Instituto
de Tecnologia
& Sociedade
do Rio

Trabalho final do IV Grupo de Pesquisa ITS Rio

Discriminação tecnológica: desmistificando a neutralidade da Inteligência Artificial em meio à crise de inclusão e de diversidade nas tecnologias emergentes

Paula Guedes

Direito

Atualmente, é impensável e talvez até impossível desassociar o ser humano da tecnologia. Quem não gosta do conforto de receber listas com indicações de conteúdo especialmente preparada para si? Ou da facilidade de desbloqueio de *smartphones* com biometria ou reconhecimento facial? Ou até a otimização de tempo com a utilização de ferramentas de busca, assistentes virtuais e dispositivos inteligentes? Certamente, a Inteligência Artificial (AI) está cada vez mais presente em nossas vidas cotidianas, desde as funções mais simples, como recomendações de produtos ou serviços, às complexas, a exemplo de otimização de processos, auxílio na descoberta de novos medicamentos e até em ferramentas antifraude, o que gera diversos benefícios para o ser humano. [1]

Por isso, a percepção da maioria da sociedade é que os sistemas baseados em Inteligência Artificial tendem a ser naturalmente neutros, objetivos e imparciais, a partir da habilidade de tomar decisões, fazer previsões e otimizar processos de forma automatizada, por meio de uma enorme disponibilização de dados, supostamente neutralizando a subjetividade humana e alcançando resultados (*outputs*) cada vez mais justos e imparciais.[2] Porém, na prática, tal presunção não se mostra verídica, uma vez que esses sistemas podem refletir os preconceitos e vieses humanos já existentes na sociedade, de forma a violar direitos humanos variados, especialmente de grupos historicamente marginalizados[3], como negros, mulheres, deficientes, pobres, membros da comunidade LGBT e até alguns grupos étnicos minoritários.

Hoje, não há mais dúvidas de que a IA, como tecnologia emergente, possui enorme capacidade de reproduzir, reforçar e até exacerbar a desigualdade já existente em diferentes contextos, já que a tecnologia é produto da sociedade, de seus valores, prioridades e, inclusive, desigualdades, o que inclui as relacionadas ao racismo, ódio e intolerância. O *design* e o uso dessas ferramentas podem, direta ou indiretamente, de forma intencional ou não, discriminar determinados grupos sociais[4]. Muitas dessas possíveis violações de direitos humanos não são novas, mas exacerbadas pela escala, volume, rápida (e descuidada) proliferação e impactos reais imediatos facilitados pela IA[5]. A marginalização e discriminação de certas camadas da sociedade são, então, refletidas nos dados e reproduzidas nos resultados que consolidam padrões históricos de preconceitos enraizados[6].

Os Estados Unidos são um exemplo atual de como as tecnologias digitais emergentes, como a IA, sustentam e reproduzem estruturas discriminatórias na justiça criminal, desde o policiamento até o processo de tomada de decisão por juízes. Vários estados do país já utilizam ferramentas de avaliação de risco em todas as etapas do processo criminal para, após processamento inicial de dados, gerar uma pontuação para determinado indivíduo e, com isso, rotular o indivíduo em baixo, médio ou alto risco de reincidência em determinado crime. A partir dessa análise, os resultados são utilizados por juízes no processo de tomada de decisão a respeito de concessão

ou não de fiança e liberdade condicional, delimitação de tempo de sentença e até aplicação de medidas de segurança[7].

