of corporations as in the participation of the civil society in the defense of human rights.

As a society, therefore, we should ask ourselves – and act – about which technologies and public policies we want to include in our possible futures and what we consider as goals and desirable results (Constanza-Schock, 2020). It's already possible to recognize that some algorithmic systems can "function perfectly, with full enrollment, complete transparency, seamless integration and exacting discriminatory power" (Abebe, 2019, p.187),

however. It grows, as a consequence, the perception that some emerging algorithmic technologies can – and should - be the objects of collective rejection.

The *status* of full humanity is multi-faceted and connected to countless everyday processes that restrain or amplify the opportunities, barriers and possibilities of concrete, intellectual and psychological action of individuals, according to their socially perceived affiliation to a racial group. In this overview, the algorithmic racism is a phenomenon directly linked to the problem of double opacity – the way in which hegemonic groups search to present the idea of "neutrality" in technology as much as to dissipate the debate about racism and white supremacy in the West. Studying, debating and acting about the relations between technology and race, thus, becomes doubly challenging in societies dictated by technochauvinism (Broussard, 2018), the racial democracy myth (Nascimento, 2016) or post-racialism (Bonilla-Silva, 2015).

Sueli Carneiro reminds us, though, that the social idea of race has a double sense when evoked as an epistemological tool because of its social transformation. In one side, as a

> methodological instrument, intends to comprehend the unequal relations between the differing human groups, more specifically the inequalities of treatment and perceived social conditions between black and whites in Brazil. As a discursive practice, the studies inspired in it aim to modify the social relationships that produce the discriminations and racial asymmetry. (Carneiro, 2005, p.52)

The dehumanization, the recovery or the maintenance of full humanity of the individuals go through understanding the counter narratives at play, as much as history as collective projects. We can connect the anti-racist thought about the technology not only as critique, but also for new emergencies (Benjamin, 2019) that have as prerogative to reject oppressive potential.

If fatalism is a colonial tool for domination (Lapa, 2018), to think horizons of alternative possibilities, to unveil the naturalization of inequalities and

strengthen paths that addres local and global impacts of algorithmic racism is a task that fortifies alternative imagery for the common good of all humanity.

References

Abdurahman, J. K. (2019, Feb 25). "FAT* Be Wilin'", @blacksirenradio. https://medium.com/@blacksirenradio/fat-be-wilin-deb56bf92539.

Abebe, R. T. (2019). *Designing algorithms for social good*. [Doctoral Dissertation, Cornell University].

Achiume, E. T. (2020). *Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis,* Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance to the United Nations Human Rights Council.

Alexis, G. Y. (2010). "Global Sullivan Principles", in: Nevin Cohen; Paul Robbins (orgs.), *Green Business: An A-to-Z Guide*, Thousand Oaks (EUA): SAGE Publications.

Bahia, S. (2019). Interview given to the author via videoconference in September 2019.

Barros, T. N. (2020). "Estamos em Marcha! Escrevivendo, Agindo e quebrando códigos", *in:* Tarcízio Silva, *Comunidades, Algoritmos e Ativismos Digitais: olhares afrodiaspóricos*, São Paulo: LiteraRUA.

Barros, Z. (2019). *"Feminismo Negro e Internet"*. https://www.academia.edu/1497162/Feminismo_negro_na_Internet acesso em 23 dez. 2019.

Benjamin, R. (2019). *Race after Technology: Abolitionist Tools for the New Jim Code*, Cambridge (UK): Polity Press.

Birhane, A., van Djik, J., (2020). "Robot Rights? Let's Talk about HumanWelfare Instead", *AIES '20*, February 7–8, 2020, New York, NY, USA.

Bonilla-Silva, E. (2015). "The structure of racism in color-blind,"post-racial" America." *American Behavioral Scients*, vol. 59, 11..

Bradshaw, K. (2017, Mar 31). "Vigil in East Palo Alto protests Amazon, Facebook policies", *The Almanac*. https://www.almanacnews.com/news/2017/03/31/vigil-in-east-palo-alto-protests-amazon-facebook-policies.

