Transparência e Explicabilidade de Sistemas de IA

Alexandre Marques Tortoza Canoa
Escola Politécnica
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, PR, Brasil
a.marquestortoza@gmail.com

Renato Mendes Tesserolli
Escola Politécnica
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, PR, Brasil
renatotesserolli@gmail.com

Paulo Henrique Perin
Escola Politécnica
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, PR, Brasil
paulo.perin@pucpr.edu.br

Brian Augusto Mendes Ferreira
Escola Politécnica
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, PR, Brasil
brian.ferreira@pucpr.edu.br

Giancarlo Perli
Escola Politécnica
Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba, PR, Brasil
giancarlo.perli@pucpr.edu.br

***Abstract*—O uso da inteligência artificial tornou-se intrínseco à vida moderna. Dezenas de milhões de pessoas utilizam essa tecnologia de forma indiscriminada e, muitas vezes, sem discernimento. Com o avanço constante da IA, diversos setores passaram a delegar tarefas críticas a essas ferramentas, como seleção de candidatos para vagas de emprego, julgamentos judiciais e diagnósticos médicos. Diante disso, é necessário refletir sobre até que ponto as decisões fornecidas por sistemas de IA são efetivamente questionadas e compreendidas? Este artigo propõe uma análise crítica da transparência e explicabilidade desses sistemas, partindo do pressuposto de que a opacidade das redes neurais e a proteção de propriedade intelectual criam barreiras fundamentais à compreensão. Considera-se, portanto, não apenas o risco de enviesamento, mas a própria impossibilidade de auditoria completa de sistemas.**

***Index Terms*—Inteligência Artificial, Transparência, Explicabilidade, XAI, Caixa-preta, Ética em IA**

I. Introdução

A inteligência artificial (IA) é um campo da ciência da computação voltado ao desenvolvimento de sistemas capazes de executar tarefas que normalmente exigiriam raciocínio humano. Segundo André A. Suave, trata-se de tecnologias que simulam capacidades cognitivas humanas, por meio de algoritmos que operam sobre dados [1]. Antes de adentrarmos nas questões de transparência e explicabilidade, é fundamental compreender alguns conceitos fundamentais relacionados à IA.

As redes neurais podem ser vistas sob uma ótica similar à do cérebro humano, em que o aprendizado e a memória ocorrem por meio das sinapses, conexões ou ligações entre neurônios, que são reforçadas conforme o uso. Para ilustrar, no contexto de 'textos', poderíamos imaginar um sistema que, munido de regras gramaticais, ortográficas e dicionários e devidamente treinado, formularia frases. Entretanto, muitos problemas não se tratam de mundos binários, onde há apenas memorização de regras e condições de verdadeiro e falso.

Um paralelo interessante pode ser feito com o processo de aprendizagem de crianças. Inicialmente, elas passam longo tempo ouvindo adultos, até começarem a formular respostas para o que desejam expressar, e então refinam seu conhecimento com o feedback recebido dos adultos. Na computação, processo semelhante é denominado aprendizado supervisionado, no qual um modelo é treinado e continuamente ajustado a partir de feedbacks.

Surge, então, o primeiro questionamento deste artigo: e se intencionalmente ensinarmos uma criança a identificar um 'gato' como 'cão'? Sempre que questionada, ela responderia com convicção, mesmo que errônea. De maneira análoga, modelos de IA podem reproduzir e reforçar erros quando alimentados com dados enviesados ou incorretos, acreditando estar corretos. Assim, tanto em humanos quanto em máquinas, o aprendizado supervisionado depende diretamente da qualidade e da intenção dos dados fornecidos.

Esse cenário nos leva ao segundo questionamento fundamental: até que ponto estamos delegando nosso discernimento a sistemas de IA, sem compreender plenamente seus critérios internos? Em situações cotidianas, desde a escrita de um simples e-mail até diagnósticos médicos, decisões automatizadas influenciam escolhas humanas. Essa delegação do julgamento levanta preocupações profundas sobre confiabilidade.

A importância deste tema se intensificou com regulamentações emergentes. No Brasil, a Lei Geral de Proteção de Dados (LGPD) "dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, [...] com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural" [2]. Esta proteção torna-se particularmente

desafiadora quando sistemas de IA opacos processam dados pessoais para decisões automatizadas que impactam significativamente os indivíduos.

II. Opacidade das Redes Neurais

As redes neurais artificiais representam uma das abordagens mais poderosas e ao mesmo tempo mais complexas da inteligência artificial contemporânea. Inspiradas no funcionamento do cérebro humano, essas estruturas são compostas por múltiplas camadas de neurônios artificiais interconectados, que processam informações de forma distribuída. Cada camada transforma os dados recebidos, ajustando pesos e parâmetros de modo a minimizar erros durante o processo de aprendizado supervisionado [3]. O avanço dessas arquiteturas, como as redes profundas (Deep Neural Networks), impulsionou resultados notáveis em áreas como reconhecimento de imagens, processamento de linguagem natural e diagnósticos médicos. Entretanto, essa mesma sofisticação introduziu uma barreira significativa à compreensão humana dos processos internos de decisão.

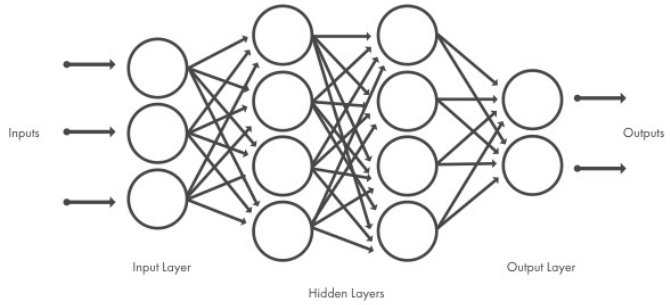

Fig. 1. Exemplo de uma arquitetura de rede neural profunda.

A. Explicação das Redes Neurais

Uma rede neural é composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Durante o treinamento, o modelo ajusta seus parâmetros internos para aprender padrões complexos a partir de grandes volumes de dados. A capacidade de generalização dessas redes está relacionada à profundidade das camadas e à quantidade de parâmetros ajustáveis, o que lhes permite extrair representações altamente abstratas [4]. Essa característica distingue as redes neurais de modelos lineares e baseados em regras, cuja lógica é diretamente observável.

De acordo com Ortigossa et al. (2024), a evolução das arquiteturas de aprendizado profundo trouxe uma nova dimensão à complexidade algorítmica, tornando cada vez mais difícil correlacionar as entradas e saídas do sistema com o raciocínio interno do modelo [5]. A explicabilidade, nesse contexto, passa a ser desafiadora porque o conhecimento aprendido não está localizado em um único ponto, mas distribuído em milhões de conexões matemáticas. Como apontado pela DARPA no programa Explainable Artificial Intelligence (XAI), há uma tensão inerente entre desempenho e interpretabilidade: quanto mais preciso o modelo, menos compreensível tende a ser [6].

B. A Natureza da Opacidade

A opacidade das redes neurais refere-se ao distanciamento entre o funcionamento interno do modelo e a capacidade humana de interpretá-lo. Em outras palavras, trata-se da dissociação entre o processo computacional de tomada de decisão e o entendimento conceitual que dele se pode extrair. Essa opacidade decorre, primeiramente, da distribuição do conhecimento entre bilhões de parâmetros interdependentes, o que impede a identificação de uma lógica central [5]. Além disso, a natureza não linear das funções de ativação faz com que as transformações internas não possuam correspondência direta com conceitos humanos [4].

Balasubramaniam et al. (2023) destacam que a transparência de um sistema de IA depende da capacidade de se compreender os critérios que sustentam suas decisões, algo comprometido pela opacidade intrínseca das redes neurais profundas [4]. Em vez de um raciocínio semântico, essas redes operam sobre abstrações estatísticas, nas quais a explicação emerge de correlações e não de causalidades compreensíveis. Ortigossa et al. (2024) denominam esse fenômeno de "opacidade estrutural", em que a alta dimensionalidade e a complexidade não linear impedem que os resultados sejam traduzidos em linguagem humana [5].

Assim, a opacidade não é apenas uma característica técnica, mas epistemológica: o modelo "sabe" algo que o ser humano não pode diretamente compreender. Essa dissociação entre cálculo e sentido é o que transforma a rede neural em um artefato matemático eficaz, porém hermético.

C. Implicações da Opacidade

A opacidade das redes neurais acarreta consequências diretas para a confiabilidade, a auditabilidade e a ética dos sistemas de IA. Primeiramente, torna-se difícil verificar ou contestar as decisões produzidas por modelos complexos, uma vez que não há uma explicação clara de como cada variável contribuiu para o resultado final. Esse problema é especialmente crítico em aplicações sensíveis, como medicina, direito ou finanças, onde decisões automatizadas podem afetar vidas humanas [7].

Além disso, a impossibilidade de rastrear as razões internas de uma previsão compromete a reprodutibilidade científica e o controle de vieses. Cheong (2024) argumenta que a falta de transparência mina a responsabilidade dos desenvolvedores e reduz a confiança social nas tecnologias inteligentes [3]. Quando o sistema não oferece justificativas compreensíveis, há também um risco jurídico: de acordo com legislações como a LGPD e o GDPR, indivíduos têm direito a explicações sobre decisões automatizadas que os afetam [4].

Em síntese, a opacidade das redes neurais representa um dos principais desafios contemporâneos da inteligência artificial. Ela evidencia a contradição entre modelos altamente eficazes e a incapacidade humana de compreender sua lógica interna. Essa opacidade estrutural constitui o núcleo do problema da "caixa-preta", discutido na seção seguinte.

III. A Caixa-Preta

O termo "caixa-preta" (black box) é amplamente utilizado para descrever sistemas de Inteligência Artificial cujos processos internos de decisão são inacessíveis ou incompreensíveis ao ser humano. Nesses sistemas, observa-se apenas o comportamento de entrada e saída, sem visibilidade sobre o raciocínio intermediário. Esse fenômeno surge principalmente em modelos de aprendizado profundo, em que milhões de parâmetros interagem de forma não linear para produzir um resultado. O problema não reside apenas na quantidade de cálculos, mas na impossibilidade de traduzir essas transformações em lógica interpretável. Como descreve Gunning (2019), quanto mais preciso o modelo, menos compreensível tende a ser, e essa tensão marca o dilema entre desempenho e explicabilidade [6].

A metáfora da caixa-preta tem implicações epistemológicas profundas. Em sistemas tradicionais baseados em regras, o raciocínio é explícito: cada etapa pode ser rastreada e compreendida. Já nas redes neurais profundas, a "lógica" é substituída por correlações estatísticas que, embora eficazes, carecem de significado semântico. Isso cria uma dissociação entre a decisão do modelo e a noção humana de causalidade. Na prática, o sistema acerta "sem saber por quê". Essa ruptura entre cálculo e compreensão representa o nascimento de uma nova forma de conhecimento técnico, em que o desempenho substitui a explicação como critério de verdade. Ortigossa et al. (2024) chamam esse fenômeno de "opacidade estrutural", pois o conhecimento é distribuído e impossível de localizar em um ponto observável [5].

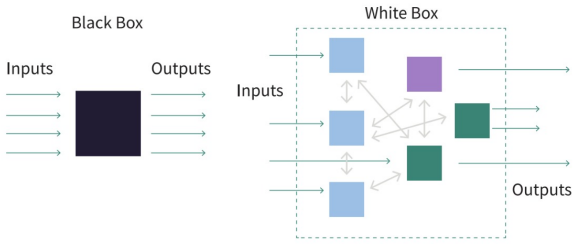

Fig. 2. Ilustração do conceito de sistema caixa-preta.

A metáfora da caixa-preta tem implicações epistemológicas profundas. Em sistemas tradicionais baseados em regras, o raciocínio é explícito: cada etapa pode ser rastreada e compreendida. Já nas redes neurais profundas, a "lógica" é substituída por correlações estatísticas que, embora eficazes, carecem de significado semântico. Isso cria uma dissociação entre a decisão do modelo e a noção humana de causalidade. Na prática, o sistema acerta "sem saber por quê". Essa ruptura entre cálculo e compreensão representa o nascimento de uma nova forma de conhecimento técnico, em que o desempenho substitui a explicação como critério de verdade. Ortigossa et al. (2024) chamam esse fenômeno de "opacidade estrutural", pois o conhecimento é distribuído e impossível de localizar em um ponto observável [5].

A presença da caixa-preta em sistemas críticos, como diagnósticos médicos, decisões judiciais ou análise de crédito, traz riscos éticos e sociais significativos. Um modelo capaz de classificar milhares de pacientes ou determinar a probabilidade de reincidência criminal pode apresentar vieses sistemáticos sem que haja meios de identificá-los. A ausência de interpretabilidade compromete princípios fundamentais de justiça e responsabilidade, tornando difícil atribuir culpa ou corrigir falhas. Além disso, a opacidade tecnológica tende a concentrar poder nas mãos de poucas empresas e especialistas capazes de controlar o funcionamento interno dos modelos. Balasubramaniam et al. (2023) argumentam que a transparência é o primeiro passo para restabelecer a confiança pública e a responsabilidade social da IA [4].

A caixa-preta também desafia a própria noção de responsabilidade técnica. Se um sistema de IA erra, quem é o responsável: o programador, o algoritmo ou o dado que o treinou? Essa pergunta, aparentemente filosófica, torna-se jurídica diante das leis de proteção de dados e das normas éticas em desenvolvimento. No Brasil, a Lei Geral de Proteção de Dados (LGPD) e, na União Europeia, o GDPR, garantem o direito à explicação de decisões automatizadas que impactem indivíduos. Na prática, porém, esse direito ainda esbarra na impossibilidade técnica de traduzir o funcionamento interno das redes neurais. Assim, a caixa-preta não é apenas uma metáfora sobre incompreensão, mas uma barreira concreta à aplicação da lei e à garantia de transparência.

Frente a esse cenário, a superação da caixa-preta tornou-se um dos grandes desafios contemporâneos da Inteligência Artificial. Surge, então, o campo da Explainable Artificial Intelligence (XAI), cujo objetivo é desenvolver mecanismos que tornem os modelos mais compreensíveis, sem sacrificar seu desempenho. A explicabilidade propõe não apenas visualizar pesos ou correlações, mas também criar uma ponte cognitiva entre o raciocínio da máquina e a interpretação humana. Como sintetiza Ortigossa (2024), explicar significa traduzir o raciocínio matemático da IA em termos compreensíveis, possibilitando confiança, auditoria e uso ético da tecnologia [5]. Essa busca pela transparência redefine a própria relação entre humanos e algoritmos, ao transformar a caixa-preta de um obstáculo técnico em uma fronteira ética e científica a ser ultrapassada.

IV. Explicabilidade e Transparência

A crescente adoção de sistemas baseados em aprendizado profundo reacendeu o debate sobre a necessidade de compreender não apenas o resultado de um modelo, mas também o raciocínio que o sustenta. Essa preocupação deu origem ao campo da Explainable Artificial Intelligence (XAI), ou Inteligência Artificial Explicável, cujo propósito é permitir que humanos compreendam e validem decisões automatizadas. A XAI não busca apenas abrir o "interior" dos algoritmos, mas traduzir seus processos para uma linguagem compreensível, promovendo confiança, responsabilidade e uso ético da tecnologia. Segundo Ortigossa, explicar significa reduzir a distância cognitiva entre o raciocínio matemático

da máquina e a capacidade humana de interpretação [5]. A explicabilidade, portanto, é tanto um requisito técnico quanto uma demanda social, necessária para a aceitação segura e consciente da IA.

A explicabilidade difere da interpretabilidade, embora os termos sejam frequentemente utilizados como sinônimos. Modelos interpretáveis são aqueles cuja lógica é transparente por natureza, como regressões lineares, árvores de decisão e redes bayesianas. Já os métodos explicáveis atuam sobre modelos opacos, revelando como variáveis ou padrões específicos influenciaram uma decisão. Entre as abordagens mais difundidas estão o LIME (Local Interpretable Model-Agnostic Explanations), que gera explicações locais simplificadas a partir de perturbações dos dados de entrada, e o SHAP (SHapley Additive exPlanations), baseado na teoria dos valores de Shapley, que quantifica a contribuição de cada atributo para o resultado final do modelo [6]. Enquanto o LIME permite entender decisões individuais de forma aproximada, o SHAP fornece uma métrica consistente de importância global e local das variáveis, sendo amplamente utilizado em aplicações críticas, como medicina e finanças.

Outros métodos complementam essa busca pela transparência em modelos complexos. O Grad-CAM (Gradient-weighted Class Activation Mapping) é utilizado em redes neurais convolucionais para gerar mapas visuais que destacam as regiões de uma imagem que mais influenciaram a classificação. Já as explicações contrafactuais oferecem uma perspectiva causal: em vez de mostrar por que uma decisão foi tomada, indicam o que precisaria mudar para que o resultado fosse diferente. Tais abordagens tornam o processo decisório mais acessível e verificável, transformando o modelo de uma entidade opaca em um sistema parcialmente auditável. De acordo com Balasubramaniam et al. (2023), a aplicação combinada desses métodos amplia a transparência, ao fornecer explicações complementares sob diferentes perspectivas, estatística, visual e causal [4].

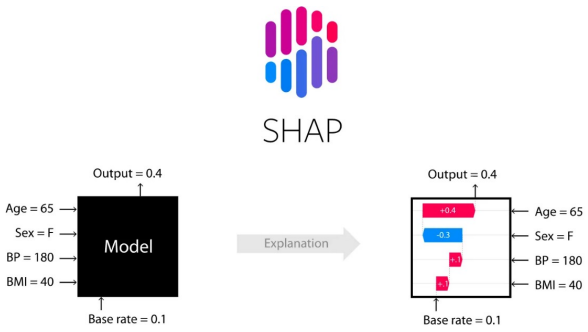

Fig. 3. Exemplo de visualização de um método de explicabilidade (XAI).

Outros métodos complementam essa busca pela transparência em modelos complexos. O Grad-CAM (Gradient-weighted Class Activation Mapping) é utilizado em redes neurais convolucionais para gerar mapas visuais que destacam as regiões de uma imagem que mais influenciaram a classificação. Já as explicações contrafactuais oferecem uma perspectiva causal: em vez de mostrar por que uma decisão foi tomada, indicam o que precisaria mudar para que o resultado fosse diferente. Tais abordagens tornam o processo decisório mais acessível e verificável, transformando o modelo de uma entidade opaca em um sistema parcialmente auditável. De acordo com Balasubramaniam et al. (2023), a aplicação combinada desses métodos amplia a transparência, ao fornecer explicações complementares sob diferentes perspectivas, estatística, visual e causal [4].

A transparência, nesse contexto, não se limita à abertura técnica do modelo, mas envolve também a comunicação clara e compreensível dos resultados. Um sistema transparente é aquele que permite rastrear as etapas de seu raciocínio, avaliar sua confiabilidade e questionar suas conclusões. Isso requer não apenas métodos computacionais, mas também o design de interfaces explicativas adequadas ao público-alvo, sejam cientistas de dados, profissionais da saúde ou usuários finais. Essa dimensão humana da transparência garante que a explicabilidade seja mais do que um requisito de engenharia: ela se torna uma prática de ética e responsabilidade tecnológica. A transparência, portanto, é a base para o desenvolvimento de sistemas de IA confiáveis, auditáveis e socialmente legítimos.

Em síntese, a explicabilidade e a transparência constituem pilares da Inteligência Artificial contemporânea. A XAI representa um esforço de convergência entre ciência, filosofia e engenharia, visando reduzir o hiato entre desempenho e compreensão. Explicar não é apenas descrever o funcionamento de um algoritmo, mas legitimar seu uso perante a sociedade. À medida que a IA assume papéis decisórios em contextos humanos críticos, compreender o "porquê" das respostas torna-se tão importante quanto a precisão com que são produzidas. Dessa forma, a explicabilidade não é o oposto da eficiência, é a condição para que ela seja eticamente aceitável.

V. Soluções e Caminhos para a Explicabilidade

A busca por soluções para o problema da opacidade algorítmica levou à consolidação de diferentes estratégias dentro do campo da Explainable Artificial Intelligence (XAI). Nenhum método isolado é capaz de garantir explicações completas e universais, mas o avanço da área demonstra um movimento conjunto entre ciência da computação, psicologia cognitiva e ética aplicada. O objetivo é que sistemas inteligentes possam justificar suas decisões de forma compreensível, mensurável e confiável. Segundo Ortigossa et al. (2024), o desenvolvimento da explicabilidade requer uma abordagem multidimensional que una engenharia, design e governança para equilibrar desempenho e inteligibilidade [5].

Do ponto de vista técnico, destacam-se duas grandes abordagens: a criação de modelos intrinsecamente interpretáveis e o uso de métodos pós-hoc. Os primeiros incluem algoritmos que já possuem estrutura explicável, como regressões lineares, árvores de decisão e redes baseadas em regras simbólicas. Já os métodos pós-hoc atuam sobre modelos complexos, gerando explicações aproximadas sem alterar o funcionamento interno do sistema. Entre esses, o SHAP e o LIME

continuam sendo as ferramentas mais difundidas, permitindo quantificar a importância de variáveis e simular cenários locais de decisão. Outras técnicas, como o Grad-CAM em redes convolucionais, e os modelos autoexplicativos (self-explaining models), ampliam a transparência ao incorporar mecanismos de visualização e ponderação interpretável [6].

Uma tendência emergente é a adoção de abordagens sensíveis ao contexto. Essas estratégias adaptam o formato da explicação ao domínio e ao perfil do usuário, tornando o entendimento mais eficaz. No mercado de trabalho, por exemplo, explicações contextuais podem ajudar candidatos e recrutadores a compreenderem por que determinados perfis são priorizados por sistemas de IA. Segundo Pinto et al. (2024), incorporar variáveis contextuais nas explicações aumenta a confiança, a satisfação e a compreensão dos usuários [8]. Essa personalização representa um avanço na direção de uma explicabilidade centrada no humano, que reconhece a diversidade de públicos e necessidades.

Além das soluções técnicas, é necessário considerar fatores humanos e organizacionais. A explicabilidade é mais efetiva quando desenvolvida por equipes multidisciplinares, que reúnem especialistas de áreas como ciência de dados, direito, psicologia e ética. Balasubramaniam et al. (2023) enfatizam que a definição de requisitos de explicabilidade deve ser colaborativa, envolvendo múltiplas perspectivas e experiências [4]. Abordagens de human-in-the-loop reforçam esse princípio, permitindo que humanos participem ativamente do ciclo de aprendizado, interpretando e validando decisões automatizadas. Essa cooperação contínua garante que o sentido das explicações permaneça alinhado aos valores humanos e aos princípios de justiça e transparência.

Por fim, os caminhos para a explicabilidade apontam para uma visão integrada da Inteligência Artificial: técnica, ética e social. A explicação não deve ser vista como um complemento opcional, mas como parte constitutiva do próprio sistema. O futuro da IA dependerá de sua capacidade de tornar-se compreensível, auditável e justa. O desafio é equilibrar a precisão matemática dos algoritmos com a clareza moral das decisões, garantindo que a inteligência artificial permaneça a serviço do humano e de seus direitos fundamentais.

VI. Dificuldade e Restrições na Compreensão

Ainda que promissora, a Interpretabilidade em IA (XAI) esbarra em grandes barreiras, pois a busca por clareza traz consigo novos problemas. Um dos maiores entraves é o dilema entre precisão e facilidade, já que descrições acessíveis costumam ser resumos que não refletem com profundidade o raciocínio do modelo, criando uma ilusão de confiança.

Além disso, há o perigo de direcionamento, onde as explicações podem ser usadas para explorar vulnerabilidades do sistema ou para fornecer justificativas superficiais que simulam conformidade ética, tática chamada de ”fairwashing” [9]. A eficácia da XAI também é afetada pela necessidade de conhecimento do usuário, já que uma descrição técnica pode não servir para quem não entende do assunto.

Por fim, a proteção e os direitos autorais colocam empecilhos, já que a clareza total pode revelar algoritmos privados e pontos fracos. Esses problemas mostram que o objetivo não deve ser a clareza total, mas sim uma explicação relevante, que equilibre clareza, exatidão e proteção de acordo com cada situação.

VII. Conclusão

A evolução da Inteligência Artificial trouxe ganhos expressivos em desempenho e automação, mas também evidenciou uma contradição fundamental: quanto mais poderosos se tornam os modelos, mais difíceis são de compreender. A opacidade das redes neurais e o fenômeno da caixa-preta expõem o desafio de traduzir processos matemáticos complexos em raciocínios acessíveis ao entendimento humano. Essa lacuna entre cálculo e sentido não é apenas técnica, mas ética e social, pois limita a capacidade de auditoria, responsabilização e confiança nos sistemas de IA.

A explicabilidade e a transparência emergem como respostas a esse problema. Elas representam o esforço conjunto de pesquisadores e legisladores em tornar a IA compreensível, auditável e justa. Ao permitir que usuários entendam as razões por trás de uma decisão algorítmica, a XAI transforma a relação entre humanos e máquinas, resgatando o princípio da responsabilidade compartilhada. Explicar não significa abrir completamente o código, mas comunicar o raciocínio do sistema de forma verificável e inteligível.

As soluções apresentadas pela XAI, como os métodos SHAP, LIME, Grad-CAM, explicações contrafactuais e modelos autoexplicativos, indicam caminhos promissores para a superação da opacidade. Ainda assim, a explicabilidade plena requer uma integração mais ampla entre engenharia, ética e design. A inclusão de contextos de uso e a colaboração entre especialistas de diferentes áreas reforçam que a transparência é um valor coletivo, e não apenas uma propriedade técnica.

Conclui-se, portanto, que a confiança em sistemas de IA depende diretamente de sua capacidade de se fazer entender. A verdadeira inteligência artificial do futuro não será apenas preditiva, mas também compreensível. Somente ao tornar visíveis seus processos internos, a IA poderá servir de forma ética, transparente e responsável à sociedade que a criou.

References

[1] Suave, André Augusto. *Inteligência Artificial*. Freitas Bastos, 2024.

[2] Brasil. *Lei Geral de Proteção de Dados Pessoais (LGPD)*, Lei nº 13.709, de 14 de agosto de 2018.

[3] Cheong, M. et al. “Transparency and Accountability in Artificial Intelligence Systems.” *Frontiers in Human Dynamics*, 2024.

[4] Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., Kujala, S. “Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements.” *Information and Software Technology*, 2023.

[5] Ortigossa, E. