Discriminação algorítmica: a relação entre homens e máquinas

Os riscos das interações disfuncionais e da irresponsabilidade organizada

Parte VII

Ana Frazão

Advogada. Professora de Direito Civil e Comercial da UnB. Ex-Conselheira do CADE.

Como já se viu ao longo da série, a transferência ou delegação total do processo decisório de agentes públicos e privados para sistemas algorítmicos é procedimento que envolve diversos riscos, considerando as limitações já apontadas na programação e nos *designs* de tais sistemas.

Um dos maiores riscos é o da irresponsabilidade organizada, pois, como igualmente já se apontou, assim como programadores podem não se sentir responsáveis pelos resultados concretos e pelas utilizações de seus sistemas, usuários podem não se sentir responsáveis quando simplesmente executam uma aplicação de terceiro.

Por outro lado, mesmo quando não ocorre a terceirização total e o sistema algorítmico tem o papel de ser apenas um auxiliar no processo decisório, permanecendo o ser humano com a "última palavra", os desafios não são banais. Afinal, pouco se sabe sobre como se comportam seres humanos diante de decisões algorítmicas, havendo o fundado receio de que tendam a concordar com os seus resultados, até porque, considerando a opacidade dos algoritmos, não conseguem compreendê-los nem questioná-los.

Como mostram Chiodo e Clifton[1], é ingenuidade achar que boas práticas de gestão poderão resolver esse tipo de problema pois, a cada estágio de separação do trabalho matemático traduzido no sistema algorítmico, perde-se parcela de significado sobre o próprio alcance da programação. Consequentemente, fica muito difícil para o tomador de decisões, ao buscar auxílio nos sistemas algorítmicos, entender todo o trabalho matemático que foi feito, assim como as suas limitações.

Ainda segundo Chiodo e Clifton[2], é da própria natureza da gestão que decorrem tais problemas. Assim como um gestor não tem condições de reproduzir todo o trabalho feito pelos seus subordinados, igualmente não faria sentido que pretendesse fazer isso com sistemas algorítmicos, ainda mais quando, diante das características destes, é grande a probabilidade de que não sejam compreendidos, mesmo após consideráveis esforços nesse intento.

Se tal cenário reforça a necessidade de matemáticos e programadores considerarem os resultados concretos de seus sistemas e incluírem parâmetros éticos nas programações, também mostra como a existência de controle humano sobre tais resultados pode ser de pouca utilidade.

Com efeito, muito se tem pensado na questão do controle por pessoa natural como solução para resolver as limitações dos julgamentos algorítmicos. Esse é um dos principais aspectos ressaltados pelas Diretrizes da União Europeia para uma Inteligência Artificial Confiável[3], documento que chega a afirmar que a supervisão humana pode exigir a capacidade de anular a decisão algorítmica.

Da mesma maneira, era essa a intenção original da LGPD ao tratar das decisões totalmente automatizadas em seu art. 20, na qual se previa o direito de explicação e revisão de decisões totalmente automatizadas por meio de pessoa natural, requisito que foi excluído posteriormente pela Lei nº 13.853/2019. Apesar das merecidas críticas à exclusão da revisão por pessoa natural, a grande questão é saber em que medida a supervisão humana é

1 CHIODO, Maurice; CLIFTON, Toby. The Importance of Ethics in Mathematics. European Mathematical Society. Newsletter No. 114, December 2019.
2 Op.cit.
3 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

realmente adequada e eficiente para resolver problemas de decisões algorítmicas enviesadas, discriminatórias ou disfuncionais.

Logo, é fundamental refletir sobre como interagem os usuários com os diagnósticos e predições de sistemas algorítmicos e qual é a medida da sua capacidade de reação a determinados resultados. Afinal, para que a supervisão humana tenha resultado prático, deve haver efetivo controle e possibilidade de divergência diante dos resultados algorítmicos.

Entretanto, tais possibilidades parecem limitadas, como demonstraram Ben Green e Yiling Chen[4] em interessante estudo no qual analisam sistemas algorítmicos como o Compas, utilizados para auxiliar juízes na dosimetria de penas por meio de cálculos sobre o potencial de reincidência dos réus sob julgamento.

Antes de testarem suas hipóteses, os autores tentam mapear os trabalhos anteriores sobre o tema, realçando a parte da literatura que aponta o quanto as pessoas são ruins para incorporar, em suas análises, predições quantitativas. O chamado fenômeno do viés da automação (*automation bias*) sugere que ferramentas de automação influenciam decisões humanas de formas significativas e geralmente ruins. Dois tipos de erros são particularmente comuns: (i) erros de omissão, nos quais as pessoas não reconhecem quando os sistemas automatizados erram e (ii) erros comissivos, nos quais as pessoas seguem os sistemas automatizados sem considerar suas informações contraditórias.

