DEMOCRACIA, INTELIGÊNCIA ARTIFICIAL E DESAFIOS REGULATÓRIOS: DIREITOS, DILEMAS E PODER EM SOCIEDADES DATIFICADAS

DEMOCRACY, ARTIFICIAL INTELLIGENCE AND REGULATORY CHALLENGES: RIGHTS, DILEMMAS AND POWER IN DATAFIED SOCIETIES

DEMOCRACIA, INTELIGENCIA ARTIFICIAL Y DESAFÍOS REGLAMENTARIOS: DERECHOS, DILEMAS Y PODER EN LAS SOCIEDADES DATIFICADAS

Sivaldo Pereira da Silva[1]

Resumo: Este artigo visa identificar questões-chaves que atravessam hoje a construção de políticas de Inteligência Artificial (IA), tendo como premissa o horizonte regulatório adequado para este campo em ascensão. A análise foi desenvolvida a partir de pesquisa documental e bibliográfica, sob a lente de princípios normativos com pressupostos em Teoria Política e Teorias da Democracia. Após sintetizar os aspectos mais proeminentes que configuram o modo de funcionamento de sistemas de IA, o trabalho elencou sete problemas-chaves de ênfase política que estão nos alicerces deste debate: (1) personificação e imputabilidade da máquina; (2) dilemas e julgamentos morais; (3) autoritarismo estatístico das métricas; (4) obscurantismo matemático nos processos; (5) pervasividade utilitarista dos sistemas-autônomos; (6) controle e fronteiras da eficiência; (7) diversidade e representatividade nos códigos.

Palavra-chave: Inteligência Artificial; Governança algorítmica; Filosofia da Tecnologia; Comunicação digital e Regulação; Políticas Públicas

Abstract: This article aims to identify key issues in the debate on Artificial Intelligence (AI) policies, keeping in mind the construction of an adequate regulatory landscape for this field. The analysis was guided by documentary and bibliographic research, taking normative principles from Political Theory and Democracy Theories. After summarizing the most prominent aspects of the modus operandi of AI systems, the work listed seven key political problems that are at the foundation of this discussion: (1) personification and imputability of the machine; (2) dilemmas and moral judgments; (3) statistical authoritarianism of the metrics; (4) mathematical obscurantism in the processes; (5) utilitarian pervasiveness of autonomous systems; (6) frontiers of efficiency and control; (7) diversity and representativeness in the codes.

Keywords: Artificial Intelligence; Algorithmic governance; Philosophy of Technology; Digital Communication and Regulation; Public policy.

Resumen: Este artículo tiene como objetivo identificar cuestiones clave en el debate sobre las políticas de Inteligencia Artificial (IA), teniendo en cuenta la construcción de un panorama regulatorio adecuado para este campo. El análisis se guió por la investigación documental y bibliográfica, tomando principios normativos de la teoría política y las teorías de la democracia. Después de resumir los aspectos más destacados del modus operandi de los sistemas de IA, el trabajo enumeró siete problemas políticos clave que están en la base de esta discusión: (1) personificación e imputabilidad de la máquina; (2) dilemas y juicios morales; (3) autoritarismo estadístico de las métricas; (4) oscurantismo matemático en los procesos; (5) omnipresencia utilitarista de los sistemas autónomos; (6) fronteras de eficiencia y control; (7) diversidad y representatividad en los códigos.

[1] Professor da Faculdade de Comunicação (FAC) e do Programa de Pós-Graduação em Comunicação da Universidade de Brasília (UnB). PhD em Comunicação e Cultura Contemporâneas pela Universidade Federal da Bahia, com estágio doutoral na University of Washington (EUA). Foi pesquisador visitante no Instituto de Pesquisa Econômica Aplicada (IPEA); consultor ad hoc da Unesco para aplicação de indicadores de desenvolvimento da mídia no Brasil. É fundador e coordenador do grupo de pesquisa Centro de Estudos em Comunicação, Tecnologia e Política (CTPol) e pesquisador do Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT-DD).

Palabras clave: Inteligencia Artificial; Gobernanza algorítmica; Filosofía de la tecnología; Comunicación digital y regulación; Políticas públicas.

1 Introdução

As inovações propiciadas por sistemas e máquinas baseadas em Inteligência Artificial (IA) configuram hoje um relevante impulso na eficiência de processos nas mais diversas áreas como comunicação, política, transportes, segurança pública, saúde, educação etc. A tendência, para as próximas décadas, é que haja saltos significativos neste campo, com o horizonte de se tornarem cada vez mais cotidianas e onipresentes. Estas tecnologias significam bem mais que ferramentas puramente instrumentais. São, na verdade, artefatos técnico-culturais que alteram de forma substancial processos de tomadas de decisão, com efeitos em diversos ramos da atividade humana, desde parâmetros de consumo até as relações de poder entre diversos atores (seja entre Estado e cidadãos; empresas e consumidores; organizações e indivíduos).

Para lidar com essas transformações, governos e organizações em diversas jurisdições em nível local, nacional e multilaterais (a exemplo da cidade de Nova York, da Federação alemã, da União Europeia, ONU, OECD) estão elaborando planos estratégicos, legislação ou políticas públicas que visam dirimir as tensões decorrentes desta conjuntura, bem como recepcionar tais inovações visando garantir seus potenciais efeitos positivos.

Paralelamente aos benefícios práticos que os sistemas algorítmicos mais avançados propiciam, isso também tende a gerar novas formas de desigualdade, violações de direitos ou ampliar concentração de poder. A proliferação de sistemas-autônomos repercute em pontos politicamente sensíveis como privacidade; liberdades; direitos individuais e coletivos; desinformação; transgressões éticas; autoritarismo etc.

Diante de tal contexto, este trabalho pretende identificar e caracterizar os principais problemas-chaves que qualquer política de Inteligência artificial precisa responder. Neste sentido, o artigo traz um estudo exploratório, baseado em pesquisa documental e bibliográfica, analisada sob a lente normativa de princípios democráticos. Para isso, o estudo está dividido em duas partes: primeiramente, faz uma abordagem conceitual sobre Inteligência Artificial (IA), suas origens e características fundamentais. A segunda parte sintetiza sete importantes problemas-chaves de ênfase política que estão nas bases do atual debate regulatório sobre IA no mundo e que são discussões determinantes para se entender a complexa relação entre democracia e sistemas-autônomos digitais.

2 Inteligência artificial: técnica para além da técnica

A expressão "Inteligência Artificial" nos leva a imaginar máquinas que pensam ou artefatos técnicos autoconscientes. Porém, como ocorre em toda metáfora, trata-se de uma terminologia generalista que nos auxilia cognitivamente em uma síntese do fenômeno, porém minimiza aspectos importantes, gerando uma definição carente de maior precisão. De todo modo, não devemos considerar a utilização dessa terminologia como um problema uma vez que já está amplamente difundida no imaginário social, em documentos governamentais, noticiário, diretrizes de empresas etc. É possível adotá-la desde que possamos contextualizar e dimensionar tal metáfora e realçar, sobretudo, os elementos que a expressão esconde.

Especificamente, Inteligência Artificial diz respeito a um conjunto de métodos lógicos que visam solucionar problemas com base em algoritmos que são treinados (através de **inputs**, entrada de dados) para compreender padrões, aprender com erros e se reconfigurar chegando a resultados (**output**) cada vez mais próximos do esperado. Então é importante notar que não estamos falando de uma máquina que pensa e sim que resolve problemas lógicos e é treinada neste sentido a partir da experiência (dados) que recebe.

Do ponto-de-vista histórico, artefatos técnicos que auxiliavam na concretização de alguma operação lógica, principalmente matemática, não é novidade. Em diversas culturas, como na Mesopotâmia e China, instrumentos como o ábaco já existiam desde a Antiguidade. Na Modernidade estes mecanismos ganharam uma nova versão com as primeiras calculadoras. Como situa Gleick (2011, p.99):

> Blaise Pascal criou uma máquina de somar em 1642, com uma fileira de discos giratórios, um para cada dígito decimal. Três décadas mais tarde, Leibniz aprimorou a obra de Pascal ao usar um tambor com dentes salientes para "reagrupar" as unidades de um dígito ao seguinte.

Porém, o autor lembra que os protótipos de Pascal e Leibniz continuaram bem próximos do ábaco, pois faziam registros passivos dos estados da memória de determinada operação matemática.