Em 2016, um estudo da *ProPublica* demonstrou que uma das ferramentas mais utilizadas nos EUA para avaliação de risco de reincidência criminal, conhecida como COMPAS *(Correctional Offender Management Profiling for Alternative Sanctions)*, apresentava viés racial, classificando pessoas brancas como menos arriscadas e negras como mais arriscadas do que de fato eram, além de identificar afrodescendentes como duas vezes mais propícios do que os brancos ao cometimento de crimes violentos.[8] Tal diferenciação ocorre principalmente em razão da utilização de dados históricos de prisões e condenações anteriores, que acabam por perpetuar práticas policiais e judiciais racistas, exacerbando as disparidades raciais enraizadas na sociedade.[9]

Ainda, mesmo que o sistema não utilize a raça como critério de análise, o uso de padrões sociais para avaliação de risco, como educação, emprego e fatores econômicos, também pode culminar em discriminação racial de forma indireta. Em razão da desigualdade socioeconômica sistêmica presente em algumas sociedades, que afeta mais fortemente indivíduos negros, esse grupo tem mais chances também de ser rotulado desfavoravelmente por critérios socioeconômicos.[10] Desta forma, essas ferramentas permitem que pessoas tenham direitos fundamentais negados por escolhas algorítmicas de que não têm acesso e dados históricos que as colocam sistematicamente em posição de desvantagem.[11]

Para além da utilização de Inteligência Artificial na justiça criminal, outros usos do nosso dia a dia já se provaram igualmente enviesados. Após anos de pesquisa sobre a ferramenta de busca do Google, a pesquisadora Safiya Noble constatou que esses sistemas discriminam meninas e mulheres, principalmente negras, que são comercializadas, sexualizadas e discriminadas em suas identidades. Por exemplo, até 2016, o resultado da busca por "meninas negras" *("black girls")* correspondia a sites pornográficos e anúncios de conteúdo sexual, mesmo sem que fossem incluídos quaisquer indicativos relacionados à pornografia ou sexo na pesquisa.[12]

De forma semelhante, na busca por imagens na plataforma do Google, os resultados para as palavras-chave "penteados não profissionais para o trabalho" traziam referências a mulheres com cabelos cacheados ou penteados afro, enquanto os resultados da pesquisa por "mulher" *("woman")* ou "menina" *("girl")* eram representados majoritariamente por mulheres e meninas brancas. Além de questões de gênero, a ferramenta também se mostrou racialmente enviesada: em 2016, um adolescente afro-americano tornou-se viral ao divulgar um vídeo de sua pesquisa no Google Imagens para "três adolescentes negros". Os resultados encontrados eram imagens associadas à criminalidade, o que não acontecia ao alterar as palavras-chaves para "três adolescentes brancos", associados a cenários felizes e saudáveis.[13]

Esses resultados demonstram as visões de mundo hegemônicas e as narrativas dominantes dos desenvolvedores que construíram tais sistemas. De acordo com o último relatório de diversidade publicado pelo Google, 67,5% da empresa é composta por homens e 43,1% de brancos, enquanto o percentual de mulheres e de negros é de, respectivamente 32,5% e 5,5%[14]. Essa sub-representação de mulheres e negros, principalmente nos níveis mais altos de decisão, é característica comum entre as principais empresas de tecnologia da atualidade, a exemplo de Google, Facebook, Microsoft, Apple e Amazon. A mesma crise de diversidade de gênero e raça é encontrada também em cursos universitários associados à tecnologia que não costumam apresentar disciplinas relacionadas à ética e aos direitos humanos.[15]

Como resultado, os valores culturais, econômicos e políticos existentes nas *big techs*, atualmente concentradas no Vale do Silício, nos Estados Unidos, são repassadas para o código e os recortes de dados utilizados[16]. Desta forma, a lacuna de diversidade é um dos grandes motivos da discriminação algorítmica, pois esses sistemas, mesmo de forma não intencional, herdam vieses e preconceitos dos desenvolvedores, que podem reproduzir *bias* enraizados no contexto da sociedade em que estão inseridos, que tendem a ser ambientes extremamente brancos, masculinos e ricos, com histórico de problemas de discriminação, exclusão e assédio sexual.[17]

