Discriminação Algorítmica: Conceito, Fundamento Legal e Tipologia

LAURA SCHERTEL MENDES[1]

Professora adjunta de Direito Civil da Universidade de Brasília (UnB) e da Pós-Graduação em Direito Constitucional do Instituto Brasiliense de Direito Público (IDP), Doutora *summa cum laude* em Direito Privado pela Universidade Humboldt de Berlim, Mestre em Direito, Estado e Constituição pela UnB. Integra o Conselho Diretor da Associação Luso-Alemã de Juristas (DLJV-Berlim) e do Instituto Brasileiro de Política e Direito do Consumidor (Brasilcon). É autora dos livros *Privacidade, Proteção de Dados e Defesa do Consumidor* (Saraiva, 2014) e *Schutz gegen Informationsrisiken und Gewährleistung einer gehaltvollen Zustimmung: Eine Analyse der Rechtmäßigkeit der Datenverarbeitung im Privatrecht (A proteção de dados pessoais no setor privado: riscos do tratamento de dados e a garantia de um consentimento substancial)* (De Gruyter, 2015).

MARCELA MATTIUZZO[2]

Mestra em Direito Constitucional na Universidade de São Paulo. Foi Membro do Grupo Constituição, Política e Instituições na mesma universidade e é sócia de VMCA Advogados. Foi pesquisadora visitante na Yale Law School (2016-2017), Chefe de Gabinete e Assessora da Presidência do Conselho Administrativo de Defesa Econômica (2015-2016).

RESUMO: O artigo visa analisar o debate teórico sobre discriminação algorítmica, com a finalidade de esclarecer o potencial discriminatório de práticas baseadas em *profiling* e decisões automatizadas. Para tanto, discutir-se-á, primeiramente, os conceitos de algoritmo e discriminação algorítmica, buscando esclarecer porque tais conceitos são relevantes na economia movida a dados. Em seguida, o texto apresenta de que forma o *Big Data,* em conjunto com os algoritmos, alterou processos decisórios cotidianos para discutir de que forma tal cenário pode apresentar desafios, especialmente no que diz respeito ao risco de condutas discriminatórias. A última seção visa expor soluções para lidar com o tema da discriminação algorítmica, apresentando a literatura de governança algorítmica, bem como os principais debates entre especialistas no assunto, enfatizando as discussões sobre os limites da transparência enquanto alternativa apta a resolver as questões colocadas.

PALAVRAS-CHAVE: Discriminação algorítmica; decisões automatizadas; *profiling*; proteção de dados; *Big Data*; governança algorítmica.

SUMÁRIO: Introdução; 1 Algoritmos e discriminação estatística – Perfilamento, classificação e *scoring*; 1.1 Algoritmos na economia orientada por dados; 1.2 Discriminação algorítmica; 2 Um debate acerca das soluções na literatura: discriminação algorítmica e governança; Considerações finais; Referências.

1 Orcid: <http://orcid.org/0000-0001-8675-4994>.

2 Orcid: <http://orcid.org/0000-0001-5641-1130>.

INTRODUÇÃO

Uma das atribuições mais importantes do processamento de dados por meio de algoritmos é a de oferecer embasamento para tomadas de decisões dos mais variados tipos, o que, por sua vez, contribui para a mitigação de riscos. A análise e a mitigação dos riscos tornam-se mais relevantes quando aplicadas a setores econômicos nos quais a assimetria de informação é abundante. Essa é a razão pela qual o processamento de dados exerce um importante papel em negócios como a concessão de crédito e o mercado de seguros, nos quais o risco é uma característica inerente. Entretanto, sua relevância também tem aumentado em outros setores, pois o processamento de informações e dados no setor privado mostra-se como um meio de simplificar decisões e de aumentar a eficiência em qualquer ambiente caracterizado por déficits de informação.

O desenvolvimento tecnológico provocou importantes contribuições nesse aspecto: não apenas é possível coletar e processar mais informação para a análise de riscos do que nunca, como também essa análise pode ser quantificada a partir de um *score* – que nada mais é do que uma prognose do comportamento futuro de um indivíduo. Esse *score* é produzido a partir de um procedimento automatizado, no qual os dados existentes são incorporados em um algoritmo e os indivíduos são alocados a uma categoria de risco específica. Desenvolvimentos recentes no campo da tecnologia da informação – que podem ser sintetizados sob o agora popular termo *data analytics* – fornecem incentivos ainda maiores para o uso de projeções no setor privado, na medida em que permitem que mais informações sejam processadas e que novas correlações entre dados e comportamentos futuros possam emergir.

Nos negócios de *credit report* e *credit score*, a automatização dos processos decisórios foi, em um primeiro momento, vista como um meio de superar as tendências de enviesamento e discriminação. No entanto, logo tornou-se claro que o método estatístico, o qual teoricamente receberia dados objetivos como *inputs* e, portanto, deveria gerar resultados objetivos como *outputs*, poderia reproduzir vieses já existentes, levando também a resultados discriminatórios. Isso porque, em primeiro lugar, nexos de causalidade e correlações são, muitas vezes, predefinidos pelos controladores dos dados, que, por sua vez, transmitem aos algoritmos os mesmos vieses presentes nos processos tradicionais de tomada de decisões.

Ou seja, se alguém acredita que as mulheres são inapropriadas de modo geral para alguns tipos de atividade – por exemplo, para a engenharia mecânica – e essa pessoa programa um algoritmo que internaliza tal lógica, o *output* de tal algoritmo poderá apresentar essas mesmas inclinações, independentemente da qualidade do *input*. Mesmo em casos em que o algoritmo seja programado para identificar suas próprias correlações a partir da colheita de dados brutos já existentes – o que deveria eliminar o problema de transferência de predisposições

do programador –, ainda assim poderia acabar reproduzindo correlações discriminatórias presentes em tais dados. Em outras palavras, os algoritmos poderiam absorver padrões discriminatórios presentes na sociedade e replicá-los como uma "verdade objetiva". Ou seja, mesmo que o *designer* do algoritmo não acredite que homens seriam engenheiros mecânicos melhores que mulheres, em havendo no conjunto de dados analisado elementos suficientes a indicar que o gênero pode ser uma variável relevante para determinar tais aptidões – por conta do maior número de homens do que mulheres no ramo da engenharia, por exemplo –, o *output* poderia reproduzir as condições discriminatórias existentes ao invés de auxiliar a superá-las (Barocas; Selbst, 2016)[3].

Ademais, na medida em que os algoritmos se baseiam, em grande parte, em discriminação estatística, isto é, na diferenciação de indivíduos baseada nas características de um grupo e na probabilidade de tal grupo agir de determinada maneira, torna-se indispensável compreender se os processos e critérios utilizados para classificar indivíduos são corretos, transparentes e, em última instância, justos.

Neste artigo, pretendemos analisar o debate teórico sobre discriminação algorítmica, com a finalidade de buscar possíveis soluções contra o potencial discriminatório de práticas baseadas em *profiling* e decisões automatizadas. Para tanto, dedicamos a Seção 1 a delimitar parâmetros e estabelecer definições que se mostrarão úteis ao longo do texto, especialmente em definir brevemente os conceitos de algoritmo e de discriminação algorítmica, bem como explicar a razão pela qual tais conceitos são particularmente significativos para a economia orientada por dados. A Seção 2 apresenta propostas para lidar com a discriminação em decisões baseadas em algoritmos. Para tanto, examina a literatura a respeito de governança algorítmica e as principais divergências presentes nesse debate.

1 ALGORITMOS E DISCRIMINAÇÃO ESTATÍSTICA – PERFILAMENTO, CLASSIFICAÇÃO E *SCORING*

1.1 Algoritmos na economia orientada por dados

Um algoritmo é comumente descrito como um conjunto de instruções, organizadas de forma sequencial, que determina como algo deve ser feito. De maneira alguma é um conceito dependente do uso do poder do computador moderno, pois é possível que alguém crie um algoritmo para auxiliá-la a se vestir, um algoritmo para pegar o ônibus para o trabalho, para fazer uma receita de bolo, ou para inúmeras outras atividades, já que um algoritmo é nada mais do que uma fórmula na qual tarefas são colocadas em uma ordem específica

3 Os autores identificam cinco mecanismos por meio do qual os efeitos nocivos da extração de dados poderiam se dar, e vão além, explicitando a razão pela qual estes resultados discriminatórios emergiriam.

para atingir determinado objetivo. Entretanto, apesar de esta ser uma descrição correta, ela não oferece informações suficientes para o propósito deste artigo. Assim, adotaremos a definição de Thomas Cormen, que é cuidadoso ao indicar que há uma diferença entre um algoritmo qualquer e aqueles – parte dos quais discutimos aqui – que operam em computadores. Computadores, diferentemente de seres humanos, não compreendem o significado de termos como "suficiente", "quase", "ruim" ou qualquer outra palavra que implique em uma avaliação subjetiva do mundo ao seu redor. É por essa razão que um algoritmo que determine que um celular reduza a luz de sua tela sempre que "quase não haja mais bateria" é inútil. Um computador é capaz de interpretar porcentagens, mas não de determinar o que "quase sem bateria" significa, a não ser que alguém explicite como fazê-lo. De acordo com Cormen:

