Accountability de Algoritmos: a falácia do acesso ao código e caminhos para uma explicabilidade efetiva

Isabela Ferrari[1]

RESUMO

Depois de esclarecer o que são e como funcionam os algoritmos, apontamos os problemas relativos à visão de que o acesso ao código-fonte seja uma via apta a garantir a compreensão dos aspectos definidores da solução apontada. Em seguida, sugerimos caminhos mais profícuos em direção à explicabilidade

ABSTRACT

After presenting what is and how algorithms work, we clarify the problems associated to defining access to code as the main solution for algorithmic accountability. We argue that this is not necessary, neither enought to guarantee te proper knowledge of the mechanisms that drive decision. After, we suggest more promising paths towards explainability.

[1] Juíza Federal, Mestre e Doutoranda em Direito Público pela Universidade do Estado do Rio de Janeiro (UERJ), Visiting Researcher pela (Harvard Law School 2016/2017). Coordenadora Acadêmica do Instituto New Law. Membro do Board da The Future Society/Brasil. Membro do Comitê de Direito Administrativo e Ambiental da Escola da Magistratura Federal da 2a Região (EMARF). Professora de Direito Administrativo do Curso Ênfase. Email: isabelarossicortesferrari@gmail.com

INTRODUÇÃO

Três histórias ilustram as preocupações que endereçamos neste artigo.

A primeira é a história de Joy Buolamwini. Começa quando nossa protagonista estava finalizando seu curso de computação na Georgia Institute of Technology, na Universidade de Oxford. Seu trabalho de conclusão de curso era relativamente simples: desenvolver um robô social[2]. Decidiu criar um software capaz de brincar de pikaboo, uma popular brincadeira infantil que consiste basicamente em cobrir o rosto e descobri-lo em seguida, quando então se diz "pikaboo".

Para atingir seu objetivo, Joy se utilizou de um software aberto de reconhecimento facial – já que identificar o rosto descoberto era crucial para que a brincadeira virtual fosse bem-sucedida. Percebeu, com curiosidade, que apesar de o programa funcionar com diversos amigos, era incapaz de reconhecer seu rosto. Joy é negra.

Descoberto o defeito do software-base e apesar do incômodo com a situação, Joy focou em terminar o trabalho – usando uma máscara branca ou sua colega de quarto para checar o funcionamento do programa.

Anos depois, e já cursando o seu Ph.d. no Massachussets Institute of Technology (MIT), Joy participou de uma competição em Hong Kong. Em visita a uma startup local, planejada pelos organizadores, novamente Joy era a única pessoa em quem um programa, que dependia de reconhecimento facial, não funcionava. Já desconfiada da razão, Joy descobriu, estupefata, que o software base utilizado em Hong Kong era o mesmo que ela havia utilizado no seu trabalho de conclusão de curso, anos antes nos Estados Unidos.

Nossa segunda história é mais singela, mas não menos relevante. Recentemente, o Estado de Nova York passou a se utilizar de softwares para avaliar os professores que trabalhavam em algumas escolas públicas e recomendar a demissão daqueles cuja performance fosse considerada abaixo

[2] Robôs sociais são aqueles criados para interagir com seres humanos.

do esperado. Os resultados indicaram a demissão de professores muito bem avaliados por pais e alunos[3].

A terceira situação, em nossa opinião, é mais sensível de todas. Trata-se do caso Loomis. Em 2013, Eric Loomis foi preso em flagrante após furtar um veículo, evadir-se de um agente de trânsito e se envolver em um tiroteio. Levado à presença de um juiz, determinou-se, inicialmente, que respondesse ao processo em liberdade. Em seu julgamento, foi condenado a seis anos de prisão. Com o seu passado de agressão sexual, a pena aplicada a Loomis não foi surpresa.

O caso, no entanto, tornou-se mundialmente conhecido, porque tanto a negativa da liberdade provisória, quanto o patamar aumentado da pena foram definidos a partir da avaliação de que Loomis apresentaria alto risco de violência, reincidência e evasão, avaliação essa feita por um software, à qual aderiu o juiz sem adicionar qualquer análise própria. A situação é ainda mais sensível porque o referido programa, denominado COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), é um software privado, que funciona a partir de um algoritmo secreto, ao qual nem os juízes que o utilizam têm acesso[4].

Loomis, então, recorreu à Suprema Corte de Winsconsin, requerendo o acesso aos critérios que levaram o software a classificá-lo como uma pessoa de alto risco. O Procurador-Geral do Estado foi contra. Ele defendeu que, como o uso de algoritmos para a tomada de decisões é muito recente, a questão ainda não estaria madura para julgamento, sustentando que Loomis estaria livre para questionar o resultado da sua avaliação e possíveis falhas, mas que não poderia acessar o código-fonte do algoritmo. Na mesma linha, os representantes legais da Northpointe Inc., desenvolvedora do COMPAS, defenderam que a forma de operação do sistema estaria protegida por segredo industrial.