Outro exemplo de tecnologia emergente com base em IA e tendência discriminatória é o reconhecimento facial. Um estudo de 2019 do *National Institute of Standards and Technology* (NIST)[18], que avaliou 189 algoritmos de reconhecimento facial pertencentes a 99 desenvolvedores ao redor do mundo, constatou que a maioria tinha de 10 a 100 vezes mais chances de identificar imprecisamente o rosto negro ou asiático em comparação com o branco, o que é agravado quando a análise é feita em mulheres.[19] No mesmo sentido, pesquisa feita pelo *Institute of Electrical and Electronics Engineers* (IEEE) concluiu que, em diferentes grupos demográficos, a ferramenta apresenta menor acurácia em pessoas do sexo feminino, negros e jovens entre 18 a 30 anos.[20] Especificamente em relação aos *softwares* de reconhecimento facial da Microsoft e IBM, análise da pesquisadora Joy Buolamwini do MIT confirmou o melhor desempenho em homens brancos (94-88% de acurácia) e pior em mulheres negras (79,2-65,3% de acurácia).[21]

Além do constrangimento enfrentado pelos indivíduos equivocadamente não reconhecidos (falsos negativos) ou reconhecidos (falso positivo) pelas ferramentas, que comprovadamente discriminam em razão de gênero e raça, as consequências práticas da falta de acurácia podem violar também outros direitos humanos além da não-discriminação. É o caso de Robert-Julian-Borchak Williams, cidadão negro norte-americano, que foi preso após sua identificação como autor do crime de furto

pelo *software* de reconhecimento facial utilizado pela polícia estadual de Michigan. O incidente foi considerado o primeiro caso conhecido de falha no reconhecimento facial que levou à prisão de indivíduo por crime que não cometeu.[22]

Além da utilização prematura das ferramentas de reconhecimento facial, postas à disposição do público antes de serem realizados todas as testagens necessárias para a garantia de segurança e precisão[23], um dos principais motivos da falta de acurácia desses *softwares* está na insuficiência de dados de entrada para os grupos discriminados, o que está em desacordo com a diversidade existente na sociedade. Em outras palavras, há uma enorme disponibilidade de dados para um determinado grupo e falta de dados para outros, especialmente aqueles já marginalizados na sociedade. Considerando que a atividade de seleção de dados para alimentação dos sistemas de Inteligência Artificial, por si só, é uma atividade subjetiva, não há dúvidas de que, apesar da neutralidade alegada por parte da sociedade, essas tecnologias não são neutras e destituídas de valores, podendo reproduzir, perpetuar e agravar padrões discriminatórios existentes.[24]

Desta forma, embora haja enorme necessidade de escrutínio e responsabilização pela qualidade técnica e precisão das ferramentas que utilizam Inteligência Artificial, o cumprimento dos princípios da igualdade e não discriminação, além de outros direitos humanos, deve iniciar com o reconhecimento de que o problema não é meramente técnico ou matemático, mas principalmente uma questão social, política e econômica. A perpetuação de vieses nos sistemas de IA não será curada apenas por modelagens tecnológicas perfeitas, mas com a união de agentes e áreas distintas da sociedade, o que inclui, além dos especialistas em tecnologia, o setor público, empresas privadas, sociedade civil e a academia,[25] em aplicação de soluções técnicas, éticas e voltadas para os direitos humanos.

Em regra, as ferramentas de Inteligência Artificial são desenvolvidas com base em métricas de desempenho, como acurácia, velocidade e eficiência, sem levar em consideração a existência de vieses. Por isso, é fundamental a criação de times heterogêneos e multidisciplinares para conduzir as pesquisas e projetos de IA, de forma a incluir também métricas baseadas em direitos humanos e ética.[26] No mesmo sentido, considerando a crise de diversidade do setor tecnológico, é extremamente necessário o recrutamento de maior pluralidade e diversidade para os cargos de cientistas de dados e demais profissionais relacionados à IA, principalmente de indivíduos de grupos sub-representados, para garantia de convivência de visões de mundo diversas e prevalência do respeito e não discriminação.[27]