> Você pode ser capaz de tolerar quando um algoritmo é descrito de maneira imprecisa, mas um computador, não. [...] Assim, um algoritmo computacional consiste em uma série de etapas para completar uma tarefa que é descrita de maneira precisa o bastante para que um computador possa realizá-la. (Cormen, 2013, p. 1)

Também é importante notar que o objetivo dos algoritmos, da maneira como são utilizados hoje e discutida aqui, é, sobretudo, solucionar problemas e auxiliar na tomada de decisões. Quando se busca, por exemplo, por voos de São Paulo para Berlim, busca-se resposta para uma pergunta. Mais ainda, o que se quer é encontrar a resposta *correta* para aquela questão. Aqui, novamente, nos deparamos com uma particularidade dos algoritmos: o programa será tanto mais útil quanto mais precisa a informação (ou *input*) fornecida, e estará correto sempre que utilizar essa informação de acordo com suas especificações. Assim, ao buscar o "melhor" voo de São Paulo para Berlim, o algoritmo precisará saber se por "melhor" queremos dizer "mais curto" ou "mais barato". Se o algoritmo é programado para encontrar a rota mais curta, em termos de quilômetros viajados, poderá considerar que o tempo gasto em um aeroporto aguardando um voo de conexão é irrelevante, e poderia, assim, oferecer uma resposta que, apesar de incorreta com relação às nossas preferências (é razoável admitir que para a maior parte das pessoas tempo gasto em conexões é um fator relevante na decisão de rota de viagem), é correta do ponto de vista do programa. O problema, nesse caso, não é com o algoritmo em si, mas sim com as especificações a ele fornecidas.

Nesse sentido, um dos objetivos fundamentais dos algoritmos é fazer previsões utilizando probabilidades. Embora algoritmos não possam fornecer respostas precisas a todas as questões, eles podem analisar os dados fornecidos (*inputs*) e oferecer "palpites" coerentes. Quanto maior a quantidade e qualidade dos dados disponibilizados ao algoritmo, maior a chance de o resultado estar próximo do real. Permita-nos utilizar o exemplo do lançamento de dados para exemplificar nosso argumento. Para qualquer dado, é possível assumir que a

probabilidade de que o número 5 seja o resultado da jogada é de uma em cada seis chances. Isso é verdade, pois, se lançarmos um dado aleatoriamente por um número infinito de vezes, as chances de que cada uma das faces seja o resultado é a mesma, e o número 5 está presente em apenas uma dessas faces. Apesar de isso se verificar para qualquer dado imaginário ideal, não é sempre verdade para um dado em particular. Isso porque os dados podem ser manipulados. Se um jogador engenhoso adiciona um dado viciado ao jogo, qual é a probabilidade de o número 5 ser o resultado? Para responder a essa questão é necessário observar jogadas sucessivas. Depois de analisar um certo número de jogadas teremos elementos suficientes para antecipar, com maior ou menor precisão, qual será o resultado das jogadas subsequentes do dado viciado.

É importante destacar que, quanto maiores os incentivos para o uso de processamento de dados por meio de algoritmos como base para tomadas de decisão, e quanto mais prontamente disponíveis e baratas as tecnologias para tornar isso possível, mais urgente se torna a discussão acerca das consequências de tais procedimentos para os indivíduos e os riscos a eles associados.

Como visto, os algoritmos necessitam de um *input* básico para oferecer respostas relevantes: dados. Não por outra razão, a quantidade crescente de informações disponíveis levou ao crescimento exponencial de sua utilização e de seu impacto em nossas vidas. O termo *Big Data* foi cunhado para traduzir esse fenômeno[4]. Como apontam Mayer-Schönberger e Cukier, *Big Data* não é somente sobre tamanho, mas especialmente sobre "a habilidade de transformar em dados muitos aspectos do mundo que nunca foram quantificados antes" (Mayer-Schonberger; Cukier, 2014, tradução livre).

A função mais importante de *Big Data* é elaborar previsões baseadas em um grande número de dados e informações: desde desastres climáticos até crises econômicas, do surto de uma epidemia até o vencedor de um campeonato de esportes, do comportamento de um consumidor até a solvência dos clientes. Assim, as análises de *Big Data* podem ser utilizadas para elaborar prognósticos, tanto com relação à economia, à natureza ou à política, quanto sobre comportamento individual. No que se refere ao assunto aqui discutido, a predição do comportamento individual é de grande interesse, na medida em que gerar informação e conhecimento sobre o comportamento de uma pessoa a partir de dados pessoais oferece base para tomada de decisões. Uma análise de *Big Data*

4 Há uma grande discussão a respeito de se este termo é apropriado e o que ele de fato significa. Boyd e Crawford apresentam uma boa síntese a respeito: "*Big Data* é relevante não devido ao seu tamanho, mas por causa de sua relação com outros dados. Devido a seus esforços para extrair e agregar dados, *Big Data* é fundamentalmente interconectado. Seu valor vem dos padrões que podem ser derivados a partir das conexões criadas entre dados, sobre um individuo, sobre indivíduos em relação a outros, sobre grupos de pessoas, ou simplesmente sobre a estrutura da informação em si" (tradução livre) (D. Boyd e K. Crawford, "Six Provocations for Big Data". Apresentado em *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011. p. 1-2. Disponível em: <https://ssrn.com/abstract=1926431>. Acesso em: 17 abr. 2018).

pode, portanto, afetar diretamente um indivíduo – e produzir resultados discriminatórios que impactem sua vida.

De acordo com Mayer-Schönberger e Cukier, não há definição precisa para *Big Data*, mas o fenômeno pode ser caracterizado por três tendências. Em primeiro lugar, a quantidade de dados e informações coletadas. As análises de *Big Data* não apenas reúnem mais dados do que nunca, mas buscam reunir todos os dados e informações referentes a uma situação em particular, não apenas uma amostra deles – como colocam os autores, em *Big Data*, n = tudo. Em segundo lugar, devido à grande quantidade de informações disponíveis, os dados podem ser imprecisos. Na medida em que a magnitude aumenta, do mesmo modo aumentam as chances de equívocos. A terceira propriedade é a de buscar correlações ao invés de causalidades. Isso significa que a relação entre dois fatos ou características é determinada de acordo com uma análise estatística.

Particularmente relevante para nossos propósitos é a substituição de causalidades por correlações[5]. Durante muito tempo, a ciência e a descoberta científica funcionaram por meio da busca por causalidades. Esse é um aspecto desenvolvido e estimulado na pesquisa científica, e geralmente é visto como o único modo por meio do qual seria possível compreender o que ocorre ao nosso redor. Com *Big Data*, contudo, a causalidade perde espaço para as correlações. Uma correlação é a probabilidade de um evento ocorrer caso outro evento também se realize. É uma relação estatística entre tais acontecimentos. Ao invés de tentar assimilar os mecanismos internos de um fenômeno, as correlações nos permitem compreender o mundo por meio de *proxies*:

> Ao permitir que identifiquemos uma *proxy* útil para determinado fenômeno, correlações nos auxiliam a captar o presente e a prever o futuro: se A geralmente ocorre juntamente com B, é preciso ficar atento a B para podermos estimar que A ocorrerá. Utilizar B como *proxy* ajuda a compreender o que provavelmente está ocorrendo com A, ainda que não seja possível mensurar ou observar A de maneira direta. (Mayer-Schonberger; Cukier, 2014, tradução livre)

Todos os temas aqui tratados adquirem maior complexidade em se considerando os desenvolvimentos recentes da ciência da computação no campo da inteligência artificial (IA). A IA volta-se ao desenvolvimento de máquinas "inteligentes", sejam elas robôs, carros ou computadores. Dentro da IA, há uma área de especial interesse para os cientistas da computação, que recebe o nome de *machine learning* (aprendizagem de máquinas ou ML em sua sigla em inglês). Como Pedro Domingos esclarece, o *machine learning* muda as regras do jogo porque:

5 Como afirmou corretamente Nate Silver, a busca por correlações está incorporada a um contexto complexo, no qual se refletem preconceitos e assunções subjetivas; isso significa que até mesmo métodos estatísticos não são completamente objetivos.

> Todo algoritmo possui um *input* e um *output*: o dado ingressa no computador, o algoritmo faz o que seu código determina com esse dado, e, então, sai o resultado. O *Machine Learning* muda essa lógica: adentram na máquina tanto o dado como o resultado desejado, e o produto é algoritmo capaz de tornar a relação entre dado e resultado verdadeira. Algoritmos inteligentes – também conhecidos como *learners* – são algoritmos que criam outros algoritmos. Com *machine learning*, computadores escrevem seus próprios programas, para que nós não tenhamos que fazê-lo. (Domingos, 2015, p. 6, tradução livre)

Como a própria ascensão do uso de *machine learning* revela, outra questão decisiva com relação aos algoritmos é a obscuridade dos algoritmos em seus processos decisórios. O tema ganha maior relevância pois soluções algorítmicas vêm sendo amplamente adotadas tanto pelo setor privado quanto pelo setor público.