---

[3] FERRARI, Isabela; BECKER, Daniel; WOLKART, Erik Navarro. Arbitrium ex machina: panorama, riscos e a necessidade de regulação das decisões informadas por algoritmos. *Revista dos Tribunais*, v. 995, set. 2018.

[4] ISRANI, Ellora. Algorithmic due process: mistaken accountability and attribution in State v. Loomis. JOLTdigest. Disponível em: [https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in Acesso em: 25.10.2017.

Durante o julgamento, algumas questões desconfortáveis foram levantadas, como o relatório da ONG ProPublica, sobre o enviesamento do Compas contra afro-americanos[5]. Apesar disso, a Suprema Corte de Winsconsin negou o pleito de Loomis, afirmando que ele teria recebido a mesma sentença a partir de uma análise humana dos fatores usuais: seu crime e seus antecedentes. Loomis recorreu à Suprema Corte Americana, que negou o *writ of certiorari*, algo semelhante a um pedido de admissão para julgamento, por ele apresentado. Loomis permanecerá preso até 2019[6].

O que aproxima as três histórias é a dificuldade de antecipar um problema decorrente do uso de algoritmos, em boa medida em razão da opacidade de sua forma de operação. Este texto destina-se a esclarecer como funciona a categoria dos algoritmos que utiliza inteligência artificial em sua operação, demonstrando, assim, porque razão existe necessariamente uma dificuldade em controlar esse processo. Em seguida, depois de rejeitar o acesso ao código fonte como solução para o problema, indicamos os caminhos que nos parecem mais promissores para atingir a desejada explicabilidade mínima relativa ao funcionamento de processos operativos decisórios por algoritmos de *machine learning.*

O QUE SÃO E COMO FUNCIONAM OS ALGORITMOS?

O que é um algoritmo? Existem várias formas de responder a essa pergunta. Neste trabalho, usaremos a definição de Pedro Domingos, valiosa por sua simplicidade: algoritmo é uma sequência de instruções que diz a um computador o que fazer[7]. Wolkart[8] explica os algoritmos comparando-os com

---

[5] ANGWIN, Julia et al. Machine Bias. Pro Publica. Disponível em: [https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing]. Acesso em: 25.10.2017.

[6] Loomis, 881 N.W.2d at 754.

[7] DOMINGOS, Pedro. Op. cit., p. 2.

[8] WOLKART, Erik Navarro. Análise econômica e comportamental do processo civil: como promover a cooperação para enfrentar a tragédia da Justiça no processo civil brasileiro. 2018. 835 f. Tese (Doutorado em Direito) – Universidade do Estado do Rio de Janeiro, Rio de Janeiro.

uma escada, que determinada pessoa utiliza para sair de um ponto inicial até o topo. O algoritmo faz o mesmo: divide determinada tarefa (chegar até o topo) em tarefas menores (passar por cada um dos degraus).

Quanto ao seu funcionamento, podemos classificar os algoritmos em duas espécies: os programados e os não programados. Algoritmos programados seguem as operações ("o caminho") definidas pelo programador. Assim, a informação "entra" no sistema, o algoritmo atua sobre ela, e o resultado (output) "sai" do sistema. O programador domina, portanto, todas as etapas operativas do algoritmo.

Ainda em 1950, referindo à operação de algoritmos, Alan Turing, no seminal Computing Machinery and Intelligence, propunha que, no lugar de se imitar o cérebro de um adulto, programando todas as operações a serem realizadas, seria mais produtivo adotar estratégia diversa: simular o cérebro de uma criança, com capacidade randômica de aprendizado[9]. Nascia aí a ideia motriz dos algoritmos não programados, aqueles que usam a técnica que ficou conhecida como aprendizagem de maquinas, ou *machine learning.*

Essa categoria de algoritmos, denominados learners, opera criando outros algoritmos. Nesse caso, os dados e o resultado desejado são carregados no sistema (input), que produz o algoritmo (output) que transforma um no outro. Como destaca Pedro Domingos, o computador escreve a própria programação, de forma que humanos não tenham que fazê-lo[10].

A técnica de *machine learning* pode ser definida, então, como a prática de usar algoritmos para coletar e interpretar dados, fazendo predições sobre fenômenos. As máquinas desenvolvem modelos e fazem predições automáticas, independentemente de nova programação[11]. Os dados, aliás, são a matéria

---

[9] TURING, Alan. Computing Machinery and Intelligence. Mind, New Series, v. 59, n. 236, p. 433-460, out. 1950.