Ainda, quando nos referimos à inclusão, é necessário que medidas sejam tomadas não apenas para inserção na sociedade de minorias por categorias isoladas, uma vez que as formas de opressão se cruzam e os esforços de diversidade que visam, por exemplo, mulheres, sem reconhecer o papel da raça e outras formas de identidade,

privilegiam implicitamente as mulheres brancas.[28] Diante dessa premissa, torna-se essencial a aplicação de políticas público-privadas de educação, inclusão interseccional e empoderamento tecnológico para capacitar os indivíduos em geral, mas principalmente os grupos minoritários, a entender melhor a IA, abrindo portas para que ascendam a cursos relacionados (e se mantenham neles) e tenham melhores condições de acesso ao mercado de trabalho.

Portanto, a diversidade e a inclusão desempenham papel fundamental no desenvolvimento dos sistemas de Inteligência Artificial no mundo real. A maioria dos desenvolvedores, na maior parte do tempo, não se considera sexista, racista, homofóbico, xenófobo ou opressor, mas está propenso a excluir ou discriminar grupos marginalizados de forma sistemática pelos sistemas de IA.[29] Por isso, as equipes que concebem, desenvolvem, testam, mantém, implantam e compram a tecnologia devem ser diversificadas, não só em termos de gênero, cultura e idade, mas também no que tange a experiências profissionais e competências no geral, de forma a possibilitar uma reflexão sobre as necessidades múltiplas e diversas dos utilizadores e da sociedade em geral, além do respeito igualitário aos direitos humanos. A Inteligência Artificial deve trabalhar em benefício do ser humano, considerado todas as suas diversidades, e não contra ele.

Notas

1. CORTIZ, Diogo. Inteligência Artificial: equidade, justiça e consequências. Panorama Setorial da Internet: Nº 1, Ano 12, Maio de 2020, pp. 1-5. Disponível em: https://www.cetic.br/media/docs/publicacoes/6/20200626161010/panorama_setorial_ano-xii_n_1_inteligencia_artificial_equidade_justi%C3%A7a.pdf. p. 1.

2. BRAGA, Carolina Henrique da Costa. Decisões Automatizadas e Discriminação: Pesquisa de Propostas Éticas e Regulatórias no Policiamento Preditivo. Dissertação de Mestrado do programa de Pós-Graduação em Princípios Fundamentais e Novos Direitos da Universidade Estácio de Sá (UNESA). Orientação do professor Nilton César da Silva Flores. Rio de Janeiro, 2019. p. 10-11; Human Rights Council. Racial discrimination and emerging digital technologies: a human rights analysis. Report of the Special Rapporteur n contemporary forms of racism, racial discrimination, xenophobia, and related intolerance (Advance Edited Version), 18 jun. 2020, A/HRC/44/57. Disponível em: https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session44/Pages/ListReports.aspx. p. 4.

3. BRAGA, Carolina Henrique da Costa. Decisões Automatizadas e Discriminação: Pesquisa de Propostas Éticas e Regulatórias no Policiamento Preditivo. Dissertação de Mestrado do programa de Pós-Graduação em Princípios Fundamentais e Novos Direitos da Universidade Estácio de Sá (UNESA). Orientação do professor Nilton César da Silva Flores. Rio de Janeiro, 2019. p. 10-11.

4. Human Rights Council. Racial discrimination and emerging digital technologies: a human rights analysis. Report of the Special Rapporteur n contemporary forms of racism, racial discrimination, xenophobia, and related intolerance (Advance Edited Version), 18 jun. 2020, A/HRC/44/57. Disponível em: https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session44/Pages/ListReports.aspx. p. 4.

5. ANDERSEN, Lindsey. Human Rights in the Age of Artificial Intelligence. Access Now. Nov. 2018. Disponível em: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf. p. 08.

6. Ibid. p. 18.

7. YAVUZ, Can. Machine Bias: Artificial Intelligence and Discrimination. Faculty of Law, Lund University, Spring term 2019. JAMM07 Master Thesis – International Human Rights Law. Supervisor: Karol Nowak. p. 58 e 59.