Talvez o caso mais famoso nesse sentido seja o do *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS). Trata-se de uma ferramenta pensada originalmente para o gerenciamento de penitenciárias a partir de "informações sobre gestão de detentos críticos", passando desde a triagem de saúde mental dos detentos até o rastreamento de gangues. Como afirma a *Equivant*[6], desenvolvedora do COMPAS, a ferramenta funciona a partir de uma árvore decisória, que classifica os detentos em um espectro de risco que varia de um a nove, sendo nove o mais alto e um o mais baixo. Embora tenha sido projetado originalmente para o monitoramento de penitenciárias, o algoritmo tem sido usado para outros propósitos nos Estados Unidos, especialmente para avaliação do risco de reincidência[7]. O caso de Eric Loomis, no Estado de Wisconsin, é um exemplo. Em 2013, Loomis foi acusado de fugir da polícia na Cidade de La Crosse ao dirigir um carro anteriormente utilizado em um tiroteio. Ele havia sido condenado previamente por agressão sexual e, após uma avaliação do COMPAS, considerou-se que havia alto risco de cometer outro crime, tendo sido condenado, assim, a uma sentença de seis anos.

Os advogados de Loomis recorreram da sentença, alegando que a defesa não teve acesso à avaliação de risco de reincidência realizada pelo COMPAS, dada sua natureza confidencial, mesmo tendo sido tal resultado instrumental na sentença judicial. O caso chegou à Suprema Corte de Wisconsin, que, por sua vez, manteve a decisão do juiz, alegando que esta não teria se baseado somente na avaliação do COMPAS. Um *writ of certiorari* foi posteriormente levado à Suprema Corte, mas negado.

---

6 Como a *Court View Justice Solutions Inc.*, a *Constellation Justice Systems Inc.* e a *Northpointe Inc.* renomearam para formar a *equivant*, em 9 de janeiro de 2017.

7 State *v.* Loomis. 881 N.W.2d 749 (Wis. 2016). Disponível em: <https://harvardlawreview.org/2017/03/state-v-loomis/>. Acesso em: 24 set. 2019.

Outro exemplo é o chamado *CrimeRadar*[8], uma ferramenta criada no Brasil e voltada para a previsão de taxas e padrões de criminalidade na Cidade do Rio de Janeiro. Até o presente momento, não há informações de que ela tenha sido utilizada pelas autoridades públicas no combate ao crime, mas essa certamente é uma tendência em outros locais. Algoritmos semelhantes, como o chamado *PredPol*[9], estão em uso pela polícia em diversos estados estadunidenses e mudaram o modo pelo qual tais departamentos operam.

O uso de ferramentas similares no setor privado também é numeroso. Em especial, os algoritmos têm sido muito utilizados por empresas para recrutamento de pessoas. Como destacado pelo *Wall Street Journal* (Gee, 2017), empresas como a Unilever estão acabando com processos seletivos tradicionais e confiando em *softwares* para selecionar seus candidatos a ofertas de emprego. O objetivo, expressamente declarado pela empresa, é o de diversificar o grupo de candidatos. Danieli, Hillis e Luca esclarecem a razão pela qual o uso de algoritmos nesta área é tão atrativo:

> Para analisar a estreita relação existente entre algoritmos e contratação, considere o simples fato de que a contratação é, essencialmente, um problema de predição. Quando um supervisor analisa currículos de candidatos, ele está implicitamente tentando prever quais candidatos terão um bom desempenho e quais não. Organizações tentam prever quais indivíduos fecharão negócios com sucesso. Escolas tentam prever quais professores serão capazes de dar vida a uma sala de aula. Departamentos de polícia buscam prever quais policiais conseguirão manter um bairro seguro. (Danieli; Hillis; Luca, 2016, tradução livre)

Além disso, o uso de algoritmos tem o potencial de levar a resultados consistentes e a ajudar as empresas a economizarem – tendo em vista que os processos de contratação consomem muito tempo e dinheiro. De fato, os seres humanos são tendenciosos e é sabido que tais predisposições influenciam processos de recrutamento[10]. Nesse sentido, o uso de algoritmos para selecionar candidatos pode minimizar discriminações. O problema, porém, é que, quanto mais recorremos a ferramentas como essas, mais difícil se torna para os candidatos "fora dos padrões" a entrada no mercado de trabalho e mais difícil é entender o que exatamente faz com que sua candidatura seja menos atrativa do que as outras.

---

8 Um *pop-up* no *site* do CrimeRadar informa, antes que a pessoa seja autorizada a utilizar a ferramenta: "As estimativas de probabilidade presentes na função FUTURO deste aplicativo são baseadas em um algoritmo preditivo. Como tal, a precisão de tal informação está sujeita a diversas incertezas e recomenda-se fortemente que os usuários não confiem somente nesta função para propósitos decisórios" (tradução livre).

9 Para mais informações a respeito do PredPol, ver: <http://www.predpol.com>.

10 Para mencionar somente um exemplo, um estudo da *Northwestern University*, conduzido por Lincoln Quillian, demonstrou que, entre 1990 e 2015, candidatos brancos receberam 36% mais telefonemas de retorno do que candidatos negros e 24% mais retorno do que latinos.

O caso de Kyle Behm é paradigmático. Ele teve problemas para encontrar um emprego depois de ser diagnosticado com transtorno bipolar, mesmo com resultados quase perfeitos em seus exames SAT (a versão estadunidense do vestibular). Behm entrou com um processo contra sete companhias pelo uso de um teste de personalidade desenvolvido pela *Kronos*, uma empresa de gerenciamento de força de trabalho, por entender que o responsável por sua dificuldade de ser recolocado no mercado era o algoritmo (O'neil, 2018).

Como veremos adiante, a falta de transparência é uma séria preocupação no que se refere às consequências legais da discriminação algorítmica. Isso ocorre, em primeiro lugar, porque, se o algoritmo é obscuro, é difícil afirmar que algum tipo de discriminação ocorreu; em segundo lugar, pois pode ser difícil prevenir que discriminações ocorram; terceiro, porque os algoritmos, se utilizados de maneira descuidada, podem acabar por reforçar resultados discriminatórios ao invés de combatê-los.

1.2 Discriminação algorítmica

Para esclarecer o que precisamente queremos dizer com discriminação algorítmica, utilizamos três fontes principais. Primeiro, os trabalhos de Frederick Schauer, especialmente seu livro *Profiles, Probabilities and Stereotypes* (2006); segundo, a teoria econômica de discriminação estatística; e, finalmente, a obra de Gabriele Britz (2008). Nessa seção, com base nos conceitos desses autores e depois de analisar o conceito de discriminação estatística na literatura econômica, discutiremos quatro formas de discriminação algorítmica.

Ao pensar no conceito de discriminação, geralmente imaginamos um cenário no qual certa pessoa é excluída de um grupo pelo fato de possuir determinada característica. Alguém é considerado menos merecedor de um emprego porque não frequentou uma universidade de primeira linha ou não é chamado para uma festa por ser considerado antissocial. Embora essa seja uma forma comum de se compreender a discriminação, o foco deste trabalho são os resultados discriminatórios que decorrem do fato de alguém pertencer a determinado grupo e ser julgado a partir das características desse grupo; um cenário no qual as características individuais de uma pessoa são desconsideradas, e aquela pessoa é vista somente como um membro de um dado conjunto de pessoas.

De acordo com Schauer, um dos problemas do conceito de preconceito – e tomamos o seu uso dessa palavra como sinônimo para nosso uso do termo discriminação – é nossa utilização linguística do termo. Para ele, certa confusão é criada pelo fato de utilizarmos a mesma expressão para descrever circunstâncias diferentes.

Schauer explica o problema ao examinar o conceito de generalização. De acordo com ele, há dois tipos principais de generalizações, as chamadas

sound ("consistentes") e as *unsound* ("inconsistentes"). As generalizações consistentes podem ser (i) universais – o exemplo mais famoso é o utilizado por Aristóteles: "Todos os humanos são mortais", o que significa que a totalidade da raça humana um dia, de fato, morre, de modo que a generalização se mostra verdadeira em 100% dos casos; e (ii) não universais – o que significa que a generalização não se presta a descrever a totalidade de um grupo, mas sim uma característica compartilhada pela maioria dos indivíduos daquele grupo. Quando alguém diz "os brasileiros possuem ascendência europeia", é evidente que a afirmação não se aplica a todos os brasileiros e que algumas das pessoas nascidas no Brasil podem não possuir raízes europeias. Ainda assim, a generalização é consistente e útil caso possa ser confirmada na maioria dos casos[11].

Em razão de nosso uso linguístico, Schauer acrescenta uma terceira e última categoria a esse grupo de generalizações: aquela que não é universal e nem descreve uma característica compartilhada pela maior parte dos membros de um grupo, mas que "retrata com precisão os membros de uma classe como possuindo uma maior prevalência de um traço do que a grande classe da qual o grupo é normalmente considerado parte, ainda que tal atributo não apareça na maioria dos membros de ambos os grupos" (2006, p. 11). Ele utiliza o seguinte exemplo para esclarecer o que quer dizer com essa categoria de generalização: quando alguém afirma que "buldogues têm quadris ruins", isso certamente não significa que todos os buldogues têm problemas nos quadris e também não significa que a maioria dos buldogues têm problemas nos quadris, significa simplesmente que buldogues, em comparação à grande categoria de cachorros, tendem a ter problemas nos quadris mais frequentemente que outras raças. Esse uso de uma generalização se mostrará estatisticamente congruente enquanto buldogues possuírem, de fato, quadris ruins em maior proporção do que a maioria dos cachorros. Em suma, esta terceira categoria de generalização depende fortemente de uma dimensão *comparativa*.

