Daniel Henrique Arruda Boeing
Alexandre Morais da Rosa

ENSINANDO UM ROBÔ A JULGAR

PRAGMÁTICA, DISCRICIONARIEDADE, HEURÍSTICAS E VIESES NO USO DE APRENDIZADO DE MÁQUINA NO JUDICIÁRIO

EMais
Editora & Livraria Jurídica
Academia

Editora-Chefe | Fernanda Pacheco Amorim
Capa e Diagramação | Carla Botto de Barros
Produção editorial | Deborah Cristina Amorim

**CIP-BRASIL. CATALOGAÇÃO NA PUBLICAÇÃO
SINDICATO NACIONAL DOS EDITORES DE LIVROS, RJ**

B655e

Boeing, Daniel Henrique Arruda
Ensinando um robô a julgar : pragmática, discricionariedade, heurísticas e vieses no uso de aprendizado de máquina no judiciário / Daniel Henrique Arruda Boeing, Alexandre Morais da Rosa. - 1. ed. - Florianópolis [SC] : Emais Academia, 2020.
118 p. ; 21 cm.

Inclui bibliografia e índice
ISBN 978-65-86439-00-7

1. Direito - Filosofia. 2. Inteligência artificial. 3. Aprendizado do computador. I. Rosa, Alexandre Morais da. II. Título.

20-64686

CDU: 340.12

Leandra Felix da Cruz Candido - Bibliotecária - CRB-7/6135

EMais
Editora & Livraria Jurídica

Todos os direitos desta edição reservados à EMais.
www.emaiseditora.com.br
euquero@emaiseditora.com.br
Florianópolis/SC

Impresso no Brasil / Printed in Brazil

DANIEL HENRIQUE ARRUDA BOEING
ALEXANDRE MORAIS DA ROSA

ENSINANDO UM ROBÔ A JULGAR

PRAGMÁTICA, DISCRICIONARIEDADE, HEURÍSTICAS E VIESES NO USO DE APRENDIZADO DE MÁQUINA NO JUDICIÁRIO

EMais
Editora & Livraria Jurídica
Academia

2020

Capítulo 3

ARTIFICIALMENTE PARCIAL: CONTEXTO E VIESES EM ALGORITMOS

Visto sob o ponto de vista da filosofia da linguagem ordinária, o processamento de linguagem natural pode ser alvo de novas indagações. Como pode, afinal, um algoritmo que opera através de uma linguagem computacional, estritamente lógica, ser capaz de "compreender", ainda que minimamente, a linguagem humana? Do ponto de vista da semiótica, técnicas de NLP estariam tentando levar a cabo a assimilação de uma linguagem que opera em três níveis: sintático, semântico e pragmático, ao passo que a linguagem por meio da qual seus algoritmos foram escritos está limitada aos dois primeiros. Talvez se trate de uma forma de "entendimento" e não de "compreensão"[199], em que a máquina entende, mas não compreende, na linha de Searle[200] e Penrose[201].

Certamente a resposta para tal questionamento é complexa e levará em conta não apenas os questionamentos da filosofia da linguagem, mas também da linguística, ciências da computação, psicologia cognitiva, dentre outros campos do conhecimento. Ainda assim, para os fins deste livro, a perspectiva filosófica é suficiente para apontar a relevância do estudo da linguagem humana para que melhor se possa compreender os desafios que envolvem a elaboração de algoritmos de NLP.

199. MENDES, Alexandre José; MORAIS DA ROSA, Alexandre; ROSA, Otacílio Izaías da. Testando a *Methodology Multicriteria Decision Aid – Constructivist* (MCDA-C) na construção de algoritmos de apoio à estabilidade das decisões judiciais. Revista Brasileira de Direito, v. 15, n. 02, 2019, p. 281-305 (ISSN 2238-0604)

200. SEARLE, J.R. O mistério da consciência. Trad. André Yuji Pinheiro Uema e Vladimir Safatle. São Paulo: Paz e Terra, 1998, p. 59.

201. PENROSE, R. Shadows of the mind: a Search for the Missing Science of Consciousness. Oxford: University Press, 1994.

Dadas as reflexões até aqui expostas, é possível afirmar que a compreensão plena da linguagem excederia as competências de uma inteligência artificial de propósito limitado (fraca), dado que o comunicar humano envolve processos bastante complexos, tais como abstrações, generalizações, bem como requer certas pré-concepções acerca dos falantes e do ambiente no qual eles estão inseridos. Tais habilidades, inerentes à inteligência humana, só seriam possíveis de serem desempenhadas por uma inteligência artificial de propósito geral, ainda inviável.

De todo modo, entre a assimilação plena da linguagem natural e sua total incompreensão, há diversas situações intermediárias, entre as quais se situa o atual estágio de desenvolvimento do processamento de linguagem natural. No que tange ao campo pragmático da linguagem, algoritmos são capazes de utilizar técnicas matemáticas para "calcular" o significado de termos a partir de sua ocorrência em determinados "contextos" textuais. A ideia subjacente a esse método é que o agregado de contextos nos quais um termo ocorre é capaz de definir características que revelam similaridades de significado entre palavras e conjuntos de palavras entre si[202].

Anteriormente, abordou-se a questão da vetorização de textos e como tal processo é utilizado para, dentre outras funções, classificar documentos. Todavia, resta explorar, ainda que sucintamente, como o aprendizado de máquina trata da questão de dar significado a diferentes termos, tarefa que permite, por exemplo, identificar relações semânticas e sintáticas, expandir conceitos e identificar estruturas argumentativas em documentos[203].

Defender-se-á que técnicas de NLP que dão maior ênfase ao contexto textual em que os termos se encontram alcançam maior êxito na compreensão/entendimento de seus significados. De todo modo, não parece possível equiparar tais processos àquilo que é comumente tido como compreensão/entendimento humano da linguagem. Ainda assim, isso não é necessário para que tais algoritmos sejam capazes de um desempenho bastante satisfatório em tarefas de processamento de linguagem natural, trazendo consigo diversas possibilidades de aplicação dentro e fora do campo jurídico.

202. ASHLEY, Kevin D., op. cit., p. 242.

203. Cf. item 1.2 deste trabalho.

3.1. ENSINANDO A MÁQUINAS O CONTEXTO DA DECISÃO

Estaria errado dizer que foi a leitura de textos sobre filosofia da linguagem por parte de cientistas da computação que permitiu avanços na área do processamento de linguagem natural, mas também é fato que uma mudança de abordagem, através da qual se tem uma maior preocupação com o contexto dos termos, possibilitou a efetiva melhora no desempenho desse tipo de algoritmo. Por conta de tais mudanças, há quem afirme a ocorrência de uma "virada pragmática" no processamento de linguagem natural, de modo que algoritmos como o Word2vec seriam uma "prova empírica" de que as ideias de Wittgenstein sobre jogos de linguagem estariam corretas, refutando a concepção de que palavras são entidades isoladas, signos que simplesmente representam um certo objeto existente no mundo real[204].

O conjunto de técnicas de NLP consistentes em adaptar palavras e frases em vetores matemáticos dotados de representações numéricas constitui o *word embedding*[205]. Acontece que redes neurais podem "aprender" somente a partir de dados numéricos, o que implica a necessidade de converter dados textuais em entidades matematicamente analisáveis, o que pode ser feito de diferentes formas. A função do *word embedding* é facilitar o aprendizado de redes neurais ao representar referidos dados de forma eficiente, o que faz através de vetores[206]. Todavia, ao criar tais representações, esses modelos permitem igualmente a identificação de novos tipos de relações entre as palavras em um determinado vocabulário.

Uma abordagem "clássica" para lidar com dados textuais, o *Bag-of-Words* (BoW), consiste em transformar cada palavra de um vocabulário de N termos em um vetor de tamanho N (chamado de vetor *one-hot*), composto apenas de "zeros", exceto por um "um", que permite diferenciá-lo dos demais termos/vetores. Assim, um vocabulário de três termos ("rei", "rainha" e

204. BELLONI, Massimo. Neural Networks and Philosophy of Language. Disponível em: <https://towardsdatascience.com/neural-networks-and-philosophy-of-language-31c34c0796da>. Acesso em: 16 nov. 2019.

205. A tradução literal de *word embedding* para o português seria algo como "embutimento" ou "acondicionamento de palavras". Todavia, trata-se de um termo que é majoritariamente utilizado em inglês, motivo pelo qual optou-se por utilizá-lo em sua forma original.

206. KARANI, Dhruvil. Introduction to Word Embedding and Word2Vec. 2018. Disponível em: <https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa>. Acesso em: 16 nov. 2019.

"carro") poderia ser representado pelos vetores: "rei [0,0,1]", "rainha [0,1,0]" e "carro [1,0,0]". Através do BoW, documentos inteiros também são representados por um vetor único, no qual cada um de seus elementos representa o número de ocorrências (ou não ocorrência) do enésimo termo naquele documento, de forma que tais dados podem ser utilizados por classificadores que levam em conta a ocorrência de palavras como *features*.

Acontece que, em tal abordagem, além da ineficiência em termos computacionais, uma vez que é necessário realizar operações matemáticas com entidades compostas essencialmente por zeros, cada palavra é tratada como apenas uma dimensão em um espaço vetorial, que não interfere nas demais, ou seja, todas são igualmente distantes entre si. Dessa forma, palavras como "rei" e "rainha" são tão diferentes entre si quanto são da palavra "carro". É bastante evidente, contudo, para um humano, que as duas primeiras palavras possuem bastante similaridade, distinguindo-se apenas quanto ao gênero, e que ambas guardam certa distância de "carro", ao passo que estão igualmente próximas de "palácio".

Ainda assim, não é fácil explicar tais relações para uma máquina. Na verdade, o que um humano quer dizer com palavras "parecidas" ou "diferentes" significa que certos termos possuem maior chance de ocorrer em um determinado contexto do que em outros. Dessa forma, ao se falar em "casa", "quintal" e "jardim", as palavras "cachorro" e "gato" terão maior número de ocorrências que, por exemplo, "elefante".

Nesse sentido, o Word2vec, uma outra abordagem de *word embedding*, desenvolvida em 2013, por um time de engenheiros da Google liderados por Tomas Mikolov, buscou trazer maior eficiência a esse processo. Ao contrário de modelos que tratam palavras como "unidades atômicas" e que não possuem, portanto, uma noção de "similaridade", o Word2vec estabelece um modelo que cria representações para cada um dos termos de seu vocabulário analisando as demais palavras que compõem o seu contexto. Essa mudança de abordagem permite que termos sejam representados em um número de dimensões bastante inferior ao número N de palavras que compõem o *corpus*, o que aumenta sua eficiência computacional e permite que o modelo seja treinado em conjuntos de dados milhares de vezes maiores que seus antecessores.

O Word2vec faz isso ao treinar de forma não supervisionada uma rede neural que tem por objetivo adivinhar uma determinada palavra com base em um conjunto de palavras contexto (*Continuous Bag of Words – CBOW)* ou realizando o processo inverso (*Skip-Gram)*[207]. Assim, a rede neural cria exemplos positivos e negativos a partir de frases do *corpus* (e.g. "o cão está deitado no jardim"), substituindo a palavra-alvo por palavras aleatórias daquela base de dados (e.g. "o golfinho está deitado no jardim" ou "o casa está deitado no jardim"), o que possibilita criar dados de treino e de teste, aprimorando as predições.

Por conta dessa metodologia, é possível que a rede neural posicione em regiões próximas de um espaço vetorial termos que normalmente são acompanhados por palavras-contexto similares e distantes de termos que quase nunca ocorrem juntos. Na prática, isso significa que o modelo "compreendeu/entendeu" a existência de certas relações sintáticas e semânticas entre os termos do vocabulário. Não apenas isso, torna-se possível realizar, inclusive, operações algébricas com palavras[208].