No século XIX, no contexto da Revolução Industrial, Charles Babbage deu um passo adiante inserindo um elemento importante nas máquinas de calcular: o automatismo. Isso abriu um precedente no desenvolvimento do computador que seria criado de fato no século seguinte. Porém, a máquina de Babbage era basicamente mecânica (não utilizava energia elétrica) e não pressupunha uma estrutura lógica versátil, como a perspectiva binária (baseada em dois dígitos, 0 e 1) ou variáveis **booleanas** (desligado/ligado). O terreno para Inteligência Artificial tal como conhecemos hoje começa de fato a ficar mais fecundo na segunda metade do século XX. De modo mais específico, suas origens estão vinculadas ao termo *"***machine intelligence**" (inteligência de máquina) difundido por Alan Turing, com primeiros registros em manuscritos

ainda em 1941 (COPELAND, 2004)[2]. O princípio básico da ideia de Turing dizia respeito à solução de problemas lógicos e matemáticos através da automatização em sistemas eletrônicos binários e a possibilidade de se construir máquinas capazes de aprender com a experiência. Em artigo pulicado em 1950 intitulado "*Computing machinery and intelligence*" na revista *Psychology and Philosophy*, Turing propõe pensarmos na seguinte pergunta: "As máquinas conseguem pensar? ("*Can machines think*?") (Turing, 1950). Embora o autor faça uma longa discussão das objeções a esta pergunta, sua preocupação está na verdade em reformular tal questionamento em direção ao que chamou de "Jogo da Imitação". Para o autor, a capacidade de imitação universalizada de uma máquina (baseada em linguagem binária) seria um dos elementos diferenciais e que mereciam especial atenção:

> This special property of digital computers, that they can mimic any discrete state machine, is described by saying that they are universal machines. The existence of machines with this property has the important consequence that, considerations of speed apart, it is unnecessary to design various new machines to do various computing processes (TURING, 1950, p. 441).

Em 1943 Warren McCulloch e Walter Pitts publicaram um artigo intitulado "*A logical calculus of the ideas immanet in nervous activity*" explicando como neurônios funcionam e modelaram uma rede neural artificial simples utilizando circuitos elétricos para demonstrar suas hipóteses (MCCULLOCH; PITTS, 1943)[3]. Somando isso às ideias de Turing, diversos pesquisadores foram estimulados a pensar redes neurais artificiais em computadores e como isso poderia ser alinhado à concepção de uma "máquina inteligente". Em 1956, no estado de New Hampshire (EUA) o termo "Inteligência Artificial" tal como conhecemos hoje apareceu pela primeira vez em uma conferência intitulada "*The Dartmouth Summer Research Project on Artificial Intelligence*", considerada por muitos como a pedra fundadora deste campo de pesquisa.

Mas se a noção de Inteligência Artificial já existia há pelo menos meio século atrás, por que somente agora estamos falando em leis e regulação para este campo como se fosse algo recém descoberto? A explicação é relativamente simples: porque um conjunto de condições técnicas que antes não estavam dadas passou a co-existir e convergir principalmente a partir das primeiras décadas deste século[4]. Neste cenário, estima-se que haverá um *boom* do uso de IA nas próximas duas ou três décadas (CATH, 2017; DAFOE, 2018; GRACE et al, 2018).

---

[2] Embora não falasse diretamente sobre inteligência artificial, em 1945, Vannevar Bush publicou um artigo intitulado "As we may think" na revista The Atlantic Monthly. Ele propunha uma máquina de memória coletiva que chamou de Memex capaz de aglutinar e processar informação transformando-a em conhecimento. Uma reprodução deste texto está disponível em: https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/. Acesso em: 4 jul 2019.

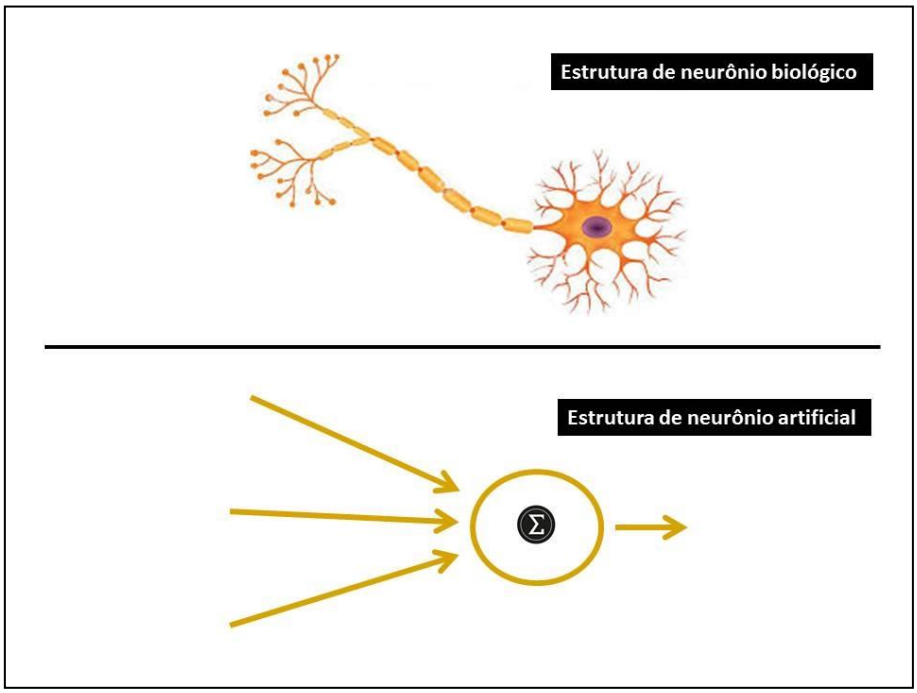
[3] Outros obras também contribuíram nesta linha: em 1949 o livro "*The organization of behavior: a neuropsychological theory*", de Donald Hebb; e em 1958 o artigo de Frank Rosenblat intitulado "*The perceptron: a probabilistic model for information storage and organization in the brain*".

[4] Como e criação de infraestruturas digitais avançadas, especialmente 5G; intensificação de mecanismos de Big Data que refletem a exponencial capacidade de coletar e processar grandes volumes de dados, de variadas fontes com velocidade antes inexistente; desenvolvimento de algoritmos mais sofisticados etc.

Para vislumbrarmos melhor a natureza deste fenômeno, convém sintetizarmos quatro aspectos que merecem especial atenção, pois representam dimensões que nos ajudam a compreender o funcionamento dos sistemas baseados em Inteligência Artificial: (a) redes neurais artificiais; (b) o sentido da noção de imitação; (c) o poder do automatismo e (d) os níveis tipológicos de IA.

Primeiramente, para um sistema ser chamado de "inteligente" isso pressupõe que seja capaz de aprender e tomar decisões baseados em lógica. Um método inovador neste sentido é a ideia redes neurais artificiais (também uma metáfora vinculada ao cérebro) que se tornou uma das concepções mais promissoras e influentes de IA, na qual estão assentadas as técnicas de *machine learning* e *deep learning.* De modo mais didático, uma rede neural artificial é uma composição de algoritmos inspirados na estrutura e no modo de funcionamento de um neurônio biológico, conforme ilustra a Figura 1:

Figura 1 – Ilustração comparativa das estruturas de neurônio biológico e artificial.

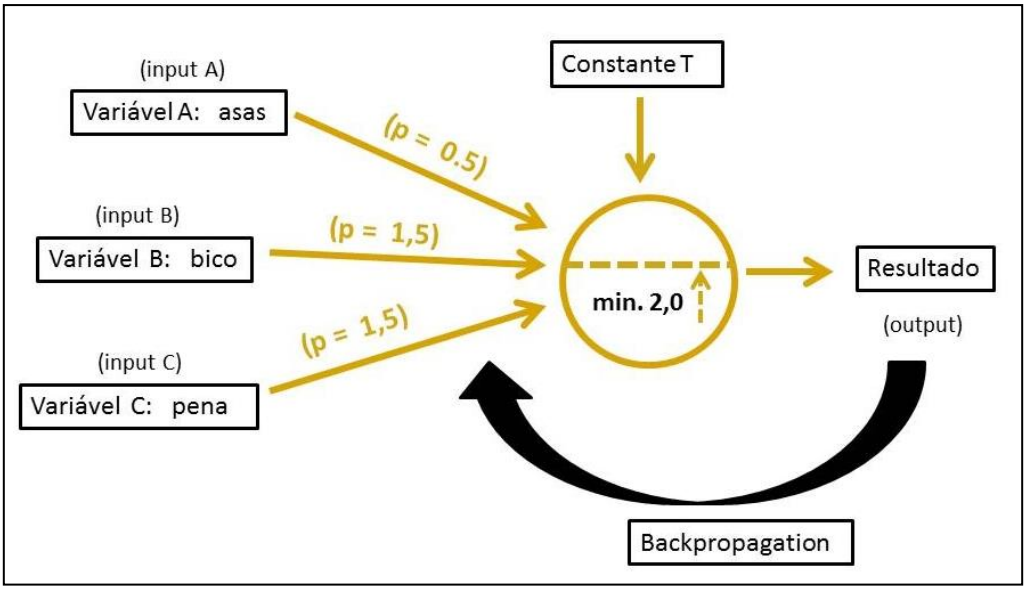


Fonte: elaboração própria a partir da proposta de McCulloch e Water Pitts (1943)