As generalizações inconsistentes, por sua vez, falham em preencher os parâmetros supracitados. Se alguém afirma que "arianos são impulsivos", por exemplo, é fácil verificar que (i) essa não é uma característica universal – nem todas as pessoas nascidas entre 21 de março e 20 de abril são impulsivas, (ii) não há evidências de que essas pessoas sejam mais impulsivas do que aquelas nascidas em outros períodos do ano, e (iii) descrever alguém como impulsivo não é um indicativo de que aquela pessoa seja ariana ou vice-versa (Schauer, 2006). Uma conhecida generalização desse tipo no campo jurídico são os estudos de Cesare Lombroso acerca do "homem delinquente". Lombroso foi um médico italiano, fundador da escola de Criminologia Positivista, cujos estudos se dedicaram a comprovar como criminosos teriam nascido dessa for-

11 Dados mais detalhados sobre a origem dos cidadãos brasileiros. Disponível em: <http://www.pnas.org/content/112/28/8696>. Acesso em: 24 set. 2019.

ma e como certas características físicas poderiam ajudar a identificar a criminalidade. A sua pesquisa concluiu que o homem criminoso reuniria determinados traços, como braços excessivamente longos, crânio e rosto assimétricos, etc. Até hoje, não existem evidências concretas de que Lombroso estava correto, de modo que o fato de uma pessoa possuir braços longos não é um indicativo de que aquela pessoa vai cometer ou já cometeu um crime.

Voltando à questão do preconceito, Schauer afirma que utilizamos o termo generalização para nos referirmos a dois cenários diferentes. Descrevemos algo como preconceituoso quando uma afirmação se baseia em generalizações estatísticas inconsistentes, mas também quando nos referimos a generalizações estatisticamente consistentes, mas não universais. Nesse sentido, dizer que arianos são impulsivos é tão preconceituoso quanto afirmar que homens homossexuais têm HIV, ainda que não haja evidências para embasar a primeira afirmação, mas haja evidências suficientes para embasar a segunda[12]. Isso porque atribuímos à palavra "generalização" uma conotação negativa, e, portanto, não nos sentimos confortáveis em aplicá-la ao segundo cenário enquanto não for verdade que todos os homens homossexuais têm HIV. Como Schauer aponta, isso decorre de um entendimento de que "todos os seres humanos [...] merecem ser tratados como indivíduos, e não simplesmente como membros de um grupo, de modo que decisões atuariais sobre seres humanos são, na maioria das vezes, moralmente erradas" (2006, p. 19).

O problema com esse raciocínio é o de que decisões (ou discriminações) ditas atuariais – ou seja, baseadas em estatísticas – a respeito de seres humanos são extremamente comuns em qualquer sistema jurídico e, em grande medida, indispensáveis. Sempre que a lei diz que somente pessoas acima de certa idade podem votar ou ingerir bebidas alcoólicas, está-se tomando uma decisão atuarial sobre seres humanos. Certamente, alguns indivíduos de 15 anos seriam aptos a votar ou a ingerir bebidas alcoólicas de maneira responsável, mas essas possibilidades são ignoradas em prol de outros valores. O mesmo vale para a determinação do limite de velocidade em uma rodovia. Naturalmente, pessoas diferentes são capazes de dirigir de maneira segura em diferentes velocidades, mas estabelecemos um limite – estatisticamente testado – de acordo com o qual o número de acidentes cai a níveis considerados aceitáveis. Vale destacar que algumas pessoas ainda assim talvez não sejam capazes de dirigir de maneira tão prudente quanto outras dentro dos limites existentes, mas nós, como sociedade, decidimos que esse risco é suportável.

Decisões atuariais não estão limitadas ao mundo jurídico, pelo contrário. Constantemente aplicamos a mesma lógica a inúmeras situações cotidianas: escolhemos dirigir ainda que saibamos que existem riscos associados a estar

12 Nesse sentido, vide <https://www.hiv.gov/hiv-basics/overview/data-and-trends/statistics>.

em um carro em alta velocidade, decidimos aplicar provas para que as pessoas entrem nas universidades, faculdades ou escolas, ainda que saibamos que essas provas não são capazes de contabilizar todas as habilidades cognitivas e acabem por deixar de fora muitos candidatos talentosos.

Além de compreender que a discriminação (nesse sentido utilizado por Schauer) é um aspecto corriqueiro de nossos sistemas legais – e da vida em geral –, outro esclarecimento importante está na melhor compreensão do que é, propriamente, a discriminação estatística para a teoria econômica. Trata-se de uma teoria cujas origens são atribuídas a Kenneth Arrow e Edmund Phelps, que tenta elucidar como a desigualdade pode ser um problema mesmo quando não se está buscando propositalmente qualquer tipo de resultado discriminatório (Phelps, 1972; Arrow, 1973)[13]. Isso porque, de acordo com os autores, por vezes – e, como veremos, algoritmos o fazem com muita frequência –, são utilizadas características de um grupo para avaliar a totalidade dos indivíduos a ele pertencentes de forma inconsequente.

Vale a pena mencionar que a teoria supõe que tal forma de discriminação é racional e decorre do fato de que, em um mundo de recursos escassos e racionalidade limitada, os agentes ainda assim precisam tomar um sem número de decisões. Consequentemente, eles "tendem a utilizar características facilmente observáveis, como gênero, raça, educação etc., como *proxies* para características produtivas" (Goodman; Bryce W., 2016, p. 3). Em outras palavras, agentes oferecem opiniões sobre outros indivíduos baseadas em características observáveis, as quais, por sua vez, são utilizadas como substitutas de outras características não observáveis. Tome-se o mercado de trabalho como um exemplo. Empregadores talvez sejam mais propensos a contratar homens ao invés de mulheres porque acreditam que o grupo "mulher" tem chances de enfrentar uma trajetória profissional mais difícil – frequentemente, elas têm de escolher entre trabalho e família, e nem sempre escolhem o trabalho. O empregador não sabe nada sobre a situação da mulher em concreto que entrevistará, mas pode adotar essa generalização em seu processo decisório. A teoria econômica, diferentemente do direito, não discute se essa forma de categorização empregada é ou não justa, ela simplesmente analisa se o seu uso é ou não racional – e, como dito, muitas vezes existe racionalidade na generalização.

Outro aspecto importante a ser observado é o de que a discriminação estatística pode ocorrer por diferenças exógenas ou endógenas entre grupos (Moro, 2009). No primeiro caso, a variável que distingue um grupo do outro é externa, ao passo que, no segundo, é interna e pode até mesmo integrar a dife-

13 Em oposição a tal forma de discriminação, a literatura econômica identifica o que é chamado de discriminação baseada em preferências, como definido por Becker em *The Economics of Discrimination*. Neste livro, ele afirma que alguns indivíduos possuem um "gosto" por discriminação, e, desse modo, inclui na função de utilidade de tais indivíduos um "coeficiente de discriminação", representando tal preferência.

renciação[14]. Utilizando o exemplo anterior do mercado de trabalho para esclarecer essa distinção, é possível afirmar que as mulheres foram, historicamente, mais envolvidas na criação dos filhos e em tarefas domésticas do que os homens. Mas esse é um *resultado* do fato de que a elas foram dadas menos oportunidades profissionais, consideradas por muito tempo como incompatíveis com as tarefas domésticas, e não uma característica inerente do sexo feminino que torna as mulheres menos capacitadas ou menos interessadas em oportunidades profissionais. A consequência da discriminação, nesse caso, leva à confirmação da hipótese inicial, pois estamos analisando uma variável endógena.

Diferentemente, tomando-se o mercado de seguro de automóveis como exemplo, é fácil observar que o seguro para motoristas jovens do sexo masculino é mais caro do que o seguro para motoristas jovens do sexo feminino. O gênero guarda forte correlação com a taxa de acidentes de trânsito e, assim, é frequentemente utilizado para precificação. Neste segundo cenário, porém, o gênero não é uma variável endógena, mas sim exógena, pois nada no fato de que os homens pagam mais por seguros leva esse grupo a efetivamente se envolver em mais acidentes.

Economicamente, essas observações são importantes, porque a discriminação estatística baseada em aspectos endógenos pode ser ineficiente. O resultado – por exemplo, ter um número menor de mulheres inseridas no mercado de trabalho – poderia ser modificado, levando a uma satisfação geral maior (maior número de mulheres qualificadas contratadas, empregador satisfeito com o trabalho realizado, maiores níveis de produtividade, etc.) se a hipótese inicial não estivesse presente.