Por exemplo, podem-se distinguir diferentes tipos de relações entre palavras ao somar ou subtrair os vetores que as representam. Sabe-se que "a palavra *grande* é similar a *maior* no mesmo sentido que *pequeno* é para *menor*", de maneira que se pode explicitar tais relações através de perguntas, que podem ser formuladas em termos matemáticos, tais como "qual palavra é semelhante a *pequeno* no mesmo sentido que *o maior* é similar a *grande?*[209]. Seguindo a mesma lógica, chega-se a "Roma" através da operação "Paris – França + Itália"[210], o que significa que Paris está para França da mesma forma que Roma está para Itália.

207. MIKOLOV, Tomas et al. Efficient Estimation of Word Representations in Vector Space, pp. 3-4.

208. MIKOLOV, op. cit., p. 2.

209. Tradução livre: "the word big is similar to bigger in the same sense that small is similar to smaller" e ""What is the word that is similar to small in the same sense as biggest is similar to big?" em MIKOLOV, op. cit., p. 5.

210. A ideia é estimar o valor numérico da relação "ser a capital de" ao se extrair do vetor "Paris" o vetor "França" e adicionar tal valor ao vetor "Itália", o que leva a algo próximo de "Roma". O mesmo raciocínio se aplica a "Rei – Homem + Mulher = Rainha" ou "Grande – Maior + Pequeno = Menor".

Figura 6 – Exemplos de cinco tipos de relações semânticas e nove tipos de relações sintáticas encontradas pelo Word2Vec[211/212]

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Dentre as contribuições de modelos que são capazes de assimilar tais relações linguísticas, figuram o aprimoramento de aplicações atuais de NLP, tais como traduções por máquina, *information retrieval* e *question answering systems*, bem como viabilizar que novas técnicas sejam criadas[213].

Após o Word2vec, foram propostas melhorias no *word embedding*, tais como o GloVe (*Global Vectors for Word Representation)*[214], da Universidade de Stanford e o fastText, do Facebook. O primeiro deles aponta que métodos que criam vetores de palavras a partir de "janelas de contexto locais", sem analisar um documento como um todo (como é o caso do Word2vec), "conseguem captar analogias entre palavras, mas não utilizam de forma satisfatória as relações estatísticas do *corpus*"[215]. Para superar essa limitação, os criadores do GloVe propõem uma matriz global que relacione as co-ocorrências de cada um dos termos do *corpus*, o que, além de uma melhoria de performance, permite entender melhor as origens das regularidades semânticas e sintáticas da representação de palavras[216].

211. MIKOLOV, op. cit., p. 6.
212. MIKOLOV, op. cit., p. 6.
213. MIKOLOV, op. cit., p. 5; Cf. Seção 1.2 deste trabalho.
214. Cf. PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher. Glove: Global Vectors for Word Representation. Proceedings Of The 2014 Conference On Empirical Methods In Natural Language Processing (emnlp).
215. PENNINGTON, Jeffrey, op. cit., p. 2.
216. PENNINGTON, Jeffrey, op. cit., p. 1.

Outro tipo de limitação do Word2vec, segundo os criadores do *fastText*, é que aquela abordagem não leva em conta as estruturas internas das palavras. Assim, o *fastText* cria n-gramas, que possibilitam representar a mesma palavra de diversas formas e relacionar n-gramas de diferentes palavras entre si, bem como criar vetores de palavras que não constavam originalmente no *corpus*[217].

Por fim, ainda que o *word embedding* represente um grande avanço no processamento de linguagem natural, deve-se fazer a ressalva de que tais tecnologias ainda estão aquém do nível de compreensão humano da linguagem. Captar o contexto textual das palavras não necessariamente significa compreender a linguagem como uma forma de vida, tampouco ser capaz de "jogar seus jogos".

A capacidade de relacionar as palavras significa a superação de um primeiro nível de complexidade, mas não consegue emular o raciocínio humano explícito, motivo pelo qual a ampliação das conexões entre os níveis continua sendo o desafio.

A importância do contexto, conforme explorado no capítulo anterior, não diz respeito somente a co-ocorrências de palavras, mas também às próprias vivências coletivas dos usuários da linguagem. A representação vetorial de palavras estaria mais próxima da atitude de um observador externo, que apenas descreve regularidades que ele verifica ao analisar a transcrição de certos jogos.

Dessa forma, um dos argumentos que serão sustentados nesta obra é que os algoritmos, por não serem capazes de compreender sozinhos a linguagem natural, terão de "importar" dos seres humanos certas pré-compreensões, que estarão presentes nos dados que são através dos quais eles são treinados. Esse processo, todavia, pode resultar no ocultamento de certas opiniões, que impactarão diretamente em seus resultados.

3.2. VIESES: COMO OPINIÕES SE TORNAM "CIÊNCIA"

3.2.1. HEURÍSTICA E VIESES

> Preconceito e resistência parecem ser mais a regra do que exceção no desenvolvimento científico avançado. Além disso, em condições normais eles caracterizam a melhor investigação, a

217. BELLONI, Massimo. Neural Networks and Philosophy of Language. Disponível em: <https://towardsdatascience.com/neural-networks-and-philosophy-of-language-31c34c0796da>. Acesso em: 16 nov. 2019.

> mais criativa e também a mais rotineira. Não está também em questão qual a sua origem. Não se tratam de características anômalas de indivíduos, mas de características da comunidade com raízes profundas no processo como os cientistas são treinados para trabalhar na sua profissão[218].

Antes de se explorar a questão da pressuposição de neutralidade das máquinas, cumpre abordar, ainda que brevemente, como as decisões humanas estão sujeitas a heurísticas e vieses. Nesse sentido, a psicologia cognitiva trouxe importantes contribuições sobre as possibilidades de engano na cognição humana e, ainda que seja parca a literatura específica sobre vieses decisórios no direito[219], não há motivos para crer que o processo de decisão judicial esteja ileso às mesmas limitações de outros tipos de decisões[220].

Partindo-se de modelos teóricos que distinguem dois modos coexistentes de pensar, também conhecidos por *teorias do processo dual*, dividem-se as atividades cognitivas humanas em dois sistemas: um mais rápido e intuitivo (apelidado de Sistema 1) e outro deliberado e devagar (Sistema 2)[221]. Ambos os sistemas (conjuntos de processos mentais) atuam concorrentemente, de forma que

218. KUHN, Thomas. A função do dogma na investigação científica, p. 24.

219. No direito brasileiro, exploram esse tema Alexandre Morais da Rosa, em Guia do Processo Penal Conforme a Teoria dos Jogos; Paola Wojciechowski, em Vieses da justiça: como as heurísticas e vieses operam nas decisões penais e a atuação contraintuitiva; Dierle Nunes et al, em Desconfiando da imparcialidade dos sujeitos processuais. Conferir: NUNES, Dierle; LUD, Natanael; PEDRON, Flávio Quinaud. Desconfiando da Imparcialidade dos Sujeitos Processuais: um estudo sobre os vieses cognitivos, a mitigação de seus efeitos e o debiasing. Salvador: JusPodivm, 2018; COSTA, Eduardo José da Fonseca. Levando a imparcialidade a sério: proposta de um modelo interseccional entre direito processual, economia e psicologia. Salvador: JusPodivm, 2018. GOULART, Bianca Bez. Análise Econômica do Litígio: entre acordos e ações judiciais. Salvador: JusPodivm, 2019. A bibliografia internacional é, ademais, abundante sobre o tema; ABIKO, Paula Yurie. Vieses da Justiça e Atuação Contraintuitiva. https://canalcienciascriminais.com.br/vieses-justica/; MACKAAY, Ejan; ROUSSEAU, Stéphane. Análise Econômica do Direito. Trad. Rachel Sztajn. São Paulo: Atlas, 2015, p. 35: "Pesquisas psicológicas mostram que os seres humanos julgam situações complexas de modo imperfeito. Aqui o espírito, mais uma vez, tende a simplifica-las, mediante heurística, para levá-las a nível em que possam ser abordadas com as faculdades mentais ordinárias de que, no momento, dispomos. Tversky e Kahneman propõem, sob a denominação de *prospect theory*, representar a decisão em duas etapas. A primeira consiste em encontrar uma moldura para o problema e enquadrá-la (*framing et editing*); a segunda etapa é a avaliação. Para o que nos interessa, é a primeira etapa que intervêm as normas: a representação obtida é função da maneira pela qual o problema é apresentado a quem toma a decisão, mais do que pelas normas, hábitos e cuidados que adote. A representação determina os aspectos do problema que considerará".

220. WOJCIECHOWSKI, Paola Bianchi; MORAIS DA ROSA, Alexandre. Vieses da justiça: como as heurísticas e vieses operam nas decisões penais e a atuação contraintuitiva, p. 48.

221. Cf. KAHNEMAN, Daniel. Rápido e devagar: duas formas de pensar; WOJCIECHOWSKI, Paola; MORAIS DA ROSA, Alexandre, op. cit., p. 22.

não podem ser entendidos como unidades autônomas, mas com características bastante distintas no que diz respeito à velocidade, controlabilidade e conteúdo de suas operações[222].

O Sistema 1 serve para tomar decisões rápidas, com pouco esforço cognitivo e quando se dispõe de pouca informação, estando comumente associado a processos intuitivos ou "natos". Especialmente em sociedades primitivas, nas quais uma resposta rápida a um estímulo ambiental poderia significar a diferença entre a vida e a morte, ele foi (e continua sendo) bastante útil.

Todavia não levar em conta influência do Sistema 1 nos processos decisórios pode levar a erros sistemáticos de pensamento[223]. Ele atua por meio de atalhos cognitivos, também denominados de heurísticas, que nada mais são que mecanismos de reconhecimento de informações, que ajudam a encontrar respostas simples, ainda que imperfeitas, para perguntas que demandariam maior reflexão e esforço cognitivo – características do Sistema 2[224].

Assim, partindo do acúmulo de informações pretéritas, o Sistema 1 responde automaticamente em caso de ocorrência de uma situação semelhante[225]. Na verdade, esse sistema está em busca de coerência, de modo que ele acaba, não raro, por confundir "familiaridade e conforto cognitivo com ilusões de veracidade", tornando-se insensível "à quantidade e à qualidade da informação que estiver na gênese das impressões e intuições". Dessa forma, a busca por consistência das histórias narradas pelo Sistema 1 é muitas vezes gerada sem levar em conta a deficiência da informação disponível[226].

O Sistema 2, por outro lado, exige esforço e atenção e está relacionado a processos lógicos, sequenciais, e conscientes, mas também lentos e ineficientes. Além disso, ele é o único que permite pensar abstratamente e "programar a memória para obedecer instruções que encarnam tarefas não habituais"[227], como, por exemplo, encontrar objetos de características deliberadamente escolhidas em meio a vários outros. Além disso, a repetição de

222. WOJCIECHOWSKI, Paola; MORAIS DA ROSA, Alexandre, op. cit., p. 23.

223. WOJCIECHOWSKI, Paola; MORAIS DA ROSA, Alexandre, op. cit., p. 25.

224. KAHNEMAN, Daniel, op. cit., p. 127; NUNES, Dierle; LUD, Natanael; PEDRON, Flávio. Desconfiando da imparcialidade dos sujeitos processuais, p. 50.

225. NUNES, Dierle; LUD, Natanael, op. cit., p. 50.

226. WOJCIECHOWSKI, Paola; MORAIS DA ROSA, Alexandre, op. cit., pp. 28-29.

227. WOJCIECHOWSKI, Paola; MORAIS DA ROSA, Alexandre, op. cit., p. 35.

tarefas que inicialmente demandavam o Sistema 2, tais como dirigir ou praticar algum esporte, acaba por transferi-las ao Sistema 1, permitindo sua execução de forma automática.

Os dois sistemas, portanto, não competem, mas cooperam, visto que há processos mentais que exigem o Sistema 1 e outros que necessitam do Sistema 2. Todavia, ocorre que, por conta do esforço exigido pelo Sistema 2 e da velocidade e eficácia do Sistema 1, é possível que, mesmo diante de questões complexas, aquele venha apenas a endossar generalizações deste. Quando ocorre tal inversão, os resultados podem ser desastrosos, pois parte-se de simplificações exageradas que levam a decisões subótimas. As heurísticas e vieses ocorrem, portanto, quando se utiliza o Sistema 1, enquanto crê-se estar utilizando o Sistema 2[228].

