UNIVERSIDADE ESTÁCIO DE SÁ

CAROLINA HENRIQUE DA COSTA BRAGA

DECISÕES AUTOMATIZADAS E DISCRIMINAÇÃO: PESQUISA DE PROPOSTAS ÉTICAS E REGULATÓRIAS NO POLICIAMENTO PREDETIVO

Rio de Janeiro
2019

CAROLINA HENRIQUE DA COSTA BRAGA

DECISÕES AUTOMATIZADAS E DISCRIMINAÇÃO: PESQUISA DE PROPOSTAS ÉTICAS E REGULATÓRIAS NO POLICIAMENTO PREDETIVO

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo programa de Pós-Graduação em Princípios Fundamentais e Novos Direitos da Universidade Estácio de Sá - UNESA.

Orientador: Prof.ºNilton César da Silva Flores

Rio de Janeiro
2019

CAROLINA HENRIQUE DA COSTA BRAGA

DECISÕES AUTOMATIZADAS E DISCRIMINAÇÃO: PESQUISA DE PROPOSTAS ÉTICAS E REGULATÓRIAS NO POLICIAMENTO PREDETIVO

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo programa de Pós-Graduação em Princípios Fundamentais e Novos Direitos da Universidade Estácio de Sá - UNESA.

Aprovado em _____ de _____ de 2019.

Banca Examinadora

_____

Prof. Drº Nilton César Flores

_____

Prof. Drº Vinicius Chaves

_____

Prof.Drª Caitlin Mulholland

Dedico esta dissertação ao Dennys, por tudo, sempre.

AGRADECIMENTOS

Um artigo nunca é escrito sozinho, este fato é ainda mais verdadeiro quando se trata de uma dissertação. Embora existam momentos de solidão durante a pesquisa, de estudo e de reflexão, é apenas com a contribuição de diversas pessoas que um texto chega ao sua real potencial.

Dessa forma o primeiro agradecimento vai para meu marido, Dennys, que tem estado ao meu lado em todos os momentos, dando sentido às minhas conquistas. Agradeço também pelas eternas discussões e debates que mudaram minhas opiniões e moldaram esta pesquisa.

Agradeço a minha família, que foram os primeiros a ler esta dissertação, revisar e comentar. Se eu pude tornar este texto o melhor possível foi devido à eles. Mas, mais importante, gostaria de agradecer pelo constante apoio, não apenas nessa pesquisa, mas em todos os momentos.

Agradeço ao meu orientador Nilton César Flores, por aceitar minha proposta de pesquisar sobre inteligência artificial quando mais ninguém no curso de mestrado pesquisava sobre o assunto, e por me incentivar a apoiar durante toda a pesquisa e redação desta dissertação.

Por fim, mas não menos importante, agradeço aos membros do grupo de pesquisa Droit, em especial à Caitlin, pelos debates e discussões, pois sem eles esta dissertação não existiria.

“I`m not worried about artificial intelligence giving computers the ability to think like humans. I`m more concerned with people thinking like computers without values or compassion.”

(Tim Cook – Apple Ceo)

RESUMO

A inteligência artificial (IA) e o aprendizado de máquina estão revolucionando indústrias e sociedades em geral. Suas capacidades estão incorporadas em nossas vidas diárias e suas aplicações estão se expandindo. Os resultados de suas decisões agora afetam vidas e, com isso, surge um risco significativo. As decisões automatizadas já vem sendo aplicadas não apenas em setores como publicidade e marketing, mas também na áreaa judicial e policial. Sua utilização com o fim de melhorar a atuação policial, denominado de policia preditiva, desperta especial preocupação devido aos direitos e liberdades que estão em jogo. Para que essas tecnologias continuem a ter um impacto positivo na sociedade, elas precisam respeitar a ética e os princípios fundamentais. Considerando que a tecnologia é tão boa quanto o seu criador, todos que trabalham no campo são responsáveis por criar uma estrutura ética para a inteligência artificial que reflita nossos valores. Esta dissertação se propõem a apontar quais são esses riscos gerados pela AI e investigar as abordagens éticas e regulatórias adequadas para tratá-los.

Palavras-Chave:Inteligência Artificial;Decisões Automatizadas;Algoritmos;Big Data; Discriminação; Polícia Predetiva; Ética; Regulação.

ABSTRACT

Artificial intelligence (AI) and machine learning are revolutionizing industries and societies in general. Their capabilities are embodied in our daily lives and their applications are expanding. The results of their decisions now affect lives, and with it, a significant risk arises. Automated decisions have already been applied not only in sectors such as advertising and marketing, but also in the judicial and police areas. Its use in order to improve police performance, called the predictive police, arouses particular concern due to the rights and freedoms that are at stake. For these technologies to continue to have a positive impact on society, they must respect ethics and fundamental principles. Since technology is as good as its creator, everyone working in the field is responsible for creating an ethical framework for artificial intelligence that reflects our values. This dissertation proposes to point out those risks generated by AI and investigate the appropriate ethical and regulatory approaches to address them.

Key-Words: Artificial Intelligence; Automated Decisions; Algorithms; Big Data; Discrimination; Predictive Policing; Ethics; Regulation.

SUMÁRIO

INTRODUÇÃO

Os algoritmos de hoje decidem muitas coisas sobre nós como quem é contratado, quem é demitido, quem recebe uma hipoteca e quem é um criminoso perigoso, o que pode tornar o que pode ser uma decisão difícil, simples. Dessa forma, defende-se que a objetividade relativa da inteligência artificial pode ajudar a neutralizar a subjetividade humana, o que de fato pode ocorrer, e vem sendo aplicado com sucesso para domínios como ciência, saúde, tecnologia e finanças.

No entanto, nos últimos anos pesquisadores vem apontando para a existência de riscos significativos decorrentes da utilização de sistemas de deicsões automatizadas, em especial quando estas envolvem direitos fundamentais humanos, como a liberdade, que devem ser mitigados antes do consumo público.

Desde tenra idade, pessoas, cultura, educação, religião, mídia e política moldam nossas perspectivas e crenças sobre o mundo. Essas experiências geram informações que nossas mentes conectam e categorizam para dar sentido a tudo isso. Essas categorias às vezes formam estereótipos, que desencadeiam associações cognitivas sobre características como gênero, idade e raça que não refletem a realidade. Isso tem impactos prejudiciais na sociedade. O problema é que este processo é imediato, automático e muitas vezes não é algo que estamos cientes de que está acontecendo. A maioria contradizia um pensamento discriminatório que entra na consciência, mas dado que o comportamento é em grande parte determinado pela mente inconsciente, as intenções positivas não são suficientes.

Considerando que sistemas de decisão automatizadas não são verdadeiramente neutros, refletindo os preconceitos e viéses humanos deve-se proceder com cautela na sua implementação de forma a garantir que a IA não cause danos. Não podemos terceirizar nossas responsabilidades éticas para as máquinas. Precisamos proteger nossos padrões éticos e incorporá-los em uma estrutura que orienta o projeto, a implementação e a utilidade da IA na sociedade.

O tema desta dissertação, portanto, busca apontar os riscos ocasionados pela utilização de sistemas de decisão automatizadas e investigar propostas éticas e regulatórias que possam gerar uma aplicação dessa tecnologia de forma compatível com o respeito a direitos e liberdades humanas.

Em outras palavras, visa-se verificar, mecanismos para regular a utilização da inteligência artificial de modo a evitar os danos gerados à sociedade e, especialmente, à grupos considerados como vulneráveis. Dessa forma, o tema trata direitamente da violação

aos princípios fundamentais da equidade e da privacidade, previstos no art.5º, caput e incisos X da Constituição da República respectivamente.

Considerando, no entanto, que a inteligência artificial tem sido aplicada nas mais diversas funções e setores o tema, conforme exposto até o momento, deve ser delimitado. Diante do extenso número de possibilidades de aplicação de sistemas de decisão automatizada, a fim de possibilitar uma análise mais aprofundada das suas dificuldades e riscos, bem como possibilidades regulatórias, optou-se por estudar a implementação dessa tecnologia pela polícia, o que vem sendo denominado de policiamento preditivo.

Dessa forma a pesquisa tem como objetivo geral, a análise das possíveis formas de regulação da inteligência artificial a fim de evitar a violação a direitos fundamentais dos seres humanos. Para tanto pretende-se no primeiro capítulo apontar o que se entende por inteligência artificial, passando por uma breve análise histórica da evolução da inteligência artificial desde os seus primórdios até um possível futuro para o qual estamos nos dirigindo, para, após, apontar os benefícios e riscos que esta pode ocasionar.

No segundo capítulo pretende-se delimitar a análise para apontar o que se entende por policiamento preditivo, as formas pelas quais esta atividade ocorre e como efetivamente é gerada a discriminação pelos sistemas de decisão automatizadas.

Após delimitar o problema: que decisões automatizadas efetivamente geram tratamentos discriminatórios que refletem preconceitos humanos e que mantêm, ou até amplificam, relações de marginalização de grupos vulneráveis, os capítulos seguintes visam enfrentá-lo.

Nesse espírito o terceiro capítulo visa investigar quais seriam as abordagens éticas que vem sendo utilizadas para tratar tais efeitos discriminatórios. Em um primeiro momento vai-se verificar qual a diretriz ética mais adequada para tratar desse problema. Após aponta-se brevemente as tentativas de introduzir parâmetros éticos nos sistemas de inteligência artificial e, por fim, analisa-se os princípios éticos que vem sendo desenvolvidos por organizações de diversos setores e países.

O quarto e último capítulo, partindo desses princípios éticos, verifica como os Estados Unidos, a Europa e a China, respectivamente, vem tratando, ou deixando de tratar, a regulação da inteligência artificial. A autora tem conhecimento de que existem diversos outros países com propostas regulatórias, como, por exemplo, o Japão. No entanto, na presente pesquisa optou-se por estudar esses dois países, e continente, no caso da Europa, por se entender que representam visões emblemáticas de regulação para a inteligência artificial.

Este capítulo também aponta quatro normas regulatórias que visam trazer uma primeira organização para o tratamento e responsabilização dos danos ocasionados pelas decisões automatizadas. O capítulo termina por levantar a necessidade de maiores estudos sobre a discriminação derivada das decisões automatizadas e sugere a utilização de instrumentos de análise de impacto para obter tais informações para que seja possível, futuramente, criar regulações mais adequadas à implementação da inteligência artificial.

1. INTELIGÊNCIA ARTIFICIAL E DECISÕES AUTÔNOMAS POR ALGORITMOS

A pesquisa em inteligência artificial é um campo da ciência da computação que vem se expandindo nos últimos anos. Trata-se, conforme definição de McCarthy (2007, p. 2), de "a ciência e engenharia de fazer máquinas inteligentes, especialmente programas de computador inteligentes", e tem como objetivo elaborar máquinas que simulem o comportamento humano.

Quando Alan Turing disse acreditar que chegaríamos ao novo milênio falando sobre 'máquinas pensantes' com naturalidade, ele não poderia estar mais certo. No entanto, encontrar uma definição precisa para a inteligência artificial é um grande desafio. No fundo, algoritmos de *machine learning* podem, grosseiramente, ser considerados ferramentas capazes de aumentar sua precisão, ou seja, "de aprender", a partir da utilização de grande volume de dados (*input*), fornecendo, em troca, algum tipo de resposta otimizada (*output*), como, por exemplo, *rankings*, avaliações ou diagnósticos, de acordo com o procedimento pré-programado.

Tecnologias que se utilizam de inteligência artificial estão em franca expansão, sendo adotadas mundialmente. Objetos e aplicativos utilizados cotidianamente como celulares Iphone e assistentes pessoais como Alexa e Siri contêm sistemas de IA, que nos permite fazer predições de modo a personalizar a experiência de cada usuário. Além da esfera pessoal, esses sistemas vêm sendo aplicados nas mais diversas áreas como agricultura, indústria, judiciário,

polícia, e finanças, onde eles são utilizados para prever desde nossos gostos musicais até a probabilidade de cometermos um crime.

1.1. EVOLUÇÃO HISTÓRICA DA INTELIGÊNCIA ARTIFICIAL

Apesar desse rápido crescimento, o conceito de inteligência artificial existe há mais de sessenta anos. Este campo de pesquisa nasceu oficialmente em 1956 em um seminário organizado por John McCarthy no *Dartmouth Summer Research Project on Artificial Intelligence*[1]. O objetivo, então, era investigar as formas como uma maquina poderia simular aspectos da inteligência humana, ideal que continua a motivar esse campo de pesquisa até os dias atuais.

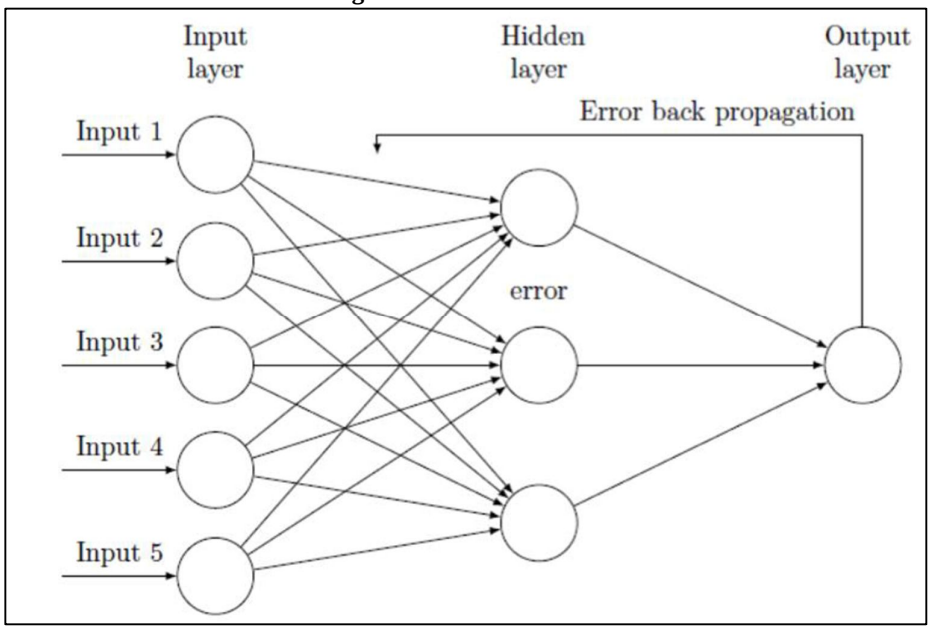
As ideias mais influentes em se tratando de ciência computacional advêm de Alan Turing, conhecido como pai da computação. Seu texto mais famoso, denominado *Computing Machinery and Intelligence*, indaga sobre a possibilidade de computadores que criados para simular inteligência, além de explorar diversos ingredientes que atualmente são associados com a inteligência artificial, incluindo como a inteligência[2] pode ser testada e como as maquinas podem aprender de forma automática.

Nos anos entre 1950 e 1970 surgiram diversos estudos focados em IA de modo que tal período ficou conhecido como a primavera da inteligência artificial. Allen Newell, John C. Shaw, e Herbert A. Simon foram pioneiros em pesquisa heurística, que consiste em um processo eficiente para encontrar soluções em amplos espaços combinatórios. Nesse mesmo período, surgiram os primeiros trabalhos na área da computação capazes de reconhecer a

[1]Apesar desse *workshop* ter criado uma identidade unificada para a área, além de uma comunidade de pesquisadores, muitas das ideias que viriam a caracterizar a inteligência artificial já existiam. Ver Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, OrenEtzioni, Greg Hager, Julia Hirschberg, ShivaramKalyanakrishnan, EceKamar, SaritKraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLeeSaxenian, Julie Shah, MilindTambe, and Astro Teller. "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report. Acessado em 22/04/2018.
[2]Conhecido como o famoso teste de Turing.

feição de pessoas, dando base para aplicações mais complexas como o reconhecimento facial. No final dos anos sessenta surgiram estudos focados no processamento de linguagem, na mobilidade de robôs e em sistemas de aprendizagem de máquinas[3].

Figura 1 – Redes Neurais



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

A primeira forma desenvolvida pelos pesquisadores da inteligência artificial, denominada de Redes neurais (com *back propagation*) (1969) trata-se de uma forma de aprendizado baseada em erros e acertos, com identificação paulatina dos caminhos e decisões mais corretas para atingir determinados objetivos.

Nesse caso, como demonstrado pelo desenho acima, há um objetivo (*output*), e vários *inputs*. Os *inputs* são testados em vários caminhos. Quando se chega ao resultado, o caminho mais assertivo recebe um peso maior (na conta matemática), ou seja, as camadas neurais

[3]"Shakey", a wheeled robot built at SRI International, launched the field of mobile robotics. Samuel's Checkers-playing program, which improved itself through self-play, was one of the first working instances of a machine learning system."Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, ShivaramKalyanakrishnan, EceKamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLeeSaxenian, Julie Shah, MilindTambe, and Astro Teller. "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report. Acessadoem 22/04/2018.

internas (*hidden layers*) passam a dominar a tarefa e a entregar resultados mais precisos na medida em que o algoritmo confere um peso maior às conexões que apresentam resultados mais próximos aos desejados[4].

Os pesos algorítmicos para os caminhos neurais do sistema computacional que obtiveram os resultados mais próximos do desejado equivalem à dopamina do cérebro humano quando percebemos que conseguimos cumprir determinadas tarefas difíceis[5].

Apesar desses significativos avanços iniciais, em 1980 os pesquisadores de IA ainda não haviam conseguido obter nenhum sucesso prático significativo. Essa distância entre teoria e prática levou à crença de que as máquinas nunca poderiam apreender a linguagem humana e tal descrença, junto com a movimentação para setores de pesquisa da neurociência, levou o campo de pesquisa a uma parada abrupta já que não havia mais interesse pela inteligência artificial e tampouco investimentos para pesquisa. Esse período ficou conhecido como inverno da inteligência artificial.

No entanto, no final dos anos 80 o entusiasmo pela inteligência artificial ressurgiu e os estudos no campo se expandiram de forma que esse novo período foi denominado de primavera. Dois fatores foram essenciais para tal desenvolvimento: *Big Data* e *Machine learning*.

1.1.1 *Big Data*

O mundo gera, diariamente, 2,5 quintilhões de *bytes* sendo que, atualmente, a cada um ano e meio se gera a mesma quantidade de informação já criada pela humanidade desde o seu surgimento. Esse volume incomensurável de dados é denominado de *Big Data*.

---

[4]RUMERLHART, David E; HILTON, Geoffrey E e WILLINANS, Ronald J. Learning Representations by back-propagating erros.In Nature, vol 323, issue 9, outubro de 1986, p. 533.

[5]Erik, citando ITO, Joi e HOWE, Jeff. Whiplash: how to survive our faster future. 2016, New York and Boston, Grand Central Publishing,p. 240-241.

Tendo surgido em meados da década de 90 e cunhado pela *National Aeronautics and Space Administration* (NASA), descreve uma quantidade massiva de dados disponíveis, além das capacidades computacionais e analíticas por processos tradicionais. Através da análise e mineração desses dados se torna possível agregar informações de origens diversas de forma a relacioná-las e gerar conclusões, auxiliando de forma cada vez mais eficiente na tomada de decisões.

Entre as diversas fontes de coleta de dados pode-se apontar:

- *Data Exhaust*: dados coletados de forma passiva de transações feitas por pessoas ao utilizarem serviços digitais como celulares, compras online, buscas na internet, entre outros;
- Informação online: conteúdo da internet como sites de notícia ou interações nas redes sociais;
- Sensores físicos: imagens de satélites ou câmeras de vigilância de paisagens, tráfego, desenvolvimento urbano entre outros;
- Informações de cidadãos ou *Crowd-sourced*[6]: informação produzida ativamente ou submetida pelos cidadãos através de celulares (principalmente) como questionários, *hotlines*, etc.

Em meados dos anos 2000 o *Big Data* ganhou a mídia através da definição cunhada por Doug Laney, membro da Gartner, de seus 3 Vs: Volume, dados massivos de inúmeras fontes; Velocidade, capacidade e tempo de processamento sustentável ao objetivo e a necessidade de processamento em tempo real para diversos segmentos; e Variedade, diversas fontes como vídeos, fotos, *hashs*, transações financeiras, etc.

[6]A palavra "*crowdsourcing*"se refere ao uso de atores não oficiais ("*the crowd*") como uma fonte de informação, conhecimento e serviços, em oposição à prática comercial de "*outsourcing*".

A empresa *Statistical Analysis System*(SAS), que possui uma família de *softwares*, utilizam mais duas dimensões para descrever o *Big Data*, além dos 3Vs de Laney, que são a Variabilidade, inconsistência de fluxos de dados variados com picos sazonais, e a Complexidade, a qual dispensa definição. Majoritariamente, porém, passou-se a ser adotado como definição os 5 Vs: Volume, Variedade, Velocidade, Veracidade e Valor.

No entanto, apenas recentemente o *Big Data* efetivamente caiu nas graças da sociedade e das empresas, passando a ser denominado como "o novo petróleo"[7]. O setor privado não demorou a descobrir diversas utilidades para a mineração massiva de dados, em especial na capacidade de gerar lucro diante de um tratamento mais individualizado ao seu consumidor. Conforme demonstrado pela pesquisa realizada pela consultoria em negócios norte-americana Bain&Company, as empresas que utilizam *Big Data* possuem 5 vezes mais chances de tomarem decisões rápidas do que seus concorrentes e 2 vezes mais chances de obterem performance superior.

Isso ocorre devido à prática da análise dos dados obtidos, que trata-se de utilizar algoritmos, dentro de condições pré-estabelecidas, para analisar grandes volumes de dados tornando-os inteligíveis de forma a determinar quais são as ações necessárias para tomar uma determinada decisão. Trata-se, portanto, de levar o fenômeno do *Big Data* ao seu máximo potencial.

1.1.2. *Machine learning*

Machine learning, por sua vez, existe desde a década de 70, mas apenas recentemente ganha força em razão da explosão de dados. Por definição, trata-se de qualquer metodologia

[7]Big Data é o novo petróleo, afirma executiva da IBM" acessado em https://olhardigital.com.br/noticia/big-data-e-o-novo-petroleo,-afirma-executiva-da-ibm/34986 e "The world'smostvaluableresourceis no longeroil, but data" acessado em https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource.

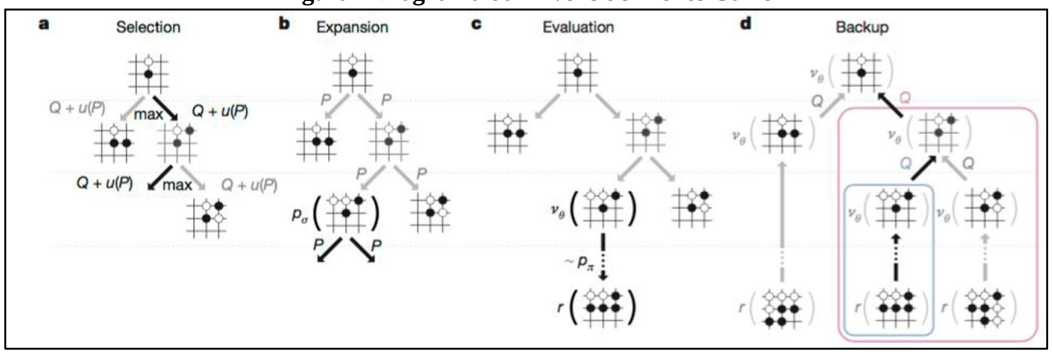
ou conjunto de técnicas que utilizam dados coletados para criar novos padrões e conhecimento gerando modelos que podem ser utilizados para realizar predições sobre os dados. Trata-se da capacidade de alterar as regras de forma autônoma. De forma simples, pode-se dizer que essa técnica coloca novos inputs no modelo de treinamento, a partir daqueles inseridos pelos humanos. O aprendizado pode ser:

- Supervisionado: Os modelos são ensinados ao algoritmo por inputs acrescidos por humanos.O caso mais conhecido é o *Deep Blue*. Trata-se de um algoritmo treinado para jogar xadrez que ficou famoso por derrotar o melhor xadrezista humano da época, Hugo Kasparov. Para tanto, o computador foi carregado manualmente com milhares de jogadas e combinações possíveis, usando sua enorme capacidade de processamento para escolher a melhor jogada.
- Não supervisionado: O próprio algoritmo decide quais os melhores modelos para serem aplicados a determinados *inputs*.Essa técnica é a utilizada pelo algoritmo do Alpha Go, criado para jogar o jogo chinês go. Devido às peculiaridades desse passatempo era impossível a análise ou escolha da melhor jogada a partir de uma prévia programação, visto que ele permite um total de $2{,}1x10^{170}$ posições possíveis. Sendo assim, usando redes neurais convolucionais e a chamada "Árvore de Monte Carlo"[8], conforme mostra a

[8]Monte Carlo TreeSearch – MCTS "é a arquitetura responsável pelo aprendizado das simulações dentro das redes neurais. Cada simulação (de uma jogada/posição) atravessa a árvore, permitindo que se selecione a escolha final que tenha o maior valor (Q) somado a um bônus (u) multiplicando esse resultado pela probabilidade de aquela posição acontecer em um jogo real. As escolhas finais (binárias, do tipo esquerda/direita) são sempre feitas para prestigiar o maior valor encontrado. Por fim, o valor final dessa ação (Q) é incorporado para definir o valor médiodessajogada, de acordo com o valor dado pela jogada em si, medido pela rede neural de valoração (v) e pelo resultado que ela proporcionaria no final do jogo (r) se implementada (rolloutpolicy). Em resumo, o sistema escolhe uma jogada, atribui-lhe um valor, simula a sua influência em um jogo completo, verifica o resultado e atualiza esse valor."disponível em https://jeffbradberry.com/posts/2015/09/intro-to-monte-carlo-tree-search/.

Figura 2, o Alpha Go utilizou uma combinação de *reinforced* e *supervisedlearning*paraderrotar os jogadores humanos.

Figura2–Diagrama da Árvore de Monte Carlo



Depois que o sistema aprende a classificar e valorar essas posições, ele passa (*policy gradient*) para uma fase mais avançada de aprendizado, não supervisionada (*reinforcedlearning* – RL), na qual o algoritmo participa sozinho de múltiplos jogos simulados aleatórios e vai aprendendo a fazer as melhores escolhas (*Reinforced Learning policy network*) e a valorá-las de modo preciso (*value network*). Dessa forma, o Alpha GO avalia muito menos posições a cada jogada que o *Deep Blue*, mas o faz de forma precisa e inteligente, selecionando e valorando suas escolhas de forma muito mais eficiente.

Dessa forma resta clara a diferença entre o aprendizado supervisionado e o não supervisionado. No estudo supervisionado é o humano quem escolhe a informação que será observada pela máquina para que ela aprenda, controlando esse processo de implementaçãorápida de jogadas e posições protagonizadas por grandes jogadores de go.

Depois que o sistema aprende a classificar e a valorar essas posições, ele passa para uma fase mais avançada de aprendizado, não supervisionada na qual o algoritmo participa sozinho de múltiplos jogos simulados aleatórios e vai aprendendo a fazer as melhores escolhas e a valorá-las de modo preciso.

Nessa segunda fase o algoritmo produz novas regras de decisão para resolver novos *inputs*. Dessa forma, o operador humano não precisa entender a lógica das regras para o algoritmo funcionar.

Isso torna as tarefas realizadas pelos algoritmos difíceis de serem previstas ou de serem explicadas, o que gera problemas caso ocorra algum dano e haja necessidade de responsabilização.

1.2. ASPECTOS POSITIVOS DAS DECISÕES AUTÔNOMAS

Considerando que dados estão sendo gerados pelas pessoas em tempo real, torna-se possível, como já percebido pelo setor privado, a sua análise em tempo real por computadores de alta performance criando, como dito anteriormente, um enorme potencial de auxílio na toma de decisões. Se as possibilidades para as empresas privadas já são inúmeras, esse potencial aumenta quando se trata do setor público.Quando se trata de políticas públicas para resolver os diversos problemas das cidades o *Big Data* surge como uma alternativa aos recursos tradicionais. Como dito por Emmanuel Letouzé, economista de desenvolvimento:

> É tempo para a comunidade que trata sobre desenvolvimento e aqueles que fazem políticas ao redor do mundo reconhecer e tomar essa oportunidade histórica para responder a desafios do século vinte e um, incluindo os efeitos da volatilidade global, mudanças climáticas e movimentos demográficos, com ferramentas do século vinte e um[9].

De fato, há uma sensação generalizada de que o mundo se tornou mais volátil nos últimos anos e quem sofre são os mais vulneráveis. As crises na economia, não apenas no Brasil, e a crise de representatividade gerados, entre outros, pela falta de empregos, alta no preço de alimentos e da gasolina, se alastraram pelo mundo devido, em parte, pela

[9]No original "It is time for the development community and policymakers around the world to recognise and seize this historical opportunity to address twenty-first century challenges, including the effects of global volatility, climate change, and demographic shifts, with twenty-first century tools." LETOUZÉ, Emmanuel, Global Pulse, "Big Data for Development: Challenges & Opportunities". Disponível em: http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

interconectividade global. Problemas locais geram impactos muito mais longe do que ocorria anteriormente.

Para solucionar essas crises, governos locais não podem mais tratar apenas dos resultados sem se preocupar com as causas. Prevenir os danos ou mantê-los em um nível mínimo é muito mais barato ao longo prazo. Para tanto, os meios tradicionais de coleta e utilização de dados, embora tenham sua importância e forneçam importantes informações, são muito custosos e incipientes, não conseguindo prever soluções adequadas, pois ainda que a causa seja conhecida não se dispõem de meios de saber ou distinguir quais os grupos afetados, onde, quando e o grau do dano causado.

Enquanto o setor público ainda lida com esses dados tradicionais, o setor privado vem continuamente sendo bem sucedido na utilização do *Big Data* na tomada de decisões[10]. No entanto, o setor público e a comunidade internacional vêm acordando para as possibilidades de utilização do *Big Data* para fins de auxílio na tomada de decisões em políticas públicas, além de diversos fins humanitários.

> É o duplo reconhecimento da promessa da revolução de dados e a necessidade por informações melhores e mais rápidas na era de aumento da volatilidade mundial que levou os líderes do G20 e a Secretaria-geral das Nações Unidas a requerer pelo estabelecimento da iniciativa Global Pulse (respondendo à crise econômica global em andamento), com o objetivo de desenvolver uma nova visão para monitoramente de impactos sociais e análise de comportamento através da construção de novas fontes de dados e novas ferramentas de análise[11].

Através de poderosos algoritmos os dados são extraídos das mais diversas fontes e analisados de forma a se encontrar correlações, padrões. Uma vez treinados (com os dados fornecidos), os algoritmos podem ser utilizados no auxílio de predições que podem encontrar padrões ou ainda anomalias, divergentes das tendências esperadas.

---

[10]Tem-se como exemplo a pesquisa feita pelo professor do MIT Erik Brynjolfsson que constatou diferenças significativas na utilização de dados—uma diferença de 5% na produtividade, considerada como uma vantage decisiva—percebida por empresas que utilizam "data-drivendecision-makingprocesses"em face daquelas que continuam a se utilizar principalmente em experiência e intuição. Traduçãolivrepelaautora.Lohr, Steve. "When There's No Such Thing as Too Much Information."The New York Times. 23 Apr. 2011.

[11]No original "It is the double recognition of the promise of the data revolution and the need for better, faster information in an age of growing global volatility that led the leaders of the G20 and the UN Secretary-General to call for the establishment of the Global Pulse initiative (in the wake of the on-going Global Economic Crisis), with the aim of developing of a new approach to "social impact monitoring" and behavioural analysis by building on new sources of data and new analytical tools."Big Data for Development: Opportunities and Challenges.UN Global Pulse's research. Disponível em https://pt.slideshare.net/unglobalpulse/big-datafordevelopment-globalpulsemay2012-13137063.

No entanto, o tipo de análise utilizado varia dependendo do objetivo que se pretende alcançar. Ela pode ser classificada em Prescritiva, Diagnóstica, Descritiva e Preditiva. Como se poderia esperar, a análise preditiva é a que mais chama a atenção do setor público diante da possibilidade de se utilizar da mineração de dados históricos para traçar tendências ou possibilidades futuras.

Um exemplo dessa utilização é o *Google FluTrends*, lançado em 2008, que se baseia na análise de pesquisas no Google sobre sintomas da gripe. Já há pesquisas indicando que "porque a frequência relativa de certas consultas é altamente correlacionada com a porcentagem de consultas médicas em que um paciente apresenta sintomas semelhantes aos da influenza",seria possível "estimar com precisão o nível atual de atividade semanal de influenza em cada região dos Estados Unidos, com um atraso de cerca de um dia"[12].

Sendo assim, seria possível a utilização dos dados de pesquisas no Google para detectar epidemias de influenza em áreas com uma alta quantidade de pessoas com acesso à internet. De forma geral, os dados online vem sendo utilizados como parte de vigilância de síndromes de forma a prevenir e combater surtos antes que afetem um grande número de pessoas. O *Google Dengue Trends* funcionava exatamente dessa forma[13].

Outro caso de utilização do *Big Data* de forma preditiva foi a captação de informações dadas pelas pessoas após o terremoto no Haiti. Foi instalado um sistema centralizado que permitia que se mandassem mensagens de texto com informações de pessoas presas em

[12]No original "because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms (…) to accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day." Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. "Detecting Influenza Epidemics Using Search Engine Query Data." Nature 457.7232(2008): 1012-1014.
Disponívelem:http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/papers/detecting-influenza-epidemics.pdf

[13]8https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_&hl=en&dl=en#!ctype=l&strail=false&bcs=d&nselm=h&met_y=flu_index&scale_y=lin&ind_y=false&rdim=country&idim=country:BR&ifdim=country&hl=en_US&dl=en&ind=false

prédios afetados. De acordo com Patrick Meier, da Ushahidi[14], esses resultados evidenciam a capacidade do sistema de prever, com surpreendente alta precisão e significância estatística, a localização e a extensão dos danos estruturais após o terremoto.

Sendo assim, percebe-se que o *Big Data* tem o potencial de diminuir as falhas humanas e a demora na produção, coleta e análise das informações, de forma a otimizar a tomada de decisões quando se trata de políticas públicas. Ainda sim, deve-se ter ciência de que a utilização do *Big Data* e dos algoritmos para sua análise, está longe de ser a solução para todos os problemas da humanidade. Ao contrário, eles próprios apresentam diversas dificuldades na sua utilização, além de colocarem em risco diversos direitos dos cidadãos, como a privacidade e a igualdade.

1.3 RISCOS NA UTILIZAÇÃO DA INTELIGÊNCIA ARTIFICIAL

Na sociedade atual decisões estão cada vez mais sendo deixadas a cargo de algoritmos. Há a promessa de que a mineração de dados vá auxiliar no entendimento da grande quantidade de dados que está surgindo com a internet das coisas. Além disso, algoritmos que se utilizam de *machine learning* podem automaticamente detectar distorções ou informações inexatas. Mas ao mesmo tempo não há garantia de que se trata de um comportamento eticamente aceitável.No entanto, antes de adentrar na questão ética em si, que será devidamente abordada no terceiro capítulo, é necessário analisar quais os malefícios que podem ser ocasionados pela aplicação de algoritmos dotados de inteligência artificial, especialmente *machine learning*.

[14]Ushahidi é uma companhia sem fins lucrativos que foi desenvolvida para mapear informações de violência no Kenya após as eleições de 2007. Ela se especializou em desenvolver softwares abertos egratuitos para a coleta de informações, vizualização e mapeamento interativo. Disponível em https://www.ushahidi.com/

Ao mesmo tempo em que traz benefícios, o uso de algoritmos apresenta riscos não evidentes, derivados especialmente: (i) de *data sets* viciados; (ii) da opacidade na sua forma de atuação, consequência das técnicas de *machine* e *deep learning*; (iii) da possibilidade de promoverem a discriminação, ainda que bem estruturados.

Diante da comprovação da existência de riscos graves, é urgente desenvolver mecanismos de governança de algoritmos a partir da colaboração entre juristas, filósofos, cientistas políticos e cientistas da computação. Essa situação parece ser ainda mais premente quando se imagina o seu emprego em substituição ou auxílio na tomada de decisões de questões sensíveis, em especial na esfera da justiça e da segurança pública.

Com base na tabela criada por Brent Daniel Mittelstadt, Luciano Floridi e outros[15] pode-se perceber seis principais riscos gerados por decisões autônomas:

- Evidência inconclusiva: As conclusões geradas por algoritmos retiradas dos dados processados são apenas prováveis, ou seja, desprovidas de qualquer certeza. Não são, portanto, suficientes para gerar uma conexão causal;
- Evidência inescrutável: Os algoritmos utilizados, assim como os dados, não são acessíveis, não podendo ser examinados, entendidos ou criticados;
- Evidência má orientada – *garbage in*, *garbage out* – o *output* não pode exceder o *input*. Em outras palavras, as conclusões alcanças serão apenas tão boas quanto os dados utilizados;
- Consequências injustas: Trata da ação advinda das decisões dos algoritmos. Uma ação pode ser discriminatória se tiver efeitos em um grupo protegido,

[15]Brent Daniel Mittelstadt, Patrick Allo, MariarosariaTaddeo, Sandra Wachter, Luciano Floridi. "The ethics of algorithms: Mapping the debate". First Published December 1, 2016 Research Article.Disponível em https://doi.org/10.1177/2053951716679679

ainda que a evidência gerada pelo algoritmo seja conclusiva, compreensível e bem orientada;

- Efeitos transformativos: Algumas decisões tomadas por algoritmos, apesar de serem eticamente questionáveis podem se passar por neutras, visto que aparentam não gerar nenhum tipo de dano. Isso ocorre porque os algoritmos podem afetar como os humanos percebem o mundo, além de modificar a sua organização social e política;

> As atividades algorítmicas, como a criação de perfis, reontologizam o mundo, compreendendo-o e conceitualizando-o de maneiras novas e inesperadas, e acionando e motivando ações com base nos *insights* que essas atividades geram[16].

- Rastreabilidade: Os algoritmos herdam os desafios éticos do design das novas tecnologias de forma que os danos gerados são difíceis de identificar e de ser apontar um responsável.

Diante desses riscos há diversos desafios que são trazidos pela inteligência artificial, e que serão enfrentados com maior profundidade nos próximos capítulos. Traçaremos aqui uma primeira abordagem de forma a elencá-los.

O primeiro seria que as correlações geradas por algoritmos, apesar de serem inconclusivas são consideradas suficientes criveis para direcionar ações (Evidências inclusivas que levam a ações injustificadas). O problema é que os algoritmos geram conhecimento sobre grupos, enquanto que as ações são sobre indivíduos. Os indivíduos, dessa forma acabam sendo descritos, e julgados, através de modelos e classes simplificadas, o que aumentam as chances desse julgamento ser errado, ou até mesmo discriminatório. Tal ocorre,

[16]No original “Algorithmic activities, like profiling, reontologise the world by understanding and conceptualising it in new, unexpected ways, and triggering and motivating actions based on the insights it generates.”Brent Daniel Mittelstadt, Patrick Allo, MariarosariaTaddeo, Sandra Wachter, Luciano Floridi. “The ethics of algorithms: Mapping the debate”. First Published December 1, 2016 Research Article.Disponível em https://doi.org/10.1177/2053951716679679

por exemplo, no sistema de polícia preditiva utilizado em Chicago (*heatlist*) no qual um cidadão pode estar na lista de risco e receber uma visita da polícia, apenas por se associar a pessoas com histórico criminal.

Um segundo problema gerado por decisões autônomas, e considerado por muitos pesquisadores do tema como o principal, é a questão da opacidade (Evidência inescrutável que levam à opacidade). Em geral prefere-se a transparência porque algoritmos imprevisíveis ou difíceis de explicar também são difíceis de controlar, monitorar e corrigir.

Entende-se por transparência a disponibilização das informações, das condições de acessibilidade e como essas informações podem sustentar o processo de decisão do usuário. Transparência, portanto, é composta pelo binômio acessibilidade e compreensão.

No entanto, para haver transparência seria necessário, em muitos casos, ter acesso a tecnologias que estão protegidas por segredo industrial. Além disso, as empresas detentoras dos dados e dos algoritmos se negam a dar transparência por receio de passar a ter uma desvantagem competitiva no mercado. Tais empresas argumentam, ainda, que eventual transparência permitiria que os consumidores se utilizassem desse conhecimento sobre os dados utilizados e as funcionalidades do algoritmo para burlar o sistema de modo a serem beneficiados injustamente (*gaming the system*). Há, ademais, o argumento de que o amplo acesso às informações geraria um risco desnecessário para a segurança nacional e para a privacidade dos titulares dos dados.

Por fim, argumenta-se que todos esses riscos seriam gerados sem nenhum benefício verdadeiro para os usuários visto que o mero acesso ao algoritmo e suas funcionalidades não garante sua compreensão pelos usuários, que são leigos. Na verdade, quando se trata de algoritmos que utilizam *machine learning*, muitas vezes nem mesmo os programadores conseguem compreender a lógica utilizada pela máquina por detrás do resultado obtido.

No entanto, “os titulares de dados mantêm interesse em entender como as informações sobre eles são criadas e influenciam as decisões tomadas em práticas orientadas por dados”[17]. Em outras palavras, as pessoas que sofrem os efeitos das decisões detêm o direito de entender como as informações sobre elas levaram aos resultados das análises de dados, sob pena de terem prejudicado o seu consentimento dado, já que não podem ter verdadeira noção dos riscos, assim como a confiança na tecnologia.

Há, portanto, um desequilíbrio na relação entre os usuários e os processadores de dados no que tange o conhecimento e o poder de decisão, favorecendo os últimos.

Considerando, então, que a informação, para ser verdadeiramente transparente deve ser, além de acessível, também compreensível, há a possibilidade de se disponibilizar as informações para terceiros treinados ou reguladores que representem o interesse público, em oposição aos próprios sujeitos de dados. Tal medida pode gerar mais impacto do que a disponibilização para o público em geral.

Há uma crença, baseada principalmente no entendimento de que regras matemáticas são imparciais, da neutralidade dos algoritmos. No entanto, considerando que o *output* é apenas tão bom quanto o *input*, os resultados obtidos pelo algoritmo serão contaminados pelas distorções dos programadores que criaram o algoritmo, assim como pelos dados com que foi alimentado e treinado (Evidência má orientada que leva a distorções). Dessa forma, as decisões tomadas pela máquina podem ser tão falíveis, ou até mais, do que as realizadas pelo ser humano.

O desenvolvimento não é um percurso linear ou neutro, já que não há uma única escolha correta, mas sim muitas possíveis alternativas, de forma que o resultado refletirá os

[17]No original “data subjects retain an interest in understanding how information about them is created and influences decisions taken in data driven practices.”Brent Daniel Mittelstadt, Patrick Allo, MariarosariaTaddeo, Sandra Wachter, Luciano Floridi. “The ethics of algorithms: Mapping the debate”. First Published December 1, 2016 Research Article.Disponível em https://doi.org/10.1177/2053951716679679

valores do autor. Além disso, as distorções podem ainda surgir das regras de decisões desenvolvidas pelo próprio algoritmo.

Além das distorções geradas pelo processo de decisão do algoritmo em si (*bias*) há ainda as discriminações que ocorrem como um efeito da decisão (consequências injustas que levam à discriminação). A ocorrência mais comum desse tipo de viés é o *profiling*, que se trata de uma construção ou inferência de padrões por meio de mineração de dados e a sua aplicação a indivíduos que combinem com esses padrões, já que os indivíduos são julgados com base nas conclusões de algoritmos sobre grupos, e não com base em seu real comportamento. Como bem explica Danilo Doneda:

> Nela [na técnica conhecida como *profiling*], os dados pessoais são tratados, com o auxílio de métodos estatísticos, técnicas de inteligência artificial e outras mais, com o fim de obter uma 'metainformação', que consistiria numa síntese dos hábitos, preferências pessoais e outros registros da vida desta pessoa. O resultado pode ser utilizado para traçar um quadro das tendências de futuras decisões, comportamentos e destinos de uma pessoa ou grupo[18].

Esses comportamentos, por sua vez, se convertem em dados que são incluídos no treinamento do algoritmo, gerando assim o que CathyO'Neil denomina de um *feedback loop* pernicioso, já que as análises discriminatórias podem levar à concretização de profecias que solidificam a estigmatização de grupos vulneráveis, prejudicando a sua autonomia e participação na sociedade.

Além disso, segundo Solon Barrocas, a mera retirada de termos sensíveis que contribuiriam para a discriminação como cor ou raça não resolve o problema visto que os dados utilizados como *proxies*, taiscomo endereço, podem também, junto com outros dados, gerar perfis relacionados com gênero, raça, preferência sexual, etc.

Para tentar controlar essa relação gerada pelos *proxies* seria necessário que os desenvolvedores: (i) Controlem as distorções do dados utilizados no treinamento da máquina;

[18]DONEDA, Danilo. *Da privacidade à proteção de dados pessoais*. Rio de Janeiro: Renovar, 2006, p. 173.

(ii) Integrem critérios anti-discriminatórios no classificador do algoritmo; (iii) haja uma pós-verificação dos modelos de classificação e, por fim; (iv) que haja uma modificação das predições e decisões para manter uma proporção justa dos efeitos entre grupos protegidos e não protegidos.

Além do *profiling*, outra prática muito discutida é a personalização, que se trata de uma prática relacionada, que pode segmentar a população de forma que apenas alguns agentes tenham acesso a determinadas informações, o que reforçaria desvantagem de certos grupos, como preços. Tal conduta levanta sérias questões éticas sobre justiça e igualdade.

Com relação aos problemas gerados pelos efeitos transformativos das decisões autônomas, pode-se dividi-los em duas principais consequências: por um lado há os desafios para a autonomia humana; e por outro, os desafios relacionados com a privacidade informacional dos indivíduos.

Com relação ao primeiro caso, as decisões tomadas por algoritmos imbuídas de valores podem prejudicar a autonomia. A personalização, já descrita anteriormente, pode gerar nudges nos indivíduos alvo dos dados, filtrando as informações recebidas gerando o denominado filtro bolha. Nesse caso, os algoritmos filtram informações, reduzindo a diversidade ao retirar aquelas que, segundo compreendido pelo próprio algoritmo, seriam irrelevantes ou contraditórias para as crenças do usuário. Logo, o usuário deixa de ter acesso a informações que permitiriam uma descoberta espontânea de novas ideias ou opiniões.

Além disso, o algoritmo pode manipular a percepção da realidade do indivíduo, de forma que ele passe a tomar decisões que em outras condições não tomaria caso não tivesse as informações providas pelo algoritmo. Dessa forma, suas decisões passam a refletir valores e interesses de terceiros, o que viola a autonomia individual.

Por outro lado, os desafios ocasionados para a privacidade informacional, ou o direito dos titulares de protegerem os seus dados pessoais, surgem quando se tenta dar respostas aos problemas anteriores.

Privacidade informacional se refere à capacidade dos indivíduos de controlar informações sobre ele próprio, em contraste com o esforço de terceiros de obter essas informações. Tal proteção comumente se baseia na possibilidade de identificação do indivíduo, como ocorre nas atuais legislações de proteção de dados europeia (GDPR) e na lei brasileira (LGPD). No entanto, tal proteção é fraca porque os algoritmos se utilizam de grande quantidade de dados que geram conclusões baseadas em grupos, de forma que a identidade do indivíduo seria irrelevante já que haveria uma anonimização. Sendo assim, tal proteção não protege efetivamente dados de grupos.

Por fim, em se tratando de responsabilidade moral, quando a tecnologia falha e gera danos deve haver sanções. Em geral, se responsabiliza o *designer* ou todos aqueles envolvidos na cadeia de produção. No entanto, tal responsabilização considera que o agente responsável tem pelo menos uma parcela de controle e intencionalidade ao realizar a ação, o que pode ser perfeitamente aplicável nos casos em que o algoritmo não utiliza aprendizagem automática.

Mas o que fazer quando esses algoritmos aprendem sozinhos, visto que nesses casos os produtores não têm controle, nem sequer conseguem prever as ações do algoritmo? Trata-se de hipótese em que há uma distância entre o dano causado e quem será efetivamente responsabilizado (*accountability gap*). Isso ocorre porque o humano pode estar impossibilitado de efetivamente analisar o processo de decisão, devido à falta de transparência discutida anteriormente. Além disso, os humanos que participam do processo de tomada de decisão (*in the loop*) podem estar mal equipados para identificar os viéses e tomar as providências corretas.

Alguns autores, como tentativa de solucionar a questão, consideram que os algoritmos podem ser considerados responsáveis desde que tenham autonomia, comportamento interativo e um papel com responsabilidade causal. Tal responsabilização, portanto, seria distinta da responsabilidade moral que necessita de intencionalidade. Por outro lado, há pesquisadores que discutem se seria aceitável, devido a todos os problemas abordados, a total substituição das decisões humanos por decisões autônomas.

Independentemente da filosofia de *design* escolhida, desenvolvedores tem a responsabilidade de projetar para diversos contextos, governados por diferentes quadros morais. Também seria possível o desenvolvimento colaborativo de requisitos para sistemas computacionais para fundamentar um protocolo ético operacional.

Tais problemas e propostas de soluções serão devidamente aprofundados e debatidos nos próximo capítulos. Antes de abordar possíveis soluções éticas e jurídicas, no entanto, é necessário realizar uma análise mais profunda de como ocorrem as violações de princípios fundamentais da privacidade e dignidade. Para tanto, no próximo capítulo vai-se analisar a aplicação da inteligência artificial no policiamento, prática que se denomina de polícia preditiva.

2. DISCRIMINAÇÃO NAS DECISÕES POR ALGORITMOS

Nos últimos seis anos, o departamento de polícia de Nova York compilou um enorme banco de dados contendo os nomes e detalhes pessoais de pelo menos 17.500 pessoas que acreditam estar envolvidas em gangues criminosas[19]. O esforço já foi criticado por ativistas dos direitos civis que dizem que é impreciso e racialmente discriminatório.

"Agora imagine casar a tecnologia de reconhecimento facial com o desenvolvimento de uma base de dados que, teoricamente, pressupõe que você está em uma gangue", disse Sherrilyn Ifill, presidente e diretora-conselheira do fundo *NAACP Legal Defense no AI Now Symposium*, realizado em Nova York, em outubro de 2018[20].

No primeiro capítulo falamos de inteligência artificial. No entanto, como visto, o termo é utilizado de forma muitas vezes ampla e vaga. Desse modo, este capítulo irá abordar

[19]AI Now Symposium. A CLOSER LOOK AT FACIAL RECOGNITION. Disponível em: https://ainowinstitute.org/symposia/videos/a-closer-look-at-facial-recognition.html

[20]No original " Now imagine marrying facial recognition technology with the development of a data base that theoretically assumes you are in a gang." AI Now Symposium. A CLOSER LOOK AT FACIAL RECOGNITION. Disponível em https://ainowinstitute.org/symposia/videos/a-closer-look-at-facial-recognition.html

especificamente decisões autônomas por algoritmos e, mais especificamente, o uso dessas decisões para o policiamento.

Controlados por algoritmos, sistemas automatizados de tomada de decisões ou de apoio à decisão são procedimentos pelos quais as decisões são inicialmente, de forma parcial ou total, delegadas para outra pessoa ou corporação, que, por sua vez, usa modelos de decisões automarizadas para executar uma ação. A automatização é da execução, portanto, e não da decisão que, na maioria dos casos atuais, permanece sendo humana.

O termo também afasta a questão levantada pelo termo “inteligência”, que remete a comportamento e pensamentos dotados de certo grau de autonomia, o que geraria problemas mais à frente quando se analisar a questão de regulação e responsabilização.

Considerando os casos já citados anteriormente é possível constatar que sistemas de decisões autônomas já vêm sendo usados em diversos países, de modo que lidar com a questão da discriminação gerada por eles é tratar de um problema atual, e não futuro.

Embora haja ampla evidência[21]de que decisões humanas sejam dotadas de viéses sendo, muitas vezes, preconceituosas, não há indícios de que as decisões por algoritmos sejam melhores ou livres desses viéses. Na verdade, o que ocorre é justamente o contrário, como se demonstrará a seguir.

2.1 Polícia Predetiva

Em maio de 2010, motivado por uma série de escândalos de alto perfil, o prefeito de Nova Orleans pediu ao Departamento de Justiça dos EUA que investigasse o departamento de

[21]Um estudo publicado no Proceedings of National Academy of Science analisou decisões judiciais de juízes israelenses que presidiram audiências para concessão de liberdade condicional descobriu que os juízes deram decisões mais brandas no início do dia e imediatamente após uma interrupção programada, como por exemplo o almoço. Eles descobriram que a probabilidade de uma decisão favorável atingiu o pico no início do dia, declinando constantemente ao longo do tempo, de uma probabilidade de cerca de 65% para quase zero, antes de voltar para cerca de 65% após uma pausa para uma refeição ou lanche. Shai Danziger, Jonathan Levav,and LioraAvnaim-Pesso. Extraneous factors in judicial decisions.PNAS April 26, 2011 108 (17) 6889-6892;Disponível em: https://doi.org/10.1073/pnas.1018033108

polícia da cidade (NOPD). Dez meses depois, o DOJ ofereceu sua análise: durante o período de sua revisão, a partir de 2005, o NOPD havia violado repetidamente as leis constitucionais e federais[22].

Havia usado força excessiva e desproporcional contra moradores negros, minorias raciais específicas, falantes não nativos de inglês e indivíduos LGBTQ, além de não ter abordado o tema da violência contra as mulheres. Os problemas, disse o assistente do procurador geral Thomas Perez na época, eram "sérios, abrangentes, sistêmicos e profundamente enraizados na cultura do departamento"[23].

Na tentativa de buscar uma solução para tais resultados, a cidade entrou em uma parceria secreta no ano seguinte com a empresa de mineração de dados Palantir, para implantar um sistema de policiamento preditivo[24]. O sistema usou dados históricos,incluindo registros de detenções e relatórios policiais eletrônicos, para prever o crime e ajudar a moldar as estratégias de segurança pública, de acordo com materiais da empresa[25] e do governo local[26]. No entanto, não consta nesses documentos qualquer indício que sugira ter ocorrido algum esforço para limpar ou corrigir os dados utlizados pelo algoritmo, que traziam em si as violações reveladas pelo DOJ. Com toda a probabilidade, os dados corrompidos foram alimentados diretamente no sistema, reforçando as práticas, já discriminatórias, do departamento.

O policiamento preditivo, sistema utilizado no caso descrito acima, geralmente descreve qualquer sistema que analisa dados disponíveis para prever onde um crime pode

---

[22]Department of Justice Releases Investigative Findings Involving the New Orleans Police Department.March 17, 2011. Disponível em https://www.justice.gov/opa/pr/department-justice-releases-investigative-findings-involving-new-orleans-police-department

[23]No original “He described the problems at NOPD as “serious, wide-ranging, systemic and deeply-rooted within the culture of the department.””.Disponível em: https://www.nola.com/crime/2011/03/yearlong_justice_department_pr.html

[24]Ali Winston. Palantir has secretely been using New Orleans to test its predictive policing technology. Feb 27, 2018. Disponível em https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd

[25]https://www.documentcloud.org/documents/4344816-NOLA-Murder-Reduction-White-

[26]https://assets.documentcloud.org/documents/4344815/Nola-hc3-Final-20140403.pdf

ocorrer em um determinado período de tempo (baseado em local) ou quem estará envolvido em um crime como vítima ou perpetrador (baseado em pessoa).

Trata-se da mais recente ferramenta de combate à criminalidade, que se utiliza de técnicas de análise de dados e, tal como o uso de sistemas de decisão autônomos adquiridos e utilizados por outros setores, o policiamento preditivo é vendido com a promessa de neutralizar os preconceitos, conscientes ou inconscientes, dos tomadores de decisão humanos, neste caso, a polícia.

No entanto, poucos fornecedores de sistemas de policiamento preditivo são totalmente transparentes sobre como seus sistemas operaram, devido ao fato de seus algoritmos estarem protegidos pelo segredo industrial, não informando os dados específicos que são usados, ou ainda quais medidas de responsabilidade que o fornecedor emprega para abordar eventuais danos gerados por viéses ou evidência de má conduta. Contudo, é notoriamente sabido que a principal fonte de dados usualmente utilizadasão as atividades passadas da polícia local e,enquanto as categorias específicas podem variar de acordo com o sistema, geralmente incluem informações sobre crimes passados (tipo de crime, tempo e localização) e detenções[27].

Tal caso não foi o primeiro nem o único quando se trata da utilização de algoritmos de decisão autônoma no policiamento, prática denominada de polícia preditiva. Cidades como Memphis e Nova York[28]já os utilizam há alguns anos e, recentemente, foi criado na Argentina

[27]No original “Few predictive policing vendors are fully transparent about how their systems operate, what specific data is used in each jurisdiction that deploys the technology, or what accountability measures the vendor employs in each jurisdiction to address potential accuracy, bias, or evidence of misconduct. Despite these looming questions, one known fact is that historical police data is the primary data source used to inform these systems, and while the specific data categories will vary by system, it can include information on past crimes (type of crime, time, and location), arrests and calls for service”Richardson, Rashida and Schultz, Jason and Crawford, Kate, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (February 13, 2019). New York University Law Review Online, Forthcoming. Disponível em SSRN: https://ssrn.com/abstract=. p.6.

[28]Com a ajuda da Microsoft, Nova York lançou em dezembro de 2012 o mapa interativo de crimes que possibilita filtrar bairros e ruas, além de verificar quais tipos de crimes possuem maior incidência em cada região. O Mapa pode ser acessado através do endereço https://maps.nyc.gov/crime/. O sistema por trás do mapa é o DAS (*Domain Awareness System*) capaz de processar os dados telefônicos das chamadas de emergência

o Observatório Nacional de Big Data que pretende, dentre outros objetivos, realizar análises de dados para previsão de eventos de qualquer natureza. Além disso, a implementação desses sistemas vem sendo discutida em países como a França e o ReinoUnido.

Em Memphis, há a iniciativa *Blue Crush* (*Criminal Reduction Utilizing Statistical History*)[29]onde foram coletados dados de todas as ocorrências policiais. Estes dados são registrados por aparelhos utilizados pelos próprios policiais, através do uso de um *software* preditivo da IBM que, em conjunto com um serviço de mapeamento da empresa ESRI, cria mapas com os possíveis locais de novos crimes. Se utilizando dessas informações, a políciadireciona o patrulhamento de forma a trazer maior precisão e poupar recursos financeiros[30].

A polícia de Chicago, por sua vez, se utiliza de uma tecnologia denominada *ShotSpotter*[31],que é capaz de registrar sons de tiros, além de sua localização aproximada. Recentemente, esse sistema foi integrado à tecnologia que utiliza algoritmos preditivos, que identificam padrões e tendências a partir dos dados do *ShotSpotter*.Há ainda a integração com o sistema *HunchLab*, que ajuda a polícia a tomar decisões de acordo com a análise massiva das informações detenções, chamadas para o 911, atividades de gangues e outros dados relevantes. Com esses sistemas integrados, a cidade de Chicago conseguiu reduzir os tiroteios em 49% e 66%, em fevereiro e março, respectivamente, de 2017, nos distritos onde foi implementada a tecnologia.

No entanto, o caso mais famoso certamente é a utilização do sistema *PredPol*, desenvolvido pelo Departamento de Polícia de Los Angeles em parceira com a Universidade

---

(911) e utilizar as imagens de câmeras de segurança de modo a reconhecer rostos, placas e radiação. Disponível em http://exame.abril.com.br/tecnologia/big-data-contra-o-crime/

[29]SAISSE, Renan. Big Data contra o crime: efeito Minority Report. Revista Direito e TI. Disponível em http://direitoeti.com.br/artigos/big-data-contra-o-crime-efeito-minority-report/

[30]How big data fights crime. Disponível em https://fcw.com/articles/2012/11/02/big-data-memphis.aspx

[31]https://nakedsecurity.sophos.com/2017/05/10/minority-report-in-chicago-as-police-aim-to-stop-crime-before-it-happens/

da Califórnia[32]. Tal sistema utiliza os dados das forças policiais na identificação dos problemas relacionados à criminalidade e às suas soluções, fornecendo ajudas na gestão de recursos e escolha de estratégias, de forma a prevenir fatos futuros. Da mesma forma que nos casos anteriores, o *PredPol* é utilizado para geoposicionar policiais em local e hora corretos para prevenir um possível crime.

No cenário brasileiro, pode-se apontar que, em 2014, o governo do Estado de São Paulo firmou uma parceria com a Microsoft para a utilização do sistema denominado Detecta, que utiliza o *Big Data* para análise dos dados da polícia e de radares e câmeras de vigilância.

Figura3–Detecta

CINTURÃO ELETRÔNICO DO DETECTA (2661 CÂMERAS)	
Câmeras OCR	
Vale do Ribeira	12 câmeras
Baixada Santista e ABC	208 câmeras
Litoral Norte	49 câmeras
Alto Tietê	67 câmeras
Vale do Paraíba	144 câmeras
TOTAL	480 câmeras OCR
Câmeras de videomonitoramento	
Praia Grande	1.500 câmeras
Santos	517 câmeras
Guarujá	72 câmeras
Cubatão	32 câmeras
Itanhaém	28 câmeras
Campos do Jordão	32 câmeras
TOTAL	2181 câmeras

Fonte: Secretaria de Segurança Pública de São Paulo

Dessa forma, há, atualmente 2.661 câmeras ligadas aos batalhões, delegacias e, inclusive, *tablets* das viaturas. Associado ao sistema Omega[33], da polícia civil de São Paulo, que coleta informações de cadastros de registros civis, criminal, de armas, de veículos

[32]Andrew Guthrie Ferguson. The rise of big data policing: surveillance, race and the future of law enforcement. New York University press. New York. 2017. p. 65.
[33]http://www.saopaulo.sp.gov.br/spnoticias/na-imprensa/sistema-secreto-da-policia-pode-rastrear-qualquer-um/

roubados e furtados, da junta comercial, disque-denúncia, delegacia eletrônica e DETRAN, além de mapas e sistema de identificação biométrica, os dados analisados podem ser utilizados para investigar, mapear e combater a criminalidade no Estado.

As Polícias Judiciárias Estaduais, de modo geral, ainda estão se modernizando, coletando dados úteis e desenvolvendo técnicas de análises diagnósticas e descritivas para investigações. Nada próximo ao que já ocorre nas cidades americanas. Poucas cidades possuemprocessos de análise preditiva como fonte investigativa para estratégia de policiamento e prevenção de crimes.

No entanto, este cenário está mudando com a implementação da tecnologia de reconhecimento facial em câmeras de trânsito e de segurança. No Rio de Janeiro as imagens captadas por estas câmeras são transmitidas para o Centro Integrado de Comando e Controle (CICC) onde serão cruzadas com a base de dados da Polícia Civil e do DETRAN para identificar pessoas que estejam com pedidos de prisão expedidos ou verificar placas de carros para saber se são roubados[34]. Sistemas semelhantes também foram implementados em Campinas e Salvador[35].

No dia 05.03.2019, em Salvador, um jovem de 19 anos foi preso após ser identificado por câmeras de reconhecimento facial instaladas nas ruas do circuito Dodô (Barra-Ondina)[36]. O jovem era procurado por homicídio e estava com mandado de prisão em aberto desde julho de 2018. Após ter sido flagrado pela tecnologia contratada pela Secretária de Segurança Pública da Bahia (SSP-BA), policiais o prenderam.