Feitas essas observações, vale esclarecer que o termo "discriminação algorítmica" é utilizado, neste artigo, para englobar tanto cenários que envolvem afirmações estatisticamente inconsistentes quanto cenários em que as afirmações, embora estatisticamente lógicas, de alguma forma tomam os indivíduos que dela são objeto não de forma efetivamente individualizada, mas apenas como parte de um grupo. Isso porque, a nosso ver, uma classificação, ainda que consistente sob o ponto de vista estatístico, pode em alguns casos se mostrar injusta. Com isso em mente, sistematizamos, a seguir, quatro tipos principais de discriminação algorítmica que auxiliam na compreensão do cenário:

Discriminação por erro estatístico – todo e qualquer erro que seja genuinamente estatístico, abrangendo desde dados incorretamente coletados, até problemas no código do algoritmo, de forma que ele falhe em contabilizar parte

14 Este resultado é o que geralmente se denomina *feedback loop*. Um *feedback* ocorre quando o *output* de um sistema – por exemplo, um algoritmo – é colocado de volta no sistema como um *input*. Em outras palavras, um dado efeito do sistema retorna como sua causa. O resultado de menos mulheres sendo contratadas retorna ao sistema decisório como um *input* para aquele que tomará a decisão e, assim, reforça a conclusão que ele mesmo cria.

dos dados disponíveis, contabilize-os de forma incorreta, etc. Basicamente, é o tipo de discriminação que decorre de um erro cometido pelos engenheiros ou cientistas de dados responsáveis pelo desenho do algoritmo;

Discriminação por generalização – embora o modelo funcione bem e seja estatisticamente correto, leva a uma situação na qual algumas pessoas são equivocadamente classificadas em certos grupos[15]. Por exemplo, se uma pessoa mora em uma vizinhança comumente associada à pobreza e o modelo não possui nenhuma outra informação além de seu endereço para decidir se ela é ou não uma boa candidata para um empréstimo, ele a classificará como pertencente a um grupo do qual ela talvez não seja parte, caso ela se apresente como um caso atípico. Isso poderia ocorrer na hipótese de essa pessoa ter uma renda superior ou inferior às pessoas de sua vizinhança, por exemplo. Desse modo, embora o algoritmo esteja correto e as informações também, ainda assim o resultado será uma generalização incorreta, na medida em que mesmo um resultado estatisticamente relevante apresentará um percentual de pessoas que não se encaixam perfeitamente naquela média. Isso se dá pela própria natureza de qualquer exercício probabilístico;

Discriminação pelo uso de informações sensíveis – a razão pela qual consideramos esta categoria como discriminatória, embora muitas vezes seja estatisticamente correta, é porque ela se baseia em dados ou *proxies* legalmente protegidos. É o que ocorre, por exemplo, quando um algoritmo utiliza informações sobre identificação religiosa de um indivíduo para designar seu *credit score* no Brasil – a Lei do Cadastro Positivo proíbe o uso desse tipo de informação para essa finalidade. Duas características são relevantes para se considerar um perfilamento como discriminatório nesse caso: além de utilizar dados sensíveis, a classificação deve se basear em características endógenas[16] ou, então, deve destacar grupos historicamente discriminados;

Discriminação limitadora do exercício de direitos – novamente, aqui falamos de uma categoria que pode apresentar resultados estaticamente corretos e relevantes, mas que ainda assim consideramos discriminatória. Ao contrário da categoria anterior, o problema advém não do tipo de dado utilizado, mas da relação entre a informação utilizada pelo algoritmo e a realização de um

15 Tais classificações incorretas podem ser o resultado de correlações espúrias, mas não necessariamente. Curiosamente, o problema também pode advir do fato de que o sistema algoritmo não detém informações suficientes a respeito do indivíduo, e, desse modo, classifica-o de acordo com as informações que possui, que são insuficientes para refletir a realidade.

16 Como mencionado anteriormente, os atributos utilizados em processos decisórios podem ser endógenos ou exógenos. Nosso argumento é o de que quando a característica sob consideração possuir efeitos endógenos, sempre levará à discriminação. Se, no entanto, a propriedade for exógena, ainda que seja sensível, o resultado talvez não seja discriminatório. Um exemplo seria a já mencionada maior propensão de motoristas jovens do sexo masculino se envolverem em acidentes de carro quando comparados a motoristas jovens do sexo feminino. Gênero talvez seja uma característica sensível, mas, neste caso, não é nem endógena e nem distingue um grupo historicamente discriminado.

direito. Se há uma conexão estrita entre ambos e se o direito em questão é demasiadamente afetado, provável que o uso seja discriminatório[17].

É muito relevante termos em mente qual é a utilidade de tal tipologia, já que fornecer um parâmetro de análise para situações concretas é o principal objetivo de qualquer categorização. No que se refere à primeira categoria, é fácil perceber porquê resultados discriminatórios podem ocorrer, seja pelo fato de as informações utilizadas na decisão serem incorretas, seja porque o procedimento estatístico mostra-se incorreto. Em realidade, informações incorretas, e, portanto, resultados equivocados, são um problema bastante comum em casos como o de *credit scoring*. A Comissão Federal do Comércio dos Estados Unidos (*Federal Trade Commission*) afirmou, em um parecer ao Congresso, que a margem de erro constatada em Relatórios de Crédito nos EUA está entre 10% e 21%, a depender da natureza do erro. Outros estudos nos EUA também demonstraram que equívocos em taxas de créditos de clientes podem levar a perdas econômicas elevadas se a pessoa é classificada na categoria de risco errada (Buchner, 2006).

Análises de risco também criam uma necessidade de proteção, devido à ameaça de generalizações como as descritas na segunda categoria: decisões de gerenciamento de riscos não são respaldadas somente por dados pessoais, que indicam o risco de crédito ou de quebra de contrato pelo indivíduo. Frequentemente, essas decisões também são baseadas em outras características, que não estão, inicialmente, relacionadas à avaliação dos riscos, mas que, de acordo com a experiência estatística, muitas vezes coincidem com o risco a ser analisado (Britz, 2008).

Como já visto, a literatura econômica chama esse fenômeno de discriminação estatística, porque a correlação entre as características aparentemente neutras e os dados-alvo é determinada por um método estatístico. De uma perspectiva dogmático-jurídica, essa prática pode, em alguns casos, levantar questões a respeito de justiça individual e desigualdade. A discriminação estatística abre margem para que tratamentos diferenciados ocorram com base em características pessoais, na medida em que essas características são, de acordo com hipóteses estatísticas, um aspecto relevante para a tomada de decisões. Uma vez que a característica que se está buscando é, normalmente, difícil de se mensurar (credibilidade, solvência, produtividade laboral, etc.), uma "característica *proxy*" é utilizada no lugar desta característica principal.

17 Na visão de Schauer, o problema, nesse caso, não é a discriminação *per se*, mas sim a exclusão. O mesmo pode ser constatado se imaginarmos que o exemplo anteriormente utilizado sobre o mercado de seguros de automóveis refere-se à saúde, e não a seguros de carro. Seria menos evidente que homens jovens deveriam pagar mais do que jovens mulheres pela cobertura, ainda que o grupo de jovens do sexo masculino não seja um grupo historicamente discriminado. A razão, alega Schauer, não advém da discriminação, mas sim de um sentimento de exclusão do acesso a uma ferramenta que auxilia a realização de um direito essencial: o direito à saúde.

O problema das injustiças causadas por generalizações surge quando uma pessoa demonstra ser um caso atípico: apesar de possuir a característica *proxy*, ela não apresenta as demais qualidades esperadas do grupo. Um exemplo é o uso do endereço de um cliente como parte de uma análise de crédito, ao assumir que características relacionadas aos ativos do cliente podem ser derivadas do local em que vive. Assim sendo, é possível que o simples fato de o cliente residir em uma área "pobre" resulte em uma avaliação negativa no procedimento de *scoring*, sem qualquer avaliação posterior de solvência real e avaliações do solicitante do crédito.

Outro problema que levanta reflexões em matéria de igualdade surge do tratamento diferenciado baseado em características discriminatórias e estereotipadas clássicas, como nacionalidade, gênero, idade ou identidade sexual (terceira categoria). Por um lado, essas características estão intimamente relacionadas ao cerne da personalidade de cada um, mostrando-se, portanto, invariáveis. De outro, esses atributos representam diferenças históricas de tratamento e de estereotipização de grupos. Ademais, sua utilização como base para processos decisórios pode trazer efeitos colaterais, isto é, a discriminação de certos grupos na sociedade pode ser intensificada.

Aqui, a necessidade de proteção difere da necessidade advinda da discriminação por generalização, na medida em que não se trata simplesmente de uma violação decorrente da categorização incorreta de uma pessoa "atípica"; ao contrário, nesse caso, todas as pessoas do grupo são afetadas. Isso porque a discriminação resultante do uso de informações sensíveis pode ser, e muitas vezes é, estatisticamente consistente. Contudo, por tratarem de grupos historicamente discriminados, é um dos tipos mais perversos de discriminação, ao reforçar o tratamento discriminatório e automatizá-lo, tornando mais difícil para os membros de tais agrupamentos superarem determinada situação prejudicial. Esse uso de características discriminatórias pode ser encontrado especialmente no setor de seguros, onde nacionalidade e gênero, por exemplo, podem ser utilizados como critérios negativos em uma avaliação de risco de crédito.

Com relação ao quarto tipo de discriminação, como dito, trata-se de um tipo de cenário em que a correição estatística da análise pode se verificar, mas ainda assim o resultado pode ser discriminatório. Um importante exemplo desse tipo de discriminação ocorreu na Alemanha, onde se verificou que os bureaus de crédito podem ter incentivos econômicos para considerar o exercício do direito da pessoa de acesso ao seu *score* de crédito como um aspecto relevante – e negativo – para esse *score*, ou seja, aquelas pessoas que efetivamente acessavam a informação do *score* tinham sua pontuação reduzida. Enquanto sob o ponto de vista do titular dos dados ele está apenas exercendo seu direito de acesso, do ponto de vista do bureau de crédito infere-se, a partir de exercícios estatísticos, que aqueles que com frequência acessam suas informações credi-

tícias têm um risco maior de inadimplemento. Quando essa prática foi constatada na Alemanha, ela ensejou inclusive uma mudança legislativa: a antiga Lei de Proteção de Dados Alemã (BDSG, versão da lei anterior à entrada em vigor do Regulamento Europeu de Proteção de Dados) passou a prever, em seu § 6, 3, que o exercício dos direitos dos titulares previstos na lei não poderia ser utilizado em seu detrimento.