Isso posto, torna-se evidente como tais falhas cognitivas podem interferir no âmbito judicial. Especialmente em uma situação de excesso de demandas, juízes e demais sujeitos processuais não têm condições e tampouco incentivos de analisar detalhadamente todas as peculiaridades dos casos que chegam até eles, de forma que criam "atalhos mentais" para decidir. Tais atalhos ou heurísticas configuram "mecanismos de decisão pré-pronta" e servem não somente para reduzir suas cargas de trabalho mental, mas também para gerar sensações de coerência e conforto cognitivo[229].

Tomando-se como exemplo o processo penal, certas situações recorrentes (i.e. os "crimes de sempre", como furto, roubo, tráfico, receptação etc.) levam a uma habituação de padrões que criam convicções antecipadas e fazem com que os demais detalhes sejam ignorados, pois tidos como supérfluos. A busca por coerência se traduz muitas vezes na utilização de brocardos, adágios, súmulas ou outros "mantras e lugares comuns"[230], que exemplificam alguns dos mecanismos de conforto cognitivo à disposição dos julgadores. Além disso, experiências passadas podem "contaminar" situações presentes, tais como a credibilidade de um depoimento de uma testemunha já conhecida de

228. WOJCIECHOWSKI, Paola; MORAIS DA ROSA, Alexandre, op. cit., p. 36; NUNES, Dierle; LUD, Natanael, op. cit., p. 52.

229. MORAIS DA ROSA, Alexandre. Guia do Processo Penal Conforme a Teoria dos Jogos, pp. 145-146.

230. MORAIS DA ROSA, Alexandre. Guia do Processo Penal Conforme a Teoria dos Jogos, pp. 762-763.

outros casos. Em conjunto, tais fatores sinalizam a inexatidão da crença do ser humano como "uma ficção ultra racional"[231].

Estar ciente das limitações dos processos cognitivos humanos é, portanto, essencial à análise dos vieses em algoritmos. Uma inteligência artificial será tão boa quanto for o material por meio do qual ela é treinada, de forma que dados tendenciosos farão com que ela chegue a resultados igualmente ruins[232]. Mais que isso, a depender da forma como são implementados, algoritmos não apenas irão reproduzir o comportamento decisório humano, mas desenvolver seus próprios vieses e, inclusive, acentuar distorções.

3.2.2. VIESES EM MÁQUINAS

> Quando pessoas que não estudaram matemática veem notações, elas se sentem intimidadas, assustadas. Há aí um certo tipo de autoridade, algum tipo de objetividade, de verdade científica, que eles não estão autorizados a questionar, pois não são *experts*. Essa "autoridade do inescrutável" é traduzida também para algoritmos[233].

Até o momento, abordou-se a questão da implementação do aprendizado de máquina em atividades legais a partir dos pressupostos e desafios relativos ao seu bom funcionamento. Esta seção, contudo, tratará daqueles modelos que deram errado ou, mais especificamente, das condições que podem vir a tornar um modelo em uma "arma de destruição matemática".[234]

Um modelo, conforme explica Cathy O'Neil, pode ser entendido como "uma representação abstrata de algum processo (...), que, independentemente de ser um programa de computador ou estar em nossas cabeças, utiliza informações que já sabemos para predizer respostas em variadas situações"[235]. Essencialmente, o que faz um modelo é utilizar dados preexistentes para descrever certas regularidades, que podem vir a ser úteis para processos de tomada de decisão futuros, à medida que tais padrões podem ser aplicados em novas situações. Dessa

231. MORAIS DA ROSA, Alexandre. Guia do Processo Penal Conforme a Teoria dos Jogos, pp. 162-163.

232. NUNES, Dierle; LUD, Natanael, op. cit., p. 148.

233. O'NEIL, Cathy. Weapons of Math Destruction: Vídeo-conferência. Personal Democracy Forum 2015. Disponível em: <https://www.youtube.com/watch?v=gdCJYsKIX_Y>. Acesso em: 16 nov. 2019. Tradução livre.

234. Cf. O'NEIL, Cathy. Weapons of math destruction. New York: Broadway Books, 2016.

235. O'NEIL, Cathy, op cit., p. 25.

forma, todo modelo conta com *inputs* (dados de entrada), *outputs* (dados de saída ou predições) e uma definição de sucesso. Um modelo dinâmico, ao seu turno, é aquele que recebe constantemente novas informações e as utiliza para aprimorar suas predições, com base em sua definição de sucesso.

Todavia, nenhum modelo é capaz de "captar toda a complexidade do mundo real ou as nuances da comunicação humana"[236]. Seu objetivo, na verdade, é ser uma simplificação, que permita separar aspectos relevantes e irrelevantes de um determinado processo com vistas a um determinado fim. Assim, um sistema de mapas de um GPS para veículos terrestres deve possuir informações detalhadas sobre estradas, pontes e túneis, ao passo que não precisa levar em conta, por exemplo, o formato de prédios, pessoas ou temperatura. Já um *software* que guia aviões deve, dentre outras medições, saber a velocidade do vento e a temperatura local, mas não necessita de informações sobre pontes ou estradas.

Dessa forma, ao se criar um modelo, escolhem-se quais aspectos da realidade ele deve levar em conta e os demais para os quais isso não é preciso. As características deixadas de lado ("pontos cegos"[237]) por modelos revelam as opiniões e prioridades de seus criadores, que são frutos de suas ideologias e ambições[238]. A isso soma-se o fato de que quanto mais complexo for o processo que o modelo busca descrever, mais variáveis ele deverá levar em conta, o que significa que exclusões arbitrárias podem resultar em simplificações errôneas.

Ao "sobresimplificar" processos complexos, um modelo pode começar a apresentar anomalias, o que já revela um dos grandes desafios de buscar-se implementá-los no âmbito judicial, uma vez que nele são discutidos os mais variados aspectos da vida humana.

Ainda assim, "se um modelo funciona ou não, também é uma questão de opinião"[239], pois sua definição de sucesso é arbitrária. Ocorre que um modelo pode estar funcionando bem aos olhos

236. O'NEIL, Cathy, op cit., p. 26.

237. MARTINS, Rui Cunha. O ponto cego do direito. Rio de Janeiro: Lumen Juris, 2011, p. 3: "Diz-se evidente o que dispensa a prova. Simulacro de auto-referencialidade, pretensão de uma justificação centrada em si mesmo, a evidência corresponde a uma satisfação demasiado rápida perante indicadores de mera plausibilidade. De alguma maneira, a evidência instaura um desamor do contraditório".

238. O'NEIL, Cathy, op cit., pp. 26-27.

239. O'NEIL, Cathy, op cit., p. 27.

de seus programadores, mas não sob a perspectiva das pessoas que por ele são afetadas. Se uma companhia de seguros utiliza um algoritmo para cobrar mais caro por apólices para pessoas de um determinado grupo, o modelo estará funcionando bem aos olhos de seus administradores, mas talvez não para aqueles que terão de pagar a mais pelos seus serviços[240].

Em modelos "inofensivos", utilizados, por exemplo, em predições desportivas, geralmente não há muito que se discutir em relação ao seu sucesso ou fracasso, uma vez que seu objetivo é bastante claro (e.g. ganhar mais jogos). Todavia, especialmente quando modelos são aplicados em atividades com grande impacto social, muitas vezes em prejuízo das pessoas por eles afetadas, haverá debates sobre o que define seu bom funcionamento. Além disso, pode-se falar em externalidades (positivas/negativas)[241] em face de terceiros não vinculados ao escopo original.

Fica evidente, destarte, que projetar modelos é uma tarefa permeada por subjetividade, ainda que seu grau possa variar de acordo com o tipo de processo a ser modelado. Em todo caso, para que um modelo tenha potencial de causar grande prejuízo para um grupo de pessoas, é preciso uma conjunção de fatores. Para Cathy O'Neil, um modelo se torna uma "arma de destruição matemática" quando ele é (i) opaco, (ii) possui a capacidade de crescer exponencialmente e (iii) é projetado para operar em prejuízo daqueles a ele sujeitos[242].

Especialmente quando utilizados pela Administração pública, modelos têm um grande potencial de se tornarem danosos[243]. Afinal, por pressuposto, seus usos serão estendidos a um grande número de pessoas e, junto disso, modelos apresentam uma "tendência à inescrutabilidade". Dessa forma, os dois

240. O'NEIL, Cathy, op cit., p. 28.

241. MANKIW, N. Gregory. Princípios de Microeconomia. Trad. Allan Vidigal Hasting; Elisete Paes e Lima. São Paulo: Cangage Learning, 2016, 184: "Uma externalidade surge quando uma pessoa se dedica a uma ação que provoca impacto no bem-estar de um terceiro que não participa dessa ação, sem pagar nem receber nenhuma compensação por esse impacto. Se o impacto sobre o terceiro é adverso, é denominada externalidade negativa. Se é benéfico, é chamado de externalidade positiva". Embora a noção de externalidade se vincule aos ganhos econômicos, pode-se adotar a compreensão dos efeitos (negativos ou positivos) do jogo processual em relação a terceiros não envolvidos diretamente no processo penal.

242. O'NEIL, Cathy, op cit., pp. 32-33.

243. FRANÇA, Phillip Gil. Ato administrativo, consequencialismo e compliance: Gestão de Riscos, Proteção de Daos e Soluções para o Controle Judicial na Era da IA. São Paulo: RT, 2029, p. 460.

primeiros requisitos estarão, em boa parte dos casos, automaticamente preenchidos ou em sua iminência. O terceiro requisito se torna, portanto, fundamental, uma vez que a forma como os modelos são projetados e os propósitos para os quais eles serão utilizados impactarão diretamente no seu potencial lesivo.

Não por coincidência, veio do Judiciário (estadunidense) o primeiro caso utilizado por O'Neil para exemplificar esse tipo de modelo[244]. Nos Estados Unidos, em um contexto de sobrerrepresentação de negros em cadeias, aos quais são aplicadas penas em média 20% maiores que a homens brancos condenados por crimes similares, vinte e quatro Estados aderiram a algoritmos que prometem um sistema mais imparcial e livre de vieses, para tornar as sentenças judiciais mais homogêneas e previsíveis. Ademais, prender pessoas injustamente por mais tempo representa maiores gastos, de modo que tais algoritmos poderiam contribuir para economizar recursos públicos[245].

O mais famoso desses modelos, o LSI-R, inventado em 1995, conta com um longo questionário que os condenados devem responder, que será utilizado para "prever" a probabilidade de que um condenado venha a reincidir, classificando tal risco em alto, médio ou baixo. Dentre as perguntas que devem ser respondidas pelos réus figuram: "quantas condenações anteriores você teve?" ou "qual a influência de drogas ou álcool na prática do crime?". O questionário também traz questões relativas a circunstâncias de nascimento e formação do condenado, inclusive perguntas sobre o histórico criminal de membros de sua família e amigos. Mesmo diminuído o fator aleatório decorrente de decisões humanas, alguém pode questionar se os vieses foram eliminados ou apenas camuflados. Ainda que não contenha perguntas que digam respeito expressamente à etnia, não é difícil imaginar como um modelo que leva em conta tais variáveis apresentará um viés racial.

É evidente que pessoas vindas de bairros de baixa renda (mais policiados) terão maior probabilidade de já terem tido prévios contatos com a polícia, se comparados com pessoas de classe média ou alta em face do processo de seleção secundária denunciados pela Criminologia Crítica[246]. Igualmente, será mais

244. Cf. O'NEIL, Cathy, op cit., Chapter 1 – Bomb parts: What is a Model?.

245. O'NEIL, Cathy, op cit., pp. 29-30.

246. BARATTA, Alessandro. Criminologia crítica e crítica do Direito Penal: introdução à sociologia do direito penal. Trad. Juarez Cirino dos Santos. Rio de Janeiro:

provável que alguém de sua família contenha algum histórico criminal ou problemas relacionados a álcool e drogas. Nesse cenário, o modelo sequer precisa analisar uma resposta relativa à raça para se tornar racista, pois tal informação é desnecessária quando as demais já apontam para questões racialmente sensíveis. O risco de correlações ilusórias[247] enviesa o resultado.