Basicamente, um neurônio artificial é constituído por diversas entradas (*inputs*) projetadas para captar informações das diversas variáveis que definem aquilo que está sendo avaliado (por exemplo, tomando as variáveis "bico" ou "asas" como indicadores relevantes para definir se uma imagem é ou não referente a um pássaro). Para cada entrada (variáveis captadas) é dado um peso. O quantitativo dos valores de cada variável com seus respectivos pesos é somado em um único valor que é então submetido a um núcleo no qual se exige que um quantitativo mínimo seja atingido para que o neurônio seja ativado (tal como um gatilho que só

será acionado se a soma chegar a determinado valor considerado significativo)[5]. Assim, para que o neurônio responda positivamente que a imagem em questão é de um pássaro é preciso que haja a combinação das variáveis e seus respectivos pesos com valores que chegue a um certo patamar capaz de indicar que há grande probabilidade da imagem ser, de fato, um pássaro. A Figura 2 traz uma ilustração simplificada deste processo e seus pesos[6], tomando como exemplo apenas 1 neurônio artificial que se baseia em 3 variáveis (asas, pena e bico) para determinar se uma determinada imagem é de um pássaro.

Figura 2 – Ilustração simplificada dos principais mecanismos de funcionamento de um neurônio artificial.



Fonte: elaboração própria, sintetizando e adaptando, para fins didáticos,a estrutura inicial proposta por McCulloch e Water Pitts (1943)

Se o neurônio detectar como positiva apenas a variável A (asas) e as demais forem negativas (que resultariam em valor zero, por não se confirmarem) a soma no núcleo será de apenas 0,5 pontos (valor do peso da variável "asa"), não atingindo assim o patamar mínimo estipulado neste neurônio (2 pontos) para considerar a imagem como a de um pássaro. Isso porque ter apenas asas não é suficiente para determinar que algo é, de fato, um pássaro pois há outros objetos que possuem asas e que não são pássaros, como um avião, um morcego ou uma borboleta. Em outra situação, se o neurônio detectar como positiva a variável A (asas) e B (bico), isso já somaria 2 pontos (0,5 da variável A + 1,5 da variável B), atingindo assim a pontuação mínima exigida. Diante disso, o neurônio seria acionado para dar como resultado a resposta positiva de que a referida imagem se trata de um pássaro. Isso porque se algo possui

[5] Esta é a estrutura básica de um neurônio artificial. Com o avanço dos estudos neste campo, outros elementos mais complexos foram inseridos para dar melhorar acuidade às respostas.
[6] Os valores dos pesos na Figura 1 são apenas estipulações ilustrativas. Na prática, os pesos são estipulados no processo de programação levando em conta a probabilidade das variáveis.

asas e bico aumenta-se estatisticamente a possibilidade de ser um pássaro. Porém, o processo não acaba aqui. O que o algoritmo realizou até agora foi uma "aposta". O que de fato faz a máquina "aprender" são as correções dos pesos após a verificação se a aposta se demonstrou "falsa" ou "verdadeira". A resposta do neurônio é comparada com a realidade para verificar se houve acerto ou erro na estimativa. Essa informação (de erro ou acerto) gera uma onda de modificações ou reforços dos pesos chamada de "*backpropagation*": se um neurônio (com determinadas configurações de peso) errou na sua estimativa da imagem pois detectou que se tratava na verdade da imagem de um avião, esse erro ao ser percebido se transformará em uma retropropagação que diminuirá os pesos dessas variáveis pois se mostraram não efetivas. O mesmo é feito para variáveis que foram julgadas inicialmente como pouco importantes, mas nos resultados se mostraram determinantes: desta vez, a retropropagação age para calibrar positivamente essas variáveis, enfatizando-as como relevantes. A diferença do valor do resultado final é reprocessado, neste caso, aumentando os pesos das variáveis subestimadas fazendo com que na próxima vez o neurônio funcione corretamente com os pesos devidamente calibrados para acertar, baseado nas experiências anteriores.

No fundo, a sofisticação do *machine learning* nada mais é do que uma correção de pesos inicialmente estimados a partir da constatação de erros aumentando, assim, a possiblidade de acertar numa próxima vez que se deparar com as mesmas variáveis. Porém, para o sistema funcionar efetivamente é preciso haver muitos neurônios artificiais interligados e, principalmente, muitos dados para que os neurônios possam "testar" os pesos das varáveis, isto é, para que as redes neurais sejam "treinadas". A chamada "fase de treinamento" dos algoritmos de IA necessita de uma grande quantidade de informação e, como percebemos, a disponibilidade de dados prévios e a ocorrência estatística é um elemento determinante. Por exemplo, um *software* de IA só conseguirá distinguir entre "pássaros" e "aviões" quando tiver recebido muitas imagens de *input* até conseguir fazer a distinção de forma automatizada. Naturalmente, o exemplo dado aqui é apenas didático. Redes neurais reais incluem centenas ou milhares de neurônios com centenas ou milhares de variáveis sendo testadas e recalibradas.

Um segundo elemento que é basilar no funcionamento dos sistemas de Inteligência Artificial, como mencionado, é a perspectiva da imitação. Com técnicas de *machine learning* os algoritmos passam a ter uma imensa capacidade de identificar (e repetir) padrões, fazendo valer o jogo de Turing (TURING, 1950). Não significa que o algoritmo tem consciência da diferença entre um pássaro e um avião, mas imita a nossa percepção das coisas dando pesos a determinadas variáveis que são testadas e definidas como determinantes, tal como nós fazemos ao olhar os elementos que compõem a nossa definição de "pássaro". Neste sentido, a imitação é baseada em probabilidades estatísticas onde aquilo que é majoritário (após o processo de *backpropagation*) passa a ser percebido pelos algoritmos e reforçado por estes. Já aquilo que foge do padrão preponderante, tende a ser menosprezado e ignorado ou se torna uma

informação destoante no sistema. Por exemplo, se um pássaro aparecer sem bico e sem asas, o algoritmo terá dificuldade em considerá-lo um pássaro, pois foge do padrão que a fase de treinamento sedimentou no código sobre o que é um pássaro. Logo, pássaros com deficiência congênita ou acidentado tendem a não ser reconhecidos como pássaros. Notemos que a fase de treinamento é dinâmica, porém após esta etapa, os sistemas tendem a se tornar estáveis (relativamente rígidos) baseados em uma perspectiva majoritária. Aqui podemos notar que a propriedade de imitação é baseada em algo que já está dado, isto é, imitar é conservar e reforçar algo pré-existente. Neste sentido, é possível afirmar que há paradoxalmente um misto de inovação e conservadorismo na sofisticação dos algoritmos de aprendizagem.

A terceira característica que devemos ter em mente na concepção de Inteligência Artificial é o automatismo. Dois séculos depois, o sonho de Babbage em busca da máquina automática foi levado a cabo e ao extremo. O automatismo é um elemento central em qualquer sistema de IA pois máquinas só são consideradas inteligentes se forem capazes de funcionar por conta própria a partir de um *start* inicial e encontrar seu próprio caminho. A evolução na produção e armazenamento de energia aliada à capacidade dos algoritmos de gerarem "*looping*" ou ciclos de repetição *ad infinitum* é uma importante combinação que implica, em última instância, em uma nova forma de poder. Um computador ou sistema pode funcionar de forma repetitiva por anos e séculos, enquanto tiver energia. Num mundo com cotidiano cada vez mais datificado temos, na prática, o aumento de poder de determinados atores (como Estado, instituições e corporações) devido à capacidade de impor a repetição de procedimentos ou formas de comportamento. Autoritarismos ou ações injustas podem ser repetidos de modo muito mais ágil, a baixo custo e de forma bem mais difícil de serem contrapostos quando executados por sistemas-autônomos. Isso também pode trazer maior rigidez na relação entre partes assimétricas onde o sistema tende a seguir procedimentos e não observar situações de exceção. Também coloca a máquina como um ente que toma decisões sendo que, por trás de uma decisão aparentemente técnica, está o valor embutido nas métricas.