Brock, A. (2020). *Distributed Blackness: African American Cybercultures*, Nova Iorque: NYU Press.

Broussard, M. (2018). *Artificial Unintelligence: How Computers Misunderstand the World*, MIT Press.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities

in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91).

Carneiro, A. S. (2005). *A construção do outro como não-ser como fundamento do ser*, Tese de Doutorado do Programa de Pós-Graduação em Educação da Universidade de São Paulo, São Paulo – SP.

Castro, S. (2020). "Surveilling Racialized Bodies", *NACLA Report on the Americas*, vol. 52, n.3, 2020.

Castro, S. (2020). "Surveilling Racialized Bodies", *NACLA Report on the Americas*, vol. 52, n.3.

Chan, R. (2019, Aug 17). "Protesters blocked Palantir's cafeteria to pressure the $20 billion big data company to drop its contracts with ICE", *Business* Insider. https://www.businessinsider.com/palantir-protest-palo-alto-activists-ice-contracts-2019-8.

Collins, P. H. (2002). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*, Londres: Routledge.

Constanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*, Cambridge (EUA), MIT Press.

Cunha Júnior, H. (2010). *Tecnologia Africana na Formação Brasileira*, Rio de Janeiro: CEAP.

D'Ignazio, C. D., & Klein, L.F. (2020). *Data Feminism*, Cambrige (USA): The MIT Press.

Daniels, J., Nkonde, M., & Mir, D. (2019). "Advancing Racial Literacy in Tech", Data & Society Fellowship Program Report.

Elias, J. (2020, Jun 22). "Google employees petition company to cancel police contracts", *CNBC*. https://www.cnbc.com/2020/06/22/google-employees-petition-company-to-cancel-police-contracts.html.

Epstein, Z et al. (2018). Closing the AI knowledge gap, *arXiv preprint arXiv:1803.07233*.

Epstein, Ziv *et al* (2018). Closing the AI knowledge gap, *arXiv preprint arXiv:1803.07233*.

Ferreira, M. (2017). "Projeto #BuscaPorIgualdade cobra representatividade negra dos bancos de imagens", *CEERT*, 02 mai. 2017, available at https://ceert.org.br/noticias/comunicacao-midia-internet/16909/projeto-buscaporigualdade-cobra-representatividade-negra-dos-bancos-de-imagens.

Ferreira, Matheus (2017, May 02). "Projeto #BuscaPorIgualdade cobra representatividade negra dos bancos de imagens", *CEERT*. https://ceert.org.br/noticias/comunicacao-midia-internet/16909/projeto-buscaporigualdade-cobra-representatividade-negra-dos-bancos-de-imagens.

Floridi, L, & Cowls, J. (2019). "A Unified Framework of Five Principles for AI in Society", *Harvard Data Science Review*, vol. 1, n.1.

Fouché, R. (2006). "Say It Loud, I'm Black and I'm Proud: African Americans, American Artifactual Culture,and Black Vernacular Technological Creativity", American Quarterly, vol. 58, n.3.

Gangadharan, S.P.; Niklas, J. (2019). "Decentering technology in discourse on discrimination", *Information Communication & Society*, vol. 22, n.7.

Golledge, R. (2017, Dec 14). "Protest at Amazon Rugeley over 'hellish' working conditions", *Express & Star*. https://www.expressandstar.com/news/local-hubs/staffordshire/rugeley/2017/12/14/protest-at-amazons-rugeley-warehouse-over-hellish-working-conditions/.

Gomes, J. (2020). Interview given to the author via email.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). "Towards a Critical Race Methodology in Algorithmic Fairness", *Conference on Fairness, Accountability, and Transparency (FAT* '20)*, Barcelona (Espanha).

Holmes, A. (2020, Sep 01). "Amazon posted — and then deleted — a job listing for an 'intelligence analyst' to monitor workers' efforts to unionize", *Business Insider*. https://www.businessinsider.com/amazon-posts-deletes-job-listing-intelligence-analyst-spy-worker-union-2020-9.