Em Estados como Colorado e Idaho, o *score* do preso é utilizado por juízes para fundamentar suas sentenças. Todavia, tanto nos Estados Unidos quanto no Brasil, se um juiz fundamenta sua sentença com base no histórico criminal de um familiar do condenado, sua decisão seria imediatamente objeto de impugnação, pois, em ambos os países, é ilegal (e inconstitucional) julgar alguém por algo que ele não fez[248]. Entretanto, a indicação de *score* acaba exercendo papel silencioso no momento de atribuição de responsabilidade e no *quantum* de aplicação da pena.

Acontece que, quando embutidas em um algoritmo, além de se revestirem de autoridade científica, tais opiniões passam geralmente despercebidas, pois não são inteligíveis para a maior parte das pessoas. Todavia, elas continuam (e sempre continuarão) a ser, simplesmente, opiniões, mas que agora estão perpetuadas e disfarçadas/embaladas, embora decorrentes de processos de avaliação de risco em que as entidades atribuem a probabilidade de ocorrência futura do evento. O ocorre, de fato, é que, em um dado momento, alguém decidiu que eram relevantes para um algoritmo que calcula a probabilidade de reincidência de um indivíduo informações relativas ao seu grau de escolaridade ou ao histórico criminal de seus pais. Nenhuma dessas conclusões é isenta de críticas e tampouco são "verdades científicas"[249].

Freitas Bastos, 1999; ANDRADE, Vera Regina Pereira de. A ilusão da segurança jurídica: do controle da violência à violência do controle penal. Porto Alegre: Livraria do Advogado, 1997; BATISTA, Nilo. Punidos e mal pagos. Rio de Janeiro: Revan, 1990; ZAFFARONI, Eugenio Raúl; BATISTA, Nilo. Direito Penal Brasileiro. Rio de Janeiro: Revan, 2003; YOUNG, Jock. A sociedade excludente: exclusão social, criminalidade e diferença na modernidade recente. Trad. Renato Aguiar. Rio de Janeiro: Revan, 2002; CARVALHO, Salo de. Antimanual de Criminologia. São Paulo: Saraiva, 2014; CIRINO DOS SANTOS, Juarez. Criminologia Radical. Rio de Janeiro: Lumen Juris, 2006; CHAVES JUNIOR, Airto. Além das grades: a paralaxe da violência nas prisões brasileiras. Florianópolis/SC: Tirant lo Blanch, 2018, p. 119-120.

247. STERNBERG, Robert J. Psicologia Cognitiva. Trad. Anna Maria Luche. São Paulo: Cengage Learning, 2012, p. 442.

248. Cf. BRASIL. Constituição (1988), art. 5º, XLV: "nenhuma pena passará da pessoa do condenado, podendo a obrigação de reparar o dano e a decretação do perdimento de bens ser, nos termos da lei, estendidas aos sucessores e contra eles executadas, até o limite do valor do patrimônio transferido".

249. SUMPTER, David. Dominados pelos números. Trad. Anna Maria Sotero e Mar-

Poder-se-ia argumentar que não se sabe ao certo qual o peso que o algoritmo dá a essas questões, até porque, caso se soubesse, os condenados poderiam simplesmente direcionar suas respostas para diminuírem seus *scores*. Contudo, qualquer que seja tal peso, ele será injusto (e, talvez, ilegal)[250]. Ainda que isso represente uma possível melhora com relação ao preconceito e aleatoriedade de juízes humanos, tais modelos continuam sendo injustos e isso não deve ser esquecido.

Ainda que se coloque momentaneamente de lado o critério ético, tais modelos apresentam mais um grave problema. Suas opiniões embutidas acabam por recriar as pressuposições que as sustentam, o que caracteriza um "ciclo de retroalimentação vicioso"[251]. Se um algoritmo como o LSI-R é utilizado no cálculo da pena, uma pessoa vinda de uma região socioeconomicamente vulnerável possuirá maiores chances de ser taxada com um alto grau de possibilidade de reincidência, o que lhe fará ser condenada por mais anos. Ao ficar mais tempo afastada do convívio social e em contato com outros prisioneiros, ela terá maiores dificuldades de voltar a se inserir na sociedade e conseguir um emprego, por exemplo. Dessa forma, caso cometesse outro crime e voltasse à cadeia, o algoritmo entenderia que realizou uma predição correta e reforçaria seu próprio entendimento. Esse *loop* vicioso, além de contribuir para que pessoas sejam punidas por conta de fatos que ainda não ocorreram, piora ainda mais o problema, uma vez que recria o ambiente que justifica suas premissas. Esse aspecto é frequentemente presente em "armas de destruição matemática" e está intimamente relacionado ao aspecto destrutivo de tais modelos.

Como, então, evitar que algoritmos se convertam em uma ameaça a toda uma sociedade? Tais modelos são preconcebidos

cello Neto Rio de Janeiro: Bertrand Brasil, 2019, p. 67: "Os acusados negros tiveram uma proporção bem maior de falsos positivos do que os acusados brancos". (...) "Se você for preso e um juiz usar um algoritmo para ajudar a avalia-lo a pior coisa que pode acontecer é você receber um resultado falso positivo. Um positivo verdadeiro é justo: o algoritmo previu que você era um risco e você era. Mas um falso positivo pode significar que lhe seja negada a liberdade condicional ou dada uma sentença maior do que você merece. Isso estava acontecendo com mais frequência com acusados negros do que com brancos. Quase metade dos acusados negros que não reincidiam tinha sido taxados como de alto risco. Por outro lado, acusados brancos recebem mais falsos negativos, em que o algoritmo diz que uma pessoa é de baixo risco, mas ela comete crimes no futuro". (...) "Para a sociedade, uma fração alta de falso positivo é um problema; isso significa que pessoas que deveriam ter sido detidas foram devolvidas à sociedade e cometeram crimes. Quase metade das pessoas brancas que reincidiram tinham sido taxadas como de baixo risco".

250. O'NEIL, Cathy, op cit., p. 30.

251. O'NEIL, Cathy, op cit., p. 30.

para serem escaláveis a toda a sociedade, o que torna incontornável a questão de seu crescimento exponencial. Com relação à sua opacidade, em um Estado Democrático de Direito, ela será sempre um desafio. O simples fato de um modelo ser escrito através de uma notação matemática e/ou por meio de uma linguagem computacional já o torna inacessível à maior parte das pessoas. Esse problema se torna ainda mais pertinente, quando se trata de modelos que utilizam aprendizado de máquina, cujos detalhes de seu modo de funcionamento não são conhecidos nem mesmo por seus programadores.

Especialmente quando utilizados pela Administração Pública, algoritmos devem seguir certos padrões mínimos de transparência e prestação de contas, bem como os princípios de direito que orientam o agir administrativo[252]. Ainda assim, isso não significa que modelos não possam ser tornados compatíveis com governos democráticos, com certo esforço de transparência, discussão pública dos critérios e responsabilidade pelos dados inseridos.

Pode-se dividir em ao menos duas etapas o processo de prestação de contas dos modelos, ainda que se queira preservar a integralidade de seus códigos propriamente ditos. Em um momento pré-elaboração dos códigos, é fundamental que a sociedade em geral tenha ciência de quais aspectos da realidade serão levados em conta pelo modelo, o que é percebido nos dados que serão utilizados como *inputs*. Dessa forma, é possível, por exemplo, contestar os motivos que levaram a certos dados serem considerados pertinentes e as razões pelas quais outros não foram sequer considerados. Após escritos tais códigos, pode-se fiscalizá-los por meio da auditoria de seus resultados, momento em que são detectados eventuais vieses ou distorções. Assim, uma vez implementado um modelo e verificado que ele penaliza um certo segmento social desproporcionalmente aos demais, sem razão justificada, pode-se pleitear sua alteração ou seu desativamento. Para tanto, mais uma vez, é necessário que tais resultados estejam acessíveis ao público, especialmente àqueles que são diretamente afetados.

Por fim, e talvez mais importante, é necessário que a própria ideia de concepção do algoritmo provenha de uma construção

252. Cf. BRASIL. Constituição (1988), art. 37: "A administração pública direta e indireta de qualquer dos Poderes da União, dos Estados, do Distrito Federal e dos Municípios obedecerá aos princípios de legalidade, impessoalidade, moralidade, publicidade e eficiência e, também, ao seguinte (...)".

conjunta entre todas as partes interessadas, a saber, gestores, público em geral e desenvolvedores. A participação de diversos setores da sociedade na elaboração de algoritmos é essencial para que suas definições de sucesso e as tarefas nas quais eles são empregados sejam estabelecidas da forma mais benéfica ao maior número de pessoas ou, quando isso não seja possível, sejam implementados de forma a mitigar ao máximo seus malefícios.

Com isso, diminui-se a chance de que tais algoritmos se tornem máquinas projetadas para arruinar vida de pessoas ou que entrem em ciclos de retroalimentação viciosos. Deve-se observar que um mesmo modelo pode ser mais ou menos danoso, a depender do uso que se faz dele. O citado LSI-R, ainda que o questionário que o fundamenta seja repleto de questões controversas, tem seu potencial danoso reduzido nos locais em que o utilizam não para fundamentar sentenças, mas para selecionar presidiários que serão incluídos em programas de prevenção de reincidência enquanto estão cumprindo suas penas.

Dessa forma, é imprescindível que se questione o fim para o qual o modelo será utilizado. Mais uma vez, a resposta só pode vir de uma construção conjunta entre todos os interessados. Isso nada mais é do que um dos pressupostos de um governo democrático e deveria ser adotado como padrão sempre que decisões que afetam um grande número de pessoas são tomadas.

3.3. FORMAS DE SE UTILIZAR O APRENDIZADO DE MÁQUINA NO JUDICIÁRIO

A partir das análises expostas, podem-se elencar alguns dos tipos de uso do aprendizado de máquina em atividades relacionadas à Jurisdição e suas respectivas vantagens e desvantagens. Antes de mais nada, deve-se observar que a complexidade do fenômeno linguístico representa por si só um entrave ao processamento de linguagem natural, sobretudo em atividades jurídicas, que possuem diversas especificidades no que diz respeito a formas de se definir conceitos e argumentar[253].

É verdade que a utilização de algoritmos contribui para minimizar fatores externos aleatórios tipicamente humanos, tais como cansaço e instabilidade emocional, mas eles também estão sujeitos a vieses estruturais decorrentes do sistema jurídico, da forma como

253. Cf. seção 1.2 deste trabalho.

eles são treinados e de sua própria programação.[254] Como visto, ainda que algoritmos como o Word2vec sejam capazes de assimilar, ao menos em parte, o contexto textual de palavras, não se pode afirmar que eles compreendam conceitos da forma como humanos fazem. Sua "compreensão" limita-se a associar uma palavra a outras que geralmente a acompanham e, ainda que se possa chegar a bons resultados através desse método, isso não é o suficiente para dar conta de todas as formas de uso da linguagem, que, assim como o Direito, configura um fenômeno social complexo. Mas se pode falar em possibilidade de "entendimento" e uso supervisionado.

Disso decorre a dificuldade de se quantificar em números certos conceitos, que, na verdade, não são computáveis, pois decorrem de um certo "jogo" que não pode ter suas regras previamente definidas. Uma vez que "computadores, a despeito de todos os seus avanços em linguagem e lógica, ainda têm muitos problemas com conceitos"[255], eles acabam tendo que utilizar dados "aproximados" para se alcançar uma possível definição de termos abstratos[256].