Por fim, um quarto aspecto basilar importante para compreender Inteligência Artificial é perceber que existem diferentes graus de desenvolvimento dessas tecnologias e isso repercute em diversas dimensões sobre o lugar dos artefatos. A literatura tem apontado para três níveis ou tipos de IA, como sintetiza Girasa (2020): **Inteligência Artificial Estreita, Inteligência Artificial Geral** e **Superinteligência Artificial**. A primeira se refere ao desempenho de uma tarefa singular. A segunda consegue realizar várias tarefas ao mesmo tempo de forma similar ao cérebro humano. A terceira consiste na superação da capacidade humana em vários aspectos. Atualmente, estamos no primeiro estágio, porém já com horizontes e protótipos promissores de sistemas do segundo nível. Em relação ao terceiro grau este só será possível com a criação de estruturas de processamento mais robustas do que a que temos atualmente, como a computação quântica, ainda em fase bem incipiente de desenvolvimento. O que é importante notarmos

nesses graus é justamente a sofisticação, capacidade de rupturas e abrangência de campo de ação que representam. Quanto maior o grau de desenvolvimento de IA, mais intensa é sua a pervasividade e poder, tendendo a ser bem mais impactante culturalmente, socialmente e politicamente bem mais disruptiva.

Todas essas questões que caracterizam o funcionamento da Inteligência Artificial devem ser pensadas a partir de parâmetros que consigam ir para além do horizonte da eficiência técnica. Há questões sociais, culturais, políticas e econômicas envolvidas. Uma boa metáfora para isso é a imagem que o filósofo alemão Martin Heidegger descreveu quando analisou a essência da tecnologia moderna com base no conceito *Gestell (*HEIDEGGER, 2001). Para o autor, no passado, construíamos pontes que eram dispositivos técnicos instalados nos rios. Com o uso da tecnociência em sua incessante busca por extrair, manipular e estocar energia, a hidrelétrica não é um elemento que está instalada no rio, como a ponte estava. Para ele, a situação se inverteu: agora o rio está instalado na usina (pois essa o submete aos seus objetivos) e o rio se tornou assim um dispositivo do sistema tecnológico. Trazendo isso para os avanços nos sistemas de IA, estamos falando diretamente em problemas de agenciamento e autonomia humana.

Dito isso, diante da expansão destes sistemas e a sua crescente centralidade na vida cotidiana, é preciso elaborar estratégias para que o Estado esteja apto a fomentar e estimular todos os benefícios da IA e, ao mesmo tempo, mitigar suas eventuais distorções, definindo papeis para que os diversos atores envolvidos possam interagir de forma harmônica, garantindo a proteção de direitos; evitando perda de autonomia e liberdades.

3 Inteligência artificial, regulação e democracia: sete problemas-chaves

Entre 2017 e 2019 diversos países, em todos os continentes, lançaram suas estratégias para Inteligência Artificial: Alemanha, Canadá, China, Dinamarca, Emirados Árabes, Estados Unidos da América, Finlândia, França, Índia, Itália, Japão, Malásia, México, Nova Zelândia, Quênia, Singapura, Coreia do Sul, Suécia, Taiwan, Reino Unido. Outros países que ainda não lançaram estratégia oficial (como a Austrália, Espanha, Polônia e Uruguai[7]), estavam criando comitês, consultas públicas ou projetando orçamento específico para o desenvolvimento desta área. Blocos regionais ou articulações pan-regionais também têm se preocupado com o tema, gerando documentos como a *Declaration on AI in the Nordic-Baltic Region*[8] (publicada por uma articulação de países nórdicos e bálticos) ou criando instâncias como *High-Level Expert*

[7] No caso brasileiro, ainda não há uma estratégia definida. O país possui uma Estratégia Digital (E-digital) lançada em 2018 que contém diretrizes genéricas para a transformação digital, mas não havia apresentado até o primeiro semestre de 2020 documento ou diretriz oficial mais específicos para IA.

[8] https://www.regeringen.se/49a602/globalassets/regeringen/dokument/naringsdepartementet/20180514_nmr_deklaration-slutlig-webb.pdf. Acesso em: 6 jul 2019.

Group on Artificial Intelligence (AI HLEG)[9] da União Europeia. Além disso, organismos multilaterais tradicionais como a ONU[10] e OECD[11] possuem ações, diretrizes ou recomendações sobre o tema. Organizações profissionais como o Institute of Electrical and Electronics Engineers (IEEE) e articulações civis, como a Declaração de Montreal[12] também são iniciativas preocupadas em estabelecer princípios para boas práticas de IA.

Este cenário demonstra que há uma evidente efervescência na elaboração de diretrizes e estratégias para IA em todo o mundo, porém ações regulatórias mais concretas, no formato de leis, ainda estão em processo de tramitação ou são incipientes, localizadas ou parciais (como é o caso da cidade de Nova York que criou uma das primeiras legislações no mundo sobre o tema e a Alemanha que aprovou lei que trata de uma parte do problema, especificamente regulando o uso de veículos-autônomos).

Embora as estratégias e políticas incipientes de IA tenham suas peculiaridades e ênfases (algumas com mais foco em aspectos econômicos, outras em questões éticas), ao analisarmos o conjunto de estratégias ou proto-regulações é possível identificar sete problemas-chaves de fundo político que têm implicações diretas para as democracias contemporâneas que merecem especial atenção, devido aos seus possíveis desdobramentos e impactos políticos. Podemos sintetizá-los nos seguintes termos: (1) personificação e imputabilidade da máquina; (2) dilemas e julgamentos morais; (3) autoritarismo estatístico das métricas; (4) obscurantismo matemático nos processos; (5) pervasividade utilitarista dos sistemas-autônomos; (6) controle e fronteiras da eficiência; (7) diversidade e representatividade na codificação.

4 Sobre personificação e imputabilidade da máquina

Na ficção científica principalmente do tipo distópica, o problema da personificação de sistemas de Inteligência Artificial é um elemento recorrente e muitas vezes central na trama que de desenvolve. Nos anos 2000 a série de televisão *Battlestar Galactica* dramatiza a dificuldade de se distinguir entre seres humanos e seres artificiais. No mundo real, esta indistinção ainda não é tão evidente e tão avançada quanto na ficção, porém isso já é um problema que começa a se despontar no horizonte regulatório dos sistemas de IA. Sobretudo porque quanto mais os algoritmos de IA são treinados (com dados dos próximos anos e décadas), mais verossímeis a um interlocutor humano ficarão interagindo com o público e tendendo a ficar cada vez mais difícil de serem percebidos como uma máquina. Por isso, uma primeira questão que surge com a

[9] https://ec.europa.eu/digital-single-market/en/artificial-intelligence. Acesso em: 11 jul. 2019.
[10] A ONU possui algumas iniciativas sobre IA, uma delas é a plataforma "AI for Good https://aiforgood.itu.int/ baseada nos encontros anuais sobre o tema (AI for Good Global Summit), mantida pela União Internacional de Telecomunicações (UIT). Também há outras iniciativas como o Centre for Artificial Intelligence and Robotics http://www.unicri.it/in_focus/on/UNICRI_Centre_Artificial_Robotics (vinculada à UNICRI), além de documentos sobre o tema: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf Acesso em: 8 jan 2020.
[11] https://www.oecd.org/going-digital/ai/. Acesso em: 11 jul 2019.
[12] https://www.montrealdeclaration-responsibleai.com/the-declaration. Acesso em: 15 jul 2019.

personificação desses sistemas é o direito do indivíduo em saber se está de fato conversando com um outro ser humano ou se consiste em um sistema de IA. Isso porque o contrato de comunicação que se estabelece é bastante distinto quando falamos com máquinas e não com subjetividades humanas diretamente. Esta diferenciação se torna importante no debate regulatório, pois requer normas que obriguem a máquina a se declarar como máquina. Por exemplo, quando uma empresa ou um órgão governamental adota um sistema de IA para o atendimento ao público é preciso que a conversa comece com a identificação do tipo de interlocutor que está do outro lado da linha ou que alguma informação neste sentido esteja clara e evidente.

A questão da personificação da máquina também tende a atravessar a própria dinâmica do sistema político. Um problema hoje emergente em campanhas eleitorais são os *chatbots*: algoritmos que simulam a conversação em linguagem natural, fabricados e encomendados (em escala industrial) para atuar massivamente em campanhas negativas ou em engajamentos discursivos em prol de um candidato. Com as evoluções dos sistemas de IA, o fenômeno dos *chatbots* tende a ser cada vez mais comum e sofisticado, difundindo-se para diversas áreas seja na forma de sistemas de atendimento ao consumidor, de consultorias jurídicas ou até mesmo serviços de conversação para apoio emocional. A questão é, como regular este novo "ente" que se apresenta entre nós e quais os limites para o seu uso sem que isso implique em transgressões dos direitos?