Kari Paul, "Protesters demand Amazon break ties with Ice and Homeland Security", *The Guardian*, 11 jul. 2019, disponível em <https://www.theguardian.com/us-news/2019/jul/11/amazon-ice-protest-immigrant-tech>, acesso em: 20 set. 2019.

Katwala, A. (2020, May 15). "An Algorithm Determined UK Students' Grades. Chaos Ensued", *Wired*. https://www.wired.com/story/an-algorithm-determined-uk-students-grades-chaos-ensued/.

Krishna, A. (2020, Jun 8). "IBM CEO's Letter to Congress on Racial Justice Reform", *IBM THINKPolicy Blog*. https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms.

Lapa, R. S. (2018). "O fatalismo como estratégia colonial", *Revista Epistemologias do Sul*, v. 2.

Lavigne, S., Clifton, B., & Tseng F. (2017). "Predicting Financial Crime: Augmenting the Predictive Policing Arsenal", *arXiv:1704.07826*.

Machado, C. E. D., & Loras, A. (2017). *Gênios da Humanidade: Ciência, Tecnologia e Inovação Africana e Afrodescendente*, São Paulo: DBA.

Matsakis, L (2020 Jun 10). "Amazon Won't Let Police Use Its Facial-Recognition Tech for One Year", *Wired*. https://www.wired.com/story/amazon-facial-recognition-police-one-year-ban-rekognition.

McIlwain, C. D. (2019). *Black Software: The Internet and Racial Justice, from the AfroNet to Black Lives Matter*, Oxford University Press, USA.

Mendes, R. (2020). Interview given to the author.

Morgan, E. (2006). "The world is watching: Polaroid and South Africa", *Enterprise & Society*, vol. 7, n.3.

Nascimento, A. (2016). *O genocídio do negro brasileiro: processo de um racismo mascarado*, São Paulo: Editora Perspectiva.

Nelson, A., Tu, T. L. N., & Hines, A. H. (orgs.) (2001). *Technicolor: Race, technology, and everyday life*. Nova Iorque (EUA): NYU Press.

Njeri, A., & Ribeiro, K. (2019). “Mulherismo Africana: práticas na diáspora brasileira”, *Currículo sem Fronteiras*, vol. 19, n.2.

Noble, S. U., & Roberts,S. (2020). “Elites tecnológicas, meritocracia e mitos pós raciais no Vale do Silício”, *Fronteiras – estudos midiáticos*, vol. 22, n.1.

Nunes, C. (2018, Jan 08). “O Algoritmo”, *Blogueiras Negras*. http://blogueirasnegras.org/o-algoritmo.

Nunes, Pablo (2020). Interview given to the author via email, May 2020.

Paes, Bárbara (2019). Interview given to the author , September 2019.

Peters, J. (2020, Abr 20). “Whole Foods is reportedly using a heat map to track stores at risk of unionization”, *The Verge*. https://www.theverge.com/2020/4/20/21228324/amazon-whole-foods-unionization-heat-map-union.

Peters, J. (2020, Aug 13). “Big Tech Pledged a Billion to Racial Justice, But it was Pocket Change”, *The Verge*. https://www.theverge.com/21362540/racial-justice-tech-companies-donations-apple-amazon-facebook-google-microsoft.

PretaLab (2023). *Website Pretalab*. https://www.pretalab.com.

Raji, I. D., & Buolamwini, J. (2019). “Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products”, in: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). “Auditing algorithms: Research methods for detecting discrimination on internet platforms”, *Data and discrimination: converting critical concerns into productive inquiry*, Seattle (USA).

Santiago, L. (2020). “Tecnologia Antirracista: a resposta das mulheres negras”, in: FOPIR, *Mapeamento da Mídia Negra no Brasil*. http://fopir.org.br/wp-content/uploads/2020/08/ebook_mapeamento_da_midia_negra-1.pdf.