Todavia, ao assim proceder, há o risco de perpetuação de um *status quo* relativo a uma certa ordem social[257], à medida que se pereniza uma certa concepção de mundo, dentre diversas possíveis. Por conta disso, nem sempre será possível conciliar eficiência com equidade, de modo que será necessário optar entre uma ou outra. Nesses casos, em um ambiente democrático, caberá à sociedade definir o que se espera do uso de algoritmos em atividades que impactam um grande número de pessoas. Sua atuação será fundamental para definir as balizas que guiarão tais práticas e deverá vir dela a palavra final sobre o que se está disposto a sacrificar em benefício da eficiência. De qualquer forma, parece impossível, pelo menos por hora, outros usos de robôs, que não o de estrito apoio à decisão humana.

3.3.1. A QUESTÃO DA CORROBÓTICA

No atual estado da arte, encaminha-se para um cenário em que a não utilização de algoritmos será um sinônimo de obsolescência e o Poder Judiciário se mostra um dos terrenos

254. BUOCZ, Thomas Julius. Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary, p. 44.

255. Tradução livre: "And computers, for all of their advances in language and logic, still struggle mightily with concepts." em O'NEIL, Cathy, op cit., p. 82.

256. O'NEIL, Cathy, op cit., p. 82.

257. O'NEIL, Cathy, op cit., p. 79.

mais propícios para a implementação do aprendizado de máquina. Isso se deve ao fato de que nele concorrem fatores como: i) grande concentração de dados potencialmente tratáveis ii) grande demanda por agilidade na prestação do serviço, isolada ou em parcerias; iii) disponibilidade de orçamento para implementar soluções inovadoras e; iv) escassez de recursos humanos para cumprir com a carga de trabalho demandada.

Antes de mais nada, deve-se levar em conta que as limitações atuais do aprendizado de máquina, inseridas dentro da lógica de uma inteligência artificial de propósito limitado, somente permitem que juízes sejam substituídos em situações específicas (IA fraca), uma vez que tais profissionais executam tarefas de diferentes naturezas, excedendo, portanto, o escopo de atuação daquele tipo de inteligência[258]. Além disso, não se pode esperar que máquinas atuem de forma perfeita, bastando que elas contem com precisão e transparência coerentes com o tipo de atividade a ser desempenhada, tendo sempre como referencial o desempenho humano em processos equivalentes.

Dessa forma, dado que, ao menos por hora, juízes humanos não poderão ser completamente eliminados, as questões relativas ao uso do aprendizado de máquina na administração da justiça dizem respeito principalmente à coexistência entre humanos e máquinas em um mesmo ambiente e as consequências daí decorrentes, o que define a corrobótica. Nesse sentido, torna-se relevante a forma como os algoritmos influenciarão o processo decisório levado a cabo por humanos, bem como saber se isso significa alguma renúncia de poder por parte do ser humano.

Todavia, a depender da abordagem escolhida, é possível contornar a questão da corrobótica. Em casos excepcionais, será buscado eliminar o fator humano da elaboração de decisões, ainda que estas tenham que ser revistas posteriormente por juízes de carne e osso (o que não remove por completo, portanto, o *homo sapiens* da equação). Isso ocorrerá nos casos em que "juízes-robô" irão decidir litígios do início ao fim e humanos se tornam uma espécie de instância recursal. Ainda assim tal tipo de uso será restrito a casos de baixa complexidade e que comportam pouco ou nenhum poder discricionário por parte dos juízes (casos que não fogem, portanto, da aplicação padrão da norma).

258. BUOCZ, Thomas Julius. Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary, p. 46.

Por exemplo, no artigo antes referido[259], foi possível ensinar a máquina a julgar ações de guarda nos termos do Estatuto da Criança e do Adolescente, tendo o algoritmo prolatado a decisão após 1.308 tentativas, mediante o uso de MCDA-C, constando da conclusão:

> Os testes autorizam afirmar que a Metodologia Multicritério de Apoio à Decisão – Construtivista (software MACBETH-SCORES), associada com a ferramenta machine learning é adequada a reunir CI e Direito, através de hipóteses válidas em ambas as ciências simultaneamente. É possível replicar sentenças desta forma e oferecer sistema de apoio à manutenção da estabilidade das decisões ao magistrado. Segundo, é possível afirmar que este método tem potencial para suplantar os desafios metodológico-jurídico-algoritmicos apontados na introdução. Especificamente para fins deste artigo, a subjetividade do programador e do magistrado não são empecilhos na medida em que alimentam a metodologia de mais informações e esta, através da infinitas tentativas e erros, aprende com a ajuda e calibragem de quem decide, até que este se dê por satisfeito e não possa distinguir uma sentença sua, da realizada pela máquina. Subjetividades são preservadas. Por outro lado, os resultados dos testes empíricos demonstraram que a hipótese está mal formulada. O magistrado não decide através de acurácia em face de precedentes, nos termos da CI, não deveria este ser o referencial. Raciocinar em termos de acurácia parte de pressuposto inadequado em face dos testes. Decisões favoráveis se dão mesmo quando a acurácia é inferior a 50% no que toca os fundamentos da sentença (Gráfico 4), em que pese, a estabilidade da amostragem se dar em 100% no que tange a parte dispositiva da sentença (no que toca a procedência do pedido). Por outro lado, variações maiores entre as sentenças, ou a utilização de argumentos que não se repetiram não impediram que o resultado da decisão fosse exatamente o mesmo, pela procedência do pedido das partes. O ideal de acurácia é substituído pela satisfação e confiança do decisor (magistrado). A vagueza e polissemia da linguagem enriquecem o algoritmo. Outro paradigma se apresenta, não se trata de superar obstáculo da subjetividade, a partir da MCDA-C esta não tem conotação negativa, mas é partícipe da otimização do sistema".

Assim é que o problema não diz mais respeito à corrobótica, mas a novas questões. Haverá uma "pressão" para que juízes humanos ratifiquem decisões das máquinas? O que ocorre em caso de divergência? Além disso, um caso dito simples para uma pessoa

259. MENDES, Alexandre José; MORAIS DA ROSA, Alexandre; ROSA, Otacílio Izaías da. Testando a *Methodology Multicriteria Decision Aid – Constructivist* (MCDA-C) na construção de algoritmos de apoio à estabilidade das decisões judiciais. Revista Brasileira de Direito, v. 15, n. 02, 2019, p. 281-305 (ISSN 2238-0604).

pode ser complexo para outras. Quem decidirá o tipo de litígio que poderá ser objeto de uma decisão automatizada?

Por consequente, a depender da abordagem utilizada, isto é, máquinas auxiliando humanos no processo decisório ou tomando decisões sozinhas, evita-se o problema da corrobótica ou se adere ao mecanismo. Ambas as abordagens, de todo modo, devem lidar com o fato de que algoritmos de aprendizado de máquina gozam de certa autonomia, aqui entendida como a capacidade de ditar (ao menos em parte) suas próprias regras.

Contudo, na primeira abordagem, torna-se especialmente importante a questão da transparência da comunicação entre o algoritmo e o humano. É certo que a transparência só ocorre ao custo da performance, pois significa forçar uma inteligência artificial a trabalhar com parâmetros humanos, os quais nem sempre são os mais eficientes.[260]

Todavia, nem mesmo um juiz humano é totalmente transparente, se por "transparência" entende-se compreender cada um dos motivos que levam certa pessoa a tomar uma certa decisão. Deve-se, dessa forma, distinguir entre transparência no que tange aos detalhes do funcionamento de uma máquina daquilo que diz respeito à capacidade de se explicar seus resultados, o que não é impossível de ser feito.

Já no caso da abordagem que evita o problema da corrobótica, os juízes-robô não necessariamente necessitam ser transparentes, uma vez que a instância humana revisora cumprirá esse papel. Todavia, não se pode ignorar que, tanto mais seja possível que robôs "expliquem" suas decisões, maior será a legitimidade do uso desse tipo de ferramenta. Transparência, nessa situação, significa que os algoritmos sejam capazes de expor termos ou frases que tiveram maior peso na classificação do caso, bem como quais normas, julgados ou precedentes foram aplicados em sua resolução. Mais uma vez, não pode se esperar aqui uma transparência no nível das minúcias do funcionamento, mas somente no que diz respeito a tornar algumas de suas razões humanamente inteligíveis, ou seja, das premissas tomadas como válidas para inferências.

260. BUOCZ, Thomas Julius. Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary, p. 49.

3.3.2. TRÊS TIPOS DE USO DO APRENDIZADO DE MÁQUINA NO DIREITO: A) ROBÔ-CLASSIFICADOR; B) ROBÔ-RELATOR E; C) ROBÔ-JULGADOR.

Dessa forma, verificam-se diversos usos potenciais do aprendizado de máquina no âmbito judicial, alguns dos quais já foram ou estão em vias de ser implementados[261]. Dentre as possíveis formas de utilização, podem-se elencar três grandes grupos (ou tipos de uso), que concentram a maior parte dessas abordagens. Para se definir tais grupos, foram escolhidos alguns critérios relativos a questões normativas e técnicas, a saber, (i) *grau de intervenção humana*, (ii) *interferência do algoritmo no processo decisório*, (iii) *complexidade do algoritmo envolvido* e (iv) *transparência da decisão*.

Antes de mais nada, deve-se notar que alguns dos critérios não são absolutos, mas apenas comparações com seus equivalentes em processos decisórios levados a cabo por seres humanos. Assim, quando se fala em um "alto grau de transparência", quer-se dizer que a transparência do processo decisório (iv) ocorre a um nível semelhante ao de uma decisão elaborada por seres humanos sem a intervenção de algoritmos. O mesmo ocorre com relação ao grau de intervenção humana (i).

a) Robô-Classificador: Um primeiro tipo, doravante denominado *classificador*, tem por função primordial encontrar materiais úteis para que humanos fundamentem suas decisões. Tais materiais compreendem, por exemplo, dispositivos normativos, precedentes judiciais e modelos de documentos que servirão de base para direcionar o pronunciamento judicial.

Além disso, tais algoritmos podem ser utilizados para poupar recursos ao localizar processos em tramitação que deveriam estar aguardando julgamentos em instâncias superiores, como ocorre na sistemática de vinculação de precedentes (Repercussão Geral, Recursos Repetitivos, Incidentes de Resolução de Demandas Repetitivas, Incidente de Assunção de Competência, etc). Dessa forma, pode-se sobrestar sua tramitação até que o respectivo tema seja estatuído e, com isso, evitar que os processos já

261. MORAIS DA ROSA, Alexandre; GUASQUE, Bárbara. O avanço da disrupção nos Tribunais Brasileiros. In: NAVARRRO, Erick; NUNES, Dierle; LUCON, Paulo (orgs.) Inteligência Artificial e Direito Processual: os impactos da virada tecnológica no direito processual. Salvador, JusPodivm, 2020, p. 95-81; PEREIRA, Sebastião Tavares. O machine learning e o máximo apoio ao juiz. Revista Democracia Digital e Governo Eletrônico, Florianópolis, v. 2, n. 18, p. 2-35, 2018. Disponível em: http://buscalegis.ufsc.br/revistas/index.php/observatoriodoegov/article/view/303.

julgados tenham de ser revistos por estarem em sentido contrário ao decidido por tribunais superiores. Para esse tipo de uso, a complexidade dos algoritmos envolvidos é relativamente baixa, desde que haja dados em quantidade e qualidade razoáveis.

Dado que humanos terão de elaborar integralmente os documentos, sua intervenção no processo decisório será máxima. Ainda assim, isso não significa que não haverá interferência da máquina, já que poderá induzir o entendimento dos humanos à medida que "filtra" a informação à qual eles terão acesso, o que é consequência das considerações expostas na seção anterior. Algoritmos, ainda que atuando "apenas" como classificadores, estarão sedimentalizando certos entendimentos e a mera indicação de casos ditos semelhantes/relevantes já configura uma interferência.