Isso também levanta hipóteses sobre a personalidade jurídica dos sistemas de IA, por exemplo, para alguns analistas, devido à complexidade desses sistemas, poderia ser adequado a um robô ter uma personalidade similar a uma empresa que é classificada como "pessoa jurídica" em contraponto à "pessoa física", isto é, ao indivíduo natural. Esta perspectiva aparece de diferentes formas em diversas estratégias governamentais. Um bom exemplo é a uma resolução do Parlamento Europeu de fevereiro de 2017 intitulada "*Disposições de Direito Civil sobre Robótica*" que trouxe uma série de recomendações à Comissão Europeia para se pensar no lugar dos robôs em uma sociedade cada vez mais permeada por autônomos baseados em IA. No item parágrafo 59, letra "f" os parlamentares recomendam:

> Criar um estatuto jurídico específico para os robôs a longo prazo, de modo a que, pelo menos, os robôs autónomos mais sofisticados possam ser determinados como detentores do estatuto de pessoas eletrónicas responsáveis por sanar quaisquer danos que possam causar e, eventualmente, aplicar a personalidade eletrónica a casos em que os robôs tomam decisões autónomas ou em que interagem por qualquer outro modo com terceiros de forma independente[13]

[13] Transcrição de trecho em português, por isso, as algumas grafias estão em formatos adotados em Portugal. Disponível em: http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_PT.html. Acesso em: 11 jul. 2019.

A resolução gerou polêmica e um grupo de especialistas chegou a publicar uma carta aberta à Comissão Europeia criticando sugerindo ignorar tão proposição[14]. Mas este ainda é um debate em aberto, pois devido à complexidade de atores envolvidos e ao crescente grau de personificação dos sistemas de IA - que rompem com a noção de autoria uma vez que o próprio autômato se transforma ao aprender com a sua experiência no mundo – analistas explicam que:

> [...] a legal conflict arises, as within the framework of the current legislation the robot cannot be held liable for actions and (or) inaction and as a result responsibility falls on the user, software developer, or manufacturer. At the same time, the EU resolution raises the issue of responsibility in the event that the robot caused damage due to the decisions made by the robot itself (based on the embedded algorithms) and the definition of the third party responsible for paying compensation is impossible (ATABEKOV; YASTREBOV, 2018 p. 779).

O mesmo código original pode assumir "personalidades" diferentes e agir de forma distinta se o algoritmo for treinado a partir de dados de população religiosa ou se for treinado com dados de população de ateus. O argumento utilizado é que os autores do código não podem ser culpados se um sistema de IA se tornou fascista após ter sido apropriado e treinado por grupos fascistas. Porém, é preciso criar mecanismos regulatórios que obriguem os programadores a desenvolverem soluções técnicas capazes de inibir que um sistema de IA se transforme em algo prejudicial em mãos alheias. É preciso estabelecer graus de responsabilidade de cada agente na complexa cadeia de produção e consumo que atravessa este tipo de ferramenta.

5 Sobre dilemas e julgamentos morais

Um dos temas mais importantes em teorias da democracia são os dilemas morais. Por outro lado, a dimensão moral também sempre esteve presente na concepção mais filosófica ou ficcional de Inteligência Artificial, cuja asserção mais conhecida são os princípios idealizados pelo escritor Isaac Asimov (conhecidos como três leis da robótica). Os julgamentos morais são indicadores relevantes para se analisar como democracias tratam as divergências, respeitando as diferenças e mantendo um procedimento racional e justo nos processos de tomadas de decisão; e qual o papel das instituições (ou do Estado) para lidar deliberativamente com conflitos complexos que envolvem desacordos morais de fundo (GUTMANN; THOMPSON, 1996). Julgamentos morais refletem tensionamentos sociais e o eventual desequilíbrio na balança de poder entre grupos de visões de mundo divergentes. Ao mesmo tempo, artefatos técnicos também nos colocam diante de problemas morais. O dilema do bonde ("*trolley problem*"), um experimento crítico sobre ética idealizado por Philippa Foot, é um bom exemplo disso. De

[14] A carta aberta pode ser acessada em < https://g8fip1kplyr33r3krz5b97d1-wpengine.netdna-ssl.com/wp-content/uploads/2018/04/RoboticsOpenLetter.pdf. Acesso em: 22 ago 2019.

forma resumida, a questão traz uma situação hipotética na qual há um trem em rota de colisão atingirá 5 pessoas que estão nos trilhos. Porém, um observador em posição privilegiada percebe o futuro acidente e tem em suas mãos a possiblidade de puxar uma alavanca e desviar o veículo para outro caminho alternativo. Isso salvará a vida de cinco pessoas, mas o desvio implicará na morte de uma outra pessoa que está no caminho desviado. Diante desta situação hipotética, a pergunta é: seria ético sacrificar a vida de uma pessoa para salvar a vida de cinco? Geralmente, a maioria das pessoas acredita que a resposta é positiva.

Por outro lado, o dilema se torna mais complexo com uma pequena variação da história: quando o mesmo trem continua na rota de colisão de cinco pessoas, porém nesta segunda versão o observador está em uma ponte sob a qual o trem passará e seu lado está um homem que se for jogado viaduto abaixo na linha do trem seu corpo serviria de freio e pararia o trem antes de atingir as pessoas no trilho, salvando assim as cinco vidas antes ameaçadas. A mesma pergunta é feita para esta segunda versão: seria ético sacrificar a vida de uma pessoa para salvar a vida de cinco? Nessa nova narrativa, o resultado final é o mesmo: sacrificar uma pessoa para salvar cinco. Porém, pessoas tem maior dificuldade em tomar uma posição neste caso pois não há um elemento técnico intermediário como na primeira situação: uma alavanca que torna "menos pessoal" o ato de sacrificar uma vida em razão de cinco.

No atual contexto de expansão de sistemas digitais ubíquos e preditivos espalhados pela vida cotidiana, o artefato de Inteligência Artificial ocupa o lugar observador privilegiado do "*trolley problem"* que antevê o futuro e possui mecanismos técnicos para tomar decisões sobre vidas. Diante do avanço da Internet das Coisas e o uso cotidiano de diversos tipos de autônomos (como carros sem motorista; *drones* que operam por piloto automático; robôs etc.) esses objetos tendem a se configurar como novos "atores" na fauna da dinâmica social e, ao mesmo tempo, significarão cada vez mais importantes instâncias de tomadas de decisão. Por exemplo, um carro autônomo baseado em IA toma a todo momento decisões sobre seu trajeto em vias públicas (se vira à esquerda, se para no semáforo vermelho ou se segue reto em uma avenida). Algumas decisões são óbvias e esperadas: não avançar ao sinal vermelho, por exemplo. Porém, outras decisões terão maior grau de complexidade e seriam difíceis mesmo para um ente humano por se tratarem de decisões morais.

Suponhamos que haja um carro autônomo com um casal de jovens passageiros sendo transportados em uma via. Inesperadamente, o caminhão que está à frente deste veículo para bruscamente o suficiente para que o freio do veículo que está atrás não consiga responder a ponto de evitar a colisão. O acidente seria fatal o casal de jovens passageiros e haveria perda total do veículo. O sistema de IA do carro autônomo será capaz de antever o problema e estará apto a tomar uma decisão sobre o que fará para evitar o menor dano possível. O problema começa quando colocamos nesta equação outros elementos (similar ao "*trolley problem*") que tornam a tomada de decisão um posicionamento moral que expressa determinada visão de

mundo ou de valores. Suponhamos que, se o automóvel optar por desviar para a direita, atingirá uma criança vinda da escola que certamente morrerá com o impacto, mas os danos ao veículo e seus passageiros serão mínimos e estes sobreviverão. Se optar por desviar à esquerda, atingirá o carro com três idosos que está na via contrária e, devido ao ângulo da colisão, os idosos sofrerão o maior impacto e certamente morrerão enquanto o carro com o casal de jovens terá dados menores e estes sobreviverão. Aqui temos uma decisão moral que precisa ser tomada pelo algoritmo.

Se a empresa que construiu o veículo desenhou o código para que a opção nesses casos seja o menor dano possível à propriedade (o automóvel) e ao seu cliente (o casal de jovens), isso pode significar a morte de uma criança ou a morte de três idosos ao invés da morte do jovem casal proprietário do veículo. Se o critério for o menor número de vítimas possível, a criança na calçada será sacrificada (e preservaria assim a vida dos dois jovens no carro ou dos três idosos na outra via). Se o critério for outro que considere a preservação da vida de pessoas mais jovens (que teriam toda uma vida pela frente) como prioritária em relação à vida de pessoas mais velhas (que já tiveram a oportunidade de maior experiência de vida) o resultado será a morte dos três idosos do outro lado da via. Seja qual for a decisão, será polêmica e cairá em um dilema moral. Em tese, a depender da visão do da regulação aplicada, a lei pode obrigar a empresa a optar, neste caso, pelo dano ao carro e ao passageiro, preservando vidas alheias ao acidente.