[10] DOMINGOS, Pedro. Op. cit., p. 6.

[11] ASSUNÇÃO, Luís. Machine learning, big data e inteligência artificial: qual o benefício para empresas e aplicações no Direito? LEX MACHINÆ. Disponível em: [

prima da aprendizagem. Por isso, um grande volume de dados é essencial para o sucesso da técnica.

Isso explica porque o advento do *big data* (o imenso volume de dados estruturados e não estruturados) na última década teve um impacto tão significativo para a aprendizagem de máquinas, que já existia desde a década de 70[12]. A rápida evolução computacional, embalada pelas exponenciais Leis de Moore[13] e de Kryder[14], trouxe uma abundância de dados jamais vista na humanidade e, portanto, matéria-prima sem limites para técnicas computacionais de inteligência artificial.

A forma mais simples dos algoritmos não programados, ou seja, daqueles que empregam *machine learning,* é aquela que emprega algoritmos supervisionados, na qual o sistema é alimentado com dados lapidados e previamente escolhidos por seres humanos. Nesse caso, o conjunto de dados rotulados e a saída desejada são carregados no sistema. Enquanto é treinado, o modelo ajusta as suas variáveis para mapear as entradas para a saída correspondente.

Um exemplo são os algoritmos utilizados pelos bancos para aprovar a concessão de empréstimos. Nesse caso, os dados analisados serão referentes ao histórico de crédito do cliente, e as informações utilizadas para treinar o sistema são dados já rotulados como positivos ou negativos para a concessão de crédito.

Uma espécie de estruturação algorítmica que funciona de forma supervisionada são as redes neurais artificiais (com *back propagation*).

---

https://www.lexmachinae.com/2017/12/08/machine-learning-big-data-e-inteligencia-artificial-qual-o-be Acesso em: 25.06.2018.

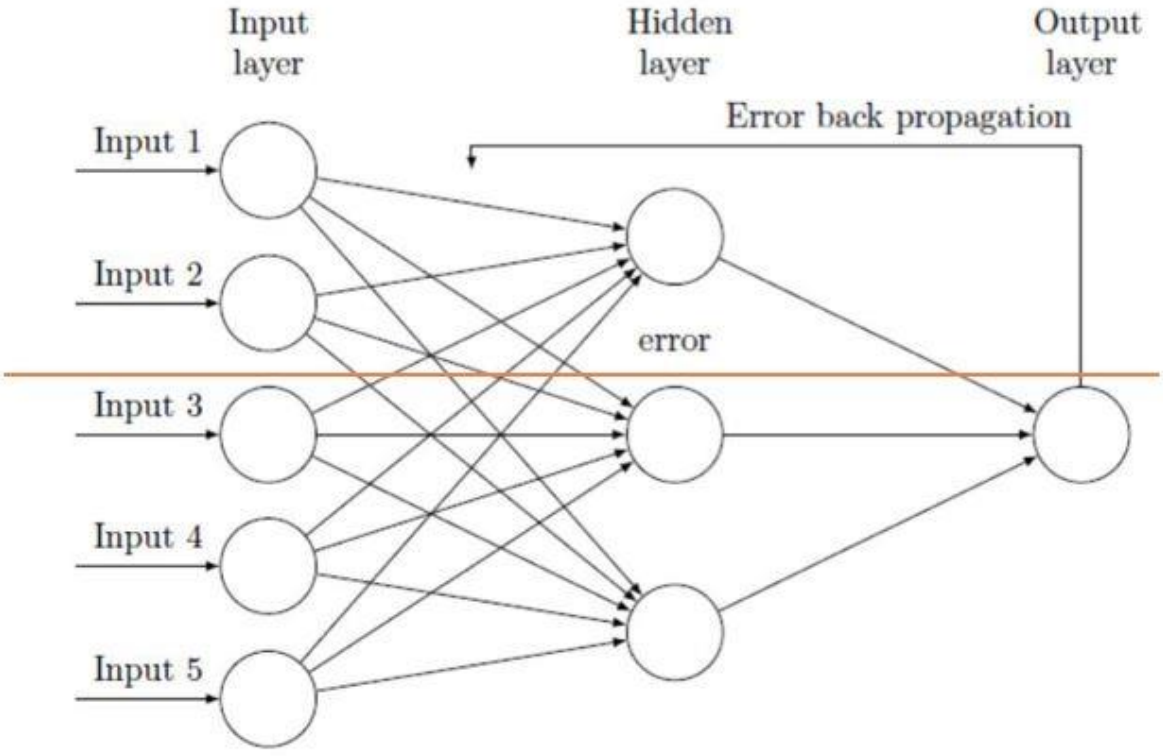
[12] BRYNJOLFSSON, Erik; MCAFEE, Andrew. A segunda era das máquinas: trabalho, progresso e prosperidade em uma época de tecnologias brilhantes. Rio de Janeiro: Alta Books, 2015. p. 84-85.