Shutterstock. [@Shutterstock]. (2017, June 06). Refine your search further with our new People Filter. *Twitter*. https://twitter.com/Shutterstock/status/872105821609570304

Silveira, S. A. (2019). "Quem governa os algoritmos? A Regulação dos Sistemas Algorítmicos no Setor Público", *Anais do 43º. Encontro da Associação Nacional de Pós-graduação em Ciências Sociais – Anpocs*.

Vieira, Carla (2019). Interview given to the author, October 2019.

Vilarino, Ramon (2020). Interview given to the author via email, July 2020.

Werneck, J. (2010). Nossos passos vêm de longe! Movimentos de mulheres negras e estratégias políticas contra o sexismo e o racismo. *Revista da Associação Brasileira de Pesquisadores/as Negros/as (ABPN)*, *1*(1), 07-17.

Werneck, Jurema (2009), "Our steps come from afar! Movimentos de mulheres negras e estratégias políticas contra o sexismo e o racismo", in: Christine Verschuur, *Vents d'Est, vents d'Ouest: Mouvements de femmes et féminismes anticoloniaux*, Geneva (Suíla): Graduate Institute Publications, 2009.

Wiewiórowski, W. (2020, Feb 21). "AI and Facial Recognition: Challenges and Opportunities", *European Data Protection Supervisor*. https://edps.europa.eu/press-publications/press-news/blog/ai-and-facial-recognition-challenges-and-opportunities_en.

Yasin, A. (2020, Aug 17). ""Fuck the Algorithm"; the Rallying Cry Of Our Youth?", *DigitalDiplomacy*. https://medium.com/digital-diplomacy/fuck-the-algorithm-the-rallying-cry-of-our-youth-dd2677e190c.

Yeginsu, C. (2018, Feb 01). "If Workers Slack Off, the Wristband Will Know. (And Amazon Has a Patent for It.)", *NY Times*. https://www.nytimes.com/2018/02/01/technology/amazon-wristband-tracking-privacy.html.

Biografias dos/as autores/as

Joaquim Paulo Serra

Doutor em Ciências da Comunicação. Professor do Departamento de Comunicação, Filosofia e Política da Universidade da Beira Interior e investigador integrado do LabCom – Comunicação e Artes.

PhD in Communication Sciences. Professor in the Department of Communication, Philosophy and Politics at the University of Beira Interior and integrated researcher at LabCom - Communication and Arts.

Doctor en Ciencias de la Comunicación. Profesor del Departamento de Comunicación, Filosofía y Política de la Universidad de Beira Interior e investigador integrado en LabCom - Comunicación y Artes.

E-mail: pserra@ubi.pt

Krishma Carreira

Jornalista. É doutora pelo programa de Pós-Graduação em Comunicação Social da Universidade Metodista de São Paulo (Umesp), com tese sobre "Reportagens Investigativas sobre Sistemas de Decisões Automatizadas", e mestre pelo mesmo programa, com a dissertação "Notícias automatizadas: a evolução que levou o jornalismo a ser feito por não humanos".

Journalist. Krishma holds a doctorate degree from the Graduate Program in Social Communication at the Methodist University of São Paulo (Umesp), with a thesis on "Investigative Reporting on Automated Decision Systems", and a master's degree from the same program, with the dissertation "Automated news: the evolution that led journalism to be done by non-humans".

Periodista. Es doctora por el Programa de Posgrado en Comunicación Social de la Universidad Metodista de São Paulo (Umesp), con la tesis "Reportaje de investigación sobre sistemas automatizados de decisión", y máster por el mismo programa, con la disertación "Noticias automatizadas: la evolución que llevó al periodismo a ser hecho por no humanos".

E-mail: krishmacarreira@gmail.com

André Lemos

André Lemos é escritor e professor titular do Departamento de Comunicação e do Programa de Pós-Graduação em Comunicação e Cultura Contemporâneas da Faculdade de Comunicação da UFBA, Diretor do Lab404 - Laboratório de Pesquisa em Mídia Digital, Redes e Espaço, pesquisador "1 A" do CNPq e Membro do Comitê Gestor do Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT-DD). Membro titular da Academia de Ciências da Bahia. Doutor em Sociologia pela Université René Descartes, Paris V, Sorbonne (1995). Esta pesquisa foi apoiada pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) sob os números 307448/2018-5 e 101235/2022-4.

