## 1. INTRODUÇÃO

Os modelos de aprendizagem profunda (*deep learning*) possibilitaram o surgimento de sistemas de inteligência artificial (IA) complexos, que desenvolvem raciocínios e tomam decisões capazes de emular o pensamento humano. Tais sistemas estão cada vez mais presentes na vida das pessoas, e suas funcionalidades práticas já são indispensáveis em nosso cotidiano. Assistentes virtuais, *smartcameras* com reconhecimento facial e tradutores de texto instantâneos para dezenas de idiomas são apenas alguns exemplos que mostram que a IA chegou para ficar. Alguns autores sugerem que, no futuro, as máquinas serão capazes de exercer profissões complexas, como o trabalho de médicos, juízes e engenheiros (SUSSKIND, 2010).

A expansão da IA nos instiga a pensar sobre inúmeras questões: privacidade, dilemas éticos e morais, responsabilização de robôs, etc. Dentre tantas implicações possíveis, este trabalho propõe um *recorte metodológico*, buscando discutir a *explicabilidade* – ou seja, a melhor compreensão humana sobre o processo decisório das máquinas inteligentes – enquanto requisito para o desenvolvimento dessas novas tecnologias. Não há dúvidas de que a explosão da IA marcará o surgimento de outra infinidade de recursos para as mais diversas tarefas. Neste cenário, *poderia a eficácia e a legitimidade desses sistemas ser limitada pela incapacidade da máquina de explicar seus pensamentos e ações aos humanos?*

A hipótese colocada é de que, se os usuários quiserem gerenciar e confiar nos sistemas artificialmente inteligentes, será fundamental oferecer mais transparência em relação aos processos internos que levaram os sistemas de IA a tomarem as suas decisões. Em última instância, explicar tais processos decisórios aumentaria a compreensão e a legitimação das ações tomadas pelos sistemas autônomos (VILLANI, 2018).

Neste sentido, o objetivo deste trabalho é investigar a *explicabilidade* enquanto requisito para a justificação e legitimação das decisões tomadas pelos sistemas de inteligência artificial. A presente pesquisa tem caráter qualitativo e descritivo, utilizando a lógica indutiva para analisar a *explicabilidade*. A fim de verificar a hipótese levantada, será feita uma revisão bibliográfica acerca do conjunto teórico disponível sobre o tema.

Conclui-se que o entendimento por trás da lógica de cada decisão produzida por algoritmos poderia facilitar processos de auditoria e responsabilização, aumentando a confiabilidade e a segurança das ferramentas da IA. Ainda, a *explicabilidade* contribui para o aperfeiçoamento contínuo destes sistemas, auxiliando na correção de vieses e preconceitos reproduzidos no interior dos algoritmos. Por consequência, decisões de máquinas mais transparentes, precisas e confiáveis terão maior legitimidade e eficácia perante a sociedade.

## 2. BREVE HISTÓRICO DA IA E CONCEITOS-CHAVE

A inteligência artificial pode ser definida como a capacidade da máquina de interpretar dados de forma racional e humana, tomando decisões autônomas com base em padrões preexistentes (NORVIG & RUSSEL, 1995). Na mesma direção, SIMONS (2016) preleciona que é a ciência de ensinar computadores a "*aprender, raciocinar, perceber, inferir, comunicar e tomar decisões como os humanos*".

As primeiras pesquisas sobre IA, feitas a partir das décadas de 1940 e 1950, visavam a solução de problemas a partir de *métodos simbólicos*, baseados em mecanismos matemáticos relativamente rudimentares, como o aprendizado por analogia/instâncias, o aprendizado por indução e o aprendizado por evolução/seleção. Nestas abordagens, as máquinas eram orientadas a manipular informações simbólicas (qualitativas), o que gerava limitações para trabalhar valores numéricos e tratar os problemas com a devida completude (OSÓRIO, 1999).

Em contraponto aos métodos de aprendizado simbólico, desenvolveu-se o estudo das redes neurais artificiais (RNAs). Elas utilizavam o chamado *método conexionista*, inspirado na estrutura dos neurônios humanos, que são conhecidamente conectados entre si e operam em paralelo. Os modelos de RNAs evoluíram ao longo dos anos, e em 1983, a agência norte-americana DARPA (Defense Advanced Research Projects Agency) fundou um departamento destinado a pesquisas em neurocomputação, impulsionando o desenvolvimento das RNAs, que enfim acabaram prevalecendo sobre os métodos simbólicos (ANDRADE, 2021).