8. ANGWIN, Julia; KIRCHNER, Lauren; LARSON, Jeff; MATTU, Surya. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23 mar. 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em 22 jul. 2020.

9. YAVUZ, Can. Machine Bias: Artificial Intelligence and Discrimination. Faculty of Law, Lund University, Spring term 2019. JAMM07 Master Thesis – International Human Rights Law. Supervisor: Karol Nowak. p. 60.

American Civil Liberties Union (ACLU). *A Tele of Two Countries: Racially Targed Arrests in the Era of Marijuana Reform*. ACLU Research Report, 2020. Disponível em: https://www.aclu.org/sites/default/files/field_document/042020-marijuanareport.pdf.

10. YAVUZ, *Can. Machine Bias: Artificial Intelligence and Discrimination*. Faculty of Law, Lund University, Spring term 2019. JAMM07 Master Thesis – International Human Rights Law. Supervisor: Karol Nowak. p. 61.

11. BRAGA, Carolina Henrique da Costa. *Decisões Automatizadas e Discriminação: Pesquisa de Propostas Éticas e Regulatórias no Policiamento Preditivo*. Dissertação de Mestrado do programa de Pós-Graduação em Princípios Fundamentais e Novos Direitos da Universidade Estácio de Sá (UNESA). Orientação do professor Nilton César da Silva Flores. Rio de Janeiro, 2019. p. 56.

12. À título de exemplo, os resultados da pesquisa levavam à sites que incluíam palavras como "sexo" ("sex"), "estrela pornô" ("porn star"), "quente" ("hot"), "hardcore", "bunda" ("ass") e "adolescentes" ("teenagers"); NOBLE, Safiya. Google Has a Striking History of Bias Against Black Girls. Time, 26 mar. 2018. Disponível em: https://time.com/5209144/google-search-engine-algorithm-bias-racism/. Acesso em 22 jul. 2020.

13. NOBLE, Safiya. Google Has a Striking History of Bias Against Black Girls. Time, 26 mar. 2018. Disponível em: https://time.com/5209144/google-search-engine-algorithm-bias-racism/. Acesso em 22 jul. 2020.

14. Google. Google Diversity Annual Report 2020. Disponível em: https://diversity.google/.

15. NOBLE, Safiya. Google Has a Striking History of Bias Against Black Girls. Time, 26 mar. 2018. Disponível em: https://time.com/5209144/google-search-engine-algorithm-bias-racism/. Acesso em 20 mai. 2020.

16. Human Rights Council. Racial discrimination and emerging digital technologies: a human rights analysis. Report of the Special Rapporteur n contemporary forms of racism, racial discrimination, xenophobia, and related intolerance (Advance Edited Version), 18 jun. 2020, A/HRC/44/57. Disponível em: https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session44/Pages/ListReports.aspx. p. 4.

17, CRAWFORD, Kate. Artificial Intelligence's White Guy Problem. The New York Times, 25 jun. 2016. Disponível em: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html. Acesso em 27 mai. 2020.

18. GROTHER, Patrick; NGAN, Mei; HANAOKA, Kayee. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. National Institute of Standards and Technology (NIST), NISTIR 8280, dez. 2019. Disponível em: https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf.

19. BUSHWICK, Sophie. How NIST Tested Facial Recognition Algorithms for Racial Bias. Scientific American, dez. 2019. Disponível em: https://www.scientificamerican.com/article/how-nist-tested-facial-recognition-algorithms-for-racial-bias/. Acesso em 31 jul. 2020.

20. BRUEGGE, Richard W. Vorder; BURGE, Mark J; JJAIN, Anil K; KLARE, Brendan F.; KLONTZ, Joshua C. Face Recognition Performance: Role of Demographic Information. IEEE Transactions on Information Forensics and Security. Disponível em: https://www.openbiometrics.org/publications/klare2012demographics.pdf.