---

[34]Rio: programa de reconhecimento facial entra em operação no carnaval, Disponível em: http://agenciabrasil.ebc.com.br/geral/noticia/2019-01/rio-programa-de-reconhecimento-facial-entra-em-operacao-no-carnaval

[35]Sistema de reconhecimento facial chinês já é testado no Brasil desde 2018. Disponível em: https://olhardigital.com.br/noticia/sistema-de-reconhecimento-facial-chines-ja-e-testado-no-brasil-desde-2018/81349

[36]Procurado por homicídio vai para o carnaval de Salvador vestido de mulher e é preso após ser flagrado por câmera. Disponível em:https://g1.globo.com/ba/bahia/carnaval/2019/noticia/2019/03/05/procurado-por-homicidio-vai-para-o-carnaval-de-salvador-vestido-de-mulher-e-e-preso-apos-ser-flagrado-por-camera.ghtml

No Rio de Janeiro, por sua vez, quatro procurados com mandado de prisão em aberto foram presos, de acordo com informações da Polícia Militar (PM)[37]. O *software* também foi responsável pela identificação e apreensão de um veículo roubado.

A Polícia Federal (PF), por sua vez, já está bem a frente nesse quesito, tendo uma gama de *softwares*, como o Nudetective[38] e o EspiaMule[39], que permitem o correlacionamento de dados de diversas fontes. Apesar de ainda se utilizar principalmente de análises descritivas e diagnósticas, a PF possui capacidade de implementar também unidades preditivas.

No entanto, foi com a operação Lava Jato que a Polícia Federal efetivamente adentrou na era do *Big Data*, ao criar um banco de dados unificado, além de procedimentos analíticos, que foi denominado como "o *Big Data* para o combate à corrupção"[40].Diante do extenso banco de dados e dos mecanismos já existente para a sua análise é possível, futuramente, a adoção de uma análise preditiva, caso as informações sejam alimentadas constantemente, de forma a se traçar perfis e tendências para o combate à fraude e outros crimes.

No entanto, há que se estar atento às possíveis violações aos direitos fundamentais. Como tratado extensamente no tópico anterior, sistemas de predição podem ser altamente discriminatórios e injustos devido aos problemas inerentes ao próprio algoritmo, nos dados utilizados ou simplesmente pelo sistema de predição em si.

2.1.1Formas de Policiamento Preditivo

[37]Câmeras de reconhecimento facial levam a 4 prisões no carnaval do Rio. Disponível em: http://agenciabrasil.ebc.com.br/geral/noticia/2019-03/cameras-de-reconhecimento-facial-levam-4-prisoes-no-carnaval-do-rio

[38]O Nudetective foi desenvolvido pelos peritos criminais Pedro Eleutério e Matheus Polastro e é utilizado na detecção de pornografia infantil por meio de análises de imagens, nomes, *hash*se vídeos. O Nudetective possui princípios que podem ser explorados para expansão conceitual e técnica instituindo uma análise forense direcionada para uma coleta massiva de dados em temporeal.

[39]O EspiaMule realiza pesquisas e coletas de informações de usuários da rede Emule, catalogando endereços e criando um mapa da distribuição de imagens de pornografia infantil, o que atualmente é realizado também na rede P2P *Utorrent*e técnicas são aplicadas também na *Deep Web*.

[40]SAISSE, Renan. Big Data contra o crime: efeito Minority Report. Revista Direito e TI. Disponível em: http://direitoeti.com.br/artigos/big-data-contra-o-crime-efeito-minority-report/

Até o momento tratou-se do policiamento preditivo como uma ação única. No entanto, há diversas formas de policiamento, que variam conforme o sistema algorítmico adotado. Embora todas essas formas tenham o potencial de gerarviéses e discriminações, como será observado com maior detalhamento no tópico seguinte, cada uma detêm certas nuanças que merecem ser analisadas.

A maioria dos sistemas de policiamento preditivo em uso são baseados em locais que visam prever quando e onde os futuros crime ocorrerá ("baseados em locais") ou que tentam prever infratores, determinar as identidades dos perpetradores ou prever vítimas em potencial ("baseados na pessoa"). À parte desses dois tipos dedicarei, ainda, um tópico para tratar da vigilância constante na sociedade que ocorre por meio de aparelhos eletrônicos, redes sociais e outros mecanismos de controle[41].

2.1.1.1 Policiamento baseado no lugar

É o tipo mais comum de policiamento, incluindo os exemplos bem conhecidos de *software* da *PredPol*[42]e o *HunchLab*. Como bem explica Andrew Guthrie Ferguson "As previsões baseadas em locais são focadas, principalmente, em pontos quentes (*hot spots*) onde haveria a maior probabilidade de ocorrerem novos crimes"[43]. Essas informações são utilizadas, por sua vez, principalmente para gestão de recursos humanos e econômicos da polícia. Em outras palavras, trata-se da tentativa de prever onde haverá um maior foco de

[41] Andrew Guthrie Ferguson. The rise of big data policing: surveillance, race and the future of law enforcement. New York University press. New York. 2017

[42] Andrew Guthrie Ferguson. The rise of big data policing: surveillance, race and the future of law enforcement. New York University press. New York. 2017. p. 65

[43]No original "Predictive policing involves a data-driven approach to identifying criminal patterns in specific geographic locations and deploying police resources to remedy those risks."Andrew Guthrie Ferguson. The rise of big data policing: surveillance, race and the future of law enforcement. New York University press. New York. 2017. p. 62

crimes, de modo que a polícia possa realocar um maior número de policiais para esta localidade.

Ocasionalmente, se houver um padrão muito específico, a polícia pode ser capaz de prever especificamente o próximo caso em uma onda de crimes, mas os sistemas de predição geralmente não são tão específicos, podendo apontar uma determinada área e período onde há maior probabilidade de ocorrer determinados tipos de delitos.

O potencial de dano decorrente desse tipo de policiamento preditivo decorre de haver um maior policiamento em localidades de residência de grupos vulneráveis como negros e latinos, nos exemplos americanos.

O sistema denominado *PredPol* utiliza como base, um sistema desenvolvido para medir abalos sísmicos que ocorrem após um terremoto. Os seus desenvolvedores descobriram que o crime segue padrões similares. O crime, ao que parece, poderia ser visualizado como tendo efeitos parecidos com ondulações e, uma vez identificado, esse padrão poderia ser mapeado e previsto.

O *HunchLab*, por sua vez, recebe dados de crimes anteriores, do censo, densidade populacional, bem como outras variáveis como localização de escolas, igrejas, bares, entre outros. O algoritmo, então,apresenta um mapa, constantemente atualizado, de locais de risco, determinando a probabilidade da ocorrência de crimes[44].

Apesar dos sistemas de predição baseados em lugar serem novos, tal estratégia já vem sendo utilizada pela polícia há décadas. O sistema de predição evoluiu da teoria da criminologia denominada “janelas quebradas”. Desenvolvida na escola de Chicago, por James Q. Wilson e George Kelling, tal teoria explica que, se uma janela de um edifício for quebrada, e não for reparada, a tendência é que vândalos passem a arremessar pedras nas outras janelas e, posteriormente, passem a ocupar o edifício e destruí-lo. Segundo esta teoria, a desordem

[44]Andrew Guthrie Ferguson. The rise of big data policing: surveillance, race and the future of law enforcement. New York University press. New York. 2017. P. 63

gera desordem, de modo que um comportamento anti-social pode dar origem a vários delitos. Por isso, qualquer ato desordeiro, por mais insignificante que possa parecer, deve ser reprimido, caso contrário, ele pode ser o difusor de inúmeros outros crimes mais graves[45].

De fato, a desordem gera a desordem, no entanto, tal teoria foi desenvolvida, e aplicada, de modo a justificar uma política agressiva de combate à criminalidade que, na verdade, não se sustenta, porque visa atacar um conflito apontando como solução um problema maior ainda: penalizar com a prisão aqueles que foram gratuitamente sancionados com a falta de estrutura física e social. O mesmo acaba por ocorrer com a utilização dos sistemas de polícia preditiva baseados em lugar.

A distorção ocorre, primeiramente, por que nem todos os tipos de crimes são analisados pelo algoritmo, por dois grandes motivos: primeiro, uma grande quantidade de crimes. Transgreções menos importantes, como consumo de drogas ou intoxicação pública nem sempre levam à prisão. Dessa forma, não são registrados pelo sistema policial. Além disso, essas prisões, quando ocorrem, são mais comuns entre pessoas da população vulnerável[46].

Além disso, para que o sistema de decisão autônomo possa chegar aos resultados desejados, seus desenvolvedores devem pegar uma pergunta abstrata "Como eu posso prevenir o crime?"e transformá-la em uma que possa ser expressa como "o valor de algum alvo variável.". Por exemplo, o *PredPol* divide um mapa em quadrados de $500 \times 500 m^2$, e para cada um deles, a variável de destino torna-se probabilidade de um determinado crime. Mas as categorias nem sempre são óbvias. Se o sistema estiver projetado para detectar crimes em um determinado quadrado em um mapa, é necessário determinar os tipos de crime.

[45]O'NEIL, Cathy . Op cit.
[46]Selbst, Andrew D., Disparate Impact in Big Data Policing (February 25, 2017). 52 Georgia Law Review 109 (2017). Disponível em SSRN: https://ssrn.com/abstract=2819182 or http://dx.doi.org/10.2139/ssrn.2819182. p.130

(violento ou não violento, crimes de propriedade ou crimes incômodos). A decisão sobre a forma de análise do problema pode ter consequências importantes para o resultado final[47].

2.1.1.2 Policiamento baseado na pessoa

Este tipo de previsão é baseado na pessoa, e não na investigação. Pode-se citar como exemplo o *software Beware*, da Intrado, que permite que a polícia aproveite dados publicamente disponíveis, incluindo os de mídia social, para verificar a "Pontuação de ameaça" de uma pessoa ou endereço quando receber uma chamada de emergência. O indivíduo é categorizado como verde, amarelo ou vermelho a depender do seu nível de risco[48]. Outros sistemas analisam a mídia social para encontrar membros de gangues[49]. Por fim, há ainda outros sistemas, como a "lista de calor" (*heatlist*) utilizada pela polícia de Chicago, na qual são colocadas as pessoas mais prováveis de se envolverem em um crime no futuro[50].Nesse último exemplo, embora o algoritmo permaneça opaco, por se tratar de segredo industrial, sabe-se que ele inclui fatores como histórico criminal, prisões, status de liberdades condicionais recebidas e se o indivíduo alvo foi identificado como parte de algum grupo suspeito. Ser selecionado para entrar na "lista" gera, em geral, uma visita da polícia no domicílio do indivíduo. Durante essa visita, a polícia entrega uma carta de notificação, onde

[47]Selbst, Andrew D., Disparate Impact in Big Data Policing (February 25, 2017). 52 Georgia Law Review 109 (2017). Available at SSRN: https://ssrn.com/abstract=2819182 or http://dx.doi.org/10.2139/ssrn.2819182. p.132.
[48]Justin Jouvenal, The New Way Police Are Surveilling You: Calculating Your Threat 'Score,' WASH. POST (Jan. 10, 2016), https://www.washingtonpost.com/local/public-safety/the-n ew-way-police-are-surveilling-you-calculating-your-threat-score/2016/01/10/e42bccac-8e15-11e 5-baf4-bdf37355da0c_story.html.
[49]LakshikaBalasuriya et al., Finding Street Gang Members on Twitter 2016 IEEE/ACM INT'L CONF. ON ADVANCES IN SOC. NETWORKS ANALYSIS & MINING (ASONAM), 685, https://arxiv.org/pdf/1610.09516v1.pdf
[50]Andrew Guthrie Ferguson. The rise of big data policing: surveillance, race and the future of law enforcement. New York University press. New York. 2017. p. 37-38

consta tudo que se sabe sobre atividades criminais passadas do indivíduo, além de um aviso sobre o que pode ocorrer no futuro[51].

Como monitoramento de seus alvos, caso ocorra um crime posterior, há maior probabilidade da polícia realizar uma investigação desses indivíduos, especialmente daqueles que constam como ameaça de alto nível. Existe, neste caso, alto risco da polícia agir violentamente ou com força desnecessária, quando essa análise for errada, ou seja, quando o indivíduo não estiver envolvido no conflito. Em outras palavras, há uma imediata manipulação na autonomia dos policiais quando lidam com indivíduos apontados por algoritmos preditivos como de risco. O dano parece ser distinto do ocorrido em sistemas preditivos de lugar.Essas pessoas passam a receber um tratamento totalmente diferenciado, de forma justificada ou não, que interfere na sua autonomia e relação com a sociedade.

2.1.1.3 Policiamento baseado em vigilância

O tipo final de sistema é o baseado em vigilância. Nesse sistema os dados coletados através das mais diversas fontes como pela câmeras de vigilância, redes sociais e outros bancos de dados públicos e privados criam modelos de possíveis criminosos, perfis daqueles que poderiam vir a cometer um determinado tipo de crime. Com base nesses perfis o modelo passa a ser utilizado para localizar suspeitos[52].

Também denominado de “vigilância e investigação em tempo real” este modelo já esta sendo adotado, por exemplo, pela polícia de Fresno, na Califórnia, que utiliza um sistema denominado *Beware* que informa em tempo real *scores* de risco de endereços e pessoas. Tal sistema realiza uma análise dos bancos de dados de consumidores e prove uma decisão

[51]Andrew Guthrie Ferguson. The rise of big data policing: surveillance, race and the future of law enforcement. New York University press. New York. 2017. p. 38

[52]Selbst, Andrew D., Disparate Impact in Big Data Policing (February 25, 2017). 52 Georgia Law Review 109 (2017). Disponível em SSRN: https://ssrn.com/abstract=2819182 ou http://dx.doi.org/10.2139/ssrn.2819182

preditiva sobre a pessoa que realizou a ligação para o 911, o endereço e a vizinhança[53]. Tendo em vista o segredo industrial, nem mesmo a polícia que o utiliza tem conhecimento sobre como o sistema calculo a pontuação de risco[54].

Este tipo de sistema aumentam a velocidade de reação da polícia, além de melhorar suas capacidades investigativas. No entanto, da mesma forma que os dados facilitam na análise e previsão de um crime eles também, da assim como nos casos anteriores, acabam por prejudicar majoritariamente grupos vulneráveis. Além disso, a própria escolha do local para instalar as câmeras também pode ser visto como discriminatória.

Pesquisas realizadas em algoritmos de reconhecimento facial apontam que estes reconhecem mais facilmente pessoas brancas, falhando o dobro de vezes em reconhecer uma foto de uma pessoa negra[55]. Segundo Andrew Ferguson esta imprecisão pode ocasionar dois resultados:

> Primeiro, porque a tecnologia de reconhecimento facial não consegue encontrar uma correspondência (mesmo que haja uma correspondência no conjunto de dados), uma pessoa que deve ser identificada não é identificada (o culpado fica livre). Em segundo lugar, como a tecnologia de reconhecimento facial não consegue encontrar uma correspondência (porque não há correspondência no conjunto de dados), o algoritmo sugere a pessoa que é a correspondência mais próxima (um inocente é sinalizado). Esta correspondência mais próxima se torna alvo de investigação e nos Estados Unidos, devido à disparidade racial embutida de quem é preso (e, portanto, quem está no sistema de dados), esse erro cairá nas minorias raciais mais do que nos brancos[56].

Por fim, ainda cabe apontar que esta vigilância constante e coleta massiva de dados apresenta um risco para a privacidade, bem como para a liberdade de expressão, visto que há um desequilíbrio na relação de poder entre os vigilantes e os vigiados.

---

[53]Andrew Guthrie Ferguson. The rise of big data policing: surveillance, race and the future of law enforcement. New York University press. New York. 2017. P. 83-83

[54]Andrew Guthrie Ferguson. Op Cit. p.84

[55]Andrew Guthrie Ferguson. Op Cit. p.94

[56]No original “First, because facial recognition technology cannot find a match (even thought there is a match in the data set), a person who should be identified is not identified (the guilty go free). Second, because facial recognition technology cannot find a match (because there is no match in the data set), the algorithm nevertheless suggests the person who is the closest match (an innocent is flagged). This closest match becames the target of investigation and in the United States, due to the embedded racial disparity of who gets arrested (and thus who is in the data system), this error will fall on racial minorities more than on whites.” Andrew Guthrie Ferguson. Op Cit. p.94

2.2. Como ocorre a discriminação

Quando se fala em utilização do *Big Data* para uma análise preditiva logo se pensa em sua utilização para fins de combate à violência, de fato um dos maiores problemas atuais em diversas cidades. No entanto, como bem observaSolon Barrocas, a utilização de *Big Data*, e algoritmos, na tomadade decisões não livra o resultado dos preconceitos e discriminações humanas. Não se trata, nem se deve ter tal expectativa, de uma decisão imparcial.

Como bem definiu Solon Barrocas:

> Aproximado sem cuidados, a mineração de dados pode reproduzir padrões de discriminação existentes, herdar prejuízos de antigos tomadores de decisão, ou simplesmente refletir as distorções que persistem na sociedade. Pode até gerar o resultado perverso de exacerbar desigualdades existentes ao sugerir que determinados grupos que sofrem desvantagens históricas na verdade merecem um tratamento menos favorável[57].

Como em regra a discriminação não é intencional, o resultado acaba por ser uma distorção velada sob a ilusão de imparcialidade. Além disso, como o mecanismo que leva a essas decisões distorcidas se encontra protegido dentro de uma caixa preta[58] a injustiça acaba por se tornar difícil de identificar. Dessa forma, é importante analisar como de fato ocorre essa discriminação.

2.2.1. As camadas de vieses

[57]No original “Approached without care, data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society. It can even have the perverse result of exacerbating existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment”. BARROCAS, Solon e SELBST, Andrew D. Big Data’s Disparate Impact. 104 CALIF. L. REV. 671 (2016). Disponível em https://scholarship.law.berkeley.edu/californialawreview/vol104/iss3/2/

[58]O termo Black Box foi cunhado por Frank Pasquale, juiz e pesquisador norte-americano, e é utilizado em seu livro “The Black Box society: the secret algorihms that control Money and information”.

A forma mais comum de discriminação gerada pelas decisões autônomas ocorre por meio dos dados utilizados em seu treinamento. A mineração de dados, desde a sua coleta até a apresentação de um resultado, pode se utilizar de cinco mecanismos que podem levar àdistorções: definição do problema, treinamento dos dados, seleção de dados, utilização de *proxiese masking*. Como explica Solon Barrocas:

> (…) definindo o alvo variável, nomeando e coletando os dados de treinamento, usando a seleção de dados e tomando decisões com base nos modelos resultantes. Cada um desses passos cria a possibilidade de um resultado final que tenha um impacto desproporcional em grupos protegidos, seja por especificar o problema a ser resolvido de maneira a afetar grupos de forma distinta, falhando em reconhecer ou endereçar distorções estatísticas, reproduzindo preconceitos históricos, ou considerando uma gama insuficiente de fatores. Mesmo em situações onde os mineradores de dados seja extremamente cautelosos, ainda há a possibilidade de surgirem resultados discriminatórios com os modelos que, ainda que de forma não intencional, se utilizem de proxies como variáveis em casos de grupos protegidos[59].

É importante considerar que a mineração de dadosé uma forma de análise estatística, o que gera alguma forma de discriminação. A própria finalidade do *Big Data* é justamente prover uma base racional em cima da qual se poderá atribuir a um indivíduo qualidades características de um determinado grupo, de forma a melhor decidir sobre esse indivíduo ou grupo.

No entanto, a realidade pode ser bem mais complexa, tendo em vista que os sistemas de decisão autônoma podem conter viéses muito antes de serem coletados os dados para seu treinamento e implementação.

---

[59]No original "(…) defining the target variable, labeling and collecting the training data, using feature selection, and making decisions on the basis of the resulting model. Each of these steps creates possibilities for a final result that has a disproportionately adverse impact on protected classes, whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors. Even in situations where data miners are extremely careful, they can still effect discriminatory results with models that, quite unintentionally, pick out proxy variables for protected classes." BARROCAS, Solon & SELBST, Andrew D., "Big Data's Disparate Impact", 104 CALIF. L. REV. 671 (2016). Disponível em https://scholarship.law.berkeley.edu/californialawreview/vol104/iss3/2/

Aponta-se três estágios, onde pode ocorrer a entrada de viéses que gerem discriminações no resultado[60]:o entendimento de justiça pelo software empregado; a qualidade dos dados utilizados e; a preparação dos dados.

A primeira camada trata da questão de se o modelo de decisão utilizado é justo. Para responder a essa pergunta é necessário estabelecer uma definição estatística de justiça, o que não é simples, visto que, como se demonstrará mais a frente, diferentes definições de justiça são muitas vezes fundamentalmente incompatíveis entre si. Muitas vezes,ao resolver o viés de uma forma, produz um tipo diferente de viés.

A segunda camada analisa se os dados coletados e utilizados pelo modelo de decisão contêm alguma forma de viés. No caso específico da polícia preditiva, o viés nos dados utilizados configura um grande problema, tendo em vista que a maior parte das informações utilizadas advém da própria polícia e adquiridos de prisões feitas anteriormente. Dessa forma, os dados contêm os viéses atribuídos aos policiais e à polícia em geral.

Por fim, a terceira camada analisa se é justo chegar a decisões sobre um indivíduo com base em dados de outras pessoas. Atualmente, a maioria dos modelos preditivos baseiam-se, fundamentalmente, no pressuposto de que pode-se usar o comportamento de outras pessoas para decidir onde deve haver policiamento e quais indivíduos representam um alto risco para a comunidade. Surge então a pergunta: Devemos usar dados de grupos para tomar decisões sobre indivíduos?

Conceitualmente, cada camada depende das outras: Se fazer julgamentos sobre indivíduos com base em grupos é injusto ou ilegítimo, a qualidade dos dados e modelos não importa; se os dados são enviesados, um modelo, ainda que justo, meramente reproduz esse viés. Para desenvolver um modelo verdadeiramente justo para decisões sobre policiamento é necessário acreditarque todas as três as camadas sejam justas.

[60]https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Nos próximos tópicos iremos analisar detalhadamente cada camada, descrevendo preocupações específicas.

2.2.1.1 Primeira camada: algoritmos justos

Para se verificar se um modelo de decisão autônoma é justo, o primeiro estágio de verificação é a análise das predições em si, em outras palavras, é verificar se o sistema de análise de risco é justo. Nesse ponto, irá se utilizar como parâmetro, a título de exemplificação, o modelo de análise de risco de reincidência de pessoas que praticaram crimes. Dessa forma, há que se verificar três fatores[61]para saber se um modelo é justo.O primeiro verifica se a pontuação gerada por um modelo significa a mesma coisa em diferentes grupos. Nesse caso, considere-se, por exemplo, que seja previsto que determinado grupo de pessoas tenha 30% de chances de cometer um novo crime no período de um ano. No caso de um modelo justo, esse valor deve se repetir em diferentes grupos. Nesse exemplo, não se pode considerar justo um modelo no qual apenas 10% de negros reincidirem, enquanto mais de 50% dos brancos, com a mesma pontuação reincidirem.

O segundo fator a ser verificado é se as pessoas que não cometeram um crime posterior obtêm pontuações semelhantes entre os grupos. Nesse caso se réus negros que não reincidem tiverem mais probabilidade de obter uma pontuação de alto risco do que réus brancos que também não reincidem, então pode-se considerar que pessoas negras são tratadas com mais rigor pelo sistema. Essa verificação denomina-se de taxa de falso positivo.

Por fim, o terceiro fator, denominado de falso negativo, refere-se à pontuação dada às pessoas que efetivamente reincidiram. Nesse caso, um indivíduo brancoreincidente deverá receber a mesma pontuação do que um indivíduo negro que também reincidiu.

[61]Eckhouse, L. *et al*. (2019) 'Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment', *Criminal Justice and Behavior*, 46(2), pp. 185–209. doi: 10.1177/0093854818811379.

O grande problema levantado pelos matemáticos é que construir um sistema justo nesses três fatores simultaneamente não é apenas difícil. “É matematicamente impossível desenvolver um modelo que seja justo no sentido de ter igual valor preditivo em diferentes grupos e justo no sentido de tratar membros de grupos da mesma forma em retrospecto”[62].

Essa questão foi levantada no caso do sistema de análise de risco denominado COMPAS, já citado no primeiro capítulo. Em 2016, um repórter da ProPublica acusou o sistema de não ser justo com indivíduos negros[63]. Segundo esta publicação, as previsões sobre os réus negros sistematicamente exageram no risco de reincidência desse grupo. De fato, a ProPublica descobriu que, daqueles que não reincidiram, 45% dos réus negros haviam sido sinalizados como de alto risco. Em comparação, apenas 23% dos réus brancos que não eram reincidenteshaviam sido considerados da mesma forma[64]. Trata-se, portanto, de um erro na taxa de falso positivo.

Em sua resposta à ProPublica, a Northpointe[65], empresa que desenvolveu o COMPAS, fez uma análise distinta: ela avançou da pontuação de risco, em vez de partir da análise dos resultados como a ProPublica. Descobriu-se que pessoas com pontuações de risco, negras ou brancas, haviam tido chances semelhantes de reincidirem. Em outras palavras, eles descobriram que daqueles classificados como de alto risco, a proporção que não reincidiram era aproximadamente equivalente entre as populações branca e negra.

Da mesma forma, eles descobriram que, daqueles que haviam sido classificados como de baixo ou médio risco, negros e brancos tinham uma chance quase igual de serem presos novamente. A Northpointe, portanto, se utilizou do primeiro fator descrito acima: o valor

[62]No original “It is mathematically impossible to develop a model that will be fair in the sense of having equal predictive value across groups, and fair in the sense of treating members of groups similarly in retrospect”. Eckhouse, L. *et al.* (2019) ‘Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment’, *Criminal Justice and Behavior*, 46(2), pp. 185–209. doi: 10.1177/0093854818811379.
[63]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[64]https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[65]COMPAS risk scale: Demonstrating Accuracy Equity and Predictive Parity. Disponível em: https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html

preditivo do modelo em diferentes grupos.Com base nessa análise, a Northpointe concluiu que estas proporções demonstrariam justiça preditiva.

Como uma possível solução para esse debate, foi apontado que os algoritmos não deveriam considerar a raça como fator decisivo, passando a ser, dessa forma, neutros. No entanto, como bem informa Solon Barrocas[66], ainda que não se utilize termos sensíveis que contribuiriam para a discriminação como cor ou raça, os dados utilizados como *proxies*, como endereço, podem também, junto com outros dados, gerar perfis relacionados com gênero, raça, preferência sexual e outros considerados como vulneráveis. Dessa forma, a análise preditiva sempre corre o risco de violar o princípio constitucional da não discriminação previsto no art.3, IV da Constituição Federal.

Equilibrar diferentes medidas estatísticas de justiça exige que os tribunais e os desenvolvedores dos sistemas de decisão escolham que tipo de justiça é mais importante: previsão precisa ou equalização de falsos positivos e falsos negativos entre os diferentes grupos. Como visto, a resposta para essa pergunta vai determinar o ponto de vista sob o qual serão analisados os resultados e se estes dados serão justos ou não.

No entanto, há sempre o risco de tratamento discriminatório em relação a grupos vulneráveis. Dessa forma, como tal escolha é inevitável, ela deve ser feita de forma explícita, de modo que os envolvidos, e seus advogados, possam ser capazes de analisar a equidade do modelo, bem como os critérios usados para medir a justiça.

Além disso, quando se trata de modelos de decisões autônomos utilizados pelo setor público, deve-se garantir ampla transparência de modo a garantir que os setores possam decidir quais modelos adotar, ou se devem sequer adotá-los.

2.2.1.2Segunda camada: qualidade dos dados

[66]BARROCAS, Solon & SELBST, Andrew D., “Big Data's Disparate Impact”, 104 CALIF. L. REV. 671 (2016).

Existem duas maneiras principais pelas quais o preconceito aparece nos dados de treinamento: ou os dados coletados não são representativos da realidade ou refletem os preconceitos existentes.

No primeiro caso, o problema advém do fato de que o *software* não tem, em sua base de treinamento, dados de pessoas representantes de determinado grupo vulnerável, o que pode resultar tanto da inexistência de dados disponíveis que representem essa população, quanto da negligência dos programadores em representá-la em seus algoritmos. Muito se falou, por exemplo, em *softwares* de reconhecimento facial que não reconhecem pessoas negras[67].Quando se trata de policiamento, pode-se apontar o caso da falta de pesquisas sobre crime de colarinho branco que são limitadasdevido ao escopo e à natureza dos crimes,que estão em constante evolução, de forma que os estudos e pesquisas realizados acabam por apresentar enfoque muito estreito e as reclamações, quando realizadas, não chegam ou não são investigadas pela polícia.

No entanto, os estudos disponíveis estimam que aproximadamente 49% das empresas e 25% das famílias foram vítimas de crimes de colarinho branco, em comparação com uma taxa de prevalência de 1,06% para crimes violentos e taxa de prevalência de 7,37% para crimes contra a propriedade[68]. Assim, embora exista uma necessidade significativa de mais pesquisas sobre crimes relacionados à corrupção, os dados disponíveis demonstram que elessão mais frequentes do que os crimes tradicionalmente visados pelos departamentos de polícia e policiamento, como propriedade e crimes violentos.

---

[67]BOULAMWINI, Joy, GenderShades. Disponível em http://gendershades.org/overview.html

[68]PWC, PULLING FRAUD OUT OF THE SHADOWS 5 (2018), https://www.pwc.com/gx/en/forensics/global-economiccrime-and-fraud-survey-2018.pdf; Gerald Cliff & April Wall-Parker, Statistical Analysis of White-Collar Crime, OXFORD RES. ENCYCLOPEDIA CRIMINOLOGY 7 (Apr. 2017), http://oxfordre.com/criminology/view/10.1093/acrefore/9780190264079.001.0001/acrefore-9780190264079-e267?print=pdf; BUREAU OF JUSTICE STATISTICS, U.S. DEP'T OF JUSTICE, CRIMINAL VICTIMIZATION, 2016: REVISITED 11 (2018), https://www.bjs.gov/content/pub/pdf/cv16re.pdf; RODNEY HUFF ET AL., NAT'L WHITE COLLAR CRIME CTR., NATIONAL PUBLIC SURVEY ON WHITE COLLAR CRIME: 2010 14 (2010), https://www.nw3c.org/docs/research/2010-national-public-survey-on-white-collar-crime.pdf?sfvrsn=e51bbb5d_8.

Na segunda hipótese, o *software* foi treinado com dados de pessoas presas anteriormente e, por isso, contêm os viéses racistas do sistema policial e judicial, como ocorreu no caso do sistema denominado PredPol[69].

Trata-se, portanto, da ocorrência de distorções no resultado apontado pelo algoritmo devido à falta de dados de determinado grupo, por um lado, ou pelo excesso de dados, por outro. Os sistemas de inteligência artificial são projetados por humanos e começam a aprender a partir de uma base de dados fornecida também por humanos. Os programadores, naturalmente, não são pessoas neutras e destituídas de valores, e, inadvertidamente, podem acabar transferindo preconceitos naturais a esses sistemas. Isso porque a própria seleção dos dados que alimentarão a máquina é uma atividade subjetiva.

Dessa forma, a análise de dados tem o potencial de prejudicar, de forma indevida ou excessiva, grupos considerados como vulneráveis ao colocá-los sistematicamente em posição de relativa desvantagem. O grande problema é que as distorções geradas pela mineração de dados não estão ligadas diretamente às atuações humanas, visto que são os dados que contêm as distorções e não o algoritmo em si. Dessa forma, cria-se a impressão de que o resultado enviesado se encontra livre de preconceitos humanos sendo, portanto, justo.

A discriminação é um efeito dessa decisão enviesada, que cria um *feedback loop*, que é a concretização da estigmatização de grupos vulneráveis, solidificando sua posição de vulnerabilidade na sociedade, como bem exemplifica Cathy O`Neil ao falar sobre a coleta de dados pela polícia americana.