André Lemos is a writer and full professor at the Department of Communication and the Postgraduate Program in Contemporary Communication and Culture at the Faculty of Communication at UFBA, Director of Lab404 - Research Laboratory in Digital Media, Networks and Space, CNPq "1 A" researcher and Member of the Management Committee of the National Institute of Science and Technology in Digital Democracy (INCT-DD). Full member of the Bahia Academy of Sciences. PhD in Sociology from the Université René Descartes, Paris V, Sorbonne (1995). This research was supported by CNPq (National Council for Scientific and Technological Development) under the numbers 307448/2018-5 and 101235/2022-4.

André Lemos es escritor y profesor titular del Departamento de Comunicación y del Programa de Posgrado en Comunicación y Cultura Contemporáneas de la Facultad de Comunicación de la UFBA, Director del Lab404 - Laboratorio de Investigación en Medios Digitales, Redes y Espacio, investigador "1 A" del CNPq y Miembro del Comité

Gestor del Instituto Nacional de Ciencia y Tecnología en Democracia Digital (INCT--DD). Miembro de número de la Academia de Ciencias de Bahía. Doctor en Sociología por la Universidad René Descartes, París V, Sorbona (1995). Esa investigación fue apoyada por el CNPq (Consejo Nacional de Desarrollo Científico y Tecnológico) bajo los números 307448/2018-5 y 101235/2022-4.

E-mail: almlemos@gmail.com

Rosa Franquet

Rosa Franquet é Professora de Comunicação Audiovisual e Publicidade na Universidade Autónoma de Barcelona (UAB) e Directora de Projectos do GRISS-UAB (Grup de Recerca en Imatge So i Síntesi). Presidente da Asociación Española de la Investigación de la Comunicación (AE-IC), ex-presidente da Societat Catalana de Comunicació (2009-2014) e foi directora académica do Programa de Doutoramento do Departamento de Comunicação Audiovisual e Publicidade da UAB. Foi PI e investigadora em vários projectos competitivos nacionais e internacionais. Participou como avaliadora externa em projectos de investigação europeus e espanhóis. Foi professora e investigadora convidada em várias universidades internacionais, incluindo: RMIT (Austrália); Universidade de Melbourne (Austrália); Universidade da Califórnia em Berkeley; Universidade da Califórnia em Davis; Universidade Estadual de São Francisco (EUA); Universidade de Londres (Goldsmiths) (Grã-Bretanha); Universidade de São Paulo ou Universidade Iberoamericana (México). É especialista em indústrias culturais e estudos de género, tendo orientado numerosas teses de doutoramento e publicado inúmeros artigos e livros.

Rosa Franquet is Professor of Audiovisual Communication and Advertising at the Universitat Autònoma de Barcelona (UAB) and Project Director of GRISS-UAB (Grup de Recerca en Imatge So i Síntesi). President of the Spanish Association of Communication Research (AE-IC), https://ae-ic.org/, former President of the Societat Catalana de Comunicació. (2009-2014) and was academic director of the PhD Program of the Department of Audiovisual Communication and Advertising at UAB. She has been PI and researcher in several national and international competitive projects. She has

participated as external evaluator in European and Spanish research projects. She has been visiting professor and researcher at several international universities including: RMIT (Australia); University of Melbourne (Australia); University of California at Berkeley; University of California at Davis; San Francisco State University (USA); University of London (Goldsmiths) (Great Britain); Universidade de São Paulo or Universidad Iberoamericana (Mexico). She is a specialist in cultural industries and gender studies, has directed numerous doctoral theses and has published numerous articles and books.