21. BUOLAMWINI, Joy; GEBRU, Timnit. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency. Disponível em: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

22. The New York Times. Wrongfully Accused by an Algorithm. 24 jun. 2020. Disponível em: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html. Acesso em 16 jul. 2020; American Civil Liberties Union (ACLU). Wrongfully Arrested Because Face Recognition Can't Tell Black People Apart. 24 jun. 2020. Disponível em: https://www.aclu.org/news/privacy-technology/wrongfully-arrested-because-face-recognition-cant-tell-black-people-apart. Acesso em 16 jul. 2020.

23. YAVUZ, Can. Machine Bias: Artificial Intelligence and Discrimination. Faculty of Law, Lund University, Spring term 2019. JAMM07 Master Thesis – International Human Rights Law. Supervisor: Karol Nowak. p. 48.

24. BRAGA, Carolina Henrique da Costa. Decisões Automatizadas e Discriminação: Pesquisa de Propostas Éticas e Regulatórias no Policiamento Preditivo. Dissertação de Mestrado do programa de Pós-Graduação em Princípios Fundamentais e Novos Direitos da Universidade Estácio de Sá (UNESA). Orientação do professor Nilton César da Silva Flores. Rio de Janeiro, 2019. p. 53.

25. Human Rights Council. Racial discrimination and emerging digital technologies: a human rights analysis. Report of the Special Rapporteur n contemporary forms of racism, racial discrimination, xenophobia, and related intolerance (Advance Edited Version), 18 jun. 2020, A/HRC/44/57. Disponível em: https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session44/Pages/ListReports.aspx. p. 5.

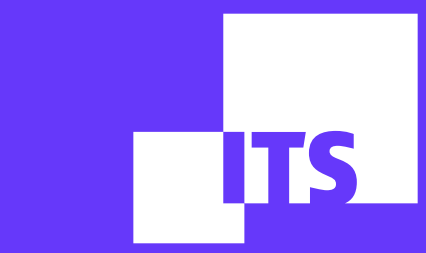
26. CORTIZ, Diogo. Inteligência Artificial: equidade, justiça e consequências. Panorama Setorial da Internet: Nº 1, Ano 12, Maio de 2020, pp. 1-5. Disponível em: https://www.cetic.br/media/docs/publicacoes/6/20200626161010/panorama_setorial_ano-xii_n_1_inteligencia_artificial_equidade_justi%C3%A7a.pdf. p. 3.

27. CRAWFORD, Kate; WEST, Sarah Myers; WHIIAKER, Meredith. Discriminating Systems: Gender, Race, and Power in AI. AI Now Institute, April 2019. Disponível em: https://ainowinstitute.org/discriminatingsystems.html. p. 17.

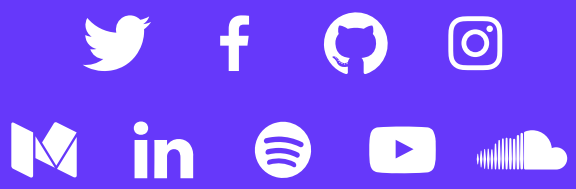
28. JOY, Erica. #FFFFFF Diversity. 7 out. 2015. Disponível em: https://medium.com/this-is-hard/ffffff-diversity-1bd2b3421e8a.

29. COSTANZA-CHOCK, Sasha. Design Justice: Community-Led Practices to Build the Worlds We Need. The MIT Press, 3 mar. 2020, 360 p. Design Values: Hard-Coding Liberation? p. 9.

30. Grupo Independente de Peritos de Alto Nível sobre Inteligência Artificial da Comissão Europeia (High-Level Expert Group on Artificial Intelligence – European Commission). Orientações Éticas para uma IA de Confiança (Guidelines on Trustworthy AI). Abril de 2019. Disponível em: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top. p. 29.



Acesse nossas redes



itsrio.org