Mas as situações podem ser ainda mais problemáticas: suponhamos que, ao invés de um carro com um casal de jovens, a colisão ocorresse com um veículo de transporte escolar com 12 crianças? Deveria o algoritmo optar pela morte dessas 12 crianças que estão diretamente envolvidas no acidente na via ou deveria sacrificar aquela criança que está na calçada (alheia ao acontecimento na via) ou ainda dos três idosos na via oposta? A única certeza neste caso é que os parâmetros para esta decisão não podem ficar a cargo apenas da empresa que desenhou os códigos. Pelo menos até 2016, a opção de empresas como a Mercedes-Benz é priorizar a vida do proprietário em caso de dilemas morais[15]. Um dos maiores desafios é justamente como regular essas novas instâncias de decisão sem que isso represente a preponderância de interesses e valores específicos (como os interesses comerciais da empresa que construiu o carro, ou interesses de um grupo que ferem direitos de terceiros).

[15] Ver em <https://www.tecmundo.com.br/mercedes/110591-sim-carro-autonomo-mercedes-atropelar-voce-salvar-condutor.htm > Acesso 20 out 2016.

6 Sobre autoritarismo estatístico das métricas

A automatização dos sistemas de IA bem como a sua efetividade estão baseadas em análises estatísticas alimentadas por um grande volume de dados e guiadas por métricas ou critérios previamente estabelecidas pelo programador. As métricas precedem os pesos, isto é, antes do sistema funcionar é preciso estabelecer quais os seus objetivos, onde se quer chegar, o que realmente importa. Como ilustra Bigonha (2018, p.6):

> Considere, por exemplo, um modelo que concede crédito e prediz a pontuação de uma pessoa, dada a probabilidade de ela pagar o empréstimo. O que representa o sucesso do sistema? Mais lucro? Maior número de pessoas recebendo empréstimo? Maior número de pagantes?

Neste caso, quem tem o poder de definir a métrica do algoritmo poderá exercer novas formas de poder codificado nos sistemas digitais. E isso pode ser bastante parcial, a ponto de gerar discriminação, acentuar desigualdades e ainda flertar com uma nova forma de autoritarismo configurado nas máquinas.

As métricas também podem ser autoritárias quando baseadas em informações que, embora tenham valor estatístico relevante e desvelem prognósticos realistas, não poderiam ser utilizadas para determinada tomada de decisão porque extrapolam sua função ou violam direitos. Nas análises de *big data*, é comum que se desconheça as razões de determinadas correlações entre variáveis, ainda que se saiba que elas existem e se utilize tal conhecimento em processos de tomada de decisão (MAYER-SCHONBERGER; CUKIER, 2013). Isso faz com que dados dos mais variados tipos e fontes sejam cruzados estatisticamente para que o sistema tome decisões lógico-estatísticas, porém autoritárias. Por exemplo:

> [...] é interessante mencionar a polêmica na Alemanha envolvendo a SCHUFA (empresa alemã que presta serviços de proteção ao crédito), a qual, no âmbito da avaliação de risco do consumidor, classificava como critério negativo o seu pedido de acesso a seus próprios dados. Isto se deve a uma correlação estatística que foi estabelecida no sentido de que consumidores que acessavam mais o seu score tinham maior chance de serem inadimplentes. A empresa sofreu inúmeras críticas em razão dessa conduta, que penalizava aquele que queria contratar um crédito com um scoring mais baixo, exclusivamente, em razão do exercício de um direito. (DONEDA et al, 2018, p. 6)

Neste caso, a regulação deve impor limites aos cruzamentos estatísticos dos sistemas de IA para o uso de determinadas informações que não dizem respeito ao tema em questão ou que podem subsidiar ações punitivas pelo simples fato de alguém exercer um direito, mas que gerou um dado que pode se voltar contra o próprio indivíduo. O desafio é como estabelecer estes limites sem criar um empecilho genérico quanto ao uso da estatística de correlações no cruzamento de informações e variáveis diversas.

É preciso definir em que situações isso é normal ocorrer e, ao mesmo tempo, em que situações isso se torna um abuso ou violação. Exigência de documentação que detalhem claramente como os códigos foram construídos e quais métricas foram utilizadas; elaboração de códigos deontológicos que estabeleçam diretrizes e limites para empresas e programadores; aplicação regular de auditorias algorítmicas independentes etc. são alguns elementos que o processo regulatório pode lançar mão.

7 Sobre obscurantismo matemático nos processos

Uma das questões mais recorrentes nas estratégias e documentos sobre políticas ou regulação de IA é o problema da opacidade dos algoritmos. Tendo como premissa que algoritmos estão cada vez mais onipresentes nos mais diversos âmbitos da atividade humana, ter a vida atravessada (e muitas vezes determinada) por regras e parâmetros que desconhecemos passa a ser um problema de autonomia e autodeterminação dos sujeitos. Por isso, de modo geral, a exigência de algum nível de controle, transparência e *accountability* das empresas que gestam projetos de IA tem sido uma tônica presente nos documentos e nas estratégias para este campo.

Em estudo que se debruçou sobre este problema da falta de transparência, Burrell (2016) identificou três tipos de opacidades que ocorrem em torno dos algoritmos. (a) Primeiro, a opacidade **como sigilo corporativo ou estatal intencional**. Empresas tendem a compreender os códigos como ativo comercial e os protegem como segredos industriais, mantendo-os longe dos olhos do público, sob a alegação de *copyright*. Governos também classificam determinados códigos como segredo de Estado ou dificultam a sua publicização. (b) Segundo, **a opacidade como analfabetismo técnico**. Algoritmos são códigos cuja transparência e entendimento é restrito àqueles que detém conhecimento da linguagem para decifrá-los. E apenas uma parcela ínfima da população possui essa "literacia". (b) Terceiro, *a* **opacidade como característica estrutural dos algoritmos de aprendizado de máquina** (*machine learning*). Este último item trata de um problema estrutural diretamente ligado à Inteligência Artificial, pois o denso treinamento das redes neurais e suas calibragens de pesos torna a reconstituição de todo o processo de como o sistema chegou a determinada decisão algo difícil de se visualizar pois quanto mais complexa a estrutura de redes neurais, mais obscuro se torna o processo (LEESE, 2014; MITTELSTADT et al 2016). Como aponta Rendtorff-Smith (2018, p.11):

> Esses sistemas, particularmente quando falamos de sistemas de aprendizado não supervisionado, são, por natureza, obscuros, o que dificulta a manutenção de princípios fundamentais de transparência, explicabilidade e accountability. Outro desafio tem sido que os governos geralmente adquirem e implantam tecnologia proprietária de IA desenvolvida por atores da indústria privada.

Para a autora, torna-se importante que os sistemas de IA sejam auditáveis, permitindo aos pesquisadores identificar a origem de um erro ou consequência adversa. Porém, auditorias tradicionais no formato que conhecemos hoje pode ser pouco efetiva pois, a olho nu, auditores terão muita dificuldade em decifrar o denso processo gerado pelas redes neurais artificiais.

Do ponto de vista regulatório, parte da solução desse obscurantismo técnico é o investimento em pesquisas e ferramentas de IA capazes de produzir auditorias automatizadas apropriadas para lidar com as camadas profundas das redes neurais artificiais. Para isso, é preciso que haja desenvolvimento de tecnologias voltadas especificamente para este fim, o que pode ser viabilizado por organizações ou agências específicas para estimular e fomentar o desenvolvimento deste campo. A criação de agências reguladoras específicas tem sido um elemento que aparece neste debate como um mecanismo necessário para este e outros problemas de *enforcement* e fiscalização junto aos sistemas de IA (SCHERER, 2016; TUTT, 2017).