É claro que o juiz pode discordar da classificação, mas a tendência será de que a sugestão do algoritmo se transforme em novas decisões judiciais, em um processo recíproco de seleção informação e sua transformação em expertise legal[262]. Uma vez publicada, tal decisão passará a compor o conjunto de processos relativos àquele tema e passará a ser sugerida em novos casos, retroalimentando um entendimento prévio. Leva-se a sério a técnica das distinções[263], apurando-se as razões fortes que vincularão, em princípio, os casos futuros, não obstante possam ser superados ou distinguidos em face do suporte fático. Diferencia-se a *ratio decidendi* (fundamentação forte) das questões satélites, denominadas de *obiter dictum* (fundamentação fraca)[264]. Confundir uma com outra é um erro comum da

262. BUOCZ, Thomas Julius. Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary, pp. 51-52.

263. DAVID, René. Os Grandes Sistemas do Direito Contemporâneo. Trad. Hermínio A. Carvalho. São Paulo: Martins Fontes, 2002, p. 411.

264. SGARBOSSA, Luis Fernando. O conceito de derrotabilidade normative: noções fundamentais e análise crítica. In: SERBENA, Cesar Antonio (coord). Teoria da Derrotabilidade: pressupostos teóricos e aplicações. Curitiba: Juruá, 2012, p. 65: "A relação entre técnica de distinções e a ideia ou conceito de derrotabilidade parece clara: o jurista do sistema da common law aprende desde o início a raciocinar por analogia, procedendo a uma comparação entre o caso a ser decidido e os precedentes potencialmente aplicáveis, e distinguindo, entre estes, entre os mais próximos e os mais distantes do caso a ser julgado". RADBRUCH, Gustav. O espírito do Direito Inglês e a Jurisprudência Anglo-Americana. Rio de Janeiro: Lumen Iuris, 2010, p.56-57: "Parte-se do caso concreto ou de uma apreciação jurídica; procura-se demonstrar que o caso atual teria outro embasamento além daquele que fundamentava o precedente, ou que o princípio jurídico aparentemente fundamentador desse precedente na verdade não estava contido nele. O princípio jurídico proferido com base em um precedente é vinculante somente o necessário para a decisão do caso citado; se, na época, ele foi compreendido mais amplamente do que teria sido necessário, não se trata de uma *ratio decidendi* (fundamento de decisão), e sim, de um *obter dictum* (exceção ocasional) ou, simplesmente, *dictum* do

prática jurídica brasileira[265]. A proposta do 'método de estudo' de caso é observar, avaliar uma situação real e, com base em fundamento teórico, efetuar um exercício para poder enxergar outras soluções[266]. O "caso", num primeiro momento, exige que seja observado, refletindo-se acerca do que está ocorrendo, como ele se dá, quais os requisitos e variáveis. No estudo de caso, portanto, deve-se escolher e limitar o fenômeno, singularizado, a ser estudado, isto é, debruçar-se sobre um contexto processual específico. A teoria do "estudo de casos" fornece as balizas teóricas e o aporte metodológico a fim de se criar um modelo de gestão de informação. Pode-se afirmar que o estudo de caso aborda um panorama vertical do fenômeno, buscando aprofundamento de dados e informações, compreensão dos significados[267] e, para

juiz. Às vezes o próprio precedente confere ao juiz a escolha entre mais de um princípio jurídico quanto, a bem dizer, os diversos juízes, em portanto, no precedente são expressas opiniões jurídicas entre si concorrentes (*concurring opinions*)".

265. MORAIS DA ROSA, Alexandre; MAIA, Maurílio Casas. Julgado não é sinônimo de Precedente: distinção que você deveria saber para evitar confusões na fundamentação dos julgados (Kanneman e os Sistemas S1 e S2): "Precedente e julgado não são sinônimos. O uso do significante 'precedente' constantemente como se referindo a julgados anteriores, no contexto brasileiro, não pode ser automaticamente acoplado ao modelo americano (...). Há distinções marcantes no modo como o Direito é construído e aplicado. (...) Afirmar que um julgado anterior deve ser aplicado a um caso concreto pode ser, muitas vezes, uma atividade envolta por conclusões precipitadas e com base em poucas evidências. (...) O respeito e observância dos elementos do precedente – como a *ratio decidendi* e o *obiter dictum* – , e assim também das chamadas técnicas de 'superação' e 'distinção' entre os precedentes, devem ser pensados à luz do modelo decisório "S2", com toda sua acuidade.". Consultado em: https://emporiododireito.com.br/leitura/julgado-nao-e-sinonimo-de-precedente-distincao-que-voce-deveria-saber-para-evitar-confusoes-na-fundamentacao-dos-julgados-kahneman-e-os-sistemas-s1-e-s2; RAMIRES, Maurício. Crítica à aplicação de precedentes no direito brasileiro. Porto Alegre: Livraria do Advogado, 2010; SAUSEN, Dalton. Súmulas, Repercussão Geral e Recursos Repetitivos. Porto Alegre: Livraria do Advogado, 2013; SANTOS JUNIOR, Rosivaldo Toscano dos. Controle Remoto e Decisão Judicial: quando se decide sem decidir. Rio de Janeiro: Lumen Juris, 2014.

266. YIN, Robert K. Estudo de Caso: Planejamento e Métodos. Trad. Cristhian Matheus Herrera. Porto Alegre: Bookman, 2015, p. 2: "A pesquisa de estudo de caso seria o método preferencial em comparação aos outros em situações nas quais (1) as principais questões da pesquisa são 'como?' ou 'por quê?"; (2) um pesquisador tem pouco ou nenhum controle sobre eventos comportamentais; e (3) o foco de estudo é um fenômeno contemporâneo (em vez de um fenômeno completamente histórico). Como primeira parte de uma definição em duas partes, um estudo de caso investiga um fenômeno contemporâneo (o 'caso') em seu contexto no mundo real, especialmente quando as fronteiras entre o fenômeno e o contexto puderem não estar claramente evidentes. A segunda parte da definição aponta para o projeto e a coleta de dados – por exemplo, como a triangulação de dados ajuda a tratar a condição técnica distintiva, por meio da qual um estudo de caso terá mais variáveis de interesses do que pontos de dados. Dentre as variações em estudo de caso, um estudo de caso pode incluir casos únicos ou múltiplos, pode ser limitado a evidências quantitativas e pode ser um método útil para fazer uma avaliação".

267. WALTON, Douglas N. Lógica Informal. Trad. Ana Lúcia R. Franco e Carlos A. L. Salum. São Paulo: Martins Fontes, 2012, p. 4: "Do ponto de vista pragmático, cada argumento tem que ser considerado no contexto de um ambiente de diálogo determinado. A sensibilidade às características especiais dos diferentes contextos de diálogo é

isso, pode ser útil perceber como ele interage com outros fenômenos, dentro do dispositivo do jogo processual singularizado.

Por fim, a transparência desse processo será máxima. Da mesma forma que um juiz pode pedir aos seus assessores que eles procurem manualmente precedentes/julgados relativos a um determinado caso, ele pode se fiar em um algoritmo para que o faça, sem que isso mude a transparência de sua decisão. Em ambas hipóteses, será necessário que ele fundamente sua decisão final da mesma forma como tradicionalmente faz. Ainda assim, é importante que o algoritmo forneça meios para que suas decisões sejam interpretáveis, uma vez que a pesquisa legal e o processo decisório são atividades interdependentes[268].

Como exemplo desse modo de se utilizar o aprendizado de máquina, tem-se o projeto Victor, do Supremo Tribunal Federal, que tem por escopo trazer maior agilidade na tramitação de processos na Corte Superior ao sugerir automaticamente vinculações de processos novos a temas de Repercussão Geral.[269] Pode-se facilmente imaginar, em um futuro próximo, que tribunais e outros órgãos vinculados à administração da justiça de todo o país implementem técnicas semelhantes ou que utilizem tecnologias similares para resgatar suas próprias decisões, na tentativa de homogeneizar sua jurisprudência.

b) Robô-Relator: Uma segunda forma de utilização do aprendizado de máquina, aqui apelidada de *relatora*, diz respeito a extrair e condensar informações relevantes de um ou mais documentos, o que pode ser utilizado para diferentes fins. Para tanto, ela deve ser capaz não apenas de encontrar documentos similares, mas ir mais a fundo em sua estrutura, diferenciando, em cada peça processual, aquilo que se refere à descrição de fatos, textos legais, jurisprudências colacionadas e estruturas argumentativas.

Nesse sentido, é necessário que o algoritmo possua habilidades em, ao menos, mineração de textos, expansão de conceitos e extração de relações[270], isso porque se torna necessário ser capaz

uma exigência para a análise racional de um argumento".

268. BUOCZ, Thomas Julius. Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary, pp. 51-52.

269. SUPREMO TRIBUNAL FEDERAL. Inteligência artificial vai agilizar a tramitação de processos no STF. Notícias STF. Brasília. 30 maio 2018. Disponível em: <http://www.stf.jus.br/portal/cms/verNoticiaDetalhe.asp?idConteudo=380038>. Acesso em: 01 nov. 2018.

270. Cf. item 1.1.1 deste trabalho

de encontrar informações que resumem um documento, sintetizar argumentos e relações entre partes e identificar relações semânticas e sintáticas entre os termos.

Utilizado dessa maneira, um algoritmo ainda estaria apenas auxiliando um juiz na tarefa de fabricar uma decisão, de forma que persiste a questão da corrobótica. Suas considerações podem ser aceitas ou recusadas, mas, frise-se, o juiz, como tendência, concordará com a máquina[271], seja por conveniência, seja pelo fato de que uma decisão amparada por um algoritmo terá menores chances de ser revista. Aproveita-se a lógica intuitiva e o atalho promovido pelas heurísticas decisórias.

Nesse contexto, a transparência se mantém a níveis próximos dos casos decididos sem o auxílio de algoritmos, pois o juiz continua tendo que dar os "toques finais" no documento (ou refazê-lo por completo). Ainda assim, se comparado ao uso anterior, é mais fácil saber a real extensão do uso da inteligência artificial.[272] A saída e a responsabilidade serão sempre humanas (julgador).

Contudo, uma vez aceitas as sugestões do algoritmo, a intervenção humana ocorre de forma significativamente mais baixa que no primeiro exemplo (classificador), dado um cenário em que o juiz terá apenas de assinar o documento, limitando sua intervenção à revisão de informações. Ao passo que o modelo aprimora sua acurácia, a tendência é que o juiz se torne uma espécie de "canal de entrega" de decisões geradas por robôs, o que aumenta nesta abordagem o grau de interferência da máquina, se comparada à anterior[273].

A maior complexidade técnica do robô-relator significa também maior versatilidade, que permite empregar esses modelos em diferentes funções. A primeira diz respeito a elaborar decisões "pré-fabricadas" para juízes: a máquina indica ao juiz as páginas em que se encontram as peças processuais, elenca os argumentos trazidos por cada uma das partes e, eventualmente, sugere uma decisão para o caso.

Uma segunda possibilidade de aplicação desses algoritmos pode ocorrer de modo próximo à atuação de um juiz leigo. Por

271. BUOCZ, Thomas Julius. Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary, pp. 54-55.

272. BUOCZ, Thomas Julius. Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary, p. 54.

273. BUOCZ, Thomas Julius. Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary, p. 55.

exemplo, em Juizados Especiais (Lei 9.099/95), após ambas as partes terem tido a oportunidade de se manifestar, não sendo necessária a produção de novas provas e saneado o processo, um algoritmo sugere um encaminhamento ao caso, baseado em decisões passadas daquela corte. Se as partes concordarem, o acordo é encaminhado a um juiz humano para ratificá-lo, caso contrário, aguarda-se o julgamento humano. Deve-se observar que, dados os atuais limites tecnológicos, tanto no primeiro caso, quanto no segundo, os processos analisados pelos algoritmos deverão ser de baixa complexidade.

Em outra forma de atuação, os robôs relatores podem atuar na predição de decisões judiciais (isto é, jurimetria). Uma vez que são capazes de distinguir diferentes estruturas textuais, é possível treiná-los para analisar tipos de argumentação legal, descrição dos fatos, bem como provas acostadas aos autos para calcular-se quão relevantes eles foram para a procedência ou improcedência de ações passadas. Assim, em um novo caso, podem-se estimar as chances de se sair vitoriosa uma determinada forma argumentativa.