[13] BECKER, Daniel; FERRARI, Isabela. A prática jurídica em tempos exponenciais. JOTA. Disponível em: [https://jota.info/artigos/a-pratica-juridica-em-tempos-exponenciais-04102017]. Acesso em: 07.06.2018.

[14] WALTER, Chip. Kryder's Law. Scientific American. Disponível em: [https://www.scientificamerican.com/article/kryders-law/]. Acesso em: 07.06.2018.

Inspiradas no cérebro humano, têm modelo de aprendizagem baseada em erros e acertos, com identificação paulatina dos caminhos e decisões mais corretas para atingir determinados objetivos.

Nesses casos, o sistema é carregado com um objetivo (output), e vários inputs. Os inputs são testados em vários caminhos. Quando se chega ao resultado desejado, o caminho mais assertivo recebe um peso maior na conta matemática. Assim, as camadas neurais internas (hidden layers) mais assertivas passam a dominar a tarefa, e a entregar resultados mais precisos na medida em que o algoritmo confere um peso maior às conexões que apresentem resultados mais próximos dos desejados[15].



Fonte: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC505667/pdf/brjopthal00011-0006

[15] RUMERLHART, David E.; HILTON, Geoffrey E.; WILLINANS, Ronald J. Learning Representations by back-propagating erros. Nature, v. 323, issue 9, p. 533, out. 1986.

Uma segunda categoria relevante é a dos algoritmos não supervisionados (*non-supervised learning algorithms*). Nesse caso, os dados que alimentam o sistema não são rotulados, deixando o algoritmo de aprendizagem encontrar estrutura nas entradas fornecidas por conta própria. Dessa forma, esses algoritmos têm a capacidade de organizar amostras sem que exista uma classe pré-definida.

O aprendizado não supervisionado é útil quando for necessário descobrir padrões em determinado conjunto de dados não rotulados, e pode ser um objetivo em si mesmo ou, ainda, um meio para atingir determinada finalidade. Essa técnica é empregada no reconhecimento e identificação de faces e de vozes, além da criação de sistemas de tomada de decisão em curto espaço de tempo, viabilizando, por exemplo, a construção de carros e drones autônomos[16].

Exemplo de estruturação algorítmica que funciona de forma não supervisionada para atingir determinada finalidade é a rede neural convolucional, utilizada com sucesso no reconhecimento de imagens e processamento de vídeo. Na área da saúde, a técnica é utilizada para o diagnóstico de determinadas doenças[17].

Finalmente, uma terceira categoria corresponde aos algoritmos de reforço (*reinforced learning algorithms*), que são treinados para tomar decisões. Nesses casos, existe um feedback sobre o sucesso ou erro do *output*, que será utilizado para aprimorar o algoritmo.

Diferentemente dos algoritmos supervisionados e não supervisionados, os de reforço não estão direcionados a gerar *outputs* "corretos", mas enfocam a questão da performance, comportando-se de forma muito semelhante aos seres

---

[16] WOLKART, Erik Navarro. Análise econômica e comportamental do processo civil: como promover a cooperação para enfrentar a tragédia da Justiça no processo civil brasileiro. 2018. 835 f. Tese (Doutorado em Direito) – Universidade do Estado do Rio de Janeiro, Rio de Janeiro, p. 683.

[17] GARDNER. G. G. et al. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. British Journal of Ophthalmology, 80, p. 940-944, 1996. Disponível em: [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC505667/pdf/brjopthal00011-0006.pdf]. Acesso em: 15.06.2018.

humanos, que aprendem com base em consequências positivas ou negativas, como uma criança que coloca o dedo na tomada e logo percebe que essa não é uma ação inteligente. Esse tipo de algoritmo é corriqueiramente utilizado em jogos, e a pontuação maior ou menor que eles atingem no processo funciona como recompensa[18-19].

Ao mesmo tempo em que se percebe que os modelos mais modernos de inteligência artificial foram inspirados na biologia e psicologia do cérebro humano[20], é evidente a perda de controle sobre os processos de aprendizagem de algoritmos.

A autonomia dos algoritmos de *machine learning* faz com que as tarefas por eles desempenhadas sejam difíceis de antever e, mesmo após a decisão, difíceis de explicar. Mesmo os *learners* mais simples, supervisionados, não permitem que se compreenda propriamente o seu funcionamento – a menos que tenham sido estruturados para tanto.