> Isso cria um *feedback loop* destrutivo. O próprio policiamento cria novos dados, que justificam mais policiamento. E nossas prisões ficam lotadas com centenas de milhares de pessoas condenadas por crimes sem vítimas, ou seja, sem grande lesividade. A maioria delas vem de bairros empobrecidos e são, em sua maioria, negros ou latinos. Então, mesmo que o modelo seja indiferente à cor, o resultado é

[69]https://www.newscientist.com/article/mg23631464-300-biased-policing-is-made-worse-by-errors-in-pre-crime-algorithms/

> tudo menos isso. Nas nossas cidades segregadas a geografia é um dado aproximado muito eficiente para raça[70].

Além disso, o *feedback loop* acaba por reforçar visões e estereótipos preconceituosos quando as pessoas observam o aumento da presença policial em comunidades marginalizadas.

> A exposição continuada ou o reforço desses estereótipos, especialmente na ausência de uma contra-tradição, podem permitir à sociedade manter uma preconceito contra grupos marginalizados, mantendo, ao mesmo tempo, um compromisso explícito igualitarismo[71].

Um estudo recente feito por pesquisadores do *AI NowInstitute*, um centro de pesquisa que estuda o impacto social da inteligência artificial, demonstrou que os sistemas de policiamento preditivo, em numerosas jurisdições, são construídos com base em dados produzidos no contexto de práticas errôneas, racistas e, às vezes, ilegais (o qual eles denominaram de "policiamento sujo")[72]. Isso pode incluir manipulação sistêmica de dados, falsificação de relatórios policiais, uso ilegal de força, evidências plantadas e buscas inconstitucionais. Tais práticas de policiamento moldam o ambiente e a metodologia pela qual os dados são criados, o que leva a imprecisões, distorções e formas de viéses sistêmicos que ficam embutidos nos dados (denominados de "dados sujos"[73]).

---

[70]No original. "Thiscreates a pernicious feedback loop. The policing itself spawns new data, which justifies more policing. And our prisons fill up with hundreds of thousands of people found guilty of victimless crimes. Most of them come from impoverished neighborhoods, and most are Black or Hispanic. So even IF the model is color blind, the result of it is anything but. In our largely segregated cities, geography is a highly effective proxy for race."O'NEIL, Cathy . Op cit. p.87

[71]No original "Indeed, continued exposure or reinforcement of these stereotypes, especially in the absence of a counternarrative, can allow society to maintain a prejudice against marginalized groups while still maintaining an explicit commitment to egalitarianism". Richardson, Rashida and Schultz, Jason and Crawford, Kate, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (February 13, 2019). New York University Law Review Online, Forthcoming. Disponível em SSRN: https://ssrn.com/abstract=. p.22

[72]Richardson, Rashida and Schultz, Jason and Crawford, Kate, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (February 13, 2019). New York University Law Review Online, Forthcoming. Disponível em SSRN: https://ssrn.com/abstract=

[73]No original Dirty Data. "Dados sujos" é um termo comumente usado na comunidade de pesquisa de mineração de dados para se referir a "Dados ausentes, dados errados e representações não padronizadas dos mesmos dados." Os autores expandiram o termo para incluir também dados derivados de ouinfluenciados por práticas corruptas, tendenciosas e ilegais, incluindo, ainda, dados que tenham sido intencionalmente manipulados ou "eliminados", bem como dados distorcidos por preconceitos individuais e sociais. Essa nova categoria de dados sujos ainda inclui dados gerados a partir da prisão de pessoas inocentesque tinha provas plantadas sobre eles ou foram falsamente acusados, além de chamadas paraserviços ou relatórios de incidentes que refletem falsas alegações de atividade criminosa. Por fim, dados sujosincorpora usos subsequentes que distorcem ainda mais os

Como bem resume Kate Crawford, cofundadora e co-diretora da *AI Now* e autora do estudo "Um sistema algorítmico é tão bom quanto os dados que você usa para treiná-lo":

> Se os dados em si estiverem incorretos, isso fará com que mais recursos policiais sejam concentrados nas mesmas comunidades superestimadas e muitas vezes racialmente direcionadas. Então, o que você fez é na verdade um tipo de lavagem de tecnologia onde as pessoas que usam esses sistemas assumem que elas são de alguma forma mais neutras ou objetivas, mas na verdade elas têm uma forma de inconstitucionalidade ou ilegalidade[74].

Os pesquisadores examinaram 13 jurisdições, concentrando-se naquelas que usaram sistemas de policiamento preditivo e foram submetidas a uma investigação comissionada pelo governo. O último requisito assegurou que as práticas de policiamento tivessem documentação legalmente verificável. Em nove das jurisdições, eles encontraram fortes evidências de que os sistemas haviam sido treinados em "dados sujos".

O problema não era apenas dados distorcidos pelo direcionamento desproporcional de minorias, como em Nova Orleans. Em alguns casos, os departamentos de polícia tinham uma cultura de propositalmente manipular ou falsificar dados, sob intensa pressão política para reduzir as taxas oficiais de criminalidade. Em Nova York, por exemplo, para desinflar artificialmente as estatísticas do crime, os comandantes das delegacias costumavam pedir às vítimas nas cenas de crimes que não apresentassem queixas. Alguns policiais chegaram a plantar drogas em pessoas inocentes para cumprir suas cotas de prisão. Nos atuais sistemas de policiamento preditivo, que dependem do aprendizado de máquina para prever o crime, esses dados corrompidos se tornam preditores legítimos.

Dessa forma, os sistemas preditivos de policiamento treinados por tais dados não podem escapar do legado de práticas de policiamento ilegais ou preconceituosas em cima dos quais são construídos.

---

registros policiais, como a manipulação sistêmicaestatísticas criminais para tentar promover relações públicas específicas, financiamento ou resultados políticos.

[74]Richardson, Rashida and Schultz, Jason and Crawford, Kate, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice (February 13, 2019). New York University Law Review Online, Forthcoming. Disponível em SSRN: https://ssrn.com/abstract=

Voltando ao caso do COMPAS, o algoritmo utilizava como parâmetro para decisão os dados sobre prisões do indivíduo para medir seu nível de risco. Ao utilizar tal dado, no entanto, presume-se que os indivíduos que cometem os mesmos delitos são presos nas mesmas proporções.

> Contudo, há muitas evidências de que as pessoas de cor, especialmente os negros, são mais propensos a serem presos do que os brancos pelo mesmo comportamento. Americanos negros são desproporcionalmente parados e revistados pela polícia, quer estejam dirigindo ou caminhando[75].

Do ponto de vista estatístico, o algoritmo não estaria errado ao apontar um indivíduo negro como de maior risco, no entanto, o mesmo não pode ser dito do ponto de vista ético e legal visto que os dados utilizados não são adequados como parâmetro para medir o verdadeiro risco de um indivíduo porque se baseiam em preconceitos.

Como bem apontam Eckhouseetal[76], o mesmo vale para outras medidas extraídas do sistema de justiça criminal. Réus de cor, especialmente homens negros e latinos, são tratados com mais rigor ao longo de todo processo criminal: são processados em maior número de vezes; há menor probabilidade de que lhes seja oferecido qualquer tipo de programa de apoio; e em geral são condenados a penas mais longas. Conforme apontado, o sistema criminal tem um histórico de sistematicamente ter como alvo grupos vulneráveis em razão de sua raça e não em razão de comportamento.

---

[75]No original "However, there is plenty of evidence that people of color, especially Black people, are more likely to be arrested than Whites for the exact same behavior. Black Americans are disproportionately likely to be stopped and searched by police, whether they are driving or walking". Eckhouse, L. *et al.* (2019) 'Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment', *Criminal Justice and Behavior*, 46(2), p.196

[76]No original "The same is true for other measures drawn from the criminal justice system. Defendants of color, especially Black and Latino men, are treated more harshly throughout the criminal justice system; disproportionately prosecuted; less likely to be offered pretrial diversion, counseling, and other supportive programming; and sentenced to longer terms. From arrest to conviction to sentencing to recidivism, criminal history is a measure of criminal justice practices that systematically targets race-class subjugated communities, not just a measure of individual behavior". Eckhouse, L. *et al.* (2019) 'Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment', *Criminal Justice and Behavior*, 46(2), p.197

Tais problemas sistêmicos nos dados não podem ser medidos com técnicas de estatística e tampouco resolvidos pelo próprio algoritmos tendo em vista que está-se pedindo que este tome decisões sobre riscos futuros com base em distorções ocorridas no passado. O sistema não está decidindo sobre comportamentos individuais, mas sim um evento determinado pelas ações anteriormente tomadas pela polícia.

2.2.1.3Terceira camada: problemas conceituais da utilização de decisões autônomas

Sistemas de decisão autônomos podem e são utilizados para substituir ou auxiliar a decisão humana em diversas esferas da sociedade. Algumas decisões são inocentes e não representam maiores riscos, como no caso de anúncios nas redes sociais. No entanto, quando se trata da esfera pública, em especial de decisões que geram consequências graves para o ser humano, como no caso de decisões judiciais ou de policiamento, há a necessidade de se analisar mais uma questão para se saber se essa decisão gera algum tipo de discriminação. É o que será analisado nesse tópico.

Mesmo que um modelo de avaliação de risco seja estatisticamente justo e baseado emdados, ainda persiste o problema fundamental de que tal sistema avalia o risco de um indivíduo usando dados sobre outras pessoas, mais especificamente sobre grupos de pessoas. O instrumento de avaliação de risco analisa informações sobre um grupo de pessoas, que não inclui o indivíduo analisado, e fornece uma pontuação com base no comportamento dos outros. Seria justo que uma pessoa seja alvo de uma decisão que a incluiu em determinado grupo e decidiu com base em informações desse grupo?

Nesse sentido, seria o mesmo que dizer que o risco atribuído a um indivíduo não se baseia em suas ações pessoais, mas sim nas ações em geral realizadas pelo grupo no qual ele foi incluído. Isso ocorre porque, ainda que a pontuação seja baseada em suas histórias

pessoais, o modelo em si e as pontuações que oferece são calibradas com base no comportamento passado de outras pessoas.

O caso fica mais grave quando se analisa a questão criminal. Do ponto de vista constitucional, o indivíduo só pode ser condenadocom base nas suas próprias ações. Dessa forma, uma decisão não poderia se basear em dados relativos a pessoas que compartilham características sociais, demográficas, ou afiliações de grupos geográficos com o indivíduo acusado.

Embora modelos estatísticos que usam médias baseadas em grupos possam produzir previsões mais precisas do que decisões humanas, tomadas com informações ou critérios inconsistentes, tal fato gera a violação de garantias constitucionais, que determinam que as pessoas devem ser tratadas como indivíduos e não com base em características como gênero, raça, religião, orientação sexual, entre outros.

Decisões autônomas que analisam risco invariavelmente incluem tais informações na sua análise, como bem explica Eckhouse, L. *et al*:

> Os fatores específicos mais frequentemente considerados exacerbam esse problema. O fato é que o risco hoje passou a significar histórico criminal, e este tornou-se um substituto para a raça. A inclusão de variáveis socioeconômicas agravam esse problema, pois as avaliações de risco tratam a raça e desigualdades de classe como uma qualidade pessoal que torna um acusado como de maior risco[77].

Alguns estudiosos apontam como possíveis soluções, que devem ser utilizadas conjuntamente[78]:

- Controlar as distorções dos dados utilizados no treinamento da máquina;

[77]No original "The specific factors most often considered exacerbate this problem. As Harcourt (2015) puts it, "The fact is, risk today has collapsed into prior criminal history, and prior criminal history has become a proxy for race" (p. 237). The inclusion of socioeconomic variables worsens this problem, as risk assessments then treat race and class inequality as a personal quality that makes a defendant riskier."Eckhouse, L. *et al*. (2019) 'Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment', *Criminal Justice and Behavior*, 46(2), p. 199

[78]MITTELSTADT, Brent Daniel. "The ethics of algorithms: maping the debate". Big Data & Society. July-December 2016

- Integração de critérios anti-discriminatórios no classificador do algoritmo;
- Pós-verificação dos modelos de classificação;
- Modificação das predições e decisões para manter uma proporção justa dos efeitos entre grupos protegidos e não protegidos.

No entanto, embora já haja diversos esforços no sentido de combater os problemas gerados pelos viéses, há que se questionar se este é verdadeiramente o caminho adequado. Possivelmente, apenas abordar os viéses gerados pela inteligência artificial não gere uma verdadeira solução para o problema. Como bem descreve Julia Powles:

> Há três problemas com esse foco no A.I. viés. A primeira é que o viés de endereçamento como um problema computacional obscurece suas causas raízes. O viés é um problema social, e procurar resolvê-lo dentro da lógica da automação sempre será inadequado.
> Em segundo lugar, até mesmo o sucesso aparente no enfrentamento do viés pode ter consequências perversas. Tomemos o exemplo de um sistema de reconhecimento facial que funciona mal em mulheres de cor devido à sub-representação do grupo tanto nos dados de treinamento quanto entre os projetistas de sistemas. Aliviar esse problema buscando "equalizar" a representação simplesmente coopta os projetistas no aperfeiçoamento de vastos instrumentos de vigilância e classificação.
> Quando as questões sistêmicas subjacentes permanecem fundamentalmente intocadas, os combatentes tendenciosos simplesmente tornam os humanos mais legíveis por máquina, expondo as minorias, em particular, a danos adicionais.
> Terceiro - e mais perigoso e urgente de todos - é o modo pelo qual a controvérsia sedutora de A.I. O viés, e o falso fascínio de "resolvê-lo", diminui as questões maiores e mais urgentes. “O preconceito é real, mas também é um desvio cativante”[79].

Pode haver uma inclinação natural para supor que os fornecedores de policiamento preditivo possam solucionar os problemas de dados sujos identificados neste artigo pela

[79]No original “There are three problems with this focus on A.I. bias. The first is that addressing bias as a computational problem obscures its root causes. Bias is a social problem, and seeking to solve it within the logic of automation is always going to be inadequate.Second, even apparent success in tackling bias can have perverse consequences. Take the example of a facial recognition system that works poorly on women of color because of the group’s underrepresentation both in the training data and among system designers. Alleviating this problem by seeking to “equalize” representation merely co-opts designers in perfecting vast instruments of surveillance and classification.When underlying systemic issues remain fundamentally untouched, the bias fighters simply render humans more machine readable, exposing minorities in particular to additional harms.Third—and most dangerous and urgent of all—is the way in which the seductive controversy of A.I. bias, and the false allure of “solving” it, detracts from bigger, more pressing questions. Bias is real, but it’s also a captivating diversion.”POWELS, JULIA. The Seductive Diversion of ‘Solving’ Bias in Artificial Intelligence.Disponívelemhttps://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

remoção de casos conhecidos. Mas os métodos são inadequados por várias razões. Primeiro, já que há poucos incentivos e quase não há requisitos para que os departamentos de polícia possam monitorar e reformar práticas ou políticas que criem dados tendenciosos ou sujos, é improvável que os departamentos de polícia identifiquem esses problemas ao fornecedor. Em segundo lugar, não há metodologia ou mecanismos atuais para identificar essas práticas e políticas problemáticas em tempo real; portanto, é impossível distinguir problemas substanciados e suspeitos.

Em terceiro lugar, como argumentado anteriormente, uma falha fundamental de dados policiais é que ele não captura todas as informações relevantes sobre crimes por causa de políticas ou práticas que ignoram certos tipos de crimes ou criminosos, relações comunitárias negativas que afetam os crimes que a polícia acompanha e práticas corruptas ou antiéticas que levam à omissão ou manipulação de registros policiais.

Além disso, ainda que se retire do sistema dados relacionados à raça (*color blind*), baseando-se em dados que buscam responder o tipo de crime, a localização e a data e horário, não importando a pessoa envolvida, o sistema ainda gera discriminações, como apontado por Cathy O'Neil:

> Quando a polícia instala seu sistema Predpol eles têm uma escolha. Eles podem focar exclusivamente em crimes denominados como Parte 1. Esses são os crimes violentos, estando ai incluídos homicídio, incêndio criminoso e roubo, que são normalmente reportados aos policiais. Mas eles, ao invés, podem expandir o foco incluindo crimes denominados como Parte 2, dentre os quais estão gandaia, pedir esmola de forma agressiva e a venda e consumo de pequenas quantidades de drogas. Muitos desses crimes, denominados como meros incômodos, iriam permanecer como não registrados caso não houvesse um policiar no local para vê-los[80].

Como a própria autora bem afirma, a maioria desses crimes menores ocorre principalmente em áreas mais pobres de uma cidade. Ao incluí-los no sistema preditivo, a

[80]No original "When Police set up theirPredPol system they have a choice. They can focus exclusively on so-called Part 1 crimes. These are the violent crimes, including homicide, arson, and assault, which are usually reported to them. But they can also broaden the focus by including Part 2 crimes, including vagrancy, aggressive panhandling, and selling and consuming small quantities of drugs. Many of these "nuisance crimes would go unrecorded if a cop weren't there to seethem." . O'NEIL, Cathy. "Weapons of Math Destruction: How Big Data increases inequality and threatens democracy" Ed. Crown. New York. 2016. p.86

análise é comprometida e os policiais tendem a ser deslocados para essas áreas, onde eles têm maior probabilidade de prender um maior número de pessoas, ainda mais considerando que um modelo matemático indicou que naquela localidade tem maiores chances de ocorrer um crime.

Não há uma maneira significativa para um fornecedor ajustar o que é desconhecido ou não registrado. A ausência de dados é tão significativa quanto a sua criação,

Em vez disso, os esforços de mitigação devem ser focados no desenvolvimento de mecanismo para avaliar os danos inerentes ao uso de dados policiais históricos, bem como dados gerados após a implementação da reforma da coleta de dados da polícia e apoiada por fortes medidas de transparência e responsabilização.

Sendo assim, há um risco de violação de diversos direitos, dentre eles a vistoria de pessoas e casas sem um devido mandado judicial ou qualquer prova ou indício de um crime.

Deve-se considerar, nesse momento, dois pontos: que esses crimes de menor lesividade também são praticados por pessoas de maior poder aquisitivo, ainda que em menor escala; e que esses indivíduos em geral praticam outros tipos de crimes ainda mais lesivos para a sociedade em geral. No entanto, esses fatos não aparecem no sistema preditivo de forma que não haverá o policiamento devido.

Ainda que não haja distorções nos resultados preditivos, como no caso do sistema PredPol, o que não se pode afirmar visto que não se sabe com clareza como os algoritmos funcionam ou quais são os dados utilizados, há uma flagrante violação da privacidade dos cidadãos.Atualmente, grande parte do debate se concentrou na mecânica do próprio sistema e se ele pode ser projetado para produzir resultados matematicamente justos. Mas, como sustentado por JuliaPowles, possivelmente separar a questão do algoritmo do sistema social ao qual ele está conectado e incorporado não gerará os resultados esperados. Nós realmente

temos que reconhecer os limites desses tipos de tentativas matemáticas baseadas em cálculos para lidar com o preconceito.

Se os mecanismos sociais e políticos que geram dados sujos não forem reformados, essas ferramentas só farão mais mal do que bem. Uma vez que as pessoas reconheçam isso, então talvez o debate finalmente se transfira para maneiras pelas quais o aprendizado de máquina e outros avanços tecnológicos podem serutilizados,para realmente combater a causa raiz do crime, tais como problemas de pobreza, desemprego e moradia usando dados do governo de uma forma mais benéfica.

Enquanto isso não ocorre, deve-se examinar formas regulatórias, éticas e jurídicas, para trazer mais transparência, responsabilidade e supervisão para o uso de ferramentas de tomada de decisão automatizadas.

3. UMA ABORDAGEM ÉTICA PARA A INTELIGÊNCIA ARTIFICIAL

Como visto nos capítulos anteriores a inteligência artificial representa uma oportunidade de aumentar o bem-estar individual e coletivo. Mas, assim como decisões autônomas podem ajudar a melhorar nossa qualidade de vida através da medicina personalizada ou da prestação de serviços de saúde mais eficientes, bem como nos ajudar a atingir os objetivos de desenvolvimento sustentável, enfrentando a mudança climática e nos ajudando a fazer melhor uso dos recursos naturais, a tecnologia também pode agravar tratamentos discriminatórios a grupos vulneráveis.

Há, portanto, a necessidade de se debater qual seria o tipo de regulação adequada para limitar os efeitos danosos da inteligência artificial, sem, contudo impedir a inovação tecnológica. No entanto, não se pode falar em regulação sem antes discutir-se um norteamento ético adequado, sem o qual qualquer regulação jurídica adotada corre o risco de se tornar inócua.

Considerando essas dificuldades é necessário criar uma resposta ética de forma a garantir a aplicação justa da inteligência artificial de modo a se promover a confiança da sociedade na tecnologia.

A confiança, portanto, é um pré-requisito para desenvolver, implantar e usar Inteligência Artificial. Como bem explicado por Luciano Floridi:

> A aceitação pública e a adoção de tecnologias de IA só ocorrerão se os benefícios forem vistos como significativos e os riscos como potenciais, evitáveis, minimizáveis, ou pelo menos contra os quais alguém pode ser protegido, por meio de gerenciamento de risco (por exemplo, seguro) ou correção[81].

Sendo tal tecnologia indigna de confiança, consequências subversivas podem decorrer, minando sua implementação pela sociedade, a qual deixaria de usufruir dos diversos benefícios econômicos e sociais oferecidos pela IA.

Dessa forma, o objetivo deste capítulo, portanto, não é fazer oposição ao uso dessas ferramentas, cujo emprego crescente parece ser uma tendência irreversível, mas mapear os problemas éticos advindos do uso de decisões automatizadas e verificar uma abordagem ética adequada para inspirar o desenvolvimento confiável, a implantação e o uso da inteligência artificial.

3.1 DIRETRIZES ÉTICAS: UTILITARISMO E DEONTOLOGIA

[81]No original “Public acceptance and adoption of AI technologies will occur only if the benefits are seen as meaningful and risks as potential, yet preventable, minimisable, or at least something against which one can be protected, through risk management (e.g. insurance) or redressing.”. Floridi, L., Cowls, J., Beltrametti, M. et al. Minds&Machines (2018) 28: 689. Disponível em https://doi.org/10.1007/s11023-018-9482-5

Antes de adentrar na discussão sobre os problemas éticos levantados pela inteligência artificial e as decisões autônomas é importante definir uma diretriz ética que definirá toda a análise futura, visto que a mesma explicitará o que se pretende alcançar e quais são as prioridades nessa análise.

A ética integra o ramo da filosofia moral que se preocupa com as escolhas sobre o que deve ser feito, e pode ser compreendida, dentro da perspectiva tradicional francesa, atribuída durante o Período das Luzes, na célebre "Enciclopédia" de D'Alembert e Diderot, como a moral submetida à crítica da razão,ou seja, "o conjunto de teorias filosóficas, racionais e reflexivas, sobre as normas e os valores em que os homens deveriam acreditar e que eles deveriam realizar em suas ações".

Um dos embates mais tradicionais envolve as correntes teleológoca e a deontológica.

A ética teleológica é consequencialista, ou seja, uma boa ação é medida pelas suas consequências, sendo que o fim do homem é a felicidade. O que verdadeiramente importa não é saber se o indivíduo age conforme de acordo com um princípio ou dever, e sim se a sua ação foi capaz de produzir a maior soma possível de prazer. Trata-se de uma ética concreta, pois demonstra como se pode atingir a felicidade.

A ética teleológica pretende conduzir o ser humano à realização de uma vida feliz, com o máximo de prazer para o máximo de pessoas possíveis. Dessa forma, as ações serão boas se forem geradas em função de um *telos*, que significa finalidade, objetivo em grego, para a qual a ação procura atingir. A "Felicidade" é este fim a ser alcançado, e a vida moral se constitui na perseguição de tal objetivo. Uma das visões teleológicas mais conhecidas é o utilitarismo.

A visão utilitarista clássica tem como seu principal defensor Jeremy Bentham[82]que busca encontrar justificação nas consequências das ações, e não em máximas absolutas. Dessa forma, tal entendimento é caracterizado pelo que muitos autores chamam de consequencialismo.

A visão utilitarista clássica, portanto, tem como um dos seus fundamentos o conceito de ato consequencialista que atrela o valor moral de toda ação a seus resultados, bons ou ruins, analisando a felicidade ou bem-estar geral que ela produz em uma perspectiva social. Dessa forma, um ato é moralmente correto quando maximiza o bem, isto é, quando o valor total de bem gerado para todos for maior do que a quantidade total de mau geral gerando, portanto, um saldo líquido positivo.

Quando se trata de decisões éticas, o exemplo mais comum é o caso do bonde desgovernado no qual dá-se a escolha para o maquinista de escolher entre matar cinco pessoas que estão nos trilhos na frente do bonde desgovernado, ou de alterar o curso do bonde e matar apenas uma pessoa. Pela lógica utilitarista descrita acima a resposta seria ade matar uma única pessoa para salvar as cincos, de modo a se garantir a maximização da felicidade e o menor custo social.

Como bem descreve John Ralws:

> Os termos apropriados de cooperação social são resolvidos pela circunstância, qualquer que seja, que alcance a maior soma de satisfação dos desejos racionais dos indivíduos. É impossível negar a plausibilidade inicial e a atratividade dessa concepção. A característica marcante da visão utilitarista da justiça é que não importa, exceto indiretamente, como essa soma de satisfações é distribuída entre os indivíduos, da mesma forma que não importa, exceto indiretamente, como um homem distribui suas satisfações ao longo do tempo. A distribuição correta em ambos os casos é aquela que produza a maior satisfação[83].

[82]BENTHAM, Jeremy. An Introduction to the Principles and Morals and Legislation.

[83]No original “The appropriate terms of social cooperation are settled by whatever in the circumstances will achieve the greatest sum of satisfaction of the rational desires of individuals. It is impossible to deny the initial plausibility and attractiveness of this conception. The striking feature of the utilitarian view of justice is that it does not matter, except indirectly, how this sum of satisfactions is distributed among individuals any more than it matters, except indirectly, how one man distributes his satisfactions over time. The correct distribution in either case is that which yields the maximum fulfillment.”. RAWLS, John. A theory of justice.P.23. Disponível em

Tal abordagem opõem-se à ótica deontológica que considera que deve-se observar certos direitos e deveres independente das consequências.

Esses dois entendimentos, portanto, ilustram duas abordagens opostas de justiça. A primeira diz que a moral de uma ação depende unicamente das consequências que ela acarreta de modo que a coisa certa a fazer é aquela que produzirá os melhores resultados, considerando-se todos os aspectos.

A segunda abordagem, por sua vez, diz que as consequências não são tudo com o que devemos nos preocupar, moralmente falando: devemos observar certos deveres e direitos por razões que não dependem das consequências sociais de nossos atos[84].

Recentemente, o dilema do bonde desgovernado foi adaptado pelo *Massachusetts Institute of Technology* (MIT) imaginando-se uma situação envolvendo um carro autônomo[85]. Tal experimento, denominado *Moral Machine*[86], demonstrou que, em uma primeira análise do estudo as pessoas tendem a optar por uma solução utilitarista, salvando o maior número de pessoas possível, de modo que parece haver uma predominância de valores e preferências. No entanto, em uma análise mais detalhada, verifica-se que há uma variação nas escolhas morais a depender da localização e da cultura, o que gera um desafio para a criação de princípios éticos que tenham um caráter universal.

Além disso, ainda que tenham optado pela visão utilitarista, as pessoas opinaram que não comprariam um automóvel com tal visão, visto que isso colocaria a vida do motorista em risco[87].

---

http://www.consiglio.regione.campania.it/cms/CM_PORTALE_CRC/servlet/Docs?dir=docs_biblio&file=BiblioContenuto_3641.pdf

[84]MICHAEL, Sandel J. "Justiça: O que é fazer a coisa certa". 22 ed. Rio de Janeiro. Civilização Brasileira, 2016.

[85]http://moralmachine.mit.edu/

[86]"(…) we designed the Moral Machine, a multilingual online 'serious game' for collecting large-scale data on how citizens would want autonomous vehicles to solve moral dilemmas in the context of unavoidable accidents. The Moral Machine attracted worldwide attention, and allowed us to collect 39.61 million decisions from 233 countries, dependencies, or territories.".Disponível em https://www.nature.com/articles/s41586-018-0637-6

[87]https://brasil.elpais.com/brasil/2016/06/22/ciencia/1466610816_591801.html

Dessa forma, embora a visão utilitarista seja por vezes atraente, pretende-se aqui expor algumas objeções à sua utilização como visão ética predominante quando se trata de decisões autônomas.

A primeira e mais óbvia objeção ao utilitarismo é de que ele não respeita os direitos individuais. Embora o indivíduo tenha importância para o utilitarismo ela se dá apenas na medida em que as preferências de cada um forem consideradas em conjunto com as de todos os demais. Segundo John Rawls, em seu livro Teoria da Justiça, tal visão permite a ocorrência de situações injustas, visto que não leva em consideração os direitos de cada indivíduo ou o que seria considerado como justo.

Em outras palavras, seria aceitável pela ótica utilitarista sacrificar direitos considerados como fundamentais em prol da maximização da felicidade. Logo, utilizando-se de sua lógica, seria possível aceitar que decisões autônomas fossem prejudiciais a grupos vulneráveis, caso tal contribuísse para a felicidade da sociedade como um todo. Ainda segundo Rawls.

> Assim, não há razão em princípio para que os ganhos maiores de alguns não compensem as perdas menores de outros; ou, mais importante, por que a violação da liberdade de alguns pode não ser corrigida pelo bem maior compartilhado por muitos[88].

No entanto, quando aplicada em casos concretos, a violação de princípios fundamentais de indivíduos sob a justificativa de assegurar a felicidade de uma maioria não é moralmente aceita pela maioria das pessoas[89], como ocorreu no exemplo dos carros autônomos.

A segunda objeção ao utilitarismo diz respeito à utilização de diferentes valores como medida comum sem distinguir pesos ou valores, como se todos tivessem a mesma natureza.

---

[88]No original "Thus there is no reason in principle why the greater gains of some should not compensate for the lesser losses of others; or more importantly, why the violation of the liberty of a few might not be made right by the greater good shared by many."RAWLS, John. A theory of justice.P.23. Op cit.

[89]SANDEL, Michael. *Justiça.*O que é fazer a coisa certa?. Civilização Brasileira. Rio de janeiro, 2009

Benthan[90]criou o conceito de utilidade para capturar, em uma única escala, a natureza discrepante de valores importantes para o ser humano, incluindo a sua vida, reduzindo tudo a uma única escala única de prazer e dor. No entanto, críticos do utilitarismo apontam que seria moralmente errado atribuir valor monetário à vida humana.

John Stuart Mill tenta, em seus trabalhos, conciliar os direitos dos indivíduos com a filosofia utilitarista, de modo a oferecer uma resposta às objeções apresentadas. No entanto, suas respostas acabam por se apoiar em ideias morais independentes da utilidade[91].

Considerando que todos os seres humanos são dignos de respeito seria errado trata-los como meros instrumentos para a felicidade coletiva[92]. Dessa forma, diante do potencial das decisões autônomas não apenas como violadoras da privacidade e das liberdades dos indivíduos, mas principalmente pelo seu potencial discriminatório, não seria aceitável a adoção de uma prática utilitarista como diretriz ética para sua utilização. Tal não contribuiria para uma regulação baseada em direitos e tampouco para o ganho da confiança dos cidadãos que serão seu alvo.

A teoria deontológica, por sua vez, foca-se na ação do agente e não em suas consequências. Conforme bem explica Eduardo Magrani:

> A deontologia enquadra-se no domínio das teorias morais que orientam e avaliam o que devemos fazer e de modo diverso das teorias utilitaristas, essas julgam a moralidade das escolhas individualmente, por um parâmetro não orientado pelos resultados[93].

Tal teoria é bem apresentada pelas ideias de Immanuel Kant, segundo o qual a moral não diz respeito ao aumento da felicidade ou a qualquer outra finalidade: ela está atrelada ao

[90]BENTHAM, Jeremy. An Introduction to the Principles and Morals and Legislation.
[91]MILL. John Stuart. Utilitarismo. Disponível em https://www.passeidireto.com/arquivo/16842751/livro-utilitarismo-de-stuart-mill
[92]MILL. John Stuart. Utilitarismo. Disponível em https://www.passeidireto.com/arquivo/16842751/livro-utilitarismo-de-stuart-mill
[93]MAGRANI, Eduardo.A Internet das Coisas: Privacidade e Ética na Era da Hiperconectividade. Tese de doutora. Puc-Rio.2018

respeito às pessoas como um fim em si mesmo[94]. Segundo Kant,a moralidade não deve se basear apenas em considerações empíricas como desejos, vontades e preferências pessoais, já que estes fatores são variáveis. A base da moralidade deve ser tal que possibilite a fundamentação de princípios morais universais, ou seja, direitos humanos universais. Para Kant esta base seria a pura razão prática.

O valor moral de uma ação não está em suas consequências, mas sim na intenção com que ela é realizada. Dessa forma, o que importa é fazer a coisa certa justamente porque é a coisa correta, e não por qualquer outro motivo exterior. Todo ser humano, segundo Kant[95], possui uma razão prática que diz sempre o que é certo e errado moralmente.

Dessa forma, toda ação é governada por algum tipo de lei. E, se nossas ações fossem governadas apenas pelas leis da física, não seríamos diferentes daquela bola de bilhar do exemplo que vimos.

Assim, se somos capazes de ser livres, devemos ser capazes de agir não apenas de acordo com uma lei que nos tenha sido dada ou imposta, mas de acordo com uma lei que outorgamos a nós mesmo. Mas de onde viria essa lei?

A resposta de Kant[96]: da razão. Não somos apenas seres sencientes, que obedecem aos estímulos de dor e prazer que recebemos dos nossos sentidos; somos também seres racionais, capazes de pensar. E, se a razão determina minha vontade, então a vontade torna-se o poder de escolher independentemente dos ditames da natureza ou da inclinação.

A razão pode comandar a vontade de duas formas distintas, estabelecendo Kant dois tipos de imperativo[97]: o primeiro seria o hipotético, que utiliza a razão instrumental. Ou seja, trata-se de uma vontade condicionada à um fator ou vontade externa; o segundo seria o

[94]KANT, Immanuel. Fundamentação da Metafísica dos Costumes (Grundlegung zur Metaphysik der Sitten, 1785). Trad: Paulo Quintela: Edições 70, 2008
[95]KANT, Immanuel. Fundamentação da Metafísica dos Costumes. Op Cit
[96]KANT, Immanuel. Fundamentação da Metafísica dos Costumes. Op Cit
[97]KANT, Immanuel. Fundamentação da Metafísica dos Costumes. Op Cit

categórico, ou incondicional, que é aquele que prevalece sem referência a nenhum outro propósito.

O imperativo categórico é apresentado por Kant em três diferentes formulações[98]:

A primeira versão é a lei universal: "Aja apenas segundo um determinado princípio que, na sua opinião, deveria constituir uma lei universal."Como se a máxima de tua ação devesse tornar-se, através da tua vontade, uma lei universal." Isso significa que só devemos agir de acordo com princípios que podemos universalizar, sem entrar em contradição. Trata-se de um teste para verificar se a ação que se pretende tomar coloca os interesses do agente acima dos de qualquer outra pessoa.

A segunda formulação determina que deve-se tratar o ser humano sempre como um fim em si mesmo: "Aja de forma que uses a humanidade, seja na sua pessoa, seja na pessoa de outrem, nunca como um simples meio, mas sempre ao mesmo tempo como fim."

Para Kant, portanto, a justiça nos obriga a preservar os direitos humanos de todos, independentemente de onde vivam ou do grau de conhecimento que temos deles, simplesmente porque são seres humanos, seres racionais e, sendo assim, merecedores de respeito[99].

Dessa forma, quando fala-se em decisões autônomas tomadas por sistemas dotados de inteligência artificial, a diretriz deontológica parece mais adequada, tendo em vista que determina o respeito ao ser humano como um fim em si mesmo. Partindo-se dessa premissa é possível justificar eticamente a não aceitação de decisões autônomas que geram prejuízos para determinados grupos ou indivíduos, como visto anteriormente.

Além disso, a visão deontológica permite se pensar em princípios éticos fundamentais que regulariam o desenvolvimento e utilização da inteligência artificial, sempre voltados para uma perspectiva centrada no ser humano.

---

[98]KANT, Immanuel. Fundamentação da Metafísica dos Costumes. Op Cit
[99]KANT, Immanuel. Fundamentação da Metafísica dos Costumes. Op Cit

3.2 DIFICULDADES DE APLICAÇÃO ÉTICA NAS MÁQUINAS

Embora cada vez mais cresça a confiança na IA, não se pode esperar que um sistema funcione por si só sem erros. Deve haver uma constante interferência humana para cada erro de atuação. Como visto anteriormente, um viés social geralmente é levado para o desenvolvimento de produtos, gerando a sedimentação de tratamentos discriminatórios à grupos vulneráveis.

Dessa forma,o art. 25 da GDPR aborda a proteção de dados desde a concepção e, por *default*, denominado internacionalmente como *data protection by design and default*. Segundo este artigo, deve o responsável pelo tratamento de dados adotar, tanto no momento de definição dos meios de tratamento como no do próprio tratamento, medidas técnicas e organizativas adequadas, a fim de aplicar os princípios da proteção de dados de forma eficaz, além de incluir as garantias necessárias no tratamento.

Por *default*, os programadores ainda devem garantir que apenas dados pessoais necessários para cada finalidade específica sejam tratados. Por sua vez, a proteção *by design*, determina que o princípio fundamental da privacidade deve ser aplicado em todo o processo de desenvolvimento de um sistema.

Geralmente, a ideia de desenho está relacionada diretamente com a forma do produto. No entanto, a sua definição é mais ampla, devendo incluir a questão ética durante a produção e avaliação do impacto do produto após o lançamento. Ou seja, todo o contexto que o produto abrange deve ser considerado[100].

Como bem descreve Eduardo Magrani:

[100]Victoria Sgarro. “What Are ‘Ethics in Design’?”. Slate. Disponível em: https://slate.com/technology/2018/08/ethics-in-design-what-exactly-does-that-mean.html.

> A fim de garantir a aplicação da ideia inerente a tal tipo de proteção, Jaap-HenkHoepman elenca algumas estratégias de proteção da privacidade by design - algumas muito similares às previstas na GDPR. São elas: (i) minimizar, estratégia pela qual a quantidade de dados processados deve ser a mínima possível; (ii) ocultar, de modo que qualquer dado pessoal deve ser ocultado da plain view; (iii) separar, de forma que dados pessoais sejam processados em compartimentos separados sempre que possível; (iv) agregar, fazendo com que os dados pessoais sejam tratados ao mais alto nível de agregação e com o mínimo detalhe possível em que (ainda) seja útil; (v) estratégias orientadas, de modo que deve-se informar sempre que dados pessoais forem processados; (vi) controle, estratégia pela qual "data subject sshould be provided agency over the processing of their personal data"; (vii) enforce, de forma que uma política de privacidade compatível com requisitos legais exista e seja aplicada e (viii) demonstrar, isto é, ser capaz de demonstrar conformidade com a política de privacidade de quaisquer requisitos legais[101].

Contudo, tomar como providencia a constante observação e intervenção sobre as ações de uma IA quebra o propósito dela ser um sistema totalmente autônomo. Embora a realização de alguns ajustes possa ser uma ação eficaz momentaneamente, não é possível fazer reparos constantes.

Além disso, não há como mensurar todos os equívocos ocorridos. Para que a intervenção seja realmente eficiente seria necessário ensinar e estabelecer uma base de dados livre de qualquer preconceito, o que, como visto no capítulo anterior, trata-se de algo praticamente impossível de ser realizado.

Atendendo aos princípios éticos e jurídicos desde a concepção, é possível desenvolver algoritmos mais responsáveis. No entanto, ainda que o *designer* consiga prever e antecipar o surgimento de consequências negativas, dificilmente ele conseguirá acabar com esse problema. Embora a aplicação de princípios de privacidade desde a concepção, bem como mecanismos técnicos, auxiliem na diminuição do risco de danos ocorrerem, os vieses fazem parte da própria natureza humana de modo que sua total erradicação seria no mínimo difícil.

Dessa forma, torna-se necessário pensar em uma regulação jurídica não apenas para impedir tais vieses de ocorrem, mas também para analisar sobre quem deve controlar os

[101]MAGRANI, Eduardo.A Internet das Coisas: Privacidade e Ética na Era da Hiperconectividade. Tese de doutora. Puc-Rio.2018

dados, quais sistemas devem ou não ser construídos e em quais situações eles devem ser efetivamente aplicados.

3.3 PRINCÍPIOS ÉTICOS

Conforme visto, a inteligência artificial possui um potencial que pode beneficiar a sociedade e seus cidadãos. No entanto, como qualquer tecnologia inovadora, também possui uma potencialidade nociva que pode gerar a violação de valores e princípios fundamentais. Dessa forma é fundamental estipular valores éticos para gerir a utilização da IA.

Para que uma IA seja confiável ela deve possuir dois componentes: (1) seu desenvolvimento, implantação e uso devem respeitar os direitos e regulamentos aplicáveis, bem como princípios e valores fundamentais, assegurando um "propósito ético", e (2) deve ser tecnicamente robusta e confiável[102] visto que, mesmo com boas intenções ou propósitos, a falta de domínio tecnológico pode gerar danos não intencionais.

Além disso, o respeito pelos direitos fundamentais, princípios e valores devem ser aplicados em todo o processo de desenvolvimento e implementação da IA.

Dessa forma, é necessário que os princípios éticos da inteligência artificial estejam embasados em princípios fundamentais. De acordo com o recente projeto de diretrizes éticas, desenvolvido pelo Grupo De Peritos de Alto Nível Sobre Inteligência Artificial da Comissão europeia (AI HLEG)[103], que utilizou como documento base a carta de direitos fundamentais da União Europeia, há cinco princípios fundamentais nos quais deve–se basear: dignidade, liberdade, igualdade, direitos dos cidadãos e justiça. Esses princípios estabelecem uma abordagem centrada no ser humano, proporcionando status de primazia nas esferas cível, política, econômica e social, além de gerar segurança jurídica no contexto ético.

[102]https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_draft_ethics_guidelines_18_december.pdf
[103]ai_hleg_draft_ethics_guidelines.Disponívelem:https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_draft_ethics_guidelines_18_december.pdf

A dignidade estabelece a ideia que de que todo ser humano possui um "Valor intrínseco", que nunca pode ser diminuído, comprometido ou reprimido por outros, muito menos por sistemas dotados de inteligência artificial. Além disso, tal princípio garante às pessoas o direito de serem respeitadas como indivíduos e não como fonte de dados. Dessa forma, a inteligência artificial deve ser desenvolvida de modo a proteger o ser humano em seu aspecto físico e moral, além de seu senso de identidade cultural e social.

A liberdade relacionada à IA refere-se, principalmente, à liberdade de escolha dos indivíduos, estando aí incluída a proteção a grupos vulneráveis de igualdade de oportunidade. Sendo assim, em relação aos sistemas de decisão autônomas,a liberdade do indivíduo requer a proteção contra a coerção direta ou indireta, vigilância, fraude ou manipulação, sejam por parte do Estado ou por entidades privadas.

No mesmo caminho que a liberdade, o dever de igualdade exige um tratamento igualitário de todos os seres humanos, independentemente de estarem em situação semelhante ou distinta. Sendo assim, a igualdade vai além do dever de não discriminação, que tolera o tratamento de indivíduos de forma desigual quando se trata de pessoas em situações distintas, desde que baseado em justificativas objetivas.

Em relação à aplicação de sistemas dotados de inteligência artificial a igualdade requer "que as mesmas regras sejam aplicadas a todos para o acesso à informação, dados, conhecimento, mercados, além de uma distribuição justa do valor gerado pelas tecnologias"[104].

Dessa forma, exige-se que haja transparência nas decisões automatizadas de modo que não haja desequilíbrio na relação entre desenvolvedores e usuário, seja em relação aos dados coletados, ou quanto ao resultado da decisão tomada pelo algoritmo.

[104]No original "that the same rules should apply for everyone to Access to information, data, knowledge, markets and a fair distribution of the value added being generated by technologies."ai_hleg_draft_ethics_guidelines. Disponível em: https://www.ospi.es/export/sites/ospi/documents/documentos/Tecnologias-habilitantes/AI_HLEG_Draft_Ethics_Guidelines.pdf

No setor público a inteligência artificial possui o potencial de melhorar a eficiência do governo na prestação de serviços públicos. Com essa possibilidade surgem também o dever e o direito dos cidadãos, de serem informados caso haja qualquer coleta ou tratamento de seus dados, de forma que possam sempre optar pela não aplicação (*opt out*). Além disso, a sociedade não deve ser sistematicamente sujeita às classificações (*scoring*) pelo governo, como vem ocorrendo na China[105].

Por fim, em relação à justiça, deve-se estabelecer a não interferência da IA em processos democráticos, de modo a garantir a pluralidade de valores e escolhas.Ainda visando garantir o dever de justiça, os sistemas dotados de inteligência artificial também devem incorporar um compromisso de cumprir leis e regulamentos, além de garantir ao indivíduo lesado um processo de revisão, feito por um ser humano, das decisões tomadas pela máquina[106].

Sendo assim, os direitos fundamentais fornecem a base para a formulação de princípios éticos. Esses princípios são normas abstratas de alto nível que os desenvolvedores, usuários e reguladores devem seguir para defender o propósito da IA humana e confiável. Os valores, por sua vez, fornecem orientações mais concretas sobre como defender os princípios éticos, ao mesmo tempo em que também sustentam os direitos fundamentais.

Muitas organizações públicas, privadas e civis inspiraram-se nos direitos fundamentais para produzir quadros éticos para a IA. Recentemente, o projeto do AI4People[107] pesquisou diversas cartas que estabeleciam princípios éticos para a inteligência artificial, resultando em uma análise de um total de 47 princípios[108]. De forma geral há um grande grau de coerência e

[105]https://www.technologyreview.com/s/611815/who-needs-democracy-when-you-have-data/

[106]Danielle Keats Citron & Frank Pasquale. “THE SCORED SOCIETY: DUE PROCESS FOR AUTOMATED PREDICTIONS”. Disponível em: https://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1318/89WLR0001.pdf

[107]L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations”, Minds and Machines 28(4): 689-707

[108]As cartas de princípios analisadas foram: “Asilomar AI Principles”, desenvolvida no “Future of Life Institute” (2017); “Montreal Declaration for Responsible AI”, desenvolvida pela Universidade de Montreal (2017), os

sobreposição entre os seis conjuntos de princípios. Além disso, conforme aponta Floridi, eles poderiam ser resumidos em quatro preceitos gerais da bioética: beneficência (definida como "fazer o bem"), não-maleficência (definida como "fazer nenhum dano"), autonomia (definida como" respeito pela autodeterminação e escolha dos indivíduos "), e justiça (definido como "tratamento justo e equitativo para todos")[109]. Estes quatro princípios foram atualizados pelo mesmo grupo para se adequar ao contexto trazido pela inteligência artificial. No entanto, eles não são exaustivos, havendo a necessidade da inclusão de um quinto princípio: o princípio da explicabilidade.

- O princípio da beneficência: Tal princípio pode ser expresso simplesmente pela determinação da inteligência artificial de sempre benéfica para o ser humano e, mais amplamente, para todo o planeta. Dessa forma sistemas dotados de inteligência artificial podem contribuir para uma sociedade justa, inclusiva e pacífica ao auxiliar no aumento da autonomia mental dos cidadãos e proporcionando uma distribuição mais igualitária de oportunidades econômicas, políticas e sociais;
- Princípio da não maleficência: Embora este princípio aparente apenas ser o contraponto do princípio anterior trata-se, sob o ponto de vista ético, de um princípio distinto, que alerta para o potencial nocivo da má ou excesso de utilização da inteligência artificial. Nos documentos analisados há uma preocupação ainda maior com a privacidade, que surge como princípio em cinco das seis cartas e como um dos direitos humanos no documento da IEEE.

princípios gerais da IEEE, segunda versão do "EthicallyAligned Design: A Vision for PrioritizingHumanWell-beingwithAutonomousandIntelligent Systems" (2017), os princípios éticos propostos pelo grupo europeu de ética na ciência e tecnologia da Comissão europeia (2018); os "fiveoverarchingprinciples for an AI code" do §417 do Comitê de Inteligência Artificial da Casa dos Lordes (2018); e os Princípios da"Partnershipon AI" (2018).

[109]L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018). Op Cit.

Tal preocupação com a privacidade e com o mal uso das tecnologias leva os documentos a levantarem a necessidade de responsabilização. No entanto, não fica claro se esse dever deve ser direcionado para a tecnologia em si ou para seus desenvolvedores;

- Princípio da autonomia: Trata-se da ideia de que indivíduos têm o direito de decidir por si mesmos sobre os tratamentos que devem ou não receber. Quando adotamos sistemas de inteligência artificial, o ser humano cede livremente parte de sua autonomia. Dessa forma, o princípio da autonomia determina que haja equilíbrio entre o poder de decisão que cedemos e o que mantemos para nós mesmos. Este princípio se encontra expressamente previsto em quatro dos seis documentos analisados, todos determinando que haja este equilíbrio de maneiras e graus distintos: que a autonomia humana seja promovida e a das máquinas restrita e reversível. Em outras palavras, o fundamental parece ser que o ser humano retenha o poder de decidir quais decisões ele deve tomar e quais podem ser delegadas;
- Princípio da justiça: Por fim, o último dos princípios clássicos da bioética pode ser interpretado sob diversos aspectos: como combate aos vieses, à estigmatização e à discriminação; criação de benefícios que possam ser compartilhados pela sociedade e; prevenção da criação de novos danos, como o enfraquecimento das estruturas sociais e políticas existentes.O princípio da justiça também determina que caso ocorra alguma dano, os sistemas devem garantir algum tipo de reparação. Da mesma forma, este princípio determina que a inteligência artificial possibilite alguma forma de justificativa (*accountability*).

Para que estes princípios sejam eficazes é necessário, portanto, que a inteligência artificial seja auditável e compreensível para o ser humano.

- Princípio da explicabilidade: Quando se trata do desenvolvimento da inteligência artificial, há uma situação de evidente desigualdade: uma pequena porcentagem de humanos desenvolvem as tecnologias que afetam a vida de uma grande quantidade de pessoas.

Surge, diante desse desequilíbrio de poder, a necessidade de um direito, e de um dever para os desenvolvedores, há uma explicação. Esse princípio se encontra expresso por diferentes nomenclaturas nos documentos analisados[110]: "*transparency*" em Asilomar; "*accountability*" na EGE e na IEEE; "*transparency*" e "*intelligibility*" na AIUK; e como "*understandable and interpretable*" no Partnership on AI. Todas essas descrições deixam claro a principal característica das tecnologias dotadas de inteligência artificial: que elas atuam de forma invisível e, muitas vezes, ininteligível para a maioria, com exceção de, na melhor das hipóteses, alguns poucos especialistas da área.

O princípio da explicação tem, portanto, o duplo sentido: "(…) no sentido epistemológico de "inteligibilidade" (como resposta à questão "como funciona?") e no sentido ético de "*accountability*" (como resposta à pergunta: "quem é responsável pela maneira como funciona?)"[111]. Em outras palavras tal princípio, "diz respeito ao direito de receber informações suficientes e inteligíveis que permita ao titular dos dados entender a lógica e os

[110]L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018). Op Cit.

[111]No original "(…)in the epistemological sense of "intelligibility" (as an answer to the question "how does it work?") and in the ethical sense of "accountability" (as an answer to the question: "who is responsible for the way it works?)"L. Floridi et al. (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", Minds and Machines 28(4): 689-707.

critérios utilizados para tratar seus dados pessoais para uma ou várias finalidades”[112], além de apontar os responsáveis pelo seu funcionamento.

Como deriva diretamente do direito à transparência considera-se que uma decisão restará devidamente explicada quando for respondida pelo menos uma das seguintes perguntas: (1) quais são os principais fatores que levaram àquela decisão?; (2) a alteração de algum dos fatores determinantes mudaria a decisão?; e (3) porque casos semelhantes receberam decisões distintas?[113]

Dessa forma, o direito à explicação complementa os quatro princípios éticos apresentados no projeto desenvolvido pelo AI4People, que resumiriam os princípios propostos nos documentos citados anteriormente: beneficência (definida como "fazer o bem"), não-maleficência (definida como "fazer nenhum dano”), autonomia (definida como"respeito pela autodeterminação e escolha dos indivíduos"), e justiça(definido como "tratamento justo e equitativo para todos").

Para que a IA seja benéfica e não promova o mal de seus usuários, estes devem ser capazes de entender o bem ou o dano que ela está causando na sociedade, e de que forma estes estão sendo gerados.

Para que a inteligência artificial possa promover, ao invés de restringir, a autonomia humana, a decisão sobre quem deve decidir deve ser informada com oconhecimento sobre como a IA vai agir, e para ser justa, deve-se garantir que as tecnologias ou, mais precisamente, as pessoas e organizações que estão desenvolvendo e implementando estas tecnologias, possam ser responsabilizadas caso haja um resultado negativo, o que exige o conhecimento sobre como essa situação negativa surgiu.

---

[112]LEITE, Renato. https://igarape.org.br/wp-content/uploads/2018/12/Existe-um-direito-a-explicacao-na-Lei-Geral-de-Protecao-de-Dados-no-Brasil.pdf

[113]Doshi-Velez, Finale, and Mason Kortz. 2017. Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper.

3.3.1 O direito à explicação no contexto europeu

Embora a criação e entrada em vigor da lei geral de proteção de dados europeia (GDPR, da sigla em inglês) seja, com razão, considerada um importante marco na defesa de dados pessoais, a Europa já vinha há décadas liderando as discussões sobre a temática. Em 1995 foi aprovada a Diretiva Europeia de Proteção de Dados (Diretiva 95/46/CE), que já abordava o tratamento de dados pessoais e sua circulação muito antes do surgimento da internet comercial e das tecnologias que necessitam de uma grande quantidade de dados.

Em seu texto já havia, inclusive, artigos que tratavam das decisões automatizadas: o artigo15 concedia ao usuário a opção de se recusar a se sujeitar à uma decisão automatizada que produza efeitos em sua esfera jurídica ou que o afete de modo significativo, salvo algumas exceções[114].

O artigo 12º, (a), por sua vez, referindo-se às decisões contempladas no artigo 15, estabelecia que o usuário teria o direito de conhecer a lógica por detrás do tratamento automatizado dos dados a seu respeito.

Devido à evolução, e crescimento, no tratamento de dados, em especial com o surgimento do fenômeno do *Big Data* e de outras técnicas, como a de aprendizado de máquinas(*machine learning*), houve a necessidade de a legislação passar por um processo de atualização que culminou com a aprovação do GDPR, que entrou em vigor em 25 de maio de 2018. Reproduzindo parte do texto da Diretiva de 1995, o GDPR adiciona, além da lealdade e

[114] "Artigo 15º Decisões individuais automatizadas 1. Os Estados-membros reconhecerão a qualquer pessoa o direito de não ficar sujeita a uma decisão que produza efeitos na sua esfera jurídica ou que a afete de modo significativo, tomada exclusivamente com base num tratamento automatizado de dados destinado a avaliar determinados aspectos da sua personalidade, como por exemplo a sua capacidade profissional, o seu crédito, confiança de que é merecedora, comportamento. 2. Os Estados-membros estabelecerão, sob reserva das restantes disposições da presente diretiva, que uma pessoa pode ficar sujeita a uma decisão do tipo referido no n. 1 se a mesma: a) For tomada no âmbito da celebração ou da execução de um contrato, na condição de o pedido de celebração ou execução do contrato apresentado pela pessoa em causa ter sido satisfeito, ou de existirem medidas adequadas, tais como a possibilidade de apresentar o seu ponto de vista, que garantam a defesa dos seus interesses legítimos; ou b) For autorizada por uma lei que estabeleça medidas que garantam a defesa dos interesses legítimos da pessoa em causa".

licitude do tratamento, já previstos no diploma anterior, a transparência, como um dos princípios do tratamento de dados[115].

Há entre os estudiosos do assunto uma divergência quanto à existência ou não da previsão de um direito à explicação no GDPR. Adotando uma corrente favorável à previsão legal, Andrew D. Selbst e Julia Powles[116]afirmam que esse direito não seria ilusório, ainda que não se encontre expressamente previsto no corpo textual da norma. Ao contrário do que afirmam os que defendem a sua inexistência[117]aqueles autores entendem que, como o GDPR estabeleceu direitos de informação sobre a lógica dos processos de decisões automatizadas (artigos 13 e 14), a mesma confere claramente o direito à explicação.

Dessa forma, o GDPR restringiria a tomada de decisões exclusivamente automatizadas que produzam efeitos significativos nos indivíduos. Os artigos 13, 14 e 15 contêm direitos relacionados à transparência de decisões automatizadas e de perfilamento (*profiling*), tratando especificamente do procedimento de coleta dos dados e dos direitos dos usuários sobre os dados que estão sendo ou não fornecidos.

Considerando-se que a explicação sobre a lógica envolvida no tratamento dos dados pessoais é uma explicação sobre o que será feito com tais dados, pode-se apontar que há, de fato, um direito à explicação. Enquanto os artigos 13 e 14 tratam dos deveres de notificação impostos aos responsáveis pelo tratamento de dados, o artigo 15 foca no direito de acesso à informação por parte do usuário.

---

[115]"Artigo 5º Princípios relativos ao tratamento de dados pessoais 1. Os dados pessoais são: a) Objeto de um tratamento lícito, leal e transparente em relação ao titular dos dados ('licitude, lealdade e transparência');"

[116]Selbst, Andrew D. and Powles, Julia, Meaningful Information and the Right to Explanation (November 27, 2017). International Data Privacy Law, vol. 7(4), 233-242 (2017). Disponível em SSRN: https://ssrn.com/abstract=3039125

[117]This would be robust but potentially disruptive and technically challenging for AI, requiring certain automated decisions to be explained to individuals. Despite a proposal by the European Parliament to guarantee a "right to explanation," this appears only in a nonbinding Recital (7). Elsewhere, individuals are guaranteed "meaningful information" about the "logic involved" in certain automated decision making through the GDPR's "right of access." Although the Regulation fails to define the scope of information to be provided in practice, only a general, easily understood overview of system functionality is likely to be required."Wachter, Sandra &Mittelstadt, Brent &Floridi, Luciano. (2017). Transparent, explainable, and accountable AI for robotics.Science Robotics. 2. 10.1126/scirobotics.aan6080.

O artigo 22, por sua vez, possibilita ao usuário recusar ser alvo de decisões exclusivamente automatizadas, desde que as mesmas o afetem na sua esfera jurídica ou, significativamente, de forma similar. O item 3 do mesmo artigo, mais adiante, prevê a possibilidade de revisão humana da decisão automatizada afim de referendar ou ajustar eventuais vieses da decisão tomada pelo algoritmo.

No entanto, o próprio artigo coloca como exceção à sua aplicação quando o procedimento automatizado for necessário para entrar ou executar um contrato, como no caso, por exemplo, do cálculo automatizado do risco de crédito para decisão sobre a concessão ou não de um empréstimo. Também se encontra aquém da proteção do Regulamento casos nos quais o processo automatizado se baseia no uso de dados pessoais coletados e tratados com o consentimento explícito do titular, o que acontece na maioria dos serviços oferecidos através da Internet, nos quais os usuários consentem com o tratamento dos seus dados por meio de políticas de privacidade.

O direito à explicação vem à tona através da interpretação do artigo 22 à luz do Considerando. 71 do Regulamento[118]. Como bem esclarecido por Isabela Ferrari e Daniel Becker:

[118]Considerando 71: "O titular dos dados deverá ter o direito de não ficar sujeito a uma decisão, que poderá incluir uma medida, que avalie aspetos pessoais que lhe digam respeito, que se baseie exclusivamente no tratamento automatizado e que produza efeitos jurídicos que lhe digam respeito ou o afetem significativamente de modo similar, como a recusa automática de um pedido de crédito por via eletrônica ou práticas de recrutamento eletrônico sem qualquer intervenção humana. Esse tratamento inclui a definição de perfis mediante qualquer forma de tratamento automatizado de dados pessoais para avaliar aspetos pessoais relativos a uma pessoa singular, em especial a análise e previsão de aspetos relacionados com o desempenho profissional, a situação económica, saúde, preferências ou interesses pessoais, fiabilidade ou comportamento, localização ou deslocações do titular dos dados, quando produza efeitos jurídicos que lhe digam respeito ou a afetem significativamente de forma similar. No entanto, a tomada de decisões com base nesse tratamento, incluindo a definição de perfis, deverá ser permitida se expressamente autorizada pelo direito da União ou dos Estados-Membros aplicável ao responsável pelo tratamento, incluindo para efeitos de controlo e prevenção de fraudes e da evasão fiscal, conduzida nos termos dos regulamentos, normas e recomendações das instituições da União ou das entidades nacionais de controlo, e para garantir a segurança e a fiabilidade do serviçoprestado pelo responsável pelo tratamento, ou se for necessária para a celebração ou execução de um contrato entre o titular dos dados e o responsável pelo tratamento, ou mediante o consentimento explícito do titular. Em qualquer dos casos, tal tratamento deverá ser acompanhado das garantias adequadas, que deverão incluir a informação específica ao titular dos dados e o direito de obter a intervenção humana, de manifestar o seu ponto de vista, de

> Os Considerandos não têm força legal se analisadas de forma independente, mas podem alargar, limitar ou explicar uma disposição com caráter normativo. Assim, o Considerando n. 71, entre outros provimentos, recomenda que o responsável pelo tratamento dos dados use procedimentos matemáticos ou estatísticos apropriados, implemente medidas para corrigir imprecisões de dados e minimize o risco de seleções adversas por parte da tomada decisões automatizadas[119].

Dessa forma, no regulamento, os responsáveis pelo tratamento dos dados têm de informar os titulares dos dados acerca da existência de decisões automatizadas, da lógica envolvida e das consequências previstas para estes titulares.

3.3.2 O direito à explicação no contexto brasileiro

De forma distinta do caso da União Europeia, o Brasil só veio a ter uma Lei Geral de Proteção de Dados (LGPD) em 2018(Lei n. 13.709/2018), apesar de a questão estar sendo debatida há quase dez anos e o projeto de lei ter passado por duas consultas públicas no Congresso Nacional.

No entanto, a lei geral contempla direitos, como é o caso do direito à transparência e de explicação, que já estavam presentes na legislação nacional de forma esparsa. Todavia, antes da aprovação da LGPD, tais direitos só eram garantidos em decisões automatizadas relativas à concessão de crédito, modelagem e cálculo de risco de crédito, o que deixa fora do escopo de proteção diversas situações, como em casos de uso de dados por planos de saúde ou pelo governo.

Estabelecendo os princípios da boa-fé e da transparência como os principais reguladores da relação de consumo, seja ela *online* ou *off-line*, a primeira lei que trouxe o que se pode chamar de direito à explicação no Direito Brasileiro foi o Código de Defesa do

---

obter uma explicação sobre a decisão tomada na sequência dessa avaliação e de contestar a decisão. Essa medida não deverá dizer respeito a uma criança."

[119]FERRARI, Isabela e BECKER, Daniel. O direito à explicação sobre decisões automatizadas: uma análise comparativa entre a União Europeia e o Brasil.Revista de Direito e as Novas Tecnologias 2018 VOL. 1 - OUT/DEZ 2018

Consumidor (CDC). Notoriamente, o setor de consumo é um dos que mais coleta e se utiliza de dados para obter informações e realizar o *profiling* de seus clientes. A prática, mais conhecida como micro *targeting*, permite que asempresas obtenham informações sobre a personalidade e hábitos do seu público alvo a um nível quase individual, podendo, inclusive, chegar a influenciar suas condutas e opiniões.

Dessa forma, o consumidor se encontra em uma posição vulnerável em sua relação com as empresas, devendo, portanto, ser protegido, o que inclui receber informações adequadas para que possa exercer seus direitos e evitar práticas abusivas e discriminatórias.

Ainda que a legislação tenha entrado em vigor apenas no início dos anos 1990, quando a Internet, como a conhecemos, e as tecnologias dotadas de inteligência artificial não eram amplamente utilizadas como atualmente, o artigo 6º, III, do CDC, já determinava um direito à informação que garantia, ainda que de forma rudimentar, a possibilidade de o usuário obter uma explicação razoável sobre uma decisão algorítmica sobre ele.

Utilizando-se o princípio da boa-fé enquanto "modelo ideal de conduta, que se exige de todos os integrantes da relação obrigacional (devedor e credor) na busca do correto adimplemento da obrigação, que é a sua finalidade"[120], como princípio guia, pode-se apontar que o dever de informação previsto no artigo 6, inciso III, do CDC, determina a prestação de informação ao consumidor de forma clara e objetiva, inclusive na fase pré-contratual, bem como o dever de transparência.

Em relação ao Código de Defesa do Consumidor, pode-se destacar, também, os artigos 43 e 46 que tratam do acesso a informações cadastrais e bancos de dados. O primeiro deixou expresso o direito de acesso do consumidor a informações a seu respeito e às respectivas fontes nesses cadastros e bancos de dados, além de determinar o dever de clareza, o direito de retificação de informações incorretas e um período máximo de armazenamento dos dados do

[120]Entendimento adotado pelo Superior Tribunal de Justiça (2012). "Teoria do adimplemento substancial limita o exercício de direitos do credor". Jusbrasil. Disponível em: https://stj.jusbrasil.com.br/noticias/100054780/teoria-doadimplemento-substancial-limita-o-exercicio-de-direitos-do-credor

consumidor de cinco anos. Além disso, o artigo determina a notificação do consumidor sobre a coleta e o uso de seus dados, ainda que o consentimento prévio não seja necessário. Tal regra não se aplica apenas nos casos de compartilhamento com terceiros, conforme entendimento do Ministério da Justiça estabelecido na Portaria nº 5, de 27 de agosto de 2002[121].

O artigo 46, por sua vez, não apenas determina o dever de informação sobre o conteúdo dos contratos, mas também que este conteúdo esteja disposto de forma inteligível de modo a garantir a compreensão do consumidor. Uma interpretação ampla dessa compreensão inclui entender como o algoritmo chegou a determinada decisão.

Contudo, a primeira vez que houve menção expressa ao direito à explicação se deu na lei do Cadastro Positivo (Lei n. 12.414/2011, LCP) que disciplina a formação e consulta a banco de dados com informações de adimplemento para fins de concessão (ou não) de crédito. A decisão para concessão do crédito é feita analisando-se o *credit scoring*.

A referida lei tem como objetivos principais reduzir a assimetria de informações e possibilitar a coleta de dados de adimplência, após o prévio consentimento do consumidor. Com isso, afirma-se, seria possível haver uma redução das taxas de juros e, como consequência, uma ampliação das relações comerciais[122]. A norma visa, também, a adequada proteção dos dados pessoais dos consumidores de modo que prevê uma série de novos direitos, estando entre eles o direito à explicação.

Encontra-se disposto no artigo 5o da lei que, dentre os direitos do cadastrado, estão os de "conhecer os principais elementos e critérios considerados para a análise de risco, resguardado o segredo empresarial; ser informado previamente sobre o armazenamento, a

[121]BRASIL. Ministério da Justiça. Portaria nº 5 de 27 de agosto de 2002. Dispõe sobre cláusulas abusivas em contratos de vendas de produtos e prestação de serviços. Diário Oficial da República Federativa do Brasil, Brasília, DF, 28 ago. 2002. Disponível em: https://www.procon.go.gov.br/legislacao/portarias/ portaria-n%C2%BA-5-27-08-2002-mj-sde-clausulas-abusivas-nome-de-consumidor-a-banco-dedados.html.

[122]PORTO, Antonio José Maristrello; FRANCO, Paulo Fernando de Mello. Por uma análise também econômica da responsabilidade civil do cadastro positivo: abordagem crítica do art. 16 da Lei 12.414/2011. Revista de Direito do Consumidor, São Paulo, v. 115, p.247 -271, jan.-fev. 2018.

identidade do gestor do banco de dados, o objetivo do tratamento dos dados pessoais e os destinatários dos dados em caso de compartilhamento", assim como o de "solicitar ao consulente a revisão de decisão realizada exclusivamente por meios automatizados; e ter os seus dados pessoais utilizados somente de acordo com a finalidade para a qual eles foram coletados."

Em outras palavras, esses direitos exigem que o consumidor seja informado sobre as fontes de dados utilizadas e as informações pessoais consideradas para o cálculo do risco de inadimplência na concessão ou não de crédito, além de garantir o direito do consumidor de requerer uma revisão da decisão resultante. A Lei também tenta limitar os tipos de dados que podem ser utilizados para cálculo do risco de crédito, estando vedado o uso de dados não relacionados com a análise do risco de crédito do consumidor, assim como dados pessoais sensíveis, conforme definidos pelo artigo 5o, inciso, II, da LGPD, tais como os pertinentes "à origem social e étnica, à saúde, à informação genética, à orientação sexual e às convicções políticas, religiosas e filosóficas".

Recentemente, em 12.11.2014, devido a diversas críticas feitas à lei em relação à coleta de dados do consumidor sem o seu consentimento para a análise de risco, o Superior Tribunal de Justiça, no REsp n. 1.419.697/RS[123], de relatoria do Ministro Paulo de Tarso Sanseverino, analisou o artigo 5º, inciso IV, da mesma, entendendo pela licitude do *credit scoring*. Porém, a corte definiu como limites a essa análise os direitos descritos acima, de forma que restaram garantidos os direitos do consumidor, entre eles o direito à explicação. Essa decisão resultou na Súmula 550:

> "A utilização de escore de crédito, método estatístico de avaliação de risco que não constitui banco de dados, dispensa o consentimento do consumidor, que terá o direito de solicitar esclarecimentos sobre as informações pessoais valoradas e as fontes dos dados considerados no respectivo cálculo"[124].

[123]RESP nº 1.419.697 RS. Disponível em: https://stj.jusbrasil.com.br/jurisprudencia/152068666/recurso-especial-resp-1419697-rs-2013-0386285-0/relatorio-e-voto-152068681

[124] Disponível em http://www.stj.jus.br/SCON/sumanot/toc.jsp?livre=(sumula%20adj1%20%20550).sub.

Em julgamento posterior (RESP nº 1.304.736/RS)[125] o STJ julgou se o direito de acesso às fontes dos dados e a explicação da lógica do seu tratamento encontravam algum fator limitador. Nesta oportunidade, a corte concluiu que há interesse de agir do consumidor que deseja conhecer os principais elementos e critérios utilizados para a análise do seu histórico de crédito, além de quais foram as informações pessoais utilizadas, desde que seja respeitado o segredo empresarial.

No entanto, o consumidor só poderia ter acesso a essa explicação caso tenha sido atingido por esses elementos e critérios ao tentar obter crédito no mercado. Seria o caso, por exemplo, do consumidor que teve seu pedido de crédito negado por uma decisão que lhe atribuiu um *score* de crédito baixo. Diante dessa decisão, tornou-se possível reconhecer o direito à explicação de decisões totalmente automatizadas quando as mesmas tiverem um impacto específico na vida das pessoas. Infelizmente esse critério só se aplica a casos de concessão de crédito. Considerando que a Lei Geral de Proteção de Dados absorveu os direitos e limites dispostos nas leis setoriais e precedentes judiciais pode-se estender a lógica aplicada pelo STJ também aos casos sob a égide da LGPD.

Conforme visto até o momento, o Código de Defesa do Consumidor e a Lei do Cadastro Positivo são leis setoriais, de modo que são insuficientes para promover uma verdadeira proteção de dados, visto que não abarcam a maioria das hipóteses de utilização de dados.

Dessa forma, a aprovação da lei geral de proteção de dados brasileira foi essencial. A LGPD complementa, harmoniza e unifica um ecossistema de mais de quarenta normas setoriais que regulam, de forma direta e indireta, a proteção da privacidade e dos dados

[125]RESP nº 1.304.736/RS. Disponível em: < https://stj.jusbrasil.com.br/jurisprudencia/178798658/recurso-especial-resp-1304736- rs-2012-0031839-3> https://stj.jusbrasil.com.br/jurisprudencia/178798658/recurso-especial-resp-1304736-rs-2012-0031839-3

pessoais no Brasil[126].Tendo sido inspirada nas discussões que culminaram no GDPR, tem como objetivo, além de conferir às pessoas maior controle sobre seus dados, fomentar um ambiente de desenvolvimento econômico e tecnológico. Sendo assim, a LGPD não visa impedir a utilização de algoritmos utilizados em decisões automatizadas, mas sim garantir sua continuidade, limitando abusos nesse processo. O direito à explicação é utilizado, portanto, para diminuir a assimetria de informações e, como consequência, de poder entre o titular de dados e os setores privado e público.

Dentre os dez princípios garantidos pela lei geral aos titulares de dados destaca-se o da transparência (artigo 6, IV), que se configura no direito de obter "informações claras, precisas e facilmente acessíveis sobre a realização do tratamento e os respectivos agentes de tratamento, observados os segredos comercial e industrial".

Tal direito dá origem ao direito de acesso a dados, previsto no artigo 19, que determina que "a confirmação de existência ou o acesso a dados pessoais serão providenciados, mediante requisição do titular" e se darão "por meio de declaração clara e completa, que indique a origem dos dados, a inexistência de registro, os critérios utilizados e a finalidade do tratamento, observados os segredos comercial e industrial, fornecida no prazo de até 15 (quinze) dias, contado da data do requerimento do titular".

Além do direito de acesso, a LGPD, de forma muito semelhante à Lei do Cadastro Positivo, prevê a possibilidade de revisão de decisões tomadas com base em tratamento automatizado de dados "que afetem seus interesses, inclusive de decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade"[127]. O objetivo é evitar que indivíduos sejam alvo de práticas discriminatórias dos algoritmos responsáveis pela decisão automatizada.

---

[126]MONTEIRO, R. L. (2017). "Proteção de dados e a legislação vigente no Brasil". Baptista Luz. Disponível em: http://baptistaluz.com.br/wp-content/uploads/2017/11/Privacy-Hub-Leis-Setoriais.pdf.
[127]Artigo 20, caput da lei 13.709/2018

Contudo, a revisão se aplica apenas a decisões que afetam os interesses dos titulares dos dados pessoais, podendo-se citar, como exemplo, àquelas utilizadas para definir perfis comportamentais de cunho pessoal, profissional, de consumo e de crédito utilizadas pela polícia para prever crimes.

No entanto, é no artigo 20, § 1º, onde o legislador previu que "o controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial", que se encontra a previsão ao direito à explicação. Nesse caso, por se tratar de uma lei geral, dá-se uma amplitude muito maior em relação ao estabelecido na Lei do Cadastro Positivo, cuja aplicação se encontra restrita a casos de *credit scoring*.

Cabe apontar que, caso o responsável pelo processamento dos dados se recuse a fornecer os dados pessoais utilizados na decisão automatizada, ou se recuse a explicar os critérios e a lógica por detrás dos algoritmos que controlam o processo de tomada de decisão, caberá à Autoridade Nacional de Proteção de Dados (ANPD)[128], caso requisitada, realizar um procedimento administrativo, após o qual poderá fazer uma auditoria nos sistemas da entidade violadora.

Como dito anteriormente, o direito à explicação e à revisão de decisões automatizadas pode ser usufruído, ainda que a decisão não gere efeitos jurídicos ao seu titular, tendo em vista a abrangência da lei, o que confere aos titulares dos dados pessoais uma ferramenta importante para coibir abusos e práticas discriminatórias em decisões que se utilizam de seus dados. Ainda, como a LGPD possibilita o exercício desse direito sempre que a decisão automatizada afetar os interesses do indivíduo, esta acabou por ter uma maior abrangência se comparado ao previsto no GDPR, que o restringe apenas às hipóteses nas quais uma decisão

[128]Embora os artigos da LGPD que tratavam da Autoridade Nacional de Proteção de Dados (ANPD) tenham sido vetados pelo ex-Presidente Michel Temer quando da promulgação da lei, a mesma foi criada pela Medida Provisória n 869/18 no dia 28 de dezembro de 2018. Atualmente a MP está aguardando ser examinada pelo Congresso Nacional.

automatizada produz efeitos jurídicos ou similarmente significantes em relação ao titular de dados.

Com a lei geral de proteção de dados, além das leis setoriais já existentes, as empresas e entidades públicas que utilizam algoritmos em decisões autônomas passam a ter que pensar, desde a concepção, a como garantir que os algoritmos estejam aptos a cumprir com os deveres de informação e explicação. Esta, muito provavelmente, será uma tarefa árdua para os desenvolvedores e programadores, tendo em vista as dificuldades já discutidas anteriormente.

No entanto, é importante ressaltar que nem o GDPR nem a LGPD definem, de forma clara e objetiva, alguns dos pressupostos básicos para a compreensão dos direitos ora discutidos: (i) o que pode-se definir como uma decisão totalmente automatizada, (ii) quais tipos de decisão automatizada afetam a esfera jurídica dos titulares de dados, e (iii) qual é o grau de transparência e explicação que será exigível em situações assim[129]. Dessa forma, torna-se necessário analisar esses conceitos de modo a se verificar quais são, na doutrina, as diferentes interpretações desses pressupostos que estão sendo debatidas.

Em relação à primeira questão, tem-se argumentando no sentido de que uma decisão não deixa de ser automatizada apenas por haver a intervenção humana em seu processo decisório. Caso contrário as previsões legais seriam facilmente neutralizadas, ou até burladas, bastando à empresa introduzir uma pessoa humana para que as mesmas não se aplicassem. Dessa forma, defende-se que a lei só poderá ser afastada caso a atuação da pessoa natural seja tal que possa efetivamente reverter o resultado automatizado[130].

No mesmo sentido estabelece uma das diretivas criadas pelo *Data Protection WorkingParty* (A29WP), ao determinar que não se pode afastar a aplicação do GDPR por meio do que se denomina de “fabricação de intervenção humana mínima no processo

[129]https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/controversias-sobre-direito-a-explicacao-e-a-oposicao-diante-de-decisoes-automatizadas-12122018
[130] FRAZÃO, Ana. Op Cit

decisório"[131]. Cabe apontar que este guia foi criado como referência interpretativa para o GDPR, mas pode ser aplicado, por analogia também, à LGPD. O mesmo alerta é encontrado no guia do *InformationCommissioner's Officer* (ICO) britânica[132].

Contudo, na prática, pode-se provar difícil definir qual o nível de intervenção da pessoa natural no processo decisório e se este é suficiente para afastar, ou não, a aplicação dos direitos previstos na lei.

Em relação à segunda questão, as *Guidelines on Automated Individual Decision Making and Profiling* do A29WP apontam como exemplos de situações que podem afetar a esfera jurídica e os interesses dos titulares de dados pessoais: as avaliações ou *scorings*, decisões automatizadas com efeitos jurídicos ou similares, monitoramento sistemático, tratamento de dados sensíveis, dados processados em larga escala, *data sets* combinados ou misturados, processamentos que impedem titulares de dados de exercerem direitos, de usarem determinado serviço ou de celebrarem determinados contratos, entre outros.

Sob outro viés, também questiona-se o grau que deve ser exigível de explicação, tendo em vista que a LGPD determina que o segredo industrial configura-se como uma limitação ao alcance do artigo 20.

Além disso, como já dito anteriormente, a explicação da mera lógica matemática não auxilia em uma verdadeira transparência visto que serão poucas as pessoas que efetivamente irão entendê-la. Deve-se acrescentar, ainda, as dificuldades naturais geradas pelo aprendizado de máquina que dificultam a reconstituição passo a passo de decisões algorítmicas.

Dessa forma, parece mais realista falar-se em uma transparência relativa, mas suficiente para assegurar um processo justo por meio de informações significativas a respeito

---

[131] *Data Protection Working Party* (A29WP).*"Guidelines on Automated Individual Decision Making and Profiling"*.Disponível em https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053

[132] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-does-the-gdpr-say-about-automated-decision-making-and-profiling/

da lógica envolvida nas decisões totalmente automatizadas. Em outras palavras, a explicação deve ser adequada de modo a tornar a decisão inteligível e compreensível.

Esta preocupação também foi expressa nas *Guidelines on Automated Individual Decision Making and Profiling*, quando afirmam que os agentes de tratamento precisam ofertar informações significativas sobre a lógica envolvida no tratamento automatizado, bem como as explicações sobre o significado dos resultados diante dos objetivos pretendidos pelo processamento.

Ademais, no caso específico dos perfis, as *Guidelines* afirmam que os controladores precisam esclarecer as informações fundamentais que lastrearam a decisão, incluindo (i) as categorias de dados usados nos perfis, (ii) as fontes de tais informações, (iii) como os perfis são criados, incluindo as estatísticas utilizadas, (iv) a razão de o perfil ser relevante para a decisão automatizada, e (v) como as informações foram utilizadas para a decisão que afetou determinado titular[133].

Dessa forma, parece claro que as leis gerais de proteção de dados europeia e brasileira, embora estabeleçam o direito à explicação na tentativa de tornar aplicável o princípio da transparência e garantir certo nível de *accoutability*,ainda deixa a desejara eficácia deste direito. Apenas a regulação jurídica não é suficiente para solucionar problemas técnicos e de *design*.

No entanto, como bem destacado por Andrew Selbst e Julia Powles, se a tecnologia não atende àquilo que a lei determina, ela não pode ser utilizada sem violar a legislação[134]. Mesmo diante de todas as dificuldades técnicas apontadas anteriormente deve a tecnologia

[133]*Data Protection Working Party* (A29WP).“*Guidelines on Automated Individual Decision Making and Profiling”*.Disponível em https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053
[134]SELBST, Andrew; POWLES, Julia.Meaningful information and the right to explanation.International Data Privacy Law, v. 7, n. 4, 2017. Disponível em [https://academic.oup.com/idpl/article/7/4/233/4762325]

respeitar o Direito sob o risco de ocorrer o contrário, passando a haver o deslocamento do direito pelo código (*code is law*) apontado por Lawrence Lessig[135].

4. COMO REGULAR AS APLICAÇÕES DE INTELIGÊNCIA ARTIFICIAL

O Estado brasileiro não consegue acompanhar de maneira adequada os avanços das novas tecnologias. O que frequentemente observamos é um grande lapso temporal separando a popularização da utilização de novas dinâmicas disrruptivas e a regulamentação legal e infra-legal de tais práticas sociais e econômicas. Três são os exemplos que ajudam a corroborar a mencionada afirmação: (i) em março do ano passado foi publicada a Lei nº 13.640/2018, que alterou as diretrizes da Política Nacional de Mobilidade Urbana para regulamentar o chamado "transporte remunerado privado individual de passageiros", com 4 anos de atraso em relação à chegada do Uber ao Brasil[136]; (ii) pouco mais de um mês depois, em abril, o Conselho Monetário Nacional (CMN) regulamentou, através das resoluções nº

[135]LESSIG, L. Code and other laws of cyberspace. New York: Basic Books, 2006

[136] Disponível em: <https://www.jota.info/opiniao-e-analise/artigos/aplicativos-de-transporte-riscos-do-novo-marco-regulatorio-22042018>.

4.656 e nº 4.657, as *fintechs* de crédito, com 3 anos[137] de atraso em relação aos primeiros empréstimos concedidos pela Creditas, uma das empresas de referência em tal segmento[138]; e (iii) o Congresso brasileiro ainda debate a respeito de uma lei de proteção de dados, ao tempo que diversos vazamentos já ocorreram no país[139]. Em paralelo, a partir do dia 25 de maio de 2018, a lei geral de proteção de dados européia (GDPR em inglês) passou a vigorar[140].

Internacionalmente, diversas aplicações da AI crescem sem que um marco legal brasileiro sequer esteja em vias de ser publicado. Assim sendo, o papel da doutrina é fundamental para pavimentar os debates que naturalmente surgirão com o desenvolvimento de tais usos. Vale lembrar que questões relacionadas à *accountability* dos algoritmos já ensejam debates relevantes nos Estados Unidos[141].

Cabe vislumbrar, desde já, que, caso não haja um marco legal único para tal tecnologia, o tema poderá acabar sendo regulado e regulamentado por diferentes instrumentos normativos setoriais e ainda sofrer reflexos de normas internacionais à exemplo do que ocorreu no âmbito da proteção de dados. Neste a ausência de um marco legal geral, devido à promulgação tardia da lei geral, oportunizou que o tema fosse regulamentado de forma setorizada e parcial, constando em diversos textos legais setoriais como o plano de internet

---

[137] Disponívelem: <https://endeavor.org.br/fintech-que-cresceu-7-vezes-em-2017-como-creditas-esta-revolucionando-o-mercado-de-creditos-brasil/>.

[138] Disponível em: <https://www.jota.info/opiniao-e-analise/artigos/banco-central-inovador-27042018

[139] Um dos exemplos mais recentes desses vazamentos ocorreu com a empresa Netshoes que teve mais de 2,5 milhões de dados de clientes do e-commerce vazados.Disponívelem: <https://www.tecmundo.com.br/seguranca/129428-vazamento-netshoes-continua-totaliza-dados-2-5-milhoes-clientes.htm>.

[140] General Data Protection Regulation - "(...) a GDPR se aplica a coleta de dados pessoais de pessoas naturais que se encontram na União Europeia, independente da sua nacionalidade, cidadania, domicílio ou residência. Adicionalmente, em suma, será verificado com este estudo, que, caso uma empresa brasileira, de uma forma ou de outra, colete, processe ou receba dados pessoais de pessoais naturais localizadas na União Europeia, independente da sua nacionalidade, incluindo dados de consumidores, colaboradores, dados financeiros, ou ofereça serviços para algum dos 28 países do bloco europeu, poderá estar sujeita a jurisdição prescrita pela norma". Disponível em:<https://baptistaluz.com.br/institucional/o-impacto-da-regulacao-geral-de-protecao-de-dados-da-ue-em-empresa-brasileira/>.

[141] Disponível em: <https://www.jota.info/opiniao-e-analise/artigos/algoritmo-e-preconceito-12122017.

das coisas[142], o decreto de transformação digital[143], o plano de segurança da informação[144] e na resolução 4.658/18 do Bacen[145]. Além disso, a ausência da lei geral levou à utilização, de forma reflexa, de legislações estrangeiras, em especial a GDPR e permitiu, diante da ausência da Autoridade Nacional de proteção de dados, a atuação ativa do Ministério Público[146].

Nesse contexto, considerando os problemas gerados por decisões autônomas, apresentados nos capítulos anteriores, não pode haver mais espaço para a noção de disrrupção sem regulação. A notória frase de Mark Zuckerberg "*Move fastand break things*" marcou, nos últimos anos, a visão da maioria das empresas de inovação, segundo a qual mais é sempre melhor. Há uma corrida para colocar os produtos nas mãos dos consumidores o mais rápido possível, sem levar em conta o mérito e a lógica dos sistemas de governança[147].

No entanto, quando se trata de sistemas de decisões autônomos e, especialmente, análise de risco, os danos ocasionados aos grupos vulneráveis deve levar em consideração a importância da regulação. Deve-se levar em conta não apenas as falhas do livre mercado de se autorregular, mas também, e principalmente, a necessidade de se garantir o respeito aos direitos humanos, que merecem ser protegidos das ações do mercado. Dessa forma, este capítulo tratará da importância de se regular a utilização de sistemas de decisões autônomase como tal regulação pode ocorrer.

---

[142] Disponível em: <https://www.bndes.gov.br/wps/portal/site/home/conhecimento/pesquisaedados/estudos/estudo-internet-das-coisas-iot/estudo-internet-das-coisas-um-plano-de-acao-para-o-brasil>.

[143] DECRETO Nº 9.319, DE 21 DE MARÇO DE 2018. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/decreto/D9319.htm>.

[144] Disponívelem: <http://bd.camara.gov.br/bd/handle/bdcamara/19863>.

[145] Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/exibenormativo?tipo=Resolu%C3%A7%C3%A3o&numero=4658

[146] Comissão de Proteção dos Dados Pessoais do Ministério Público do Distrito Federal e Territórios (MPDFT) instituída pela Portaria Normativa PGJ nº 539, de 12 de abril de 2018.Disponívelem: <http://www.mpdft.mp.br/portal/index.php/conhecampdft-menu/nucleos-e-grupos/comissao-de-protecao-dos-dados-pessoais>.

[147]Disponível em: https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over?fbclid=IwAR2jAbyGNcl1FTiIxze2vvh8A9hk_qGR90bU58DXqgegLAPV4bAacxPhiU0

EUA preparing_for_the_future_of_artificialintelligence.Executive Office of the President National Science and Technology Council Committee on Technology.2016. Disponível em: <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>.

4.1 PROJETOS DE REGULAÇÃO NOS ESTADOS UNIDOS, EUROPA E CHINA

Diante desse panorama, alguns países iniciaram as discussões sobre como planejar o desenvolvimento da aplicação dessa nova tecnologia. Em que pesem os países divergirem entre si sobre como tratar o tema, a análise das propostas dos ordenamentos estrangeiros pode ser importante farol para se refletir o tema no Brasil.

Neste sentido, nota-se que os Estados Unidos apresentam grande preocupação com a utilização responsável da tecnologia[148]. O principal objetivo da proposta norte americana parece ser alcançar um equilíbrio entre a aplicação da inteligência artificial de forma segura, evitando assim riscos e violações a princípios e aos direitos humanos, sem com isso bloquear a inovação. Para alcançar este objetivo, o relatório "*Preparing for the Future of Artificial Intelligence*"[149]aponta que serão necessários investimentos crescentes na educação da sociedade para lidar com a nova tecnologia e para desenvolver e recrutar talentos. Assim, haverá grande destaque para pesquisadores de AI, bem como na capacitação em geral da força de trabalho, para lidar com os impactos econômicos gerados.

O sistema norte americano tradicionalmente tende a confiar na força do mercado quando se trata do assunto regulação. O cerne da presente discussão norte-americana está na adequação das regulações já existentes aos riscos que serão gerados pelo uso de AI. A única regulação atualizada até o momento é a de Nova York[150](Lei Nº 1696-A[151]), que almeja garantir a transparência dos algoritmos usados para tomada automatizada de decisões da

[148]EUA preparing_for_the_future_of_artificialintelligence.Executive Office of the President National Science and Technology Council Committee on Technology.2016. Disponível em: <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>.

[149]EUA preparing_for_the_future_of_artificialintelligence.Executive Office of the President National Science and Technology Council Committee on Technology.2016. Op cit.

[150]Disponível em: <http://irisbh.com.br/inteligencia-artificial-e-regulacao-o-caso-de-nova-york/>.

[151]Disponível em: <http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0>.

polícia, do Judiciário e de outros órgãos governamentais da cidade. Infelizmente, o escopo da lei é limitado. Trata apenas dos algoritmos utilizados pela Administração Pública, silenciando quanto à sua utilização pela iniciativa privada, além de meramente prever a criação de uma força tarefa temporária que tem como missão desenvolver um relatório com sugestões de possíveis ações a serem tomadas. Verifica-se, portanto, que se trata de uma iniciativa ainda bastante tímida em termos regulatórios.

Nessa temática, ainda é muito cedo para saber se o escândalo gerado pelo caso da *Cambrigde Analytica*, que, por meio de práticas abusivas, conseguiu coletar dados pessoais de 50 milhões de usuários do *Facebook* nos EUA, cruzando-os de modo a identificar e influenciar pretensões de votos[152], poderá alterar esse posicionamento.

Nos Estados Unidos, os progressos nos estudos de AI são feitos principalmente por universidades, ou iniciativas privadas. Dessa forma, utiliza-se uma boa parcela das normas do livre mercado e da livre concorrência. Por outro lado, pesquisas de longo prazo, que atraem menos interesse por parte da iniciativa privada, tendem a ser financiadas pelo governo.

De modo geral, em solo norte americano, observa-se expressiva preocupação com questões de responsabilidade, justiça e transparência. Estas limitam boa parte da aplicação da AI na sociedade. O relatório apresentado pela Casa Branca, no ano de 2016[153],preocupa-se energicamente coma resolução destes problemas, de forma a futuramente permitir ampla aplicação da AI. Tais questões reforçam a relevância da educação ética dos programadores e da sociedade. Além disso, enfatiza-se a necessidade de inclusão de maior diversidade nos ambientes tecnológicos, de modo a refrear o surgimento de distorções, incluindo as possíveis discriminações de gênero, raça ou nacionalidade, nos sistemas de AI.

---

[152]Disponível em: <https://brasil.elpais.com/brasil/2018/03/20/tecnologia/1521582374_496225.html>.

[153]EUA preparing_for_the_future_of_artificialintelligence.Executive Office of the President National Science and Technology Council Committee on Technology.2016. Disponível em: <https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf>.

Feitas essas breves considerações, pode-se concluir que os Estados Unidos visualizam os benefícios econômicos e sociais que podem advir da aplicação de AI, mas permanecem tímidos quanto a sua efetiva aplicação devido aos riscos.

A Europa, por sua vez, embora não tenha ainda apresentado planejamento específico para o uso de AI, vem demonstrando uma tendência para a regulação precoce das novas tecnologias, em consonância com a tentativa de proteger os denominados "valores europeus". Essa tem sido a abordagem tomada pela França[154], em seu plano de inteligência artificial, e pela Europa como um todo, ao tratar a questão dos robôs[155].

Quando se trata da formulação de princípios próprios para combater os riscos gerados pelas novas tecnologias, constata-se que os europeus são mais protetivos e mais ativos do que os americanos. Além disso, diferentemente dos americanos, o desenvolvimento de diretrizes relacionadas à aplicação de AI e ao uso de robôs vem sendo liderado pela iniciativa pública, a qual demanda uma agenda de pesquisas que parte tanto dos setores privados, quanto das universidades.

Passando para uma análise da experiência asiática, há que se frisar o planejamento chinês[156], apresentado em 2017, que pretende desenvolver e aplicar AI a partir de metas ambiciosas até 2030. O plano estabelece o uso de AI como fundamental para o desenvolvimento econômico e para a segurança nacional das próximas décadas. A China pretende se tornar uma potência neste segmento, liderando essa transformação.

O governo chinês vê claramente os benefícios que a inteligência artificial pode trazer para diversos problemas sociais, tais como o envelhecimento populacional e questões

---

[154]VILLANI, Cédric, "FOR A MEANINGFUL ARTIFICIAL INTELLIGENCE: TOWARDS A FRENCH AND EUROPEAN STRATEGY". Disponível em: <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>.

[155]Disponível em: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//PT>.

[156]"A Next Generation Artificial Intelligence Development Plan". Disponível em: <https://chinacopyrightandmedia.wordpress.com/2017/07/20/a-next-generation-artificial-intelligence-development-plan/>.

ambientais. Entende-se, portanto, que o uso de AI pode melhorar substancialmente a qualidade de vida da sociedade.

No entanto, diferente do planejamento americano e europeu, no chinês não há qualquer apresentação de preocupações relacionadas à proteção de princípios ou valores, – como a privacidade ou liberdade –, de seus cidadãos. O planejamento demonstra apenas a ânsia de se tornar o líder mundial no tratamento da AI, deixando de lado a cautela com potenciais violações ideológicas advindas do uso de tal tecnologia.

Verificadas as apostas e parâmetros de diferentes países, ou continente no caso europeu, passa-se a analisar, mais especificamente, quais seriam essas leis que poderiam regular a inteligência artificial.

4.2 QUEM DEVEMOS REGULAR?

Quando se trata de regulação de máquinas dotadas de autonomia, o cenário da ficção traz as três leis de Isaac Asimov. Tais leis eram embutidas nos cérebros positrônicos de todos os robôs:

Primeira Lei: “um robô não pode ferir um ser humano ou, por meio da inação, permitir que um humano se machuque”[157].

Segunda Lei: “um robô deve obedecer a ordens dadas por seres humanos, exceto quando tais ordens entrem em conflito com a Primeira Lei”[158].

Terceira Lei: “um robô deve proteger sua própria existência, desde que tal proteção não entre em conflito com a Primeira ou Segunda Leis”[159].

[157]2 ISAAC ASIMOV, Runaround, in I ROBOT 37 (1950)
[158]2 ISAAC ASIMOV, Runaround, in I ROBOT 37 (1950)
[159]2 ISAAC ASIMOV, Runaround, in I ROBOT 37 (1950)

Durante o desenvolvimento das histórias, o robô DaneelOlivaw cria uma quarta lei que denomina de Lei zero, pois precede todas as demais: “Um robô não pode ferir a humanidade, ou através da inação, permitir que a humanidade venha a ser prejudicada”.

Tais leis detêm tanta influência que mesmo atualmente pesquisadores imaginam como seria, ou mesmo se seria possível, de inseri-las em robôs, como no caso dos carros autônomos[160],enquanto outros argumentam como essas leis não seriam necessárias[161]visto que refletem ansiedades culturais sobre robôs e é desnecessário construi-las em autômatos reais.

No entanto, seu real valor se encontra no fato de que elas trazem uma nova perspectiva sobre a visão ocidental tradicional do robô, que perde o controle e se torna mau, muito divulgada em filmes de ficção científica como Exterminador do Futuro. Asimov escreveu suas histórias para lutar contra o que ele denominou de "Complexo de Frankenstein”, que é a ideia de que os robôs são inerentemente ameaçadores ou malignos e que inevitavelmente se virarão contra seus criadores[162].

Em muitas de suas histórias as pessoas começam com preconceito contra robôs e, ao longo da trama, acabam por ver o seu valor. Um exemplo claro disto é o caso do protagonista de vários de suas histórias, o detetive ElijahBailey, que é inicialmente cético em relação aos robôs, mas que eventualmentetorna-se o melhor amigo de R. DaneelOlivaw, seu parceiro robótico.

Com a introdução das três leis, Assimov torna os robôs alvo de interpretação e regulação, ao invés de algo a ser temido. No entanto, tais leis são muto vagas e incompletas, o que faz sentido no contexto das histórias, mas não tem muita validade na prática. Contudo,

---

[160]Boer Deng, Machine Ethics: The Robot’s Dilemma, NATURE, (July 1, 2015) Disponível em http://www.nature.com/news/machine-ethics-the-robot-s-dilemma-1.17881 [https://perma.cc/5HT6-EEFA];

[161] Ulrike Barthelmess and Ulrich Furbach, Do We Need Asimov's Laws?, CORNELL U. LIBRARY 11 (2014), https://arxiv.org/ftp/arxiv/papers/1405/1405.0961.pdf [http://perma.cc/YSC2-WZYJ]

[162]LEE MCCAULEY, ASS’N FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE, THE FRANKENSTEIN COMPLEX AND ASIMOV’S THREE LAWS 9, 9 (2007), Disponível em https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-003.pdf

permanece a ideia de regular os robôs, e a inteligência artificial em geral. Dessa forma, cabe questionar como podemos utilizar as ideias de Assimov.

Como visto nos capítulos anteriores há cada vez mais o uso de algoritmos pela sociedade de modo que estamos saindo da era da internet e entrando no que Jack Balkin denominou de Sociedade Algorítmica:

> De fato, estamos nos movendo rapidamente da era da Internet para a Sociedade Algorítmica. Em breve, olharemos para a era digital como a precursora da Sociedade Algorítmica. O que quero dizer com Sociedade Algorítmica? Quero dizer uma sociedade organizada em torno da tomada de decisão social e econômica por algoritmos, robôs e agentes dotados de inteligência artificial; que não só tomam as decisões, mas também, em alguns casos, também as executam[163].

Nessa nova sociedade movida pela inteligência artificial o *Big Data* se apresenta como uma parte essencial da sua existência tendo em vista que este é o combustível que a movimenta[164], bem como o seu produto. Os sistemas de decisão autônoma geram tais análises edecisões através da coleta e tratamento de dados. Tais decisões, por sua vez, geram novos dados que voltam a treinar os algoritmos, aprimorando seu desempenho.

Dessa forma, uma regulação que trate dessa nova sociedade deve regular não apenas os algoritmos, mas também o fenômeno do *Big Data*. No entanto, é importante, antes de adentrar na questão regulatória em si, verificar quem deve ser regulado. Em outras palavras, para quem devem ser dirigidas as regras a serem criadas: para os seres humanos ou para asmáquinas?

Como visto no capítulo anterior há quem sustente um determinado grau de autonomia da inteligência artificial, justificando retirar do ser humano a responsabilidade pelos danos gerados pelo robô, tendo em vista que os desenvolvedores não teriam condições de prever,

---

[163]No original “Indeed, we are rapidly moving from the age of the Internet to the Algorithmic Society. We will soon look back on the digital age as the precursor to the Algorithmic Society. What do I mean by the Algorithmic Society? I mean a society organized around social and economic decision making by algorithms, robots, and AI agents; who not only make the decisions but also, in some cases, carry them out.”Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem: https://ssrn.com/abstract=2890965

[164]Fuel of the future - Data is giving rise to a new economy, THE ECONOMIST, May 6, 2017, disponível em: https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise -new-economy

durante o desenvolvimento do algoritmo, todas as suas ações. Tal fato se agrava quando se trata de um algoritmo dotado de aprendizado de máquina.

No entanto, como se apontou no terceiro capítulo, e será defendido mais à frente, assim como os princípios éticos que a norteiam, a futura regulação deve ser voltada para os seres humanos, e não para a máquina, diferentemente do que ocorre com as leis de Asimov. Suas leis eram direcionadas para robôs, isto é, eram instruções de programação inseridas no código dos próprios robôs. Tratam-se, portanto, de leis que os robôs tinham que seguir e não os seus desenvolvedores[165].

No entanto, o que essa Sociedade Algorítmica emergente necessita é de leis direcionadas para as pessoas que programam e usam robôs, agentes dotados de inteligência artificial e algoritmos. O conceito da Sociedade Algorítmica é o aproveitamento de dados ealgoritmos para governar a sociedade de forma mais eficiente.

No entanto, como visto até o momento, o uso de decisões autônomas gera diversos prejuízos e custos, especialmente para grupos vulneráveis: danos à reputação de um indivíduo,discriminação, modificação na própria autonomia e manipulação. Todos esses danos, ademais, ocorrem sem que seja dada qualquer transparência e, portanto, qualquer possibilidade de correção. Tais danos já foram extensivamente tratados nos capítulos anteriores, no entanto, para fins didáticos, tornaremos a aborda-los de forma resumida.

Existem duas formas centrais pelas quais os algoritmos afetam a reputação:por meio da classificação e da avaliação de risco. Ao realizar uma análise de risco um algoritmo classifica um indivíduo em um determinado nível de risco. Sendo assim, um algoritmo inclui um indivíduo (ou as pessoas que vivem em uma determinada área) em um grupo com um determinado risco financeiro, umrisco de cometer um crime futuro, um risco de gastar muitos serviços sociais,um risco de devolver itens ou ser um cliente caro e assim por diante. Dessa

[165]2 ISAAC ASIMOV, Runaround, in I ROBOT 37 (1950)

forma, o indivíduo é classificado e passa a ser tratado pela sociedade conforme este estigma, seja positiva ou negativamente.

Devido a essa análise de risco determinados indivíduo passam a ser sistematicamente discriminados, tendo negadasoportunidades que são oferecidas aos demais (cartão de crédito, empréstimo, oportunidade de emprego,promoção), ou sofrendo a imposição de custos especiais (maior suscetibilidade de ser parado pela polícia, vigilância, preços mais altos, exclusão da posse de armas ou acesso a viagens aéreas,etc.) que não são impostos a outras pessoas.

Como essa discriminação não é explícita, e por vezes sequer esperada por quem utiliza o algoritmo, ela acaba por ser normalizada na sociedade,de modo que o indivíduo acaba por internalizar suas classificações e avaliações de risco. Para evitar a vigilância e ser classificado como de risco, o indivíduo altera seus comportamentos e, portanto, sua identidade.

Em outras palavras, seres humanos e organizações podem usar algoritmos paralevar um indivíduo a fazer escolhas (mais ou menos) previsíveis que os beneficiem, mas que não aumentam o seu bem-estar. Além disso, a análise algorítmica torna mais fácil para as empresasdescobrir quais pessoas são mais suscetíveis à manipulação e como podemmais fácil e eficazmente serem manipuladas: trata-se de *profiling* e micro *targeting* já discutidos anteriormente.

Por fim, todos esses problemas ocorrem, como visto, sem que haja transparência, de como ou porque eles ocorrem, o que impede a monitorização e melhoria das operações, e impossibilita a refutação ou correção de inadequações. Em outras palavras, as decisões são tomadas e geram consequências sem que haja ummétodo de manter o algoritmo responsável.

Como bem resume Jack Balkin:

> Podemos resumir essa discussão dizendo que algoritmos (a) constroem identidade e reputação através de (b) classificação e avaliação de risco, criando aoportunidade

> para (c) discriminação, normalização e manipulação, sem (d)transparência, prestação de contas, monitoramento ou devido processo[166].

Dessa forma, fica claro que o problema central da regulação não são os algoritmos, mas sim os seres humanos que os usam, e que permitem seremgovernado por eles. Resumidamente, portanto, pode-se explicar a governança algorítmica como a governança dos seres humanospor seres humanos usando uma tecnologia particular de análise e tomada de decisão.

Portanto, a necessidade não é para leis dirigidas à robôs, como no caso das três leis de Asimov, mas sim de leis dirigidas a quem usa autômatos para analisar, controlar e exercitarpoder sobre outros seres humanos.

Para explicar esta escolha Jack Balkin se utilizou da lenda do Golen de Praga[167]na qual um Golen é criado pelo maharal, um sábio do século XVI amplamente reverenciado por seu aprendizadoe piedade, que o instrui a resolver problemas da comunidade judaica. Após ter cumprido com a sua missão, o Golen retorna ao seu criador e volta a ser argila.

O ponto desta lenda não é o que ocorre, mas sim o que não ocorre: o Golen não sai de controle nem é usado para fins maliciosos ou danosos. Isso ocorre porque quem o criou, e o controla, é um maharal, uma pessoa sábia. Como bem conclui Balkin:

> Quando falamos sobrerobôs, ou agentes de IA, ou algoritmos, geralmente nos concentramos em saber se eles causam problemas ou ameaças. Mas na maioria dos casos, o problema não é dos robôs. É dos humanos[168].

---

[166]No original “We might sum up this discussion by saying that algorithms (a) construct identity and reputation through (b) classification and risk assessment, creating the opportunity for (c) discrimination, normalization, and manipulation, without (d) adequate transparency, accountability, monitoring, or due process.”Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem SSRN: https://ssrn.com/abstract=2890965

[167]The Golem of Prague, in A TREASURY OF JEWISH FOLKLORE 603 (Nathan Ausubel ed., 1948)

[168]No original “When we talk about robots, or AI agents, or algorithms, we usually focus on whether they cause problems or threats. But in most cases, the problem isn’t the robots. It’s the humans.”Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem https://ssrn.com/abstract=2890965

São os seres humanos que projetam os algoritmos, os programam, conectambancos de dados e que decidem como usá-los, quando e para qual propósito.Além disso, também são os seres humanos que produzem, coletam e inserem os dados nos algoritmos. Dessa forma, pode-se considerar que todos os vieses e discriminações examinados anteriormente tem sua origem no ser humano.

Sendo assim, embora se fale sobre o que os robôs ou sistemas de decisão autônomas fizeram há, nesse caso, o que Balkin denomina de "falácia do homúnculo".

> A falácia homúnculo é a crença de que há uma pequena pessoa dentro do programa queestá fazendo funcionar - tem boas intenções ou más intenções e faz com queprograma faça coisas boas ou ruins.Mas, na verdade, não há pessoa dentro do algoritmo. Há simprogramação - código - e há dados. E o programa usa os dados para executar, comefeitos bons ou ruins, alguns previsíveis, alguns imprevisíveis[169].

Dessa forma, quando criticamos os algoritmos, estamos realmente criticando a programação, ouos dados a interação destes dois. Isso quer dizer, então, que estamos criticando, ainda que indiretamente, o usopara o qual eles estão sendo colocados pelos humanos que os programaram, coletaram os dados ou que empregaram os algoritmos e os dados para executardeterminadas tarefas.

A tecnologia apenas age na mediação das relações sociais entre os seres humanos, embora ela seja incorporada de modo a muitas vezes disfarçartais relações. Dessa forma, quando algoritmos discriminam ou fazem coisas ruins, eles estão apenas reproduzindo, e em alguns casos agravando, relações e problemas sociais já existentes. São as pessoas que verdadeiramente produzem e reproduzem justiça e injustiça, poder e impotência, status superior e subordinação.

---

[169]No original "The homunculus fallacy is the belief that there is a little person inside the program who is making it work—it has good intentions or bad intentions, and it makes the program do good or bad things. But in fact there is no little person inside the algorithm. There is programming—code—and there is data. And the program uses the data to run, with good or bad effects, some predictable, some unpredictable." Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592. Op cit.

Em outras palavras, os sistemas de decisão autônoma apenas são os dispositivos através dos quais essasrelações sociais são produzidas, através dos quais formas particulares de poder são processadas e transformadas. Há, portanto, uma substituição, que Balkin denominou de “efeito de substituição”.

> O efeito de substituição diz respeito aos efeitos sobre a sociedade produzidos pelo fatoque robôs, agentes de IA e algoritmos substituem os seres humanos. Eles operam comopessoas de propósito especial. A noção de robô ou algoritmo como substituto transmitequatro ideias diferentes: (1) O substituto é, em alguns aspectos, melhor que o original; (2)o substituto é de outras formas mais limitado que o original; (3) as pessoas tratam osubstituir como se estivesse vivo - eles se envolvem em animismo ou antropomorfismo; e(4) o substituto age como um fetiche ou deflexão longe das bases sociais do poderentre seres humanos e grupos de seres humanos[170].

Substituição significa superioridade: sistemas dotados de inteligência artificial sãomais poderosos e mais rápidos que os seres humanos. Eles podem ver, fazer e analisar com mais eficiência do que humanos. Logo, eles detêm a capacidade de tomar decisões muito melhor do que seres humanos. Além disso, eles nunca se cansam nem sofrem distrações ou remorso gerados por emoções[171].

No entanto, substituição também significa limitação ou deficiência. Sistemas de decisão autônomos também têm habilidades limitadas[172]. Embora possam realizar muito bem uma tarefa, eles só conseguem executá-la, pois carecem de muitas das características do julgamento humano. Em outras palavras, um algoritmo pode ser muito melhor que um humano em jogar xadrez, mas este mesmo algoritmo, diferente do humano, não consegue executar outras tarefas, mesmo que se trate de outro jogo.

---

[170]No original “The substitution effect concerns the effects on society produced by the fact that robots, AI agents, and algorithms substitute for human beings. They operate as special purpose people. The notion of robot or algorithm as substitute conveys four different ideas: (1) The substitute is in some ways better than the original; (2) the substitute is in other ways more limited than the original; (3) people treat the substitute as if it were alive—they engage in animism or anthropomorphism; and (4) the substitute acts as a fetish or deflection away from the social bases of power among human beings and groups of human beings.” BALKIN, Jack M., The Path of Robotics Law, 5 CALIF. L. REV. CIRCUIT 45, 46, 55-59 (2015). p. 57/59.

[171]Jack M. Balkin, The Path of Robotics Law, 5 CALIF. L. REV. CIRCUIT 45, 46, 55-59 (2015)

[172]Anupam Rastogi, Artificial Intelligence—Human Augmentation is what’s here and now, MEDIUM, January 12, 2017, at https://medium.com/reflections-by-ngp/artificial-intelligence-human-augmentation-is-wh ats-here-and-now-c+5286978ace0

A substituição também envolve a projeção da vida, agência e intenção paraprogramas e máquinas[173], o que gera a projeção de responsabilidades para o próprio algoritmo, o que nos leva novamente à falácia do homúnculo, já discutida.

Por fim, a substituição envolve um fetiche ou deflexão ideológica. Da mesma forma que sociedades antigas acreditavam quetotens, que eram objetos inanimados, estavam imbuídos de poderes mágicos, Marx argumentou que as pessoas em uma sociedade de mercado tratam a mercadoria como se ela tivesse valor. Na verdade, segundo Marx, o que lhes dá valor é justamente o fato de estarem inseridas em um sistema social relações sociais[174].

Em outras palavras as mercadorias não teriam valor em si mesmas, apenas o valor que lhes é atribuído pelo mercado. Os mercados, por sua vez, são relações sociais que tanto capacitam as pessoas quanto permitemque elas exerçam poder umas sobre as outras.

Embora exista uma considerável literatura antecipando[175], ou até mesmo esperando, por uma tecnologia "totalmente autônoma" na qual sistemas de *software* não serão monitorados, ou controlados por qualquer pessoa através da tecnologia *blockchain*, considerando o estágio atual dos sitemas de inteligência artificial e seu grau de autonomia, a lógica proposta por Marx pode ser aplicada para o uso de substitutos na forma de sistemas de decisão autônomos.

Essas tecnologias tornam-se parte das relações sociais de poder entre indivíduos e grupos. Não devemos confundir o Golem com o mahali. Os efeitos da tecnologia são sempre sobre as relações de poder entre seres humanos ou grupos de seres humanos.

---

[173]Jack M. Balkin, The Path of Robotics Law, 5 CALIF. L. REV. CIRCUIT 45, 46, 55-59 (2015)

[174]KARL MARX, CAPITAL: A CRITIQUE OF POLITICAL ECONOMY 81(Fredrick Angels ed., Samuel Moore & Edward Aveling, trans., Dover Publications, Inc. 2011)

[175]SAMIR CHOPRA & LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS (2011); VitalikButerin, Cryptographic Code Obfuscation: Decentralized Autonomous Organizations Are About to Take a Huge Leap Forward, BITCOIN MAG. (Feb. 8, 2014),https://bitcoinmagazine.com/articles/cryptographiccode-obfuscation-decentralized-autonomous-organizations-huge-leap-forward-1391849871

4.3. AS LEIS DA SOCIEDADE ALGORÍTMICA

Considerando que a regulação deve ser sobre os seres humanos que desenvolvem e utilizam sistemas de inteligência artificial, cabe agora questionar que leis serão essas.

Considerando ainda os princípios éticos discutidos no capítulo anterior, pode-se afirmar que as leis que precisamos são obrigações de negociação justa, não manipulação e não dominação entre aqueles que fazem e usam oalgoritmos e aqueles que são governados por eles.

Nesse sentido Jack Balkin propõe três leis para a Sociedade Algorítmica[176]:

- Quando houver uma relação direta de consumo, desenvolvedores e entidades que utilizam os algoritmos são fiduciários deinformações;
- Quando não houver uma relação direta de consumo, os desenvolvedores e entidades que utilizam os algoritmostêm deveres públicos. Se eles são governos, isso se aplica automaticamente. Se eles sãoatores privados, seus negócios são afetados por um interesse público;
- Oprincipal dever públicodas entidades que desenvolvem e se utilizam de algoritmos é o de não externalizar os custos edanos de suas operações. Nesse ponto, a melhor analogia para os danos da tomada de decisão algorítmica não é a regra aplicada à discriminação intencional, mas sim à poluição socialmente injustificada.

Ao analisar esta proposta, Frank Pasquale acrescentou mais uma lei[177]:

[176]Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592. Available at SSRN: https://ssrn.com/abstract=2890965

- Um robô deve sempre indicar a identidade de seu criador, controlador ou proprietário.

Trataremos dessas leis de forma aprofundada nos próximos tópicos.

4.3.1 Primeira lei: desenvolvedores e operadores de algoritmos são fiduciários de informações

A Sociedade Algorítmica é uma maneira de governar populações. Governança, por sua vez, significa o modo como as pessoas que controlam os algoritmos entendem, analisam, controlam, direcionam, ordenam e moldam as pessoas que são os titulares dos dados. Em outras palavras, as pessoas usam algoritmos para classificar, selecionar, compreender e tomar decisões sobre grupos de pessoas.

Essa relação não é simplesmente uma relação de mercadoe lucro. É,acima de tudo, uma relação de poder informacional. A inteligência artificial, e, portanto, a entidade que a utiliza, sabe muito sobre os titulares dos dados. No entanto, o contrário não é verdadeiro. Além de não saber praticamente nada sobre a inteligência artificial, também não é possível monitorar suas atividades e resultados.

Sendo assim, há uma assimetria de poder e de informação entre os operadores e os titulares dos dados. Verifica-se, portanto, que tal relação muito se assemelha à relação fiduciária.

Fiduciários são profissionais, a exemplo dos médicos e advogados, que estabelecem com o cliente uma relação na qual há uma assimetria significativa no conhecimento e

---

[177] Pasquale, Frank A., Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society (July 14, 2017). Ohio State Law Journal, Vol. 78, 2017; U of Maryland Legal Studies Research Paper No. 2017-21. Disponível em: https://ssrn.com/abstract=3002546

capacidade entre as partes. Além disso, o cliente não pode monitorar facilmente o que o fiduciário está fazendo em seu nome[178].

Na tentativa de balancear essa relação e garantir confiança, a lei exige que os fiduciários ajam com base na boa fé de modo a evitar conflitos de interesses com o cliente ou paciente. Fiduciários, frequentemente, também coletam informações pessoais confidenciais sobre seus clientes, que poderiam ser utilizadas para gerar danos. Por isso, a lei exige que eles protejam a privacidade de seus clientes. Cabe apontar que quando os fiduciários coletam e processam informações sobre seus clientes eles passam a ser denominados de fiduciários de informação[179].

Dessa forma, os fiduciários têm dois deveres para com seus clientes: o de cuidar, o que significa que devem agir de modo a não prejudicar os seus clientes; e o dever de fidelidade, segundo o qual o fiduciário não deve agir de modo a gerar conflitos de interesse com seus clientes.

Conforme bem aponta Balkin:

> A era digital criou um novo conjunto de entidades com muitos recursos semelhante aos fiduciários tradicionais. Eles incluem grandes empresas on-line como a Google, Facebook e Uber. Essas empresas coletam, reúnem, analisam e usam informações sobre nós. Na verdade, eles coletam enormes quantidades de informações sobre nós, que poderiam, em teoria, ser usadas em nosso detrimento. Essas empresas tornaram-se bastante importantes, em alguns casos indispensáveis, para nossa vida cotidiana. Há também uma assimetria de conhecimento entre as empresas e seus usuários finais e clientes[180].

[178]Jack M. Balkin, Information Fiduciaries and the First Amendment, 49 U.C DAVIS L. REV. 1183 (2015)

[179]Jack M. Balkin, Information Fiduciaries and the First Amendment, 49 U.C DAVIS L. REV. 1183 (2015), p. 1208.

[180]No original “The digital age has created a new set of entities that have many features similar to traditional fiduciaries. They include large online businesses like Google, Facebook and Uber. These businesses collect, collate, analyze and use information about us. Indeed, they collect enormous amounts of information about us, which could, in theory, be used to our detriment. These businesses have become quite important, in somecasesindispensable, to our everyday lives. There is also an asymmetry of knowledge between businesses and their end-users and clients.”Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem https://ssrn.com/abstract=2890965

Dessa forma, segundo o autor, essas entidades deveriam ser tratadas como fiduciários, recebendo, portanto, obrigações semelhantes perante seus clientes.

Obviamente, há algumas diferenças que devem ser levadas em consideração. Primeiramente, a monetização de dados pessoais é uma parte fundamental dos serviços prestados por essas empresas, sendo a principal fonte de lucro e o que lhes permite oferece-los gratuitamente. Dessa forma, deve-se considerar que o uso dessas informações para obtenção de lucro não viola, por si só, seu dever fiduciário.

Em segundo lugar, como dito acima, essas empresas utilizam os dados coletados como fonte de lucro. Dessa forma é do interesse delas, diferentemente dos profissionais tradicionais,que seus usuários continuem a expor o máximo possível as suas informações pessoais, produzindo um fluxo constante de conteúdo e links.

Por fim, as pessoas esperam que agentes fiduciários tradicionais, como médicos, façam muito mais do que simplesmente deixar de prejudicá-las:eles devem cuidar de seus interesses e informá-las sobre potenciais riscos. Esses mesmos deveres não são esperados de empresas que realizam análise de dados, como sistemas de pesquisa e redes sociais.

Considerando as diferenças apontadas, os fiduciários de informações digitais devem terobrigações diferentes dos fiduciários tradicionais e condizentes com os tipos de serviços que prestam. Como bem explica Balkin:

> A obrigação central dos fiduciários da informação digital é que eles não podem agircomo vigaristas - induzindo a confiança em seus usuários finais para obter informações pessoaise, em seguida, usando essa informação de maneiras que traem essa confiança e trabalhar contra ointeresses dos seus utilizadores finais[181].

Em outras palavras, as empresas online têm um dever de boa-fé, não podendo se comprometer em garantir a segurança dos dados e privacidade de seus clientes para, em

[181]No original “The central obligation of digital information fiduciaries is that they cannot act like con artists — inducing trust in their end-users to obtain personal information and then using that information in ways that betray that trust and work against the interests of their end-users”. Jack M. Balkin, Information Fiduciaries and the First Amendment, 49 U.C DAVIS L. REV. 1183 (2015). P.1224/1225

seguida,agir de maneira contrária a esses deveres manipulando e discriminando os titulares de dados com base em suas informações pessoais. Da mesma forma, tais deveres devem se estender a todos para quem essas empresas transferirem os dados de seus clientes. Pode-se considerar, portanto, que os deveres fiduciários estão ligados aos dados e caminham junto com eles.

Dessa forma, a primeira lei da robótica para a Sociedade Algorítmica é que asentidades, aqui incluídas empresas privadas, pessoas físicas e órgãos governamentais, queusarem robôs e algoritmos para fornecer serviços,têm deveres de boa fé e confiança para comseus usuários finais e clientes[182].

4.3.2Segunda lei: deveres públicos para com a sociedade em geral

As obrigações dos agentes fiduciários, e de terceiros que recebem dados e informações de agentes fiduciários, são para com seus clientes. Dessa forma, há que se questionar a quais obrigações devem responder entidades públicas e privadas em relação àqueles indivíduos e grupos com os quais não há uma relação de clientela.

Um exemplo claro desse problema é a polícia, que certamente não tem uma relação de clientela nem para com que visa proteger, nem tampouco com seus investigados. Pode-se dizer que a empresa que desenvolveu o algoritmo de análise preditiva e o vendeu para a polícia tem obrigações fiduciárias para com a própria polícia, sua cliente. No entanto, a polícia em si não tem nenhuma clientela.

Dessa forma, a lei que estabelece as obrigações fiduciárias para entidades que desenvolvem e se utilizam de sistemas de inteligência artificial não supre totalmente a necessidade de regulação. As empresas que empregam algoritmos em suas operações ainda

[182]Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem https://ssrn.com/abstract=2890965. P.25

podem causar danos à pessoas que não são seus clientes, com quem não têm nenhuma relação contratual. Conforme explica Jack Balkin:

> Se quisermos articular as regras da Sociedade Algorítmica, precisamos de algocomo MacPhersonv. Buick Motor Company para a era algorítmica. Isto é, precisamosreconhecer que o uso de algoritmos pode prejudicar não apenas o usuário final de um serviço,mas também muitas outras pessoas na sociedade[183].

MacPherson v. Buick Motor Company[184] tornou-se um caso emblemático porque foi a partir dele que as cortes mudaram seu entendimento, passando a entender que empresas detêm deveres públicos não apenas para com seus clientes, mas também perante terceiros que fossem prejudicados pelos produtos ou serviços defeituosos.

Deve-se considerar, então, que as empresas detêm deveres para com o público quando empregamsistemas de inteligência artificial. No entanto, quais deveres seriam esses, considerando que não podem ser descritos como quebra de confiança?

4.3.3Terceira lei: dever de não gerar "poluição algorítmica"[185]

Como visto nos capítulos anteriores a falta de transparência e o aprendizado de máquinas dificulta a determinação do processamento decisório de um algoritmo, de modo que não seria viável imputar a responsabilidade aos seus desenvolvedores.

No mesmo sentido, há que se considerar que o algoritmo em si não apresenta intenções de modo que não pode agir de má-fé ou com negligência, por exemplo. Dessa

---

[183]No original "If we are to articulate the rules of the Algorithmic Society, we need something like MacPherson v. Buick Motor Company for the algorithmic age. That is, we need to recognize that the use of algorithms can harm not only the end-user of a service, but many other people in society as well"Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem: https://ssrn.com/abstract=2890965. P.28

[184] 111 N.E. 1050 (1916)

[185]algorithmic nuisance no original

forma, não há como se aplicar a teoria da responsabilidade por ato de terceiros (prevista no art.932 do Código Civil).

Atualmente, ativistas acusam processos algorítmicos de apresentar vieses[186], enquanto seus proprietários ou programadores se desviam de tais acusações, insistindo que a empresa não tinha a intenção de discriminar[187]. No momento, o conceito de "Impacto desigual"[188] (*disparateimpact*, no original) permitiu uma paz desconfortável entre os dois lados: *designers* responsáveis pelo desenvolvimento de sistemas algorítmicos estão se comprometendo a tentar evitar distribuições de benefícios ou ônus em relação a categorias historicamente vulneráveis, como raça, gênero ou orientação. No que diz respeito às determinações financeiras e de crédito, a regulamentação já assegurou que "Intenção" não é uma condição *sine qua non* para responsabilidade.

Esse afastamento da necessidade de intenção é importante visto que a "agência orientada por dados", comum em sistemas algorítmicos, baseia-se em "informação e comportamento, não em significado e ação"[189]. O contraste trazido por MireilleHildebrandt[190]entre informação e significado é crítico: intençãotem sido crucial para as determinações legais, especialmente no direito penal e no direito civil, sendo utilizado para estabelecer o grau de culpabilidade atribuído ao réu e, em alguns casos, até mesmo para a definir se uma ação foi ilegal em si.

---

[186] Julia Angwin et al., Machine Bias, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[187] PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 39 (2015) (discutindo o trabalho de Latanya Sweeney); vertambém WILLIAM DIETERICH ET AL., COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY (Julho 8, 2016), http://go.volarisgroup.com/rs/430- MBX-989/images/ProPublica_Commentary_Final_070616.pdf (respostaàsacusações de Angwin et al., nota 50); Brief of Defendant-Appellant, State v. Loomis, 881 N.W.2d 749 (Wis. 2016) (No. 2015AP157-CR), 2015 WL 9412098; Brief of Plaintiff-Respondent, Loomis, 881 N.W.2d 749 (No. 2015AP157-CR), 2016 WL 485419.

[188] Solon Barocas& Andrew D. Selbst, Big Data's Disparate Impact, 104 CALIF. L. REV. 671, 674 (2016)

[189]Mireille Hildebrandt, Law as Information in the Era of Data-Driven Agency, 79 MODERN L. REV. 1, 2 (2016).

[190]Op cit

No entanto, não se pode argumentar que o algoritmo em si tenha más intenções: o algoritmo é usado por seres humanos para alcançar alguma finalidade específica, mas que, no processo, acaba prejudicando vários grupos de pessoas. Algumas dessas vítimas são fáceis de identificar, mas por vezes, os danos são difusos, de modo a não ser possível identificar um indivíduo prejudicado. Um exemplo claro é justamente o caso da polícia preditiva, pois embora possa haver um indivíduo que foi parado ou preso injustamente, o dano à coletividade é muito mais sério e difícil de identificar.

Em essência, estamos falando sobre o uso socialmente injustificado de algoritmos que externalizam custos para outros. No direito civil, podemos chamar essa externalização de um incômodo (*nuisance*), seja ele público ou privado.

Sendo assim, a regulação deve focar nos efeitos sociais provocados pelos algoritmos, verificando se esses efeitos são razoáveis ou não. Nesse sentido, o melhor é aplicar como analogia a questão da poluição no direito ambiental[191].

Além disso, os danos causados por algoritmos são questões de grau. Tais danos resultam dos efeitos cumulativos da coleta de dados, análise e tomada de decisão. Dessa forma, como bem ressalta Andrew Selbst[192], os danos causados pela discriminação algorítmica não se encaixambem em uma simples categorização binária de sim ou não, ocorre discriminação ou não. Pelo contrário, há sempre inúmeros compromissos no *design*, nos dados coletados, na forma como os programadores formulam o problema a ser resolvido, entre diversas outras variáveis vistas anteriormente.

---

[191]Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem https://ssrn.com/abstract=2890965. P.34/35

[192]Selbst, Andrew D., Disparate Impact in Big Data Policing (February 25, 2017). 52 Georgia Law Review 109 (2017). Disponívelem: https://ssrn.com/abstract=2819182 or http://dx.doi.org/10.2139/ssrn.2819182. p.46/47; vejatambémAbe Gong, Ethics for Powerful Algorithms (1 of 4), MEDIUM (July 12, 2016), https://medium.com/@AbeGong/ethics-for-powerfulalgorithms-1-of-3-a060054efd84#.35pjrg22k.

Dessa forma, pode ser difícil determinar um parâmetro de ação não discriminatória no qual se basear as operações do algoritmo, além de ser difícil, se não impossível, isolar uma única causa que responsa pelos efeitos das operações do algoritmo. Sendo assim, torna-se relevante responder se os custos impostos sobre a sociedade ao utilizar o algoritmo são injustificáveis. Discriminação algorítmica, portanto, assim comoa poluição, é uma questão de grau.

Ao se aplicar a ideia de *nuisance* passa-se a poder explicar como os danos gerados pela Sociedade Algorítmica surgem das tomadas de decisão cumulativas por uma ampla gama de atores públicos e privados.

Empresas eos governos usam *Big Data* e algoritmos para fazer julgamentos que criam identidades, características e associações de pessoas. Essas construções e remodelação das identidades afetam as oportunidades dos indivíduos, tais como para receber um emprego ou crédito, bem como as suas vulnerabilidades, passando determinados indivíduos a sofrerem um aumento na vigilância,discriminação e, como resultado, na sua exclusão da sociedade.

Outras empresas, público e privadas, bem como o governo, coletam os novos dados gerados por essas classificações e voltam a aplicá-los na sociedade, o que agrava tanto as oportunidades como as vulnerabilidades dos indivíduos. Trata-se, portanto, de um ciclo infindável e autossustentável, visto que as empresas e governos estão sempre empregando todas essas informações de maneira criativaem novos contextos, produzindo sempre novos *insights*, julgamentos eprevisões. Desta forma, a vida das pessoas está sujeita a uma cascata de julgamentos por algoritmos.

O problema central, portanto, não são os vieses em si, como bem dito por JuliaPowles[193], mas esse dano cumulativo gerado pelas relações sociais preconceituosas e agravado pela utilização de algoritmos.

[193]POWELS, JULIA. The Seductive Diversion of ‘Solving’ Bias in Artificial Intelligence.Disponívelem https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

De fato, em alguns casos, o dano pode ser causado por uma programação descuidada ou a erros no desenvolvimento do algoritmo. Podem ainda resultar de dados coletadosdiscriminatórios. No entanto, em muitos outros casos, a empresa que desenvolveu e utiliza o algoritmo pode, de forma plausível, como visto no caso do sistema COMPAS, alegar que seu modelo inicial é razoável, dada a tarefa em questão, e que os dadosanalisados e os pressupostos de fundo do modelo são válidos. Contudo, ainda assim o modelo pode gerar danos para a sociedade, especialmente para grupos vulneráveis.

Considerando a teoria da *nuisance*, do direito ambiental, mesmo que o algoritmo tenha sido desenvolvido e esteja sendo utilizado com os devidos cuidados pela empresa, caso seja gerado um nível não aceitável de dano à sociedade, a empresa deve ser penalizada. Como bem explicou Jack Balkin:

> Neste mundo, concentrar-se em intencionalidade, ou mesmo em negligencia,na construção e supervisão de algoritmos pode ser inadequado. Em vez disso, a melhoranalogia na teoria do direito civil pode ser a verificação dos custos sociais que surgem de níveis de atividade socialmenteinjustificados. Níveis aumentados de atividade produzem aumento decustos, mesmo quando uma atividade é conduzida com o devido cuidado. Mesmo supondo quea empresa exerça o devido cuidado - o que, naturalmente, pode não ocorrer - os efeitos cumulativosdo aumento da atividade pode, no entanto, gerar muitos danospara a sociedade. Estas são situações características de *nuisance*[194].

Dessa forma, a terceira lei estabelece que operadores de algoritmos têm o dever público de não "poluir", ou seja, de não externalizarinjustificadamenteos custos da tomada de decisão algorítmica de modo a gerar danos à sociedade.

Uma questão crítica que deve ser apontada para este regime é a forma pela qual identificamos efeitos tão negativos quemereçam intervenção regulatória ou responsabilidade.

---

[194]No original "In this world, focusing on intentional tort or even on the negligent construction and supervision of algorithms may be inadequate. Instead, the best analogy in tort law theory may be to the social costs that arise from socially unjustified levels of activity. Increased activity levels produce increased social costs, even when an activity is conducted with due care. Even assuming that the firm exercises due care—which, of course, it may not—the cumulative effects of increased activity may nevertheless throw too much harm onto the rest of society. These are characteristic situations of nuisance."Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem https://ssrn.com/abstract=2890965. P.38

Balkin sugere uma forma de análise de custo-benefício: ele compara o incômodo algorítmico com a “poluição socialmente injustificada” e insta os formuladores de políticas aimpedir que os usuários de algoritmos externalizem “os custos e danos de suas operações”[195]. Ele argumenta, ainda,que os formuladores de políticas devem se concentrar em métodos robóticos “que não se justificam na análise de custo benefício sob o ponto de vista da sociedade como um todo”.

Ainda que se aceite que o processamento algorítmico de informações possa impor custos indevidos a terceiros. A análise custo-benefício dever ser apenas um dos métodos que podem ser usados paraidentificar incômodos (*nuisance*) algorítmicos. A análise de custo-benefício é manipulável e pode ocultar tanto quantorevelar sobre julgamentos de valores importantes[196]. Além disso, voltando aos argumentos levantados no início do terceiro capítulo, éimportante manter os padrões deontológicos de justificação no mundo da tecnologia,de modo a complementar o utilitarismo da análise de custo-benefício.

Considere-se, por exemplo o caso de análise preditiva utilizado pela polícia, que se utiliza de dados provenientes de prisões anteriores. Nesse caso, como visto, a análise preditiva provavelmente irá reproduzir os preconceitos contidos nesses dados. Embora seja difícil quantificar o custo de tais preconceitos, isto não torna a discriminação menos objetável, independentemente de seus efeitos.Como bem explica Frank Pasquale:

> Para complementar a história de Balkin sobre o Golem, eu me lembro do Evangelho de Lucas,onde Jesus relata a parábola da ovelha perdida. Deixar 99 ovelhas em “campo aberto”, para encontraruma não é necessariamente uma decisão economicamente racional. Da mesma forma, a Lei de Espécies Ameaçadas tem sido criticado por tecnocratas por colocar valor excessivo em espécies incomuns. Mas cada decisão reflete a espontaneidade e o idealismo totalmente humanos. Esses

[195]Balkin, Jack M., The Three Laws of Robotics in the Age of Big Data (August 27, 2017). Ohio State Law Journal, Vol. 78, (2017), Forthcoming; Yale Law School, Public Law Research Paper No. 592.Disponívelem https://ssrn.com/abstract=2890965. P.35

[196]Pasquale, Frank A., Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society (July 14, 2017). Ohio State Law Journal, Vol. 78, 2017; U of Maryland Legal Studies Research Paper No. 2017-21. Available at SSRN: https://ssrn.com/abstract=3002546. P.7

> valores não devem ser sacrificados em nome de alguma monetização cientificista do valor dos serviços ecossistêmicos, ou cálculos semelhantes[197].

Ainda que se argumente em prol dos possíveis ganhos em eficiência econômica decorrentes de tais classificações, como no caso, por exemplo de patrulhasfeitas por robôs, que poderiam facilmente reduzir ferimentos e aumentar a ordem no policiamento de manifestações, tais ações devem ser evitadas em prol da proteção de direitos considerados pela sociedade como superiores. No caso da patrulha,uma sociedade justa poderia decidir proibi-las, mesmo que seja provável que tal decisão diminua o bem-estar geral, sob o argumento de que é mais importante evitar um tipo de opressão automatizada, cuja probabilidade de ocorrência era impossívelcalcular a *priori*.

Deve-se, também, manter sempre como alternativa a possibilidade de proibições do uso de sistemas de decisões autônomas que contenham certos tipos de dados, como, por exemplo, no caso de sistemas de pontuação para concessão de crédito que utilizam qualquer forma de dados de saúde.

4.3.4Quarta lei: rastreabilidade dos algoritmos

Uma das ideia centrais das leis criadas por Jack Balkin é a garantiada existência de explicação para as decisões autônomas e, através dessa explicação, um agente responsável pelos danos ocasionados para o indivíduo e para a sociedade. Um elemento-chave da explicabilidade é a garantia de que a história do algoritmo esteja clara, ou seja, que esteja evidente como esse algoritmo foi programado, para qual fim será utilizado, e como a

[197]No original “To complement Balkin’s story of the Golem, I am reminded of the Gospel of Luke, where Jesus relates the parable of the lost sheep. Leaving 99 sheep in “open country,” to find one, is not necessarily an economically rational decision. Similarly, the Endangered Species Act has been decried by technocrats as placing inordinate value on unusual species. But each decision reflects altogether-human spontaneity and idealism. Such values should not be sacrificed in the name of some scientistic monetization of the value of ecosystem services, or similar calculations.”Pasquale, Frank A., Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society (July 14, 2017). Ohio State Law Journal, Vol. 78, 2017; U of Maryland Legal Studies Research Paper No. 2017-21. Disponívelem https://ssrn.com/abstract=3002546.p.8

interação entre *hardware,software* e o ambiente externo resultaram no seu comportamento atual. Dessa forma, no centro das leis de Balkin há a preocupação em ligar indivíduos ou empresas, que atuariam no papel de mahali, o criador do golem, que seriam responsáveis por suas criações.

Considerando-se que as leis foram desenvolvidas para regular pessoas, e não algoritmos ou robôs, é essencial que haja algum monitoramento do que os proprietários e programadores estão criando e codificando. Dessa forma, torna-se necessário estabelecer algumas regras básicas, ou pré-regulamentação, das interações que os algoritmos terão com o resto do mundo para garantir a eficácia de tal monitoramento.

Sendo assim, Frank Pasquale[198] adiciona uma quarta lei às três já desenvolvidas por Balkin, segundo a qual deve-se garantir que robôs e agentes algorítmicos sejam identificados e rastreados até os seus criadores.Já há avanços nesse sentido: analistas propuseram uma licença para drones, para ligar qualquer voo imprudente ou negligente ao dono ou controlador, e já existe o registro de drones nos EUA[199].

Com base no meta-princípio introduzido pela lei zero de Asimov que estipula que Robôs não devem prejudicar a humanidade, esta quarta lei gera a presunção de que qualquer robô ou sistema algorítmico tenha um criador, controlador ou proprietário.

A vanguarda dos campos de IA, aprendizado de máquina e robótica enfatiza a autonomia, de modo que há uma noção generalizada de que tais sistemas poderiam escapar ao controle dos humanos, o que levaria à ideia de que também seus criadores deveriam escapar da responsabilidade uma vez que a fuga ocorreu. Uma exigência de que qualquer sistema de inteligência artificial tenha alguém designado como responsável por suas ações ajudaria a sufocar tais ideias.

---

[198]Pasquale, Frank A., Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society (July 14, 2017). Ohio State Law Journal, Vol. 78, 2017; U of Maryland Legal Studies Research Paper No. 2017-21. Disponívelem https://ssrn.com/abstract=3002546

[199]Joseph Lorenzo Hall, 'License Plates' for Drones, CDT BLOG (Mar. 8, 2013), https://cdt.org/blog/license-plates-for-drones/

Sendo assim, o que quer que afete a evolução de tais máquinas será de responsabilidade do criador original, que deve ser obrigado a construir certasrestrições na evolução do código para a) registrar influências eb) evitar resultados ruins. Possivelmente, alguns robôs e algoritmos irão evoluir para longe dos ideaisprogramado neles por seus proprietários, como resultado de interações com outras pessoas emáquinas. Nesses casos, deve-se considerar que haverão várias partes potencialmente responsáveisque poderão responder.

Segundo Caitlin Mulholland[200], em havendo falta de clareza do nexo causal entre os agentes, pode-se atribuir a responsabilidade ao grupo econômico como um todo, possibilitando a reparação dos danos causados através da facilitação do ônus probatório para a vítima. Dessa forma, todos os envolvidos no desenvolvimento do sistema de inteligência artificial podem ser responsabilizados por sua ação.

> Dentre as possibilidades teóricas explicadoras do fenômeno da presunção da causalidade, surgem, da teoria da causalidade adequada, outros fundamentos que utilizam a idéia da socialização dos riscos como base para que se considere a existência presumida do nexo causal. Um destes fundamentos é a chamada repartição dos riscos que busca criar uma doutrina de imputação do dever indenizatório a partir da idéia de desenvolvimento e geração de um risco irrazoável como fator de atribuição de responsabilidade. Esta teoria foi primeiramente baseada em casos de causalidade múltipla em que, (...), é virtualmente impossível a identificação exata da contribuição causal de cada um dos potenciais agentes danosos (causalidade alternativa)[201].

Após esta explicação a autora segue com a argumentação, delimitando esta aplicação à existência de alguns parâmetros:

> A responsabilidade baseada neste método probabilístico de análise do risco (chamada também de responsabilidade estocástica), deve fundamentar-se, portanto, nos seguintes parâmetros: (i) impossibilidade objetiva da prova do nexo de

[200]MULHOLLAND, Caitlin Sampaio. A responsabilidade civil por presunção de causalidade. Rio de Janeiro: GZ, 2010.
[201] MULHOLLAND, Caitlin Sampaio. A responsabilidade civil por presunção de causalidade. Rio de Janeiro: GZ, 2010. p.290/291

> causalidade; (ii) desenvolvimento de atividade altamente arriscada; e (iii) verificação de dano tipicamente associado à atividade realizada[202].

Considere-se, então, para fins de exemplo, o caso do *chatbot* Tay que gradualmente aprendeu, através de interações no *Twitter*, certos padrões de diálogo, passando rapidamente a adotar os padrões de fala de um simpatizante nazista. A *Microsoft* não o programou para ter essa conduta e possivelmente não previa este resultado, ainda que pudesse prever os possíveis riscos de expor um *bot* à uma plataforma notória por sua fracamoderação sobre discursos de ódio e assédio.

No entanto, na medida em que registrou de onde os discursos se originavam, poderia ter reportado a conta ao *Twitter*, que por sua vez, poderia ter tomado algumas medidas para suspender ou retardar tais discursos. Dessa forma, não há como apontar objetivamente o nexo de causalidade e caso tenha ocorrido dano este está diretamente relacionado com a atividade. Restaria apenas analisar se a atividade de implementação de novas tecnologias poderia ser considerada como de risco.

Considerando as leis apontadas até o momento pode-se entender que o risco deverá variar de acordo com o grau de de *nuisance*, ou de discriminação, no caso específico do policiamento preditivo, que uma decisão automatizada pode gerar.

4.4. A NECESSIDADE DE APROFUNDAR OS ESTUDOS

Muitas forças tendem a regular nossa conduta, variando de governo a corporações a profissionais. No mesmo sentido, regimes jurídicos também criam salvaguardas para assegurar que a regulamentação não torne-se muito opressiva, o que se trata de modos de

[202]MULHOLLAND, Caitlin Sampaio. A responsabilidade civil por presunção de causalidade. Rio de Janeiro: GZ, 2010. p.293

regular essa regulamentação. O mesmo deve ocorrer com a implementação de sistemas de decisões automatizadas.

Sempre que os algoritmos se movem para qualquer nova área, novas formas de regulação da regulação devem ser criadas para complementar as antigas. Sem essa modernização, uma regulação arbitrária poderia induzir a implementação de formas de inteligência artificial sem o devido cuidado, o que poderia suplantar prematuramente formas mais responsáveis de agir.

Conforme esposto até o momento, os sitemas de decisões autônomos podem gerar discriminação porque são contaminados com preconceitos e viéses humanos. No entanto, pouco se sabe efetivamente como esta discriminação ocorre, bem como sua extenção e potencial lesivo. Da mesma forma, pouco se sabe sobre a real efetividade desses sistemas de policiamento preditivo.

Isso ocorre porque a grande maioria desses sistemas são opacos, protegidos pelas alegações de segredo industrial de seus desenvolvedores. Sendo assim, a polícia não tem conhecimento sobre o sistema que ela adota e utiliza.

Neste capítulo, até o presente momento se apresentou hipóteses de regulações das regulações de modo que a implementação de sistemas de inteligência artificial possam ocorrer evitando ou diminuindo os danos à sociedade.

Em relação ao policiamento preditivo cabe, no entanto, devido à falta de conhecimento que se tem sobre os sistemas de deicisões automatizadas, é difícil dizer abstratamene quais tipos de regulação podem necessárias, ou quão grande é o problema que a tecnologia efetivamante apresenta. Dessa forma, apenas regulações de regulações não são suficientes como solução. Necessita-se de mais informações sobre o a implementação da tecnologia e, no momento, há pouco incentio para que os diversos atores envolvidos possam entende-la.

Diante deste cenário Andrew Selbst propõem que, antes de aplicar sistemas de policiamento preditivo, a polícia deva ser obrigada a criar "Análises de impacto algorítmico" (AIS na sigla em inglês)[203], baseadas nas declarações de impacto ambiental (EIAs na sigla em inglês) da Lei Nacional de Política Ambiental americana (NEPA).

> O objetivo desta proposta de legislação não é necessariamente restringir o uso de novas tecnologias de policiamento preditivo. Os AISs têm dois propósitos. Primeiro, eles assegurariam que a polícia, ao chegar a sua decisão, terá disponível, e poderá cuidadosamente considerar, informações detalhadas sobre impactos [discriminatórios]. Segundo, tais análises garantiriam que as informações relevantes serão disponibilizadas para a sociedade, que também pode desempenhar um papel, tanto no processo de tomada de decisão, quanto na implementação dessa decisão[204].

A implementação de análises de impacto já vem sendo aplicada em diversos setores, sendo, inclusive, previsto nas leis gerais de proteção de dados brasileira e européia (na qual se denomina *Data Protection Impact Assessements* – DPIA) sempre que o processamento de dados "possa resultar em um alto risco aos direitos e liberdades das pessoas"[205].

Cabe, por fim, esclarecer que a presente dissertação não pretende realizar qualquer julgamento sobre qual seria o melhor tipo de regulação relacionada ao uso de AI, mas apenas estudar as diferentes formas de se tratar as aplicações de AI, as quais variam conforme a cultura e os valores de cada sociedade. As sociedades orientais vêm demonstrando grande otimismo em relação à tecnologia em geral, e à inteligência artificial em específico, muito distinto do pessimismo e temor sentido pelas sociedades ocidentais. Possivelmente, essa distinção em relação aos efeitos de tal tecnologia decorre de diferenças culturais, filosóficas e ideológicas.

---

[203]Selbst, Andrew D., Disparate Impact in Big Data Policing (February 25, 2017). 52 Georgia Law Review 109 (2017). Disponível em SSRN: https://ssrn.com/abstract=2819182 or http://dx.doi.org/10.2139/ssrn.2819182. p.168

[204] No original "The goal of this proposed legislation is not necessarily to curtail the use of new predictive policing technologies. The AISs would have two purposes. First, they would ensure that the agency, in reaching its decision, will have available, and will carefully consider, detailed information concerning significant [discriminatory] impacts. Second, they would guarantee that the relevant information will be made availableto the larger audience that may also play a role in both the decision making process and the implementation of that decision." Selbst, Andrew D., Disparate Impact in Big Data Policing. Op Cit. p. 169

[205]Art.35 da GDPR

Espera-se que a implementação das leis aqui estudadas contribua para afastar este temor ocidental de uma futura dominação pela Inteligência artificial e contribuam para seu desenvolvimento de forma ética e consciente.

O Brasil, como país que ainda não iniciou seu planejamento e desenvolvimento da regulação relacionado ao uso de AI pode e deve se aproveitar dessa aparente desvantagem. É fundamental o estudo das abordagens feitas por outras nações, utilizando tais pontos com o intuito de encontrar o caminho mais adequado à nossa realidade.

Além disso, é essencial encurtar os espaços existentes entre o uso de novas tecnologias e respectivas regulações e regulamentações nacionais. O Direito brasileiro precisa acompanhar esta nova realidade, oportunizando maior proteção de seus cidadãos, sem deixar de estimular a inovação e o empreendedorismo.

CONCLUSÕES

O policiamento preditivo está sendo adotado rapidamente em diversos países, inclusive no Brasil, embora ainda não esteja claro se as tecnologias até mesmo oferecer qualquer benefício tangível sobre o policiamento tradicional, além de não haver pesquisas suficientes e conclusivas sobre seus efeitos discriminatórios.

Decisões automatizadas apresenta novos desafios regulatórios em diversos setores da sociedade, mas quando se trata da polícia em particular, o histórico de discriminação clama por novas formas de supervisão e transparência. Os departamentos de polícia devem, portanto, garantir que não estejam adotando uma tecnologia que produza benefícios limitados, igualando "criminoso" a "negro".

Se eles permanecerem desregulados, os sistemas de policiamento preditivo devem endurecer e perpetuar a discriminação racial que permeia o sistema de justiça criminal. A menos que a sociedade reconheça a urgência e aja, nos tornaremos acostumados com a discriminação tóxica emitidas por estes sistemas de policiamento preditivo.

A atração narrativa de “confiar nos dados” irá codificar a discriminação racial na tecnologia, gerando um “*feedback loop*” e tornando ainda mais difícil erradica-la posteriormente. Considerando o histórico de policiamento discriminatório, nenhuma tecnologia tecnologia deve ser adotada sem que haja uma investigação sobre como esta afeta populações minoritárias. A sociedade não pode permitir que o fascínio de novas tecnologias a cegue em relação às desigualdades sistêmicas que estas podem perpetuar.

Para que a sistemas de decisões automatizadas tenham uma influência positiva e sustentável sobre a humanidade, sua aplicação deve ser guiada por diretrizes éticas, além de normas regulatórias direcionadas para quem as desenvolvem e aplicam. Possivelmente ainda é muito cedo para se propor normas específicas, visto que a tecnologia é muito nova e ainda esta em pleno desenvolvimento.

É necessário, portanto, que haja princípios éticos e normas que possam direcionar esta futura regulação, bem como que viabilizem o estudo dos resultados ocasionados pela utilização de deicsões autônomas. Devido à falta de transparência pouco se sabe sobre os sitemas de deicsões autônomas já utilizados de modo que é dificil analisar suas reais consequências e propor melhorias, ou até a sua não implementação.

O puro poder e as capacidades dos sistemas autônomos eliminaram gradualmente a necessidade de os humanos se envolverem em vários estágios do processo de tomada de decisões. No entanto, o objetivo final dessas tecnologias é servir ao humano; não o contrário. Dessa forma, as comunidades que as constroem devem ser responsáveis por seguir princípios éticos, pois somente através da inteligibilidade, precisão e imparcialidade a humanidade pode alavancar o poder, e mitigar os riscos, que o uso da inteligência artificial continua a desenvolver. A dívida técnica nos sistemas de decisão automatizada não deve levar à dívida ética na sociedade.

Considerando que o verdadeiro perigo é, e sempre foi, as pessoas, as organizações e empresas que adotam e empregam esses sistemas de decisões automatizadas, e usam-os para afetar, controlar e manipular outros seres humanos os princípios éticos e as normas regulatórias devem ser direcionados para eles, pelo menos até que a inteligência artificial alcance o grau de autonomia que se espera dela.

REFERÊNCIAS BIBLIOGRÁFICAS

MCCARTHY, John. “What is Artificial Inteligence?”.2007, Disponível em: http://www-formal.stanford.edu/jmc/whatisai/

RUMERLHART, David E; HILTON, Geoffrey E e WILLINANS, Ronald J. Learning Representations by back-propagating erros.In Nature, vol 323, issue 9, outubro de 1986

Erik, citando ITO, Joi e HOWE, Jeff. Whiplash: how to survive our faster future. 2016, New York and Boston, Grand Central Publishing.

FLORIDI, Luciano et al. (2018). AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines 28(4): 689-707.

LEITE, Renato. https://igarape.org.br/wp-content/uploads/2018/12/Existe-um-direito-a-explicacao-na-Lei-Geral-de-Protecao-de-Dados-no-Brasil.pdf

DOSHI-VELEZ, FINALE, AND MASON KORTZ.2017. Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper.

Selbst, Andrew D. and Powles, Julia, Meaningful Information and the Right to Explanation (November 27, 2017). International Data Privacy Law, vol. 7(4), 233-242 (2017).DisponívelemSSRN: https://ssrn.com/abstract=3039125

Wachter, Sandra &Mittelstadt, Brent &Floridi, Luciano. (2017). Transparent, explainable, and accountable AI for robotics.Science Robotics. 2. 10.1126/scirobotics.aan6080.

Jusbrasil. Disponível em: https://stj.jusbrasil.com.br/noticias/100054780/teoria-doadimplemento-substancial-limita-o-exercicio-de-direitos-do-credor

FERRARI, Isabela e BECKER, Daniel. O direito à explicação sobre decisões automatizadas: uma análise comparativa entre a União Europeia e o Brasil.Revista de Direito e as Novas Tecnologias 2018 VOL. 1 - OUT/DEZ 2018

BRASIL. Ministério da Justiça. Portaria nº 5 de 27 de agosto de 2002. Dispõe sobre cláusulas abusivas em contratos de vendas de produtos e prestação de serviços. Diário Oficial da República Federativa do Brasil, Brasília, DF, 28 ago. 2002. Disponível em: https://www.procon.go.gov.br/legislacao/portarias/ portaria-n%C2%BA-5-27-08-2002-mj-sde-clausulas-abusivas-nome-de-consumidor-a-banco-dedados.html.

PORTO, Antonio José Maristrello; FRANCO, Paulo Fernando de Mello. Por uma análise também econômica da responsabilidade civil do cadastro positivo: abordagem crítica do art. 16 da Lei 12.414/2011. Revista de Direito do Consumidor, São Paulo, v. 115, p.247 -271, jan.-fev. 2018.

RESP nº 1.419.697 RS. Disponível em: https://stj.jusbrasil.com.br/jurisprudencia/152068666/recurso-especial-resp-1419697-rs-2013-0386285-0/relatorio-e-voto-152068681

RESP nº 1.304.736/RS. Disponível em: < https://stj.jusbrasil.com.br/jurisprudencia/178798658/recurso-especial-resp-1304736- rs-2012-0031839-3>https://stj.jusbrasil.com.br/jurisprudencia/178798658/recurso-especial-resp-1304736-rs-2012-0031839-3

MONTEIRO, R. L. (2017). "Proteção de dados e a legislação vigente no Brasil". Baptista Luz. Disponível em: http://baptistaluz.com.br/wp-content/uploads/2017/11/Privacy-Hub-Leis-Setoriais.pdf.

FRAZÃO, Ana. 2018. Disponível em https://www.jota.info/opiniao-e-analise/colunas/constituicao-empresa-e-mercado/controversias-sobre-direito-a-explicacao-e-a-oposicao-diante-de-decisoes-automatizadas-12122018

Data Protection Working Party (A29WP). "*Guidelines on Automated Individual Decision Making and Profiling*".Disponívelem: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053

https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-does-the-gdpr-say-about-automated-decision-making-and-profiling/

SELBST, Andrew; POWLES, Julia.Meaningful information and the right to explanation.International Data Privacy Law, v. 7, n. 4, 2017. Disponível em [https://academic.oup.com/idpl/article/7/4/233/4762325]

LESSIG, L. Code and other laws of cyberspace. New York: Basic Books, 2006

O'NEIL, Cathy. "Weapons of Math Destruction". Ed. Crown. New York.2016 2.

PASQUALE, Frank. "The Black Box society: the secret algorithms that control money and information". Harvard University Press. 2016

BARROCAS, Solon & SELBST, Andrew D., "Big Data's Disparate Impact", 104 CALIF.L.REV. 671 (2016).Disponível em: file:///C:/Users/Dennys%20Game/Downloads/Big%20Data%20disparate%20im pact%20-%20Solon%20Barrocas%20(1).pdf

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. "Detecting Influenza Epidemics Using Search Engine Query Data." Nature 457.7232 (2008): 1012-1014. Disponível em: http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/p apers/detecting-influenza-epidemics.pdf

LETOUZÉ, Emmanuel, Global Pulse, "Big Data for Development: Challenges &Opportunities". Disponível em

http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf

Lohr, Steve. “When There’s No Such Thing as Too Much Information.” The New York Times. 23 Apr. 2011.

“Big Data é o novo petróleo, afirma executiva da IBM” acessado em https://olhardigital.com.br/noticia/big-data-e-o-novo-petroleo,-afirma-executiva- da-ibm/34986

“The world’s most valuable resource is no longer oil, but data” acessado em https://www.economist.com/news/leaders/21721656-data-economy-demands-new- approach-antitrust-rules-worlds-most-valuable-resource

SAISSE, Renan. “Big Data contra o crime: efeito Minority Report”. Disponível em http://direitoeti.com.br/artigos/big-data-contra-o-crime-efeito-minority- report/

https://www.jota.info/opiniao-e-analise/artigos/o-debate-sobre-personalidade-juridica-para-robos-10102017

https://www.sentient.ai/blog/understanding-black-box-artificial-intelligence/

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/new-york-city-takes-algorithmic-discrimination

https://tecnoblog.net/197933/tesla-model-s-piloto-automatico-acidente-morte/

https://www.publico.pt/2018/02/27/tecnologia/noticia/california-autoriza-circulacao-de-carros-autonomos-sem-condutor-ao-volante-1804631

https://exame.abril.com.br/tecnologia/jogo-do-mit-com-carro-autonomo-deixara-voce-em-dilema-etico/

https://g1.globo.com/carros/noticia/carro-autonomo-da-uber-atropela-e-mata-mulher-nos-eua.ghtml

https://bdjur.stj.jus.br/jspui/bitstream/2011/113250/responsabilidade_civil_acidentes_silva.pdf

https://jeffbradberry.com/posts/2015/09/intro-to-monte-carlo-tree-search/

“Big Data é o novo petróleo, afirma executiva da IBM” acessado emhttps://olhardigital.com.br/noticia/big-data-e-o-novo-petroleo,-afirma-executiva-da-ibm/34986

" One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report. Acessado em 22/04/2018.