Nessa linha, pesquisadores conseguiram prever com acurácia média de 79% decisões do Tribunal Europeu de Direitos Humanos – TEDH[274]. O experimento consistiu na elaboração de um modelo de classificação binária que, tomando como *inputs* exclusivamente dados textuais, dizia se houve ou não violação de alguns dos artigos da Convenção Europeia dos Direitos Humanos – CEDH. Levando em conta a estrutura dos julgados da Corte, os pesquisadores extraíram de decisões prévias informações relativas a questões procedimentais, fatos (circunstâncias do caso) e à letra da lei e utilizaram-nas na predição da parte dispositiva dos mesmos casos[275]. Com isso, concluíram que não apenas é possível estabelecer uma correlação entre dados textuais de um caso e sua decisão, mas também que as circunstâncias fáticas tinham mais peso na predição que os argumentos legais[276].

c) Robô-Julgador: O terceiro e último tipo de uso do aprendizado de máquina no Judiciário, a saber, o robô-julgador, apresenta características muito próximas do modelo anterior, no que diz respeito às suas funcionalidades. Sua diferença reside

274. Cf. ALETRAS, Nikolaos et al. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective.

275. ALETRAS, Nikolaos et al, op. cit., pp. 6-8.

276. ALETRAS, Nikolaos et al, op. cit., pp. 15-16.

principalmente no tratamento que se dá ao resultado gerado pelo algoritmo, que será tido como a própria decisão judicial. Aplica-se, principalmente, em demandas repetitivas.

Dessa forma, ocorre um ato completamente automatizado, através do qual se decide um processo judicial. No caso de as partes discordarem do conteúdo decisório, apela-se à instância humana revisora, que poderá manter ou reformar a decisão artificialmente gerada e o processo segue normalmente seu curso. Há ao menos duas formas de se entender tal abordagem.

Por um lado, pode-se considerar que a interferência algorítmica no processo decisório é eliminada por completo, ao mesmo tempo que não há mais sentido falar-se em grau de intervenção humana, uma vez que se separa o componente humano da máquina. Dessa forma, ao contrário do que acontece com o Robô-relator, evita-se certa perda de poder decisório por parte do humano, que deriva da dificuldade de se discordar de uma certa "autoridade" que os algoritmos possuem por conta de suas pressupostas precisão e neutralidade.

Além disso, elimina-se o problema de justificação da decisão, pois o algoritmo apenas gerará um resultado, cujas explicações ficarão a cargo do juiz de carne e osso, se for o caso de inconformismo de uma das partes[277]. Nessas situações, a transparência seria máxima, pois as justificativas da decisão seriam elaboradas por humanos.

Sob outra perspectiva, a influência da máquina no processo decisório poderá ser entendida como bastante elevada. Será muito mais conveniente aos juízes humanos ratificarem a sentença artificial, haja vista que eles estarão possivelmente tratando de casos massificados e de baixa complexidade, além de que será simplesmente mais trabalhoso discordar e explicar a falha da inteligência artificial.

Além disso, não seria surpreendente se decisões automaticamente elaboradas tivessem índices altíssimos de recorribilidade, especialmente no caso de elas não conterem em si mesmas justificativas satisfatórias e compreensíveis para humanos. Ainda assim, tais índices dependeriam da matéria em questão e da logística processual em vigor. Em todo caso, seria necessária uma análise minuciosa do custo-benefício da

277. BUOCZ, Thomas Julius, op. cit., pp. 55-56;

implementação desse tipo de sistema e dos resultados práticos esperados. Um exemplo seria a criação de um regime de custas e honorários nos casos de recursos improvidos, na forma do art. 55, da Lei 9.099/95, em que o recorrente vencido arcaria com uma sanção pecuniária. Ainda assim, corre-se o risco de impor barreiras ao princípio do duplo grau de jurisdição, o que seria especialmente problemático no caso de algoritmos enviesados.

Outro problema diz respeito a quais casos poderão ser objeto de decisões automáticas. Se há discordância entre o humano e a máquina, algum deles errou, ambos erraram ou o caso apresenta mais de uma solução. O que difere caso simples (*plain/easy cases*) de casos complexos (*hard cases*) é se a aplicação da norma se dá de forma automática e sua solução se dá de maneira simples ou se ela configura uma daquelas situações em que há, ao menos aparentemente, mais de uma decisão juridicamente válida[278].

Dessa forma, em *hard cases* não será possível dizer qual dos dois errou, pelo menos enquanto não houver uma decisão de uma corte hierarquicamente superior para definir a matéria. Mesmo assim, nesses casos, importa mais a explicação das razões que levaram a uma decisão que a precisão (ou acurácia) da decisão em si. Por conta disso, *hard cases* se mostram incompatíveis com a aplicação do aprendizado de máquina[279].

3.3.3. DIFERENTES TIPOS DE ERROS

Dadas as diferentes maneiras de se utilizar modelos no Poder Judiciário, faz-se necessário distinguir como cada uma delas estará sujeita a erros sistemáticos. É certo que somente uma análise caso a caso poderia dar conta de explicar o que levou um modelo a apresentar resultados indesejados, mas, de modo geral, vieses decorrem de uma má interpretação da realidade (sempre em perspectiva) por parte de seus criadores / aplicadores, seja na escolha dos dados que irão treiná-lo, seja na formulação das perguntas / interpretação das respostas que ele fornece.

Máquinas, de fato, estão sujeitas a bugs, mas ao observar modelos que "deram errado", escolhas humanas normalmente foram um agravante ou mesmo uma condição determinante para o erro. Tome-se como exemplo, saindo um pouco do âmbito jurídico, a

278. BUOCZ, Thomas Julius, op. cit., p. 56.
279. BUOCZ, Thomas Julius, op. cit., p. 56.

robô nazista da Microsoft[280] ou o modelo racista do Google Photos[281]. Houve erros, mas de onde eles vieram? Parece inapropriado afirmar que a falha foi dois algoritmos, até porque estes podem ter permanecido inalterados. O que, sem dúvidas, teve de mudar foi o material utilizado para treiná-los, fruto direto de escolhas humanas. Modelos estatísticos (o que inclui o aprendizado de máquina) buscam padrões que melhor descrevem certos dados. Entretanto, nesses casos, os resultados indesejados não advieram de falhas ao detectar certas regularidades, mas justamente encontrar nos dados padrões indesejados e, até um certo momento, ignorados.

Não obstante, para além da qualidade dos dados, importa também a maneira de se formular as perguntas que o modelo deve responder, tendo em conta, sobretudo, suas limitações tecnológicas. Dito de outro modo, bons dados são de pouca utilidade se são feitas perguntas erradas[282]. Grandes quantidades de dados podem gerar uma sensação de completude de informação, o que quase nunca é o caso e leva a ignorar a informação que não está disponível. Isso é especialmente relevante quando modelos são aplicados na predição de situações sociais complexas.

O modelo citado no início deste capítulo (LSI-R) ilustra mais uma vez o problema. É bastante pretensioso, para dizer o mínimo, afirmar que respostas de um questionário socioeconômico seriam capazes de definir a probabilidade que uma pessoa venha a reincidir em práticas criminosas. Independentemente de sua quantidade, a qualidade desses dados está muito aquém da complexidade da pergunta e, certamente, eles só serão capazes de respondê-la se inúmeras outras pressuposições (escolhas humanas) forem adotadas. Até porque operam numa lógica criminológica que desconsidera a seleção e o etiquetamento denunciados pela Criminologia Crítica, por exemplo.

Não obstante, um mesmo modelo, ou seja, um mesmo algoritmo treinado com os mesmos dados, pode ter seu potencial danoso bastante reduzido a depender de como suas respostas são encaradas. Seria o caso em que a pergunta fosse "quais pessoas vêm de piores condições socioeconômicas?" em vez de "fulano vai

280. https://revistagalileu.globo.com/blogs/buzz/noticia/2016/03/microsoft-criou-uma-robo-que-interage-nas-redes-sociais-e-ela-virou-nazista.html
281. https://brasil.elpais.com/brasil/2018/01/14/tecnologia/1515955554_803955.html
282. https://www.ibm.com/blogs/digital-transformation/br-pt/nao-adianta-termos-bons-dados-se-nao-fizermos-boas-perguntas/

voltar a cometer um crime?" e o modelo utilizado para direcionar programas de reinserção/engajamento social e não para aumentar penas. Feitas essas ponderações, torna-se mais fácil entender como vieses (em humanos e máquinas) podem ser mitigados ou acentuados, a depender das funções desempenhadas e da etapa do processo decisórios em que modelos são inseridos.

Seguindo a mesma ordem anteriormente apresentada, serão tecidas algumas análises sobre possibilidade de ocorrência de erros sistemáticos em robôs-classificadores, relatores e julgadores. Com relação aos primeiros, a classificação, como dito, não é uma tarefa isenta de subjetividade e tampouco é "inofensiva", no quesito interferência em decisões humanas. Em um modelo supervisionado, as classes são definidas por humanos e, dentre as diversas maneiras de se separar os dados, será necessário optar por uma delas[283]. Ademais, é necessário indicar exemplares de cada classe, outro processo que envolve optar entre diferentes possibilidades. Feito isso, surgem novos questionamentos. Imagine-se que, dado um caso concreto, um juiz humano utiliza uma ferramenta de inteligência artificial para encontrar julgados/precedentes pertinentes àquele assunto. Qual critério será utilizado para definir tal similaridade? Como garantir que a amostra representará fidedignamente a toda a população de processos similares? O que acontecerá com posicionamentos minoritários?

Não há uma resposta exata para todos esses questionamentos. Por conta disso, a abordagem do robô classificador é a que mais apresenta o risco de gerar ao humano uma falsa sensação de que todo o material relevante para um certo caso está à disposição. Em um tal cenário, não raro esta máquina estará essencialmente corroborando preconcepções (ocasionando um viés de confirmação nos julgadores humanos). Ademais, robôs-classificadores podem criar um ambiente de retroalimentação, em que sugestões levam a decisões, que também passarão a ser sugeridas para novos casos. A tendência, à primeira vista, é que posicionamentos divergentes da maioria se tornassem ainda mais minoritários.

Já no caso de um robô-relator, sua complexidade faz com que persistam potenciais problemas da abordagem anterior e traz

283. No aprendizado não supervisionado, no qual humanos não definem previamente as classes, o problema persiste, ainda que de outra forma, pois ainda será necessário interpretar os *clusters* estabelecidos pelo modelo e, possivelmente, ter de adaptá-los às classes humanas.

à tona novas preocupações. Isso porque, para além da classificação, robôs-relatores vão sugerir desfechos para um caso. Além disso, é de se esperar que esses modelos sejam utilizados em conjunto com outras ferramentas de inteligência artificial, visto que, se há tecnologia suficiente para implementá-los, também haverá para utilizar modelos menos complexos.

Ao passo que a assertividade do modelo aumenta, juízes humanos terão maiores chances de serem induzidos pela máquina e se tornarem um "canal de entrega" de decisões algorítmicas. Caminha-se, portanto, em direção a um ambiente em que o modelo recria as premissas das quais ele parte, uma vez que ele será alimentado por suas próprias sugestões. Mais uma vez, o risco é grande no sentido de perda de qualidade de decisões que se atenham a peculiaridades de casos concretos, ainda que, em um primeiro momento, o aumento da isonomia para com os jurisdicionados soe como uma vantagem.

Ademais, tais modelos são junções de vários algoritmos que desempenham diferentes funções (classificação, mineração de texto, extração de relações, etc.), o que significa ter de treinar cada um deles com diferentes tipos de dados e avaliá-los individualmente antes de implementá-los no modelo final. Por conta disso, escrutínios de seus resultados deverão ser feitos ainda mais cuidadosamente, pois em cada uma de suas "etapas" haverá a chance de serem utilizados dados enviesados e, com isso, afetar os resultados finais. Uma boa prática seria utilizar diferentes fontes de dados (e.g. decisões de diferentes tribunais) para treinar cada uma das "partes" do modelo, e assim evitar que uma única fonte "contaminada" influencie o todo.