A partir da década de 1980, começou a emergir, no campo das RNAs, o aprendizado de máquina (*machine learning*). Este ramo da computação passou a desenvolver algoritmos que se aprimoram automaticamente por meio da experiência e do uso de dados, construindo modelos baseados em dados de amostra, conhecidos como 'dados de treinamento' (*training data*), a fim de fazer previsões ou decisões sem serem explicitamente programados para isso (SURDEN, 2014). No *machine learning*, o computador é desenvolvido para "se autoprogramar" com base em sua própria experiência. Ele reúne dados, interpreta essas informações e toma decisões diferenciadas, trabalhando com padrões cognitivos similares aos usados por humanos (ARENS, 2017).

Mais recentemente, sobretudo na última década, vem sendo desenvolvido o aprendizado profundo (*deeplearning*), uma classe sofisticada de algoritmos de aprendizado de máquina, que utiliza múltiplas camadas para extrair progressivamente recursos de nível superior da camada de entrada.

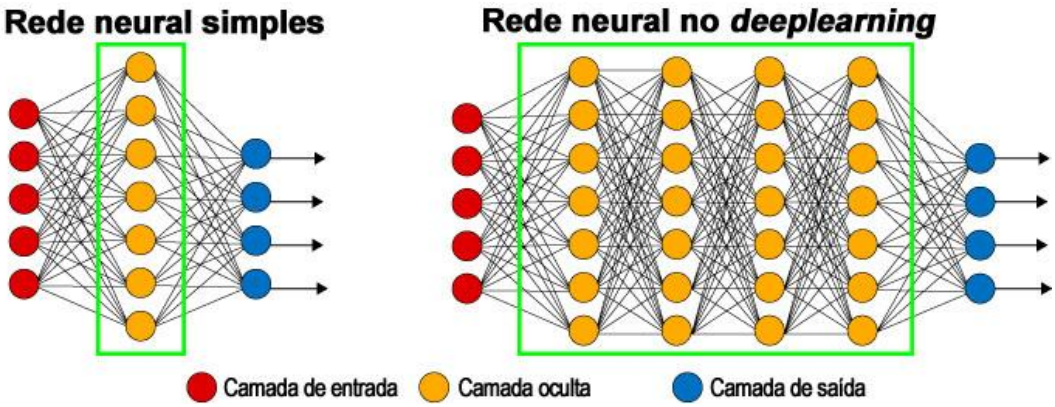


*Figura 1: diferença entre a rede neural simples e a rede neural no deeplearning.*
*Adaptado de https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning*

Enquanto o aprendizado de máquina usa algoritmos para analisar dados, evoluir com esses dados e tomar decisões informadas com base no que aprendeu, o aprendizado profundo estrutura algoritmos em camadas para criar uma rede neural artificial que pode aprender e tomar decisões inteligentes por conta própria. O *deeplearning* foi aplicado, por exemplo, em um robô projetado para detectar o câncer de pele através de fotografias, obtendo taxas de sucesso iguais às de 21 dermatologistas renomados (ESTEVA et al, 2017). Outro exemplo é o algoritmo da Google 'AlphaGo', que aprendeu a jogar um complexo jogo de tabuleiro chamado Go, derrotando grandes mestres humanos do game, depois de estudar, aprender e reverter suas técnicas mais complexas enquanto jogava (SILVER et al, 2017).