Rosa Franquet es catedrática de Comunicación Audiovisual y Publicidad de la Universitat Autònoma de Barcelona (UAB) y Directora de Proyectos del GRISS-UAB (Grup de Recerca en Imatge So i Síntesi). Presidenta de la Asociación Española de la Investigación de la Comunicación (AE-IC), ex Presidenta de la Societat Catalana de Comunicació. (2009-2014) y fue directora académica del Programa de Doctorado del Departamento de Comunicación Audiovisual y Publicidad de la UAB. Ha sido IP e investigadora en varios proyectos competitivos nacionales e internacionales. Ha participado como evaluadora externa en los proyectos de investigación europeos y españoles. Ha sido profesora e investigadora invitada de diversas universidades internacionales entre otras: RMIT (Australia); University of Melbourne (Australia); University of California at Berkeley; University of California at Davis; San Francisco State University (EE.UU.); University of London (Goldsmiths) (Gran Bretaña); Universidade de São Paulo o Universidad Iberoamericana (Mexico). Es especialista en industrias culturales y en estudios de género, en su larga trayectoria, ha dirigido numerosas tesis doctorales y ha publicado numerosos artículos y libros.

E-mail: rosa.franquet@uab.cat

Johanna K Monagreda

Possui graduação em Ciências Políticas e Administrativas - Universidade Central de Venezuela, mestrado e doutorado em Ciência Política pela Universidade Federal de Minas Gerais. É consultora na Data Privacy Brasil, pesquisadora do Grupo de pesquisa sobre gênero, raça, cultura e sociedade da Universidade do Estado da Bahia - CANDACES, e pesquisadora do Núcleo de Estudos e Pesquisas sobre a Mulher

(NEPEM-UFMG). Desenvolve pesquisas na área de proteção de dados pessoais, privacidade, inteligência artificial, raça e racismo, movimentos sociais afrolatinoamericanos, gênero e feminismo. É co-organizadora do livro "Construindo caminhos para a justiça de dados no Brasil: o papel das Defensorias Públicas na proteção de dados pessoais".

She holds a degree in Political and Administrative Sciences from the Central University of Venezuela, a master's degree and a doctorate in Political Science from the Federal University of Minas Gerais. She is a consultant at Data Privacy Brasil, a researcher at the Research Group on Gender, Race, Culture and Society at the State University of Bahia - CANDACES, and a researcher at the Center for Studies and Research on Women (NEPEM-UFMG). She conducts research into personal data protection, privacy, artificial intelligence, race and racism, Afro-Latin American social movements, gender and feminism. She is co-organizer of the book "Building paths to data justice in Brazil: the role of Public Defenders in the protection of personal data".

Es licenciada en Ciencias Políticas y Administrativas por la Universidad Central de Venezuela, máster y doctora en Ciencias Políticas por la Universidad Federal de Minas Gerais. Es consultora de Data Privacy Brasil, investigadora del Grupo de Investigación sobre Género, Raza, Cultura y Sociedad de la Universidad Estatal de Bahía - CANDACES, e investigadora del Centro de Estudios e Investigaciones sobre la Mujer (NEPEM-UFMG). Investiga sobre protección de datos personales, privacidad, inteligencia artificial, raza y racismo, movimientos sociales afrolatinoamericanos, género y feminismo. Es coorganizadora del libro "Construyendo caminos para la justicia de datos en Brasil: el papel de los Defensores Públicos en la protección de datos personales".

E-mail: johanna.monagreda@gmail.com

Tarcízio Silva

Tarcizio Silva é Tech Policy Fellow na Fundação Mozilla, Mestre em Comunicação (UFBA) e realiza pesquisa de Doutorado sobre regulação de IA (UFABC). Autor de "Racismo Algorítmico: inteligência artificial e discriminação nas redes digitais"

(Edições Sesc, 2022), relatórios e coletâneas sobre tecnologia e sociedade, disponíveis em desvelar.org.

Tarcizio Silva is a Tech Policy Fellow at the Mozilla Foundation, has a Master's degree in Communication (UFBA) and is doing doctoral research on AI regulation (UFABC). He is the author of "Algorithmic Racism: Artificial Intelligence and Discrimination in Digital Networks" (Edições Sesc, 2022), reports and collections on technology and society, available at desvelar.org.

Tarcizio Silva es Tech Policy Fellow en la Fundación Mozilla, tiene un Máster en Comunicación (UFBA) y está realizando una investigación doctoral sobre la regulación de la IA (UFABC). Es autor de "Racismo algorítmico: inteligencia artificial y discriminación en las redes digitales" (Edições Sesc, 2022), informes y colecciones sobre tecnología y sociedad, disponibles en desvelar.org.

E-mail: eu@tarciziosilva.com.br

Adriana Gonçalves

Doutoranda em Ciências da Comunicação na Universidade da Beira Interior, com uma tese sobre os efeitos da produção automática de notícias nas rotinas jornalísticas. Realiza a sua investigação de doutoramento no Labcom - Comunicação e Artes. É mestre em Jornalismo (2020) e licenciada em Ciências da Comunicação (2018) pela Universidade da Beira Interior. Foi Professora Assistente na Escola Superior de Comunicação Social (ESCS, Instituto Politécnico de Lisboa) em 2021, onde lecionou a unidade curricular de Comunicação e Linguagem.

PhD student in Communication Sciences at the University of Beira Interior, with a thesis on the effects of automatic news production on journalistic routines. She is carrying out her doctoral research at Labcom - Comunicação e Artes. She holds a master's degree in Journalism (2020) and a degree in Communication Sciences (2018) from the University of Beira Interior. She was an Assistant Professor at the School of Social Communication (ESCS, Polytechnic Institute of Lisbon) in 2021, where she taught the Communication and Language course.

Doctoranda en Ciencias de la Comunicación por la Universidad de Beira Interior, con una tesis sobre los efectos de la producción automática de noticias en las rutinas periodísticas. Realiza su investigación doctoral en Labcom - Comunicação e Artes. Es máster en Periodismo (2020) y licenciada en Ciencias de la Comunicación (2018) por la Universidad de Beira Interior. Fue profesora asistente en la Escuela Superior de Comunicación Social (ESCS, Instituto Politécnico de Lisboa) en 2021, donde impartió la asignatura Comunicación y Lenguaje.

E-mail: adriana.goncalves@ubi.pt

Luísa Torre

É investigadora do projeto MediaTrust.Lab - Local Media Lab for Civic Trust and Literacy, desenvolvido no LabCom - Comunicação e Artes da Universidade da Beira Interior (Covilhã, Portugal). Mestre em Ciências da Comunicação pela Universidade do Porto (Portugal) e licenciada em Comunicação Social - Jornalismo pela Universidade Federal do Espírito Santo (Brasil). Foi por 10 anos jornalista de meios de comunicação regionais onde atuou como repórter, colunista, editora e fotojornalista. Investiga desertos de notícias, desinformação, jornalismo, democracia e plataformas de redes sociais.

Luísa is a researcher on the project MediaTrust.Lab - Local Media Lab for Civic Trust and Literacy, developed at LabCom - Communication and Arts at the University of Beira Interior (Covilhã, Portugal). She has a Master's degree in Communication Sciences from the University of Porto (Portugal) and a degree in Social Communication - Journalism from the Federal University of Espírito Santo (Brazil). For 10 years she was a journalist for regional media, where she worked as a reporter, columnist, editor and photojournalist. She researches news deserts, disinformation, journalism, democracy and social media platforms.

Es investigadora del proyecto MediaTrust.Lab - Local Media Lab for Civic Trust and Literacy, desarrollado en el LabCom - Comunicación y Artes de la Universidad de Beira Interior (Covilhã, Portugal). Es licenciada en Ciencias de la Comunicación por la Universidad de Oporto (Portugal) y en Comunicación Social - Periodismo por la

Universidad Federal de Espírito Santo (Brasil). Durante 10 años fue periodista en medios regionales, donde trabajó como reportera, columnista, redactora y fotoperiodista. Investiga los desiertos informativos, la desinformación, el periodismo, la democracia y las plataformas de medios sociales.