Por fim, no caso de um robô-julgador, vez que o paradigma da corrobótica é superado e humanos e máquinas passam a trabalhar separadamente, problemas como a indução de resultados são reduzidos, quando não eliminados. Todavia, algumas ponderações são necessárias. Como dito, o uso de algoritmos tecnicamente mais complexos é quase uma garantia de que outros de menor robustez tecnológica também estejam em operação, uma vez que o desenvolvimento dessas ferramentas ocorre de forma gradual e o sucesso de uma abordagem serve de incentivo para o desenvolvimento de novas funcionalidades. Dessa forma, um cenário de separação total entre humano e máquina é bastante

improvável, no momento em que robôs-julgadores estiverem agindo. Por conta disso, é necessário que o juiz humano, atuando como instância revisora das decisões da máquina, não seja afetado por informações seletivas vindas de outros algoritmos ou, pelo menos, esteja ciente de que isso pode ocorrer.

Além disso, a questão da retroalimentação do modelo é potencializada pelo robô-julgador. Nas duas primeiras abordagens, ainda que o problema também exista, ele continua em uma escala humana, pois juízes de carne e osso terão de, ao menos, apertar um botão. No caso do robô-julgador, a mudança de paradigma implica igualmente uma mudança de proporções, já que milhares ou mesmo milhões de processos poderão ser julgados em frações de segundo. O problema ocorre se essa "jurisprudência algorítmica" passa a servir indiscriminadamente como fundamento para novas decisões automáticas. Isso porque, quanto mais dados apontarem em uma mesma direção, mais "certeza" terá o algoritmo de que suas "decisões" estão corretas. O que não se deseja é que tais direções sejam indicadas de forma artificial, ou seja, não reflitam a realidade humana – ou espelhem uma realidade criada pela própria máquina. Mostra-se imprescindível, portanto, que a automatização não ocorra de forma total. É necessário que decisões humanas continuem existindo, mesmo em matérias de baixa complexidade e grande número de processos, justamente para que a máquina tenha fontes fidedignas para ajustar seus parâmetros e, inclusive, adaptar-se às mudanças sociais.

É de se salientar, contudo, que o uso de ferramentas de inteligência artificial no Poder Judiciário, apesar de todas as suas limitações, promete trazer muito mais benefícios que malefícios. Ademais, conter seu avanço nos Tribunais não parece mais uma questão de escolha. Agora resta, especialmente aos juristas, capacitar-se e buscar entender tanto quanto possível sobre o modo de funcionamento dessas tecnologias. As estatísticas somente estarão do nosso lado se tivermos profissionais que consigam transitar com confiança entre dois mundos, quais sejam, o jurídico e o tecnológico. Queremos crer que esta obra tenha contribuído para introduzir as bases de uma discussão multidisciplinar e que, esperamos, esteja apenas no começo.

Conclusão

Buscou-se, com este livro, fornecer as bases de uma discussão multidisciplinar sobre o uso de ferramentas de aprendizado de máquina em atividades relacionadas à administração da Justiça. Para tanto, foram levados em conta os aspectos técnicos de tais ferramentas, assim como a própria natureza da atividade interpretativa da linguagem natural, tarefa esta que se mostra um grande desafio, tanto para humanos, quanto para máquinas.

Em primeiro lugar, pôde-se observar que o aprendizado de máquina apresenta características peculiares, se comparado a outros modelos estatísticos. Assim é que ele possui certa autonomia, pois prescinde de que humanos expliquem previamente como se dão as relações entre as variáveis que compõem o modelo, ao mesmo tempo que pode ajustar seus parâmetros de maneira autônoma para melhor descrever os dados que tem à sua disposição. Conjuntamente, tais características permitem que técnicas de *machine learning* executem tarefas cujas minúcias são enigmáticas mesmo para humanos.

Todavia, seu melhor desempenho é acompanhado de um custo interpretativo. Não é possível descrever passo a passo como tais algoritmos chegam a uma determinada decisão, aspecto que lhes confere a alcunha de "caixa-preta". Ainda assim, isso não significa que é impossível interpretá-lo, visto que, em vários casos, pode-se estimar quais variáveis tiveram maior peso para se chegar a um dado resultado.

No âmbito do Direito, o aprendizado de máquina é utilizado principalmente para ensinar computadores a "ler" textos escritos por humanos, o que configura o processamento de linguagem natural, um dos subcampos da inteligência artificial. Máquinas, contudo, não interpretam textos da mesma forma que humanos. Sua forma de assimilar a linguagem natural consiste em criar representações matemáticas de palavras, através das quais é possível

identificar relações semânticas e sintáticas entre termos e similaridades entre documentos.

Ainda assim, a leitura e a redação jurídicas compõem tarefas mais complexas que simplesmente extrair informações de textos longos ou formular frases gramaticalmente bem escritas. O raciocínio legal (*legal reasoning*) requer, por exemplo, fundamentações detalhadas acerca dos motivos que fizeram um argumento (e não outro) ser acatado, bem como levar em conta a hierarquia de leis e decisões judiciais prévias.

Por conta disso, torna-se imprescindível que máquinas possuam habilidades relativas à mineração de textos, tais como extrair informações de textos longos, encontrar respostas a perguntas específicas, minerar de argumentos, expandir conceitos, dentre outras. Por meio do aprendizado de máquina, tornou-se possível extrair tais informações de modo mais eficiente e aplicá-las a grandes quantidades de documentos, através de modelos que funcionam conjuntamente com a atuação humana, o que leva ao paradigma da corrobótica ou da computação cognitiva.

Ainda assim, a discussão sobre o uso do ML em Tribunais não se restringe somente aos seus aspectos técnicos. Defendeu-se que a forma como se pensa a linguagem impacta diretamente no modo de se conceber o fenômeno jurídico. Como exemplo, podem-se citar os casos de Hans Kelsen e Herbert Hart, que, a partir do pano de fundo das discussões, respectivamente, do Círculo de Viena e da filosofia da linguagem ordinária, propuseram diferentes considerações acerca do fenômeno jurídico. Ainda assim, ambos os jusfilósofos chegam a um ponto comum, ainda que por caminhos diferentes, a saber, que há casos em que a imprecisão da linguagem cria uma margem de discricionariedade àqueles que aplicam as normas. A inafastabilidade da discricionariedade pode ser entendida como uma consequência dos pressupostos da filosofia analítica, dentro dos quais Kelsen e Hart estão inseridos.

Uma vez que o fenômeno jurídico é encarado não como uma "ciência exata", mas como uma atividade sujeita à discricionariedade, surgem novos questionamentos sobre o uso do aprendizado de máquina e do processamento de linguagem natural. Em primeiro lugar, deve-se ter em mente que linguagens de programação operam em apenas dois níveis semióticos, isto é, sintático e semântico, enquanto linguagens naturais possuem, para além destes,

a dimensão pragmática. Ainda assim, o âmbito pragmático da linguagem ordinária não é completamente inacessível a algoritmos, uma vez que estes podem ser treinados de forma a levar em conta o contexto textual em que um termo ocorre, o que lhes confere um melhor desempenho, tanto linguístico, quanto computacional. Em segundo lugar, a existência de certo grau de discricionariedade na interpretação de termos, e, consequentemente, de normas jurídicas, levanta indagações sobre o modo como isso ocorre em Tribunais e de que forma isso interfere na elaboração de algoritmos. Nesse sentido, são valiosas as contribuições da psicologia cognitiva, que explica de que forma operam heurísticas e vieses em processos mentais decisórios. Grosso modo, pode-se dizer que o cérebro humano cria atalhos cognitivos, por meio dos quais é reduzido o esforço necessário para tomar decisões complexas, mecanismo bastante útil no dia a dia das pessoas. Todavia, em outras situações, tais como decisões judiciais, esse mecanismo pode levar à tomada de decisões subótimas, ao passo que muitas vezes ignora a quantidade e a qualidade das informações disponíveis.

Dado que os algoritmos do judiciário são treinados com base em decisões humanas, eles podem reproduzir ou, ainda, acentuar tais vieses. Isso se torna especialmente perigoso quando eles são vistos como ferramentas imparciais e revestidas de cientificidade. Para evitar que os mesmos se tornem "armas de destruição matemática", é necessário velar para que tais modelos sejam elaborados de forma transparente e conjunta com todos aqueles que serão afetados por suas decisões, bem como que seus resultados sejam cuidadosamente auditados.

Por fim, foram elencadas três abordagens mais recorrentes do uso do aprendizado de máquina no judiciário. A primeira delas corresponde ao uso robô-classificador, através do qual algoritmos auxiliam humanos em tarefas básicas, tais como encontrar e classificar processos. A abordagem do robô-parecerista consiste em condensar informações relevantes de um processo em um único documento, que eventualmente pode ser utilizado para sugerir decisões a um caso concreto. A terceira e última, robô-julgador, consiste em um passo adiante, já que seus resultados são considerados vinculativos e elimina-se do processo decisório completamente o componente humano, que se torna uma instância revisora. Entende-se que para gerar informação relevante para um decisor, o trabalho associado de máquinas e humanos é

a forma mais vitoriosa, possível tecnologicamente e, principalmente, democrática.

Antes de você acordar hoje os algoritmos do Google, Facebook, Amazon, Spotify, Netflix, do seu carro, enfim, estavam atualizando as informações sobre você, procurando melhorar a sua experiência como consumidor. Os algoritmos se valem de conjuntos de informações obtidos de você mesmo e do grupo de análise com o qual você foi identificado. O duplo movimento é significativo para acurácia do algoritmo. Nossa proposta foi a de demonstrar que isso pode ser implementado no campo do Poder Judiciário. O tempo dirá o nível do nosso otimismo.

Referências

ABIKO, Paula Yurie. Vieses da Justiça e Atuação Contraintuitiva. https://canalcienciascriminais.com.br/vieses-justica/.

ALCHOURRÓN, Carlos; BULYGIN, Eugenio. Análisis lógico y derecho. Madrid: Centro de Estudios Constitucionales, 1991.

ALETRAS, Nikolaos et al. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. Peerj Computer Science 2: Https://doi.org/10.7717/peerj-cs.93, 2016.

AMORIM, Fernanda Pacheco. Respeita as mina: inteligência artificial e violências contra a mulher Florianópolis: EMais, 2019.

ANDRADE, Vera Regina Pereira de. A ilusão da segurança jurídica: do controle da violência à violência do controle penal. Porto Alegre: Livraria do Advogado, 1997.

AROSO LINHARES, José Manuel. Entre a reescrita pós-moderna da modernidade e o tratamento narrativo da diferença ou a prova como um exercício de passagem nos limites da juridicidade. Coimbra: Editora Coimbra, 2001, p. 809-810.

ASHLEY, Kevin D.. Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. University Of Pittsburgh School Of Law: Cambridge University Press, 2017. (ISBN: 9781107171503).

ATIENZA, Manuel; MANERO, Juan Ruiz. Ilícitos atípicos. Madrid: Trotta, 2000, p. 125.

BARATTA, Alessandro. Criminologia crítica e crítica do Direito Penal: introdução à sociologia do direito penal. Trad. Juarez Cirino dos Santos. Rio de Janeiro: Freitas Bastos, 1999.

BARILI, Raphael Jorge de Castilho. Teoria do Caso e sua aplicabilidade ao Processo Penal Brasileiro. Curitiba: CRV, 2019.

BATISTA, Nilo. Punidos e mal pagos. Rio de Janeiro: Revan, 1990.

BAYÓN MOHÍNO, J. C. ¿Por qué es derrotable el razonamiento jurídico? Doxa. Cuadernos de Filosofía del Derecho, n. 24, p. 35 – 62, 2001.

BELLONI, Massimo. Neural Networks and Philosophy of Language. 2019. Disponível em: <https://towardsdatascience.com/neural-networks-and-philosophy-of-language-31c34c0796da>. Acesso em: 16 nov. 2019.

BENAVENTE CHORRES, Hesbert. La construcción de los interrogatórios