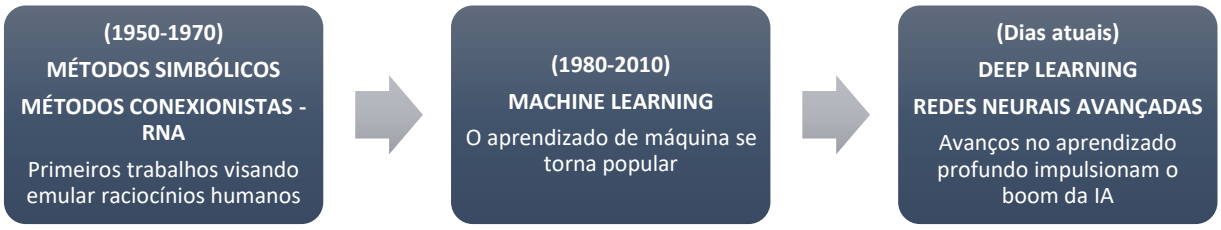


*Figura 2: A evolução da inteligência artificial. Elaborado pelo autor.*

## 3. O PROBLEMA DA OPACIDADE

O avanço das tecnologias de inteligência artificial levanta uma série de desafios éticos e jurídicos sobre os processos decisórios inerentes a estes sistemas. Um dos aspectos mais preocupantes é a dificuldade em compreender o fluxo de raciocínio no interior de um sistema de IA, uma vez que geralmente conhecemos somente o *resultado* de suas ações, mas sabemos pouco ou quase nada sobre a complexa sequência de processamento que levou a elas. É o que

se convencionou chamar de “opacidade”. Um sistema de IA “opaco” é aquele em que é difícil para um ser humano comum ter uma completa visão de *como* este sistema decide e *porquê* decide um certo curso de ações (SURDEN, 2016, p. 158).

BURRELL (2018, p. 4-5) descreveu três formas de opacidade: a primeira tem a ver com o sigilo institucional (pode ser de uma empresa ou Estado) sobre seus sistemas inteligentes; a segunda relaciona-se ao “analfabetismo técnico”, pois entender os códigos de um algoritmo geralmente requer uma habilidade especializada que a maioria dos humanos não possui. Contudo, para BURRELL, essas duas formas de opacidade não são tão preocupantes quanto a terceira: a crescente complexidade intrínseca dos algoritmos. O desafio aqui não consiste na dificuldade de acesso ou leitura do código, mas na incapacidade de entender o curso de ações de alta complexidade do algoritmo. Isso porque, além de processar um volume incomensurável de dados, o algoritmo pode alterar constantemente sua lógica de decisão interna, à medida que “aprende” com os dados de treinamento. Portanto, o acesso e a leitura do código podem não ser suficientes para entender como a rede neural está operando, pois a quantidade de dados é cada vez maior e a natureza de operações fica sempre mais heterogênea, atraindo grande opacidade. Assim, a aprendizagem profunda pode gerar falta de conhecimento: é possível observar os dados de entrada e os dados de saída, mas o funcionamento do sistema é mal compreendido, tornando-o uma espécie de “caixa-preta”.

## 4. EXPLICABILIDADE DA INTELIGÊNCIA ARTIFICIAL

Diante deste cenário, alguns estudiosos têm se preocupado em desenvolver mecanismos para reduzir ou mitigar a opacidade da IA. Uma das principais ideias é agregar, à inteligência artificial, maior *‘explicabilidade’*, ou seja, melhorar a compreensão humana sobre o processo decisório das máquinas inteligentes. De acordo com a DARPA, a *explicabilidade* ocorre quando os *“novos sistemas de aprendizado de máquina são capazes de explicar seus fundamentos, caracterizar seus pontos fortes e fracos e transmitir uma compreensão acerca das suas condutas futuras”.*

A *explicabilidade* certamente relaciona-se com o aumento da transparência das máquinas inteligentes, mas não só: também envolve estratégias para justificar, em linguagem humana, a cadeia de decisão do algoritmo; desenvolver mecanismos internos para detectar e solucionar preconceitos e vieses; identificar quais foram os responsáveis pela programação de um sistema e até criar demonstrações visuais que “ilustram” as linhas de raciocínio seguidas pela máquina.