E-mail: luisa.torre@ubi.pt

Paulo Victor Melo

Investigador integrado do Instituto de Comunicação da Universidade Nova de Lisboa, com pesquisa sobre tecnovigilância no espaço público. Professor da Faculdade de Design, Tecnologia e Comunicação - IADE/Universidade Europeia. Coordenador do Centro de Comunicação, Democracia e Cidadania da Universidade Federal da Bahia (UFBA)/Brasil. Doutor em Comunicação e Cultura Contemporâneas pela UFBA. Integrante de associações científicas de Comunicação no Brasil, a exemplo da Sociedade Brasileira de Estudos Interdisciplinares da Comunicação (Intercom), como Diretor de Projetos, e da Associação Brasileira de Pesquisadores em Comunicação e Política (Compolítica), em que atua como Secretário.

Integrated researcher at the Communication Institute of the New University of Lisbon, with research on technovigilance in the public space. Professor at the Faculty of Design, Technology and Communication - IADE/European University. Coordinator of the Center for Communication, Democracy and Citizenship at the Federal University of Bahia (UFBA)/Brazil. PhD in Contemporary Communication and Culture from UFBA. He is a member of scientific communication associations in Brazil, such as the Brazilian Society for Interdisciplinary Communication Studies (Intercom), as Project Director, and the Brazilian Association of Researchers in Communication and Politics (Compolítica), where he serves as Secretary.

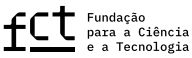
Investigador integrado en el Instituto de Comunicación de la Universidad Nueva de Lisboa, con investigaciones sobre tecnovigilancia en el espacio público. Profesor de la Facultad de Diseño, Tecnología y Comunicación - IADE/Universidad Europea. Coordinador del Centro de Comunicación, Democracia y Ciudadanía de la Universidad Federal de Bahía (UFBA)/Brasil. Doctor en Comunicación y Cultura Contemporáneas

por la UFBA. Es miembro de asociaciones científicas de comunicación en Brasil, como la Sociedad Brasileña de Estudios Interdisciplinarios de Comunicación (Intercom), como Director de Proyectos, y la Asociación Brasileña de Investigadores en Comunicación y Política (Compolítica), de la que es Secretario.

E-mail: paulomelo@fcsh.unl.pt

DOI FCT - MediaTrust.Lab
http://doi.org/10.54499/PTDC/COM-JOR/3866/2020

DOI FCT - LABCOM
https://doi.org/10.54499/UIDB/00661/2020



Esta obra reúne contributos que discutem o papel da Inteligência Artificial e dos algoritmos na sociedade. O debate foi mobilizado por um conjunto de questões, entre elas: Quais os papéis desempenhados pela IA e pelos algoritmos nas sociedades democráticas? Que questões éticas devem ser consideradas na adoção dessas tecnologias por instituições públicas e privadas? Como garantir a aplicação da IA e dos algoritmos sem que discriminações e vieses sejam reforçados? De que modo o jornalismo é impactado pela produção de conteúdos automatizados? Quais as implicações do uso de algoritmos nas notícias? Como a IA contribui para a disseminação e combate à desinformação? A IA e os algoritmos continuarão a fazer parte das práticas comunicativas na sociedade contemporânea. O seu impacto estende-se ao ecossistema informativo, às relações sociais e à qualidade das democracias. Resta aos investigadores problematizar as lógicas algorítmicas e procurar aprofundar o conhecimento sobre as tecnologias emergentes.

Este livro conta com o apoio do MediaTrust.Lab - Laboratório de Media Regionais para a Confiança e Literacia Cívicas, executado no LabCom - Comunicação e Artes, unidade de investigação da Universidade da Beira Interior, participado pela Universidade de Coimbra e financiado pela Fundação para a Ciência e a Tecnologia (PTDC/COM-JOR/3866/2020).

Editora LabCom
www.labcom.ubi.pt