8 Sobre a pervasividade utilitarista dos sistemas-autônomos

Um quarto problema que perpassa estratégias e políticas regulatórias para a Inteligência Artificial, e que merece atenção, é o caráter pervasivo e utilitarista inerentes a estas tecnologias. Pervasivo porque os sistemas de IA estarão, em um futuro não muito distante, inseridos em todas as áreas de vivência humana, agindo sobre os mais diversos tipos de serviços; intermediando inclusive nossa utilização de outros dispositivos e máquinas (como automóveis, aviões e casas); atravessando relações sociais, políticas, culturais e econômica e até mesmo afetivas; chegando primeiro onde os seres humanos em carne-e-osso não estiveram (como em missões espaciais e colonizações de planetas). Utilitarista porque há um pressuposto *benthamiano* no uso e difusão generalizada dessas tecnologias que são capazes de nos envolver ontologicamente, pois há uma aprofunda relação entre eficiência técnica, conforto e bem-viver. Observemos o que Bentham dizia no século XVIII e pensemos como isso relaciona com dispositivos de Inteligência Artificial nos dias atuais:

> By utility is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness, (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered: if that party be the community in general, then the happiness of the community: if a particular individual, then the happiness of that individual (BENTHAM, p. 14-15, 2000)

Como se pode perceber, a moralidade utilitarista é bastante alinhada com as tendências de funcionamento dos sistemas de Inteligência Artificial. Sobretudo, porque o discurso de exaltação nas estratégias, produtos e serviços neste campo versam sobre o horizonte de se melhorar a vida no futuro próximo para o máximo de pessoas possível e o majoritarismo

implícito nas métricas das redes neurais artificiais, como vimos anteriormente, reforçam esta perspectiva. Por exemplo, quando um sistema de IA capta o desejo majoritário de uma determinada população gerando benefícios específicos adequados para esta maioria, aqueles que ficaram à margem dos padrões estatísticos tendem a ser considerados desviantes, logo, a exclusão gerada pelo artefato estaria justificada com bases em uma moralidade utlilitarista que foi capaz de gerar felicidade para o máximo possível de pessoas.

Ao mesmo tempo, como ressaltam Helbing e colegas (2017), essas tecnologias serão cada vez mais capazes de se conectar, adentrar na vida e acompanhar todos o percurso de nossas ações e a tendência é que da "programação de computadores se avance para a programação de pessoas" que gradativamente estarão vinculadas e dependentes destes sistemas, em parte por opção (por consideramos mais cômodo e eficiente), em parte por haver uma pressão cultural e social, típica de sociedades datificadas:

> É possível perceber que em um mundo cada vez mais datificado com a abundância de oferta e uso massivo de aplicativos lógicos para ajudar em rotinas, não ceder os dados e decidir não alimentar *dataveillance* (ou vigilância digital) significa optar por uma vida de exclusão das diversas comodidades que o sistema oferece. Implicará em uma vida mais dura, com menos conforto, cuja execução de rotinas seria realizada em mais tempo e com mais energia a ser desprendida. No final, o grande problema da privacidade hoje está atrelado ao viés prático-lógico da cultura digital que, em última instância, significa um viver-melhor, isto é, com mais conforto e praticidade, ainda que com menos autonomia (SILVA, 2019, p. 164).

Para agravar o quadro, Nemitz (2018) lembra que a pervasividade dos sistemas de IA governará o grosso das funções de uma sociedade (saúde, educação, segurança, economia, justiça etc.) tendendo a ficar na mão de poucas empresas que de fato concentrarão o *know how* e infraestrutura necessária. E isso exige um papel balanceador do Estado, com a observância de princípios constitucionais, através de leis e mecanismos de contrapeso capazes de evitar o aumento da concentração de poder e exacerbação de desigualdades existentes e o surgimento de outras novas (NEMITZ, 2018).

Observando como estratégias nacionais de IA tem tradado essa relação entre utilidade e salvaguarda de direitos, Cath e colegas (2018) alertam que há um foco equivocado quando se prioriza o aspecto utilitário e não há ênfase social destas tecnologias. Os autores analisaram três relatórios sobre estratégias e regulação de IA publicadas nos EUA, Reino Unido e União Europeia e descobriram que há uma característica comum (que também é recorrente em documentos de outros países): a ausência de uma efetiva política que pense uma "*good AI society*". Os documentos estudados não são propositivos neste sentido e estão mais preocupados com questões de aplicabilidade (como economia, negócios etc.) e não com um efetivo plano para preparar a sociedade para uma sinergia positiva com essas inovações: " In short, we need a social strategy for AI, not mere tactics" (CATH, 2018, p. 3).

9 Sobre controle e fronteiras da eficiência

Como vimos anteriormente, um dos aspectos fundamentais dos sistemas de IA é a chamada fase de treinamento dos algoritmos. Com algumas variações, em geral existem duas formas fundamentais de treinamento: (a) aprendizagem supervisionada e (b) não-supervisionada. A primeira forma, implica em dar ao sistema dados que apontam exatamente onde se quer chegar, mostrando ao código os parâmetros desejados. Tomando nosso exemplo inicial, seria como dar ao algoritmo o objetivo de identificar se uma imagem é de um pássaro e, para isso, o código aprende a partir de um grande volume de imagens de pássaros de todos os tipos para que o sistema entenda quais os padrões de variáveis são mais prováveis de acerto para se identificar pássaros. Ou seja, treinamos o algoritmo para um horizonte específico e oferecemos a este, informações supervisionadas para tal.

Já a forma de aprendizagem não-supervisionada é quando não damos ao código instruções mais específicas e deixamos o sistema consumir aleatoriamente um montante de informação sendo capaz de identificar padrões ou correlações ocultas ou dificilmente perceptíveis a olho nu. Neste modelo, o algoritmo pode tomar caminhos diversos e descobrir coisas que não estavam previstas. Retomando nosso exemplo, seria como desenhar um código que consumisse imagens de todos os tipos até que se tornasse capaz de perceber padrões e diferenciar as imagens de um pássaro, um avião, uma borboleta. Especialmente neste último caso, a forma não-supervisionada, há uma grande preocupação sobre a perda de controle do elemento humano no sistema. Uma vez que o código é gerado e não é devidamente supervisionado a depender da área em que atua o algoritmo pode implicar em resultados ou ações danosas não previstas pelos seus desenvolvedores.

Uma outra dimensão importante que envolve ambos os casos é o risco de perda do controle, mas desta vez não no resultado final e sim no caminho que máquina optou por fazer em busca de seu objetivo. Russell (2019) argumenta que um dos principais problemas de IA é justamente o foco desmensurado na eficiência que, para ser alcançada, pode significar um trajeto eticamente contestável:

> As machines designed according to the standard model become more intelligent, however, and as their scope of action becomes more global, the approach becomes untenable. Such machines will pursue their objective, no matter how wrong it is; they will resist attempts to switch them off; and they will acquire any and all resources that contribute to achieving the objective (RUSSELL, 2019, p. 171).

Neste sentido, as diretrizes regulatórias e princípios éticos no desenvolvimento de códigos tem debatido a importância de restringir este tipo de aplicação em áreas mais sensíveis (que afetem de forma mais direta e impactante a vida dos indivíduos) e, ao mesmo tempo, criar mecanismos mais rígidos de acompanhamento nos processos capazes de tornar os sistemas de IA mais *accountable*, incluindo a possibilidade de desligamento manual no sistema em caso de

desvio ou perda de controle.

10 Sobre diversidade e representatividade nos códigos

Uma última questão-chave que tem atravessado os debates regulatórios e normativos sobre Inteligência Artificial diz respeito à dimensão inclusiva que o uso massificado de algoritmos para diversas aplicações cotidianas requer. Especificamente, implica em pensarmos nos valores que são embutidos (ou que foram esquecidos) no código. Isso nos remente a uma forma peculiar de representatividade tendo em vista que algoritmos são representações de valores incorporados na máquina que, embora pareçam falsamente neutros, são na verdade o resultado de uma visão de mundo ou de subjetividades objetivadas:

> Os programadores podem criar algoritmos que têm pressuposições tendenciosas ou limitações incorporadas neles. Eles podem inconscientemente enunciar uma questão de forma tendenciosa. Os preconceitos dos programadores individuais podem ter um efeito largo e acumulativo porque, em um sistema de software complexo composto por subsistemas menores, o viés real do sistema pode ser um composto de regras especificadas por diferentes programadores (CITRON, 2008, p. 1262).

Isso se aplica tanto aos sistemas de aprendizagem de máquina supervisionado como não-supervisionado pois a própria escolha dos dados que serão utilizados pelos algoritmos na fase de treinamento pode ser uma escolha enviesada ou pode, os próprios dados ou ambiente de treinamento, conter uma série de problemas de viés. Diversos estudos têm demonstrado o efeito discriminatório que algoritmos podem reforçar (GRAHAM, 2004; CPL, 2017; LEURS; SHEPHERD, 2017; WEST et al 2019) e isso ocorre tanto de forma direta e deliberada como também de modo indireto, como reflexo do processo produtivo de *design* dos sistemas, isto é, no nível das equipes de programadores:

> Discrimination and inequity in the workplace have significant material consequences, particularly for the under-represented groups who are excluded from resources and opportunities. For this reason alone the diversity crisis in the AI sector needs to be urgently addressed. But in the case of AI, the stakes are higher: these patterns of discrimination and exclusion reverberate well beyond the workplace into the wider world. Industrial AI systems are increasingly playing a role in our social and political institutions, including in education, healthcare, hiring, and criminal justice. Therefore, we need to consider the relationship between the workplace diversity crisis and the problems with bias and discrimination in AI systems (WEST et al, 2019, p. 15)

Por isso, diversas organizações e especialistas tem atentando para o problema da representatividade no processo de construção dos sistemas. Isso passa tanto pela existência de códigos deontológicos que programadores devem seguir para incorporar maior diversidade e visões de mundo - mesmo aquelas ausentes, indo para além do seu círculo social - quanto para romper com tendências de predominância étnica, cultura, racial daqueles que desenham código e constroem sistemas de IA que serão, em última instância, reflexo ampliado dos anseios humanos, para o bem ou para o mal.

11 Conclusão

Este artigo teve como objetivo central sintetizar e caracterizar os principais problemas-chave que uma boa política de Inteligência artificial precisa responder, no contexto das democracias contemporâneas. Inicialmente, a preocupação foi sintetizar o surgimento e a evolução da Inteligência Artificial enquanto campo de estudos e desenvolvimento, bem como destacar os principais aspectos que nos ajudam a compreender melhor o seu modo de funcionamento. Neste sentido, inicialmente discutiu-se o papel estatístico das redes neurais artificiais; a força e os problemas por trás da ideia de "imitação"; o poder da repetição nos mecanismos de automação e as tipologias ou níveis de IA.

Em seguida, a partir da perspectiva do "boom" de documentos governamentais e não-governamentais sobre estratégias e políticas para IA publicados principalmente entre 2017, elecou-se sete importantes problemas-chaves de ênfase política que estão nas alicerces deste debate: (1) personificação e imputabilidade da máquina; (2) dilemas e julgamentos morais; (3) autoritarismo estatístico das métricas; (4) obscurantismo matemático nos processos; (5) pervasividade utilitarista dos sistemas-autônomos; (6) controle e fronteiras da eficiência; (7) diversidade e representatividade nos códigos.

Longe de esgotar todas as dimensões que este complexo fenômeno levanta, a proposta foi destacar um conjunto de problemas que tem implicações políticas e que merecem especial atenção por se configurarem como elementos determinantes para a boa relação entre democracia e os sistemas de Inteligência Artificial em plena ascensão. A ideia é colaborar com uma compreensão mais ampla e política deste fenômeno, que seja útil para fomentar o aprofundamento do debate sobre estratégias e políticas regulatórias sobre IA em construção ou consolidação, fortalecendo uma perspectiva mais humanista que vá para além da eficiência técnica. Sobretudo, observou-se os dilemas que este cenário apresenta, a concentração de poder que se insinua e o possível aumento de assimetrias em sociedades cada vez mais datificadas. Todo este cenário requer que o Estado assuma seu devido papel de protetor de direitos individuais e coletivos, colocando em prática planos, políticas e regulação adequada para este campo.

Referências

ATABEKOV, A.; YASTREBOV, O. Legal Status of Artificial Intelligence Across Countries: Legislation on the Move. **European Research Studies Journal**, v. 21, n. 4, p. 773-782, 2018.

BENTHAM, Jeremy. **An Introduction to the Principles of Morals and Legislation**. Kitchener: Batoche Books, 2000.

BIGONHA, Carolina. Inteligência Artificial em perspectiva. **Panorama Setorial da Internet**, n.2. São Paulo: NIC.Br, 2018, p. 1-9.

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society,** 2016, p. 1-12;

CATH, C. *et al.* Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. **Science and Engineering Ethics**, v. 24, n. 2, 2018, p. 505-528.

CITRON, Danielle Keats. Technological Due Process. **Washington University Law Review**, v. 85, n. 6, p. 1249-1313, 2008.

COPELAND, Jack. Artificial Intelligence. *In:* COPELAND, Jack (org).**The Essential Turing**: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life: Plus The Secrets of Enigma. Oxford: Oxford University Press, 2004, p. 353-261.

CPL. Centre for Public Impact. **Destination unknown:** Exploring the impact of Artificial Intelligence on Government - Working Paper.Londres: CPL. Centre for Public Impact, 2017. Disponível em: https://resources.centreforpublicimpact.org/production/2017/09/Destination-Unknown-AI-and-government.pdf. Acesso em: 15 jul. 2019.

DONEDA et al. Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal. Pensar: **Revista de Ciências Jurídicas**, v. 23, n. 4, p. 1-17, 2018.

GIRASA, Rosario. **Artificial Intelligence as a Disruptive Technology**: Economic Transformation and Government Regulation. Cham: Palgrave Macmillan, 2020

GLEICK, James. **A informação**: Uma história, uma teoria, uma enxurrada. São Paulo: Compahia das Letras, 2011.

GRACE, Katja *et al*. When Will AI Exceed Human Performance? Evidence from AI Experts. **Journal of Artificial Intelligence Research**, n. 62, p. 729-754, 2018.

GRAHAM, Stephen. The woftware-sorted city: rething the "digital divide". *In:* GRAHAM, S. (org.). **The cybercities reader**. Londres: Routledge, 2004, p. 324-332.

GUTMANN, Amy; THOMPSON, Dennis. **Democracy and Disagreement**. London: Cambridge; Massachusetts: Harvard University Press, 1996.

HEBB, Donald O. **The organization of behavior**: a neuropsychological theory. Oxford: Wiley, 1949.

HEIDEGGER, Martin. **Ensaios e conferências**. Petrópolis: Vozes, 2001.

HELBING, Dirk et al. Will Democracy Survive Big Data and Artificial Intelligence? **Scientific American**, n. 98, p.73-98, 2017. Disponível em: www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence. Acesso em: 15 jul. 2019.

LEURS, Koen; SHEPHERD, Tamara. Datafication & Discrimination. *In:* SCHÄFER, Mirko Tobias; ES, Karin van (org). **The Datafied Society**: Studying Culture through Data. Amsterdam: Amsterdam University Press, 2017, p. 211-231.

McCULLOCH, Warren S.; PITTS, Walter H. A logical calculus of the ideas immanet in nervous activity. **Bulletin of Mathematical Biophysics**, n. 5, p. 115-133, 1943.

LEESE, Matthias. The new profiling: Algorithms, black boxes, and the failure of ant)i-discriminatory safeguards in the European Union. **Security Dialogue**, v. 45, n.5, p. 494-511, 2014.

MAYER-SCHONBERGER, Viktor; CUKIER, Kenneth. **Big Data:** como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro: Editora Campus, 2013.

MITTELSTADT, B. D. *et al*. The ethics of algorithms: Mapping the debate. **Big Data & Society**, v. 3, n. 2, p. 1-21, 2016

MONTRÉAL DECLARATION, For a **Responsible Development of Artificial Intelligence 2018**. Montreal: Université de Montréal, 2018. Disponível em: https://www.montrealdeclaration-responsibleai.com/the-declaration. Acesso em: 22 jul. 2019.

NEMITZ, Paul. Constitutional democracy and technology in the age of artificial intelligence. **Philosophical Transaction Real Society**, p. 1-14, 2018.

RENDTORFF-SMITH, Sara. Desafios de Governança em Inteligência Artificial. Entrevista. In: **Panorama setorial da Internet**, n.2, São Paulo: NIC.Br, 2018, p. 10-12.

ROSENBLAT, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychologial Review**, n. 65, p. 386-408, 1958.

RUSSELL, Stuart. **Human Compatible**: artificial intelligence and the problem of control, 2019.

SCHERER, Matthew U. Regulating artificial intelligence systems: reisks, challenges, competencies, and strategie. **Harvard Journal of Law & Technology**, v. 29, n. 2, p. 354-400, 2016.

SILVA, Sivaldo P. da. Comunicação digital, Economia de Dados e a racionalização do tempo: algoritmos, mercado e controle na era dos bits. **Revista Contracampo**, v. 38, n. 1, p. 157-169, 2019.

TURING, A.M. Computing machinery and intelligence**. Mind**: a Quarterly Review of Psychology and Philosophy, v.49, n. 236, p. 433-460, 1950.

TUTT, Andrew. An FDA for Algorithms. **Admin Law Review**, n. 83, p. 83-123, 2017.

WEST, S.M., Whittaker, M. and Crawford, K. **Discriminating Systems:** Gender, Race and Power in AI. Nova York: AI Now Institute. 2019. Disponível em: https://ainowinstitute.org/discriminatingsystems.html. Acesso em: 8 fev. 2020.

Artigo submetido em: 2020-05-02

Artigo aceito em: 2020-05-31