Pois bem, feitas essas considerações e essa tentativa de tipologizar a discriminação algorítmica, passaremos agora a debater formas de mitigar o problema, com foco especial na literatura de governança que se desenvolveu ao longo dos últimos anos sobre o tema.

2 UM DEBATE ACERCA DAS SOLUÇÕES NA LITERATURA: DISCRIMINAÇÃO ALGORÍTMICA E GOVERNANÇA

Para lidar com a questão da discriminação algorítmica e sua regulamentação, é necessário apresentar a literatura de governança algorítmica. Os autores que discutem como a discriminação pode ser um fator determinante em processos decisórios algorítmicos perceberam que seriam necessárias soluções regulatórias e de política pública para enfrentar esse problema e, portanto, focam em oferecer uma visão geral a respeito de quais poderiam ser essas soluções. Esta subseção analisará os debates mais importantes que surgiram nesta área, focando (i) nas soluções mais comumente discutidas, (ii) nas discordâncias entre os autores a respeito dessas propostas, e (iii) em nossa visão a respeito do assunto.

A literatura a respeito de governança algorítmica, embora recente, é bastante vasta. Diversos autores debatem quando e como (ou se) algoritmos devem ser regulados. Um dos primeiros esforços coletivos que objetivavam colaborar para o debate foi encabeçado por grupos de acadêmicos, os quais elaboraram princípios que poderiam reger os processos decisórios algorítmicos. A *Fairness, Accountability and Transparency in Machine Learning Organization* ("FAT-ML em seu acrônimo em inglês) é uma dessas instituições. A organização relacionou em uma lista o que acredita serem os princípios-chave que deveriam ser observados pelo setor privado e pelo governo ao lidar com algoritmos: responsabilidade (em inglês, *accountability*), explicabilidade (em inglês, *explainability*), precisão, auditabilidade e justiça (em inglês, *fairness*)[18]. Nos Estados Unidos, a *Association for Computing Machinery* (Associação de Máquinas de Computação, ou ACM, em seu acrônimo em inglês) seguiu o mesmo caminho e desenvolveu seus próprios princípios, acrescentando também à lista:

18 The Fairness, Accountability and Transparency in Machine Learning Organization. Disponível em: <https://www.fatml.org/resources/principles-for-accountable-algorithms>. Acesso em: 17 abr. 2018. Os autores que subscrevem os princípios são: N. Diakopoulos, S. Friedler, M. Arenas, S. Barocas, M. Hay, B. Howe, H. V. Jagadish, K. Unsworth, A. Sahuguet, S. Venkatasubramanian, C. Wilson, Cong Yu e B. Zevenbergen.

conscientização; acesso e reparação; proveniência dos dados; validação e experimentação[19].

Responsabilidade (ou *accountability*), sob o ponto de vista da FAT-ML, está ligada à ideia de que, ao projetar sistemas algorítmicos, é preciso ter em mente que pessoas serão afetadas pelo processo decisório e que, dessa forma, é necessário, em certa medida, oferecer alternativas para eventual reparação de danos – tanto a nível individual quanto a nível coletivo. Essa ideia está relacionada aos princípios da ACM de *conscientização* – o qual se concentra, principalmente, em tornar os engenheiros e usuários de algoritmos conscientes das possíveis consequências de seu uso, especialmente dos enviesamentos que podem surgir a partir deles – e de *fiscalização e reparação* –, de acordo com os quais os reguladores deveriam adotar mecanismos que permitam que os indivíduos afetados pelas decisões tomadas pelos algoritmos questionem e reparem os possíveis danos causados.

Como abordado por Doshi-Velez e colaboradores, a ideia de explicabilidade, quando aplicada ao processo decisório, geralmente se refere às "razões ou justificativas para aquele resultado em particular, e não a uma descrição do processo decisório em geral" (2017, p. 2). Assim, o que os autores definem como explicabilidade, e é esta definição que adotamos aqui, é uma "descrição, compreensível por humanos, do processo por meio do qual aquele que toma a decisão, ao utilizar um certo grupo de inputs, atinge uma dada conclusão"[20] (2017, p. 2-3). É importante notar que explicabilidade não é o mesmo que transparência, na medida em que ser capaz de entender o processo por meio do qual uma decisão foi tomada não é o mesmo que conhecer todos os passos tomados para se atingir aquela decisão.

O princípio da *precisão*, de acordo com Diakopoulos e Friedler, significa que "as fontes de erros e incerteza de um algoritmo e suas fontes de dados precisam ser identificadas, registradas e comparadas" (2016). De maneira objetiva, os autores afirmam que é somente compreendendo de onde vêm os erros que é possível mitigá-los. A ACM expressa uma ideia similar por meio do princípio da *proveniência dos dados*, de acordo com o qual "os designers dos algoritmos deveriam manter uma descrição do modo pelo qual os dados utilizados foram coletados, acompanhada de uma explanação das potenciais tendências induzidas pelo processo de coleta".

19 Association for Computing Machinery US Public Policy Council (USACM), "Statement on Algorithmic Transparency and Accountability", 12 January 2017. Disponível em: <http://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf>. Acesso em: 17 abr. 2018.

20 Eles vão além, para dizer que uma explicação deveria ser apta a responder ao menos uma das três perguntas seguintes: (i) quais são os principais fatores em uma decisão; (ii) a alteração de um fator teria mudado a decisão; e (iii) por que dois casos similares recebem decisões diferentes ou vice-versa.

O princípio da *auditabilidade* também está sempre presente nas discussões a respeito de governança algorítmica. Ele pressupõe a ideia de que um terceiro deve ser capaz de avaliar o método utilizado pelo algoritmo para chegar àquela conclusão[21]. Como veremos a seguir, há muita discussão a respeito de como tal divulgação seria possível e se ela deveria de fato ocorrer em algumas situações.

Justiça (fairness) talvez seja o mais óbvio, mas o menos claro de todos os princípios propostos. A ideia por trás desse conceito é a de que os algoritmos não podem levar a resultados discriminatórios. No entanto, de acordo com o que foi visto antes, analisar o que é um resultado discriminatório pode, por vezes, mostrar-se desafiador, e há discordâncias razoáveis com relação a este assunto. A ACM, embora não tratando especificamente desse princípio, estabelece o requisito de *validação e experimentação*, de acordo com o qual "[instituições] deveriam realizar testes rotineiramente para avaliar e determinar se o modelo gera danos discriminatórios".

Embora não esteja presente de maneira explícita em nenhum destes manifestos, um princípio amplamente discutido na literatura e nos círculos daqueles responsáveis pela elaboração de políticas públicas é o da *transparência*. Os algoritmos têm reconhecidamente sido chamados por Frank Pasquale de "caixas pretas" (*black boxes*) devido à obscuridade dos processos decisórios, que acaba por transformá-los em uma fonte de desconfiança. Como tais, ferramentas voltadas para desvelar algoritmos foram consideradas por alguns como um passo essencial em qualquer solução regulatória proposta (2013).

Não há consenso com relação a qual precisamente seria a combinação ideal dessas propostas, e os autores muitas vezes discordam a respeito de sua importância e utilidade. A discordância mais perceptível vem daqueles que acreditam que a transparência não seria a ferramenta mais adequada ou suficiente para lidar com algoritmos, defendendo que a auditabilidade seria uma opção melhor em boa parte dos casos.

Pasquale e Citron estão entre os autores que acreditam que a transparência é uma solução substancial para a questão da discriminação algorítmica, especialmente quando aplicada ao *credit scoring*. Em suas palavras:

> Nós acreditamos que cada titular dos dados deveria ter acesso a todos os dados a ele pertencentes. Idealmente, a lógica dos sistemas de *scoring* preditivos também

21 "Embora a complexidade dos algoritmos faça com que eles pareçam impossíveis de se compreender, estudos de auditoria podem quebrar o código por meio de tentativa e erro: pesquisadores podem aplicar conhecimentos especializados aos resultados destes testes de verificação. Ao monitorar de perto estas plataformas *online*, podemos descobrir interações entre algoritmos e dados. Em suma, a verificação destes algoritmos demanda um terceiro que possa combinar uma avaliação profissional com uma cotidiana, testando algoritmos em nome do interesse público e investigando e reportando situações em que o algoritmo talvez tenha dado errado." (tradução livre) (Sandvig et al., 2014)

> deveria ser aberta à inspeção pública. Há poucos indicativos de que a inabilidade de manter tais sistemas em segredo diminuiria a inovação. (2014, p. 26)[22]

Eles são claros ao afirmar que as "ameaças à dignidade humana" seriam suficientes para justificar a abertura não apenas da base de dados e do funcionamento do sistema em geral para as autoridades, mas também do código e da modelação dos algoritmos ao público em geral (Pasquale; Ciron, 2014, p. 30-31).

Outro autor que segue um raciocínio semelhante é Zarsky, que apresenta a discussão no contexto das previsões automatizadas em iniciativas governamentais. Em suma, ele afirma que "a justificativa mais básica e difundida para a transparência é que ela facilita a fiscalização das ações governamentais" (2013, p. 1533). Ele alega que, neste contexto, *accountability* e transparência são normalmente colocadas como sinônimos, mas argumenta que os termos são fundamentalmente diferentes, na medida em que a *accountability* se refere à ideia de que os indivíduos são eticamente responsáveis por seus atos, ao passo que a transparência é uma ferramenta – e não a única ferramenta – voltada para facilitar a responsabilização.

Alguns especialistas, contudo, apontaram limitações ao uso de soluções de transparência. Lawrence Lessig manifestou esse ponto de vista a respeito da transparência governamental ao afirmar que direcionar o panóptico aos governantes, assim como construir uma onisciência cívica, teria seus problemas. Ele desenvolve seu argumento explicando as ideias apresentadas por Brandeis na obra *Other People's Money*, especialmente a alegação de que a total transparência das informações auxiliaria o público a julgar a qualidade e, desse modo, permitiria que as próprias pessoas regulassem o mercado. No entanto, como Lessig adverte, "nem todos os dados são uma informação que os consumidores podem utilizar, apresentados de modo que eles possam utilizar" (2009). Embora o objetivo do artigo de Lessig não seja discutir a discriminação algorítmica, muitas das observações do autor mostram-se válidas para este contexto.

Referindo-se especificamente a algoritmos, Kroll e colaboradores oferecem quatro argumentos centrais tratando da razão pela qual a transparência não é uma proposta de política suficiente (2017): primeiro, porque a transparência pode simplesmente ser inatingível – haverá ou razões públicas fundamentadas que afastam o direito à divulgação, como, por exemplo, a segurança nacional, ou motivos privados voltados à prevenção de comportamentos estratégicos vol-

22 Eles também afirmam que "os técnicos especializados da FTC poderiam testar os sistemas de *scoring* para verificar a presença de vieses, arbitrariedades e caracterizações injustas. Para tanto, eles precisariam analisar não apenas os bancos de dados utilizados pelos sistemas de *scoring*, mas também o código-fonte e as anotações dos programadores descrevendo variáveis, correlações e inferências presentes nos algoritmos do sistema de *scoring*" (tradução livre).

tados a burlar o sistema[23]; segundo, porque a transparência talvez seja insuficiente – mesmo que uma decisão seja pública, "[os] métodos geralmente são insuficientes para verificar as propriedades dos sistemas de *softwares* se esses sistemas não tiverem sido projetado considerando avaliações e prestações de contas futuras" (2017, p. 633), como muitas vezes é o caso; terceiro, sempre que um algoritmo incorpora algum tipo de aleatoriedade – o que é, indiscutivelmente, uma função fundamental dos sistemas informatizados –, a presença de transparência total nada garante[24]; quarto, sistemas "inteligentes" que mudam com o tempo e se adaptam ao contexto, como os algoritmos que fazem uso de *machine learning*, não podem ser devidamente compreendidos por meio de soluções de transparência. Seguindo este raciocínio, os autores afirmam que a prestação de contas, na forma de regularização procedimental, é uma proposta de política melhor e que deveria ser estudada de maneira mais aprofundada, na medida em que poderia oferecer o resultado desejado.

Edwards e Veale também destacam as limitações da transparência e da explicabilidade, focando especialmente no Regulamento Geral de Proteção de Dados (ou GDPR em seu acrônimo em inglês). Eles apontam a inutilidade da transparência tanto por sua inviabilidade quanto pela sua inadequação às necessidades dos usuários – notadamente, sua falta de habilidade para reparar a justiça substantiva –, e sugerem uma estrutura focada na construção de soluções algorítmicas melhores desde o início (soluções que estão efetivamente relacionadas aos objetivos da política púbica) e em dar às agências e instituições o poder de supervisionar a integridade algorítmica[25].

Sunstein e colaboradores trazem valorosa contribuição ao debate ao identificar que os algoritmos têm o potencial de dar passos importantes no combate à discriminação e modificar a forma como o sistema jurídico a entende há muito tempo (2019, p. 52). Embora os algoritmos forneçam novos caminhos para que as pessoas incorporem a discriminação passada ou expressem seus preconceitos, a implantação de um sistema regulatório adequado não limita simplesmente a possibilidade de discriminação de algoritmos, mas tem também o potencial de transformar algoritmos em um poderoso contrapeso à discriminação humana e uma força positiva para o bem social.

---

23 Os autores utilizam o exemplo da evasão de divisas. Se os evasores soubessem exatamente como se identificam possíveis cenários de fraude, o algoritmo utilizado seria inútil.

24 Em suas palavras, "uma loteria oferece um excelente exemplo: um algoritmo perfeitamente transparente – utiliza um gerador de números aleatórios para atribuir um número para cada participantes e permite que os participantes com os números mais baixos vençam – produz resultados que não podem ser reproduzidos ou verificados, porque o gerador produzirá novos números aleatórios quando novamente solicitado" (tradução livre) (J. Kroll et al., 2017).

25 "Como a história das indústrias como as financeiras e de crédito demonstram, direitos à transparência não necessariamente garantem justiça substantiva ou remédios efetivos. Nós corremos o perigo de criar uma 'transparência inócua', a exemplo do que já conhecemos como 'consentimento inócuo'." (tradução livre) (Edwards; Veale, 2017, p. 22-23)

Isso porque, segundo os autores, se a regulação garantir as salvaguardas apropriadas, as perspectivas para se detectar discriminação em um mundo de design de algoritmos é muito maior do que em um contexto de decisões tomadas por humanos. Para os autores, um processo bem regulado envolvendo algoritmos se destaca por sua transparência e especificidade, pois não é obscurecido pela ambiguidade que muitas vezes ofusca a tomada de decisão humana. O acesso ao algoritmo nos permite fazer perguntas que não podemos fazer significativamente aos seres humanos (Sustein et al., 2019, p. 4).

CONSIDERAÇÕES FINAIS

A governança algorítmica é um campo extremamente disputado e complexo. Como esperamos ter demonstrado ao longo deste artigo, o assunto é ainda pouco discutido no Brasil (algo que deve mudar com o advento da Lei Geral de Proteção de Dados); no entanto, mesmo em jurisdições onde o debate a respeito da discriminação algorítmica é mais presente, os problemas estão longe de serem resolvidos. As soluções geralmente apontadas pelo meio acadêmico – transparência e fiscalização – tratam de algumas das preocupações, mas não todas. A ascensão do aprendizado de máquinas trouxe muitas dúvidas sobre o quão efetivas essas soluções podem ser em certos cenários.

Sem a pretensão de esgotar o debate, entendemos que, independentemente da solução concreta a ser adotada – seja ela a transparência, o desenvolvimento de ferramentas de *accountability* ou uma combinação dos diversos mecanismos aqui apresentados –, o caminho a ser trilhado deve sempre guiar-se pelo papel humano no processo de automação. Isso não quer dizer apenas a possibilidade de revisão de decisões automatizadas por pessoas naturais, mas também a centralidade do elemento humano em todo o processo de desenho dos mecanismos.

Esse papel precisa ter início na própria montagem das equipes responsáveis pelo design dos sistemas algorítmicos. Como ressalta Cathy O'Neil, é essencial que as pessoas que desenharão tais sistemas sejam capacitadas não só para compreender seus aspectos técnicos, mas também para visualizar os efeitos do uso daquele mecanismo no mundo real. Treiná-las para compreender aspectos éticos e morais de sua tomada de decisão, portanto, é fundamental. Diversas universidades nos Estados Unidos, muitas das quais foram responsáveis pelo desenvolvimento do Vale do Silício e das tecnologias hoje utilizadas pela inteligência artificial, estão seguindo esse caminho, oferecendo cursos sobre ética e governança da inteligência artificial e, inclusive, estruturando centros de pesquisa voltados à temática[26].

26 Para mencionar apenas algumas iniciativas, as universidades Harvard e MIT passaram a oferecer, conjuntamente, um curso sobre o tema: <https://www.media.mit.edu/courses/the-ethics-and-governance-of-

Fei-Fei Lin, ex-chefe de IA no Google e professora de Stanford, aponta no mesmo sentido e ressalta também a relevância de diversidade na constituição dos grupos de cientistas de dados. Segundo ela, se os times não forem compostos por um grupo diverso de engenheiros – de diferentes etnias, origens, religiões, etc. –, a chance de que os vieses presentes na sociedade sejam transportados para os algoritmos é muito maior (podemos nos ver frente a frente com uma rede neural treinada apenas com indivíduos brancos, e que operaria de forma insuficiente quando debruçada sobre rostos negros)[27].

Evidente que a centralidade humana também precisa estar refletida no processo de revisão de decisões automatizadas. Essa preocupação deve se apresentar tanto no momento de teste do sistema – e, portanto, antes que ele efetivamente seja utilizado para decidir concretamente a respeito de situações reais, a fim de que sejam auferidos os reais impactos dos resultados nas pessoas que serão objeto do sistema – quanto na análise de decisões tomadas e posteriormente questionadas por indivíduos que se sentiram por elas prejudicados. Para garantir que esses resultados sejam efetivamente revisados por seres humanos, é importante pensar que essa revisão deve ser feita por pessoas que realmente compreendem o processo algorítmico em análise, têm capacidade de efetivar mudanças em uma decisão concreta e idealmente estimulem uma segunda análise sobre a eventual necessidade de adaptação do sistema (caso se trate de um resultado que tem grande potencial de ocorrer novamente ou que decorre de erro do sistema).

Com isso em mente, entendemos que qualquer debate sobre discriminação algorítmica deve se centrar na seguinte ideia de que os valores que orientam a sociedade e o direito não podem ser deixados de lado quando falamos em automação e inteligência artificial. Afinal, como destacou Fei-Fei Lin, "*there are no independent machine values. Machine values are human values*". O desafio, portanto, é imaginar formas de traduzir para sistemas aquilo que vem sendo construído nas ciências humanas há milênios. A tarefa é certamente complexa e, por isso mesmo, tão relevante.

REFERÊNCIAS

AFONSO DA SILVA, V. *A constitucionalização do direito*: direitos fundamentais e relações entre particulares. São Paulo: Malheiros, 2005.

ARROW, K. The theory of discrimination. In: ASHENFELTER, O.; REES, A. (Ed.). *Discrimination in Labor Markets*. Princeton: Princeton University Press, 1973.

artificial-intelligence/>. A Universidade de Stanford, por sua vez, inaugurou um centro de pesquisa: <https://hai.stanford.edu>. Há diversos outros exemplos no país.

27 HEMPEL, J. Fei-Fei Li's Quest to Make AI Better for Humanity. Artificial intelligence has a problem: The biases of its creators are getting hard-coded into its future. Fei-Fei Li has a plan to fix that – by rebooting the field she helped invent. Disponível em: <https://www.wired.com/story/fei-fei-li-artificial-intelligence-humanity/>. Acesso em: 24 set. 2019.

BAROCAS, S.; SELBST, A. Big Data's Disparate Impact. *California Law Review*, v. 104, p. 671-732, 2016.

BENJAMIN, A. Herman et al. *Código brasileiro de Defesa do Consumidor comentado pelos autores do anteprojeto*. Rio de Janeiro: Forense Universitária, 2005.

BEWARE Spurious Correlations. *Harvard Business Review*, June 2015. Disponível em: <https://hbr.org/2015/06/beware-spurious-correlations>. Acesso em: 17 abr. 2010.

BOYD, D.; CRAWFORD, K. Six Provocations for Big Data. In: *A decade in internet time*: symposium on the dynamics of the internet and society, 2011, 17 p. Disponível em: <https://ssrn.com/abstract=1926431>. Acesso em: 17 abr. 2018.

BUCHNER, B. *Informationelle Selbstbestimmung im Privatrecht*. Tübingen: Mohr Siebeck, 2006.

BRITZ, G. *Freie Entfaltung durch Selbstdarstellung*. Tübingen: Mohr Siebeck, 2007.

BURKELL et al. Facebook: public space, or private space? *Information, Communication & Society*, v. 17, p. 974-985, 2014.

BUTLER, D. When Google got flu wrong. *Nature*, 13 February 2013. Disponível em: <https://www.nature.com/news/when-google-got-flu-wrong-1.12413>. Acesso em: 27 abr. 2018.

CORMEN, T. H. *Algorithms Unlocked*. MIT Press, 2013.

DANIELI, O.; HILLIS, A.; LUCA, M. How to Hire with Algorithms. *Harvard Business Review*, 17 October 2016. Disponível em: <https://hbr.org/2016/10/how-to-hire-with--algorithms>. Acesso em: 25 abr. 2018.

DIAKOPOULOS, N.; FRIEDLER, S. How to Hold Algorithms Accountable, 17 november 2016. Disponível em: <https://www.technologyreview.com/s/602933/how-to-hold--algorithms-accountable>. Acesso em: 17 abr.2018.

DOMINGOS, P. *The Master Algorithm*: How the Quest for the Ultimate Learning Machine Will Remake our World. New York: Basic Books, 2015.

DONEDA, D. *Da privacidade à proteção de dados pessoais*. Rio de Janeiro: Renovar, 2006.

_____; MENDES, L. Data Protection in Brazil: New Developments and Current Challenges. In: GURWIRTH, S.; LEENES, R.; DE HERT, P. (Ed.). *Reloading Data Protection*: Multidisciplinary Insights and Contemporary Challenges. Springer, 2014.

DOSHI-VELEZ, F. et al. Accountability of AI Under the Law: The Role of Explanation. *Harvard Public Law*, n. 18-07, 15 p., 2017.

EDWARDS, L.; VEALE, M. Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For. *Duke Law and Technology Review*, v. 16, n. 1, p. 18-84, 2017.

FEDERAL TRADE COMISSION. Report to Congress Under Section 319 of the Fair and Accurate Credit Transactions Act of 2003, December 2012. Disponível em: <https://www.ftc.gov/sites/default/files/documents/reports/section-319-fair-and-accurate-credit--transactions-act-2003-fifth-interim-federal-trade-commission/130211factareport.pdf>. Acesso em: 14 maio 2018.

FINLEY, S. I didn't even meet my potential employers. *BBC News*, 6 February 2018. Disponível em: <http://www.bbc.com/news/business-42905515>. Acesso em: 27 abr. 2018.

GEE, K. In Unilever's Radical Hiring Experiment, Resumes are out, Algorithms are in'. *The Wall Street Journal*, 26 June 2017. Disponível em: <https://www.wsj.com/articles/in-unilevers-radical-hiring-experiment-resumes-are-out-algorithms-are--in-1498478400>. Acesso em: 25 abr. 2018.

GINSBERG, J. et al. Detecting Influenza epidemics using search engine query data. *Nature*, v. 457, p. 1012-1014, 2009.

GOODMAN, Bryce W. Economic Models of (Algorithmic) Discrimination. *29th Conference on Neural Information Processing Systems*, Barcelona, Spain, p. 3, 2016.

HURLEY, M.; ADEBAYO, J., Credit Scoring in the Era Of Big Data. *Yale Journal of Law and Technology*, v. 18, n. 1, p. 148-216, 2016.

KROLL, J. et al. Accountable algorithms. *University of Pennsylvania Law Review*, v. 165, p. 633-705, 2017.

LAZER, D. et al. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, v. 343, p. 1203-05, 2014.

LESSIG, L. Against Transparency. *The New Republic*, 9 October 2009. Disponível em: <https://newrepublic.com/article/70097/against-transparency>. Acesso em: 27 abr. 2018.

MARQUES, C. L. *Contratos no Código de Defesa do Consumidor.* O novo regime das relações contratuais. São Paulo: Revista dos Tribunais, 2011.

MAYER-SCHÖNBERGER, V.; CUKIER, K. *Big Data*: A Revolution That Will Transform How We Live, Work, and Think. New York: First Mariner Books, 2014.

MENDES, L. *Privacidade, proteção de dados e defesa do consumidor*. Linhas gerais de um novo direito fundamental. São Paulo: Saraiva, 2014.

_____. *Schutz vor Informationsrisiken und Gewährleistung einer gehaltvollen Zustimmung*. Eine Analyse der Rechtmäßigkeit der Datenverarbeitung im Privatrecht. Berlin: De Gruyter, 2015.

MORO, A. Statistical Discrimination. In: DURLAUF, N. S.; LAWRENCE, E. (Ed.). *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, 2009.

O'NEIL, C. Personality Tests Are Failing American Workers. *Bloomberg View*, 18 January 2018. Disponível em: <https://www.bloomberg.com/view/articles/2018-01-18/personality-tests-are-failing-american-workers>. Acesso em: 27 abr. 2018.

PAPACHARISSI, Z. *A private sphere*: democracy in a digital age. Cambridge: Polity Press, 2010.

PASQUALE, F. The Emperor's New Codes: Reputation and Search Algorithms in the Finance Sector. *NYU "Governing Algorithms" Conference*, 2013.

_____; CITRON, D. The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, v. 89, p. 1-33, 2014.

PHELPS, E. The Statistical theory of Racism and Sexism. *American Economic Review*, v. 62, p. 659-61, 1972.

PINHO MELLO, J. M.; MENDES, M.; KANCZUK, F. Cadastro Positivo e democratização do crédito. *Folha de São Paulo*, mar. 2018. Disponível em: <https://www1.folha.uol.com.br/opiniao/2018/03/joao-manoel-pinho-de-mello-marcos-mendes-e-fabio-kanczuk-cadastro-positivo-e-democratizacao-do-credito.shtml>. Acesso em: 17 abr. 2018.

SANDVIG, C. et al. An Algorithm Audit. In: GANGADHARAN, S. P. (Ed.). *Data and Discrimination*: Collected Essays. Open Technology Institute, 2014.

SCHAUER, F. *Profiles, Probabilities, and Stereotypes*. Cambridge: Harvard University Press, 2006.

SILVER, N. *The Signal and the Noise*. The Art and Science of Prediction. London, 2012.

STATE *v.* Loomis. 881 N.W.2d 749 (Wis. 2016). Disponível em: <https://harvardlawreview.org/2017/03/state-v-loomis/>. Acesso em: 28 jan. 2019.

TEPEDINO, G. As relações de consumo e a nova teoria contratual. In: TEPEDINO, G. *Temas de direito civil*. Rio de Janeiro: Renovar, 1999. p. 199-216.

ZARSKY, T. Transparent Predictions. *University of Law Review*, v. 2013, n. 4, p. 1503-1570, 2013.

Artigo Convidado