ILUSTRAÇÃO DE UM AMBIENTE DE *EXPLICABILIDADE* DE IA

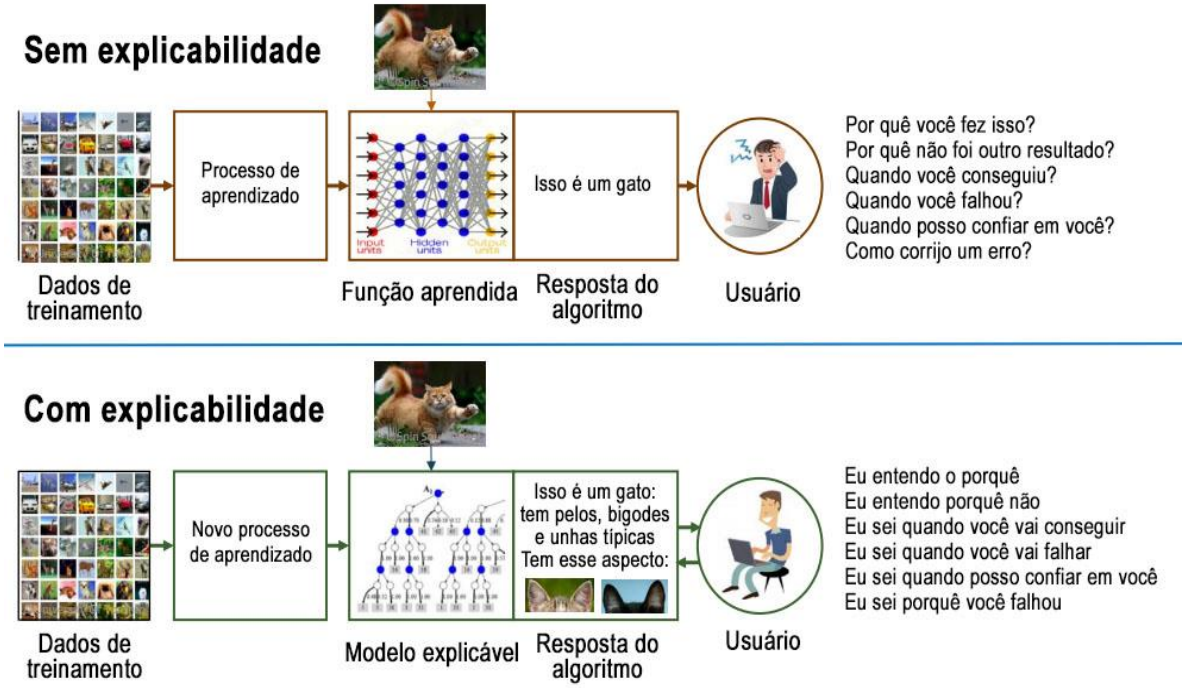


*Figura X: No cenário sem explicabilidade, o algoritmo dá a resposta sem oferecer informações sobre seu processo de aprendizado, ao passo que num ambiente de explicabilidade, o usuário tem mais informações sobre as operações algorítmicas, permitindo inclusive corrigir falhas e compreender a razão dos resultados. Adaptado de GUNNING, David, 2016*

Explicar as decisões auxiliadas por inteligência artificial traz benefícios significativos, tanto para a sociedade como para as empresas que detêm ferramentas algorítmicas. No âmbito empresarial, a *explicabilidade* garante, por exemplo, a melhoria da conformidade legal das ferramentas utilizadas, reduzindo os riscos jurídicos associados ao descumprimento de normas regulatórias sobre IA. A *explicabilidade* também aumenta a confiança dos funcionários e clientes na inteligência artificial, na medida em que permite a melhor compreensão dos processos autômatos, demonstrando uma postura respeitosa da empresa para com seus *stakeholders* (INFORMATION COMMISSIONER'S OFFICE & ALAN TURING INSTITUTE, 2020, P. 16).

De igual forma, a *explicabilidade* é de importância fundamental para os indivíduos e para a sociedade. Primeiro, porque o maior conhecimento público acerca dos processos algorítmicos possibilita um debate mais instruído e consciente sobre a adoção e o desenvolvimento das novas tecnologias. Em segundo lugar, porque ela permite a otimização das decisões de inteligência artificial, ajudando a mitigar resultados discriminatórios e a

eliminar vieses algorítmicos (INFORMATION COMMISSIONER'S OFFICE & ALAN TURING INSTITUTE, 2020, P. 17).

## 5. CONCLUSÃO

Muito embora o aprendizado de máquina possibilite a realização e o aperfeiçoamento de uma infinidade de tarefas humanas, ele nem sempre gera decisões perfeitas, equânimes e imparciais. Acreditamos que nos casos de decisões imperfeitas, a *explicabilidade* pode ajudar a identificar vieses e preconceitos, determinar responsabilidades e corrigir falhas dos sistemas de inteligência artificial.

Mesmo nos casos onde o algoritmo produz decisões aparentemente "perfeitas", é relevante entender a cadeia de raciocínio utilizada no interior do sistema. Isso poderia ser comparado à uma espécie de "*accountability*", que já está presente na sociedade, como é o caso da prestação de contas por agentes públicos ou privados. Ou seja, ainda que uma decisão algorítmica seja "acertada", sua *explicabilidade* perante os usuários permanece útil e desejável.