---

[18] Em 2015, por exemplo, experimentos na universidade de Toronto, no Canadá, levaram um único algoritmo, com a mesma rede neural e os mesmos hiperparâmetros, a ter um desempenho de alto nível, 49 jogos diferentes de Atari[28]. Os algoritmos não haviam sido especificamente desenvolvidos para nenhum desses jogos: os únicos inputs recebidos pelo sistema foram os pixels dos jogos captados por sensores, além das recompensas determinadas pela pontuação de cada jogo.

[19] Em algumas situações, algoritmos supervisionados e algoritmos de reforço são utilizados de forma combinada visando melhores resultados. Uma rede neural associando as duas técnicas foi o que permitiu o sucesso do AlphaGo. O jogo chinês de Go se assemelha ao de xadrez mas, enquanto este possui 64 casas, o tabuleiro de Go possui 361. Isso tornou impossível que se replicasse a estratégia utilizada por ocasião da programação do Deep Blue, que fora carregado manualmente com milhares de jogadas e combinações possíveis, usando sua enorme capacidade de processamento para escolher a melhor jogada. Como seria inviável programar previamente todas as 2,1x10170 posições possíveis do jogo, decidiu-se utilizar no AlphaGo uma combinação de reinforced e supervised learning. O início de seu processo de aprendizado correspondia ao estudo supervisionado: humanos escolhiam a informação a ser observada pela máquina – jogadas e posições protagonizadas por grandes jogadores de Go –, controlando esse processo. Depois de o sistema aprender a classificar e valorar essas posições, ele passava para uma fase mais avançada de aprendizado, não supervisionada (reinforced learning), na qual o algoritmo participava sozinho de múltiplos jogos simulados aleatórios e aprendia a fazer as melhores escolhas e a valorá-las de modo preciso (value network). Com isso, o AlphaGO avaliava muito menos posições por jogada do que o Deep Blue, mas o fazia de forma precisa e inteligente, selecionando e valorando suas escolhas de modo muito mais eficiente, graças à sua policy network (responsável pelos critérios de seleção) e sua value network (responsável pelos critérios de valoração das posições escolhidas).

[20] ITO, Joi; HOWE, Jeff. Whiplash: How to survive our faster future. New York; Boston: Grand Central, 2016. p. 240-241.

Quando se passa aos algoritmos não supervisionados ou de reforço, sequer há controle sobre os *inputs* utilizados na aprendizagem de máquinas. Ademais, à medida em que os algoritmos se tornam mais complexos e passam a interagir uns com os outros, a tendência é a de que esse desafio se agrave[21].

OPACIDADE, ACESSO AO CÓDIGO E A FALÁCIA DA TRANSPARÊNCIA

A dificuldade humana de compreender os mecanismos de funcionamento dos algoritmos que empregam *machine learning*, sejam eles supervisionados, não supervisionados ou de reforço, explica os problemas que foram apresentados no inicio deste artigo.

Caso os fatores que influíram na decisão fossem claramente perceptíveis, teria sido possível identificar rapidamente: no caso de Joy, a incompletude dos dados que foram utilizados pelo *learner*, que gerou a incapacidade de reconhecimento de um espectro mínimo de faces; no caso das escolas americanas, os critérios equivocados de classificação que levaram ao resultado inadequado; e, no caso Loomis, o uso inadmissível do critério étnico como fator que contribuiu de forma relevante para a análise de risco.

Por essa razão, já destacamos[22] que a maior preocupação relativa ao emprego dos *learners* em processos decisórios não se refere a problemas associados aos *data sets* utilizados para treiná-los, nem a eventual efeito discriminatório que possam gerar, por piores que possam ser essas situações e seus efeitos.

O que mais chama a nossa atenção é a opacidade inerente à sua operação, decorrente da já referida lacuna entre a atividade do programador e o

---

[21] TUTT, Andrew. An FDA for Algorithms. Administrative Law Review, 83 (2017). Disponível em: [https://ssrn.com/abstract=2747994]. Acesso em: 07.06.2018.

[22] FERRARI; BECKER; WOLKART. Arbitrium ex Machina: panorama, riscos e a necessidade de regulação ds decisões informadas por algoritmos. Revista dos Tribunais, vol. 995, Set / 2018

comportamento dessa espécie de algoritmo, que cria a própria programação. Vimos que o algoritmo modifica de forma autônoma sua estrutura enquanto opera, de acordo com os dados, lapidados ou não, que recebe.

Assim, pela complexidade de sua operação, a mera observação do *output* por um ser humano – ainda que seu próprio programador – dificilmente poderia conduzir a alguma conclusão sobre os processos internos que conduziram os *inputs* até lá, tornando o algoritmo uma verdadeira caixa-preta[23].

E a essa dificuldade de entender o seu funcionamento usualmente está associada, por razões culturais, à percepção de que os resultados apontados por eles são "científicos". A opacidade dos algoritmos, o pouco questionamento dos resultados por ele produzidos e a sua capacidade de aplicação em escala global (como ilustra a historia de Joy Buolamwini), levaram Cathy O'Neil a referir-se a eles como "*weapons of* math *destruction*", em tradução livre, "armas de destruição matemática"[24].

Por vezes, a resposta à preocupação sobre o *accountability* de algoritmos se encaminha no sentido de uma defesa do acesso ao código fonte. Surge, então, uma falsa questão: o pretenso conflito entre o atendimento a um dever de transparência em relação ao algoritmo, que implicaria a abertura de seu código-fonte, e a noção de sigilo industrial. Embora a doutrina perca tempo e energia nessa discussão, denominamos o argumento de "falácia da transparência".

Nesse sentido, como bem ressaltam Mittelstadt et al.[25], a transparência deve ser entendida sob dois aspectos fundamentais: acessibilidade e compreensibilidade. Apesar de a discussão doutrinária se voltar para a primeira,

---

[23] BURRELL, Burrel. How the machine 'thinks:' understanding opacity in machine learning algorithms. Big Data & Society, 3 (1), p. 1-12, 2016.

[24] O'NEIL, Cathy. Weapons of math destruction: how big data increases inequality and threatens democracy. Nova York: Crown, 2016.

[25] MITTELSTADT, Brent Daniel et al. The ethics of algorithms: Mapping the debate. Big Data & Society, 1-21, jul.-dez. 2016.

ou seja, para a defesa ou não de um direito a acessar o código-fonte, parece-nos que o ponto fulcral para o debate se refere ao segundo componente.

Isso porque, diante da estrutura cada vez mais complexa dos algoritmos que empregam *machine learning*, a mera abertura do código-fonte, por si só, tende a não auxiliar a compreensão da forma como operam, já que o referido código só expõe o método de aprendizado de máquinas usado, e não a regra de decisão, que emerge automaticamente a partir dos dados específicos sob análise.

Como salienta Burrell[26], a opacidade dos *learners* é consequência da alta dimensionalidade de dados, da complexidade de código e da variabilidade da lógica de tomada de decisões. Por empregarem centenas ou milhares de regras, por suas predições estarem combinadas probabilisticamente de formas complexas[27], pela velocidade no processamento das informações, e pela multiplicidade de variáveis operacionais[28], parece estar além das capacidades humanas apreender boa parte – senão todas – as estruturas decisórias que empreguem a técnica de *machine learning*. Assim, o mero acesso ao código comunica muito pouco, remanescendo a dificuldade de compreender o processo decisório[29].

Como já destacamos em trabalho anterior, "algoritmos apenas podem ser considerados compreensíveis quando o ser humano é capaz de articular a lógica de uma decisão específica, explicando, por exemplo, a influência de determinados inputs ou propriedades para a decisão"[30].

---

[26] BURRELL, Burrel. Op. cit.

[27] MARTIJN, Van Otterlo. A machine learning view on profiling. HILDEBRANDT, Mireille; DE VRIES, Katja (eds.). Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology. Abingdon: Routledge, 2013. p. 41-64.

[28] MATTHIAS, Andreas. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175-183, 2004

[29] KROLL, Joshua A. et al. Accountable Algorithms. University of Pennsylvania Law Review, v. 165, p. 633-705, 2017.

[30] FERRARI; BECKER; WOLKART. Arbitrium ex Machina: panorama, riscos e a necessidade de regulação ds decisões informadas por algoritmos. Revista dos Tribunais, vol. 995, Set / 2018

O cenário é preocupante, e o acesso ao código fonte não responde adequadamente ao problema. Existem, entretanto, outros caminhos que podem ser trilhados no sentido de uma explicabilidade possível. São essas possibilidades que passamos a explorar na próxima seção.

CAMINHOS PARA UMA EXPLICABILIDADE EFETIVA

Conferir explicabilidade aos algoritmos não é tarefa fácil. A questão não é *se* ela é necessária – existe um consenso razoável nesse sentido. A questão é *como* fazê-lo. Essa resposta não só está necessariamente está fora do direito, como ainda não foi encontrada.

Embora diversos textos discutam princípios aplicáveis à inteligência artificial e exponham princípios, *standard*s, e imponham deveres como explicabilidade, transparência, confiança, etc, não há no direito comparado um panorama jurídico capaz de provê-la explicabilidade apto a ser replicado ou importado.

Uma abordagem que busque trilhar um caminho concreto na direção da explicabilidade dos algoritmos precisa desbordar do campo exclusivamente jurídico e atentar a questões relativas aos desenhos de políticas públicas, além de observar possibilidades e ferramentas que a ciência da computação provê.

Com relação ao tema, a primeira recomendação daqueles que têm familiaridade com o tema costuma ser no sentido de que, para que a operação do algoritmo seja controlável, é necessário que essa preocupação com o *accountability* esteja presente desde o seu desenvolvimento. Nesse sentido, destacam a dificuldade de controlar um sistema que empregue *machine learning* que não tenha sido desenvolvido para ser controlado. Essa situação também denota a urgência de avançar nessa pauta.

Além disso, é preciso perceber que existem diferentes técnicas conhecidas como *machine learning*, com níveis de controlabilidade diferentes. Enquanto

algumas são fortemente opacas, outras, como a inteligência artificial semântica, podem ser estruturadas para justificar as escolhas feitas.

A partir dessa constatação e, como sempre, atentando a questões técnicas, seria possível começar a refletir sobre as técnicas que deveriam ser priorizadas na tomada de certas decisões. Por exemplo, seria legítima a escolha governamental pelo emprego de um algoritmo de *machine learning* do tipo *black box* para escolhas sensíveis quando seria viável estruturar um sistema fundado em inteligência artificial semântica?

A atenção a ferramentas técnicas também permite encontrar caminhos interessantes em direção à explicabilidade. Em um dos melhores artigos que abordam o assunto, escrito a muitas mãos, Kroll et al. destacam alguns mecanismos aptos a garantirem o que denominam *regularidade procedimental*.

Trata-se de ferramentas que, embora não garantam que o resultado derivado do emprego do algoritmo seja justo, atestam que não houve, por exemplo, uma falha no procedimento adotado, ou que a mesma política decisória foi adotada em casos diferentes. Esses instrumentos, portanto, garantem algum nível de *accountability*, ainda que aspectos do funcionamento do algoritmo sejam mantidos em sigilo.

A primeira ferramenta que apontam é chamada *verificação de software*. Diferente da análise de código, que é estática, a verificação de software é dinâmica, e examina o programa enquanto ele opera. Essa análise garante que, ao operar, o sistema sempre apresentará certas propriedades, denominadas invariantes.

A segunda são os *acordos criptográficos*, equivalentes digitais a um documento selado por uma terceira parte, ou à manutenção de um documento em local seguro. Os acordos criptográficos asseguram que o programa não foi alterado nem revelado, e são muito utilizados para programas que devem ser mantidos em sigilo por determinado tempo.

Acordos criptográficos podem ser utilizados para ocultar, por dado período, critérios utilizados pelo algoritmo em seu processo de tomada de decisão quando a divulgação imediata dos mesmos poderia possibilitar que aqueles agentes

sobre cujos interesses atua tentassem “enganá-lo”. Poderiam ter por objeto, por exemplo, os critérios empregados em sistemas de análise de declaração de imposto de renda, para lançar os sinais de alerta que levam a uma revisão da declaração ou analise mais aprofundada.

Assim, passado certo tempo, os acordos criptográficos dão certeza sobre os critérios utilizados, e a partir daí pode-se seguir análise sobre a legitimidade de sua operação pretérita. A certeza de que haverá *disclosure* futuro tem o efeito de refrear a tendência a usar critérios inadequados, discriminatórios, etc.

O terceiro instrumento indicado são as chamadas *zero-knowledge proofs*, ferramentas criptográficas que permitem que de pronto se prove que a política decisória utilizada apresenta certa propriedade, sem revelar como se sabe disso ou que política decisória é.

Em aulas de criptografia é comum explicarem essa ferramenta através do exemplo de dois milionários em um restaurante, que acordam que o mais rico entre eles deve pagar a conta, mas ao mesmo tempo não desejam informar ao outro quanto têm. Nesse caso, seria possível criar uma *zero-knowledge proof* para descobrir quem deve pagar a conta sem que haja *disclosure* do patrimônio de cada um.

Finalmente, as *fair random choices* são estratégias aptas a garantir que, quando o sistema possuir algum nível de aleatoriedade, esta será justa, e não poderá haver intromissão indevida de agentes internos na aleatoriedade do sistema. É mecanismo cuja aplicação Kroll et al. defendem no sistema de loteria de vistos americanos, que, segundo alguns programadores, não é exatamente segura, podendo ser fraudada (internamente), ainda que hipoteticamente.

Essas ferramentas específicas, além de mostrarem caminhos para melhorar o controle de certos sistemas, têm o valor de denotar a necessidade de um olhar mais cuidadoso dos criadores de politicas públicas para a área.

CONCLUSÕES

Após apresentarmos o conceito de algoritmo não programado e a expormos forma como operam, destacamos a opacidade que caracteriza o seu processo decisório.

Em seguida, demonstramos os motivos pelos quais o mero acesso a seu código-fonte não responde como o programa parte dos inputs para chegar ao resultado apontado, ou seja, não permite apreender o seu processo decisório.

Ressaltamos, na linha do defendido por Mittelstadt et al., que a noção de transparência não se esgota na ideia de acessibilidade (ao código), mas desborda para a noção de compreensibilidade, que faz referência ao efetivo entendimento de aspectos fundamentais de sua forma de operação.

Finalmente, destacamos algumas estratégias promissoras que podem ser adotadas para prover uma explicabilidade mínima para os *learners*.

REFERÊNCIAS

ANGWIN, Julia et al. Machine Bias. Pro Publica. Disponível em: [https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing]. Acesso em: 15.06.2018.

BAROCAS, Solon; SELBST, Andrew D. Big Data's Disparate Impact, n. 104, California Law Review 671, 2016.

BUOLAMWINI, Joy. How I'm fighting bias in algorithms. TED Talks. Disponível em: [https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms/transcript#t-74828 Acesso em: 20.06.2018.

BURRELL, Burrel. How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data & Society 2016, 3 (1), p 1–12.

DATTA, Anupam et al. Algorithmic transparency via quantitative input influence. Proceedings of 37th IEEE Symposium on Security and Privacy, San Jose, USA.

Disponível em: [http://www.ieee-security.org/TC/SP2016/papers/0824a598.pdf]. Acesso em: 20.06.2018.

DOMINGOS, Pedro. The master algorithm: how the quest for the ultimate machine learning will remake our world. Nova York: Basic Books, 2015.

EPIC.ORG. Algorithms in the Criminal Justice System. Eletronic Privacy Information Center. Disponível em: [https://epic.org/algorithmic-transparency/crim-justice/]. Acesso em: 23.06.2018.

ESTADOS UNIDOS DA AMÉRICA. Big data: seizing opportunities, preserving values, maio de 2014. Gabinete Executivo do Presidente dos Estados Unidos da América. Disponível em: [http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf]. Acesso em: 15.06.2018.

GARDNER. G. G. et al. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. British Journal of Ophthalmology, 80, 1996, p. 940-944. Disponível em: [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC505667/pdf/brjopthal00011-0006.pdf]. Acesso em: 15.06.2018.

ISRANI, Ellora. Algorithmic due process: mistaken accountability and attribution in State v. Loomis. JOLTdigest. Disponível em: [https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in Acesso em: 25.10.2017.

ITO, Joi; HOWE, Jeff. Whiplash: How to survive our faster future. New York; Boston: Grand Central, 2016.

KROLL, Joshua A. et al. Accountable Algorithms. University of Pennsylvania Law Review v. 165, p. 633-705, 2017.

LIPTAK, Adam. Sent to prison by a software program's secret algorithms. NY Times.
Disponível em: [https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algor Acesso em: 15.06.2018.

MACNISH, Kevin. Unblinking eyes: The ethics of automating surveillance. Ethics and Information Technology 14(2), p. 151-167, 2012.

MITTELSTADT, Brent Daniel et al. The ethics of algorithms: Mapping the debate. Big Data & Society, 1-21, jul.- dez. 2016.

MNIH, Volodymyr et al. Human-level control through deep reinforcement learning. Nature, v. 518, p. 529-533, 26.02.2015.

O'NEIL, Cathy. The era of blind faith in big data must end. TED Ideas. Disponível em: [https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end/transcript]. Acesso em: 22.06.2018.

O'NEIL, Cathy. Weapons of math destruction: how big data increases inequality and threatens democracy. Crown: Nova York, 2016.

RUMERLHART, David E.; HILTON, Geoffrey E.; WILLINANS, Ronald J. Learning Representations by back-propagating erros. Nature, v. 323, issue 9, out. 1986.

TURING, Alan. Computing Machinery and Intelligence. Mind, New Series, v. 59, n. 236 p. 433-460, out. 1950.

WALTER, Chip. Kryder's Law. Scientific American. Disponível em: [https://www.scientificamerican.com/article/kryders-law/]. Acesso em: 07.06.2018.

WOLKART, Erik Navarro. Análise econômica e comportamental do processo civil: como promover a cooperação para enfrentar a tragédia da Justiça no processo civil brasileiro. 2018. 835 f. Tese (Doutorado em Direito) – Universidade do Estado do Rio de Janeiro, Rio de Janeiro.