Consequentemente, uma forte confiança em sistemas automatizados pode alterar as relações das pessoas com as suas tarefas, criando uma espécie de "para-choque" entre as decisões e os seus impactos, com a consequente perda do senso de responsabilidade e da *accountability*. Daí o receio de que decisões algorítmicas, embora tenham sido desenhadas para reduzir os erros humanos, possam causar ainda mais problemas.

Por outro lado, os autores igualmente ressaltam estudos que mostram que (i) mesmo diante de algoritmos com maior acurácia, as pessoas

4 GREEN, Ben; CHEN, Yiling. In FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10. 1145/3287560.3287563. https://scholar.harvard.edu/files/19-fat.pdf

nem sempre os incorporam para melhorar suas decisões, preferindo confiar no seu próprio julgamento ou no julgamento de terceiro, (ii) as pessoas não sabem distinguir entre predições confiáveis e predições não confiáveis e (iii) existe uma aversão ao algoritmo, de forma que as pessoas são menos tolerantes aos erros dos algoritmos do que aos erros das pessoas.

Outra parte dos estudos sobre o assunto mostra a importância dos vieses dos julgadores humanos como verdadeiros filtros informacionais, de forma que mesmo informações que supostamente ajudariam as pessoas a tomar melhores decisões podem falhar ao ser incorporadas de forma enviesada no processo decisório.

Como se pode observar, há grande controvérsia a respeito de como os seres humanos formalmente responsáveis por determinadas decisões agem e reagem diante de resultados algorítmicos, existindo riscos tanto de que os sigam de forma irrefletida como de que os apliquem de acordo com seus próprios vieses, o que pode gerar resultados ainda mais disfuncionais.

Todo esse cenário mostra o quanto o sucesso da utilização de sistemas algorítmicos nos processos de decisão depende da prévia reflexão sobre como se dá a complexa relação entre homens e máquinas, a fim de evitar resultados indesejáveis, que vão da falta de supervisão humana à própria deturpação do resultado algorítmico.

Voltando ao estudo de Ben Green e Yiling Chen[5], os autores formulam três hipóteses que, ao final, são confirmadas pela pesquisa empírica: (i) participantes que recebem uma avaliação algorítmica de risco farão predições com menor acurácia do que a avaliação de risco, (ii) participantes serão incapazes de avaliar a sua própria performance e também a performance do algoritmo e (iii) diante da forma como os participantes interagem com a avaliação de riscos, suas posturas serão desproporcionalmente prováveis de aumentar as previsões de risco em relação a réus negros do que réus brancos.

Logo, o estudo mostra o quanto se tem a avançar na compreensão dos reais efeitos de avaliações ou diagnósticos algorítmicos sobre as decisões humanas que deles se utilizam, inclusive para avaliar o risco de que a

5 Op.cit.

intervenção humana não seja capaz de identificar falhas nem de realizar propriamente qualquer controle sobre eventuais problemas dos algoritmos.

Por mais que ainda seja cedo para formular conclusões sobre o assunto, já foi possível verificar que um resultado algorítmico pode inclusive alavancar a discriminação contra um determinado grupo. Não é sem razão que Ben Green e Yiling Chen[6] defendem a necessidade de inserir as avaliações algorítmicas de risco no contexto social e tecnológico, a fim de que seus impactos possam ser identificados e valorados.

Ademais, ficou igualmente claro, no exemplo concreto estudado pelos autores, que a introdução de análises algorítmicas de risco no sistema criminal não elimina a discricionariedade nem cria julgamento mais objetivos, mas simplesmente transfere a discricionariedade para outras "áreas", o que inclui a interpretação judicial da avaliação algorítmica e também a decisão sobre quão fortemente se pode confiar nela.

Seja porque a interação entre homem e máquina pode levar a uma total confiança nos sistemas, seja porque pode levar a interpretações que criem novas áreas de discricionariedade, que podem inclusive distorcer o sistema, é urgente que se entenda minimamente essa interação nos casos concretos, sob pena de tornar inexequível a proposta de que o controle humano possa ser suficiente para resolver problemas de discriminação.

Dessa maneira, tão importante quanto um bom *design* algorítmico, que possa considerar os resultados práticos e os desdobramentos éticos da sua aplicação, é a qualidade da interação entre os usuários e os sistemas algorítmicos, a fim de se evitar resultados disfuncionais e mesmo a irresponsabilidade organizada, efeitos que têm sido amplamente verificados na prática recente, como se verá no próximo artigo da série.

---

6 Op.cit.

Publicado em 28/07/2021

Link:https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/discriminacao-algoritmica-a-relacao-entre-homens-e-maquinas-28072021