Por fim, com o gradual implemento da *explicabilidade*, é esperado que a cadeia de benefícios por ela produzida (transparência, *accountability*, segurança, precisão dos sistemas, identificação de vieses, etc.) influencie positivamente a percepção dos cidadãos sobre os sistemas de inteligência artificial. Neste sentido, a mitigação das "caixas-pretas" algorítmicas através da *explicabilidade* pode contribuir, de modo geral, para sedimentar a confiabilidade e a segurança dessas ferramentas, oferecendo maior legitimidade e eficácia aos sistemas de IA.

## REFERÊNCIAS BIBLIOGRÁFICAS


ANDRADE, Otávio Morato de. **"Doutor Robô"? A substituição dos advogados pelas máquinas e outras considerações sobre o Direito na era pós-digital** *In* O futuro das profissões jurídicas: você está preparad@? Reflexões e caminhos. Fundação Getúlio Vargas. São Paulo, 2021.

ARENS, Bob. **Cognitive computing: Under the hood.** Thomson Reuters. Jan 2017. Disponível em https://blogs.thomsonreuters.com/answerson/cognitive-computing-hood/. Acesso em: 15/09/2020.

BURRELL, Jenna. **How the machine 'thinks': Understanding opacity in machine learning algorithms.** Big Data & Society Jan–Jun 2016: 1–12.

ESTEVA, Andre; KUPREL, Brett; NOVOA, Roberto A.; KO Justin; SWETTER Susan M.; BLAU Helen M.; THRUN, Sebastian. **Dermatologist-level classification of skin cancer with deep neural networks.** Nature, 542, p. 115–118, 2017.

GUNNING, David. **Explainable Artificial Intelligence (XAI) DARPA/I2O.** Website da Defense Advanced Research Projects Agency (DARPA), 2016. Disponível em: https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

INFORMATION COMMISSIONER'S OFFICE; ALAN TURING INSTITUTE. **Explaining decisions made with AI.** Londres, 2020.

OSÓRIO, Fernando. **Redes Neurais - Aprendizado Artificial.** Forum de I.A. 1999. Disponível em: http://www2.ic.uff.br/~labic/conteudo/textos/osorio-rn.pdf

RUSSEL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. New Jersey: Prentice-Hall, 1995.

SILVER, David; SCHRITTWIESER, Julian, SIMONYAN, Karen, ANTONOGLOU, Ioannis, HUANG Aja; GUEZ Arthur; HUBERT Thomas; BAKER Lucas; LAI Matthew; BOLTON Adrian; CHEN Yutian, LILLICRAP, Timothy; HUI Fan; SIFRE Laurent, VAN DEN DRIESSCHE George; GRAEPEL Thore; HASSABIS, Demis. **Mastering the game of Go without human knowledge.** Nature. 19 de outubro 2017.

SIMONS, John. **Tomorrow's Business Leaders Learn How to Work with A.I.** The Wall Street Journal. Nov. 2016. Disponível: https://www.wsj.com/articles/tomorrows-business-leaders-learn-how-to-work-with-a-i-1480517287. Acesso em 15/09/2020.

SURDEN, Harry. **Machine Learning and Law.** Washington Law Review, Vol. 89, No. 1, 30 mar 2014

SURDEN, Harry; WILLIAMS, Mary-Anne. **Technological Opacity, Predictability, and Self-Driving Cars.** 38 Cardozo L. Rev. 121, 2016.

SUSSKIND, Richard. **The End Of Lawyers: Rethinking The Nature Of Legal Services.** Oxford Univ. Press 2010).

TUREK, Matt. **Explainable Artificial Intelligence (XAI).** Site da Defense Advanced Research Projects Agency (DARPA). Disponível: https://www.darpa.mil/program/explainable-artificial-intelligence. Acesso em: 15/09/2020.

VERGARA, Sylvia Constant. **Projetos e relatórios de pesquisa em administração.** São Paulo, 1988.

VILLANI, Cédric. **Donner uns sens à li'intelligence artificielle: pour une stratégie nationale et européenne.** Disponível em: https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf. Acesso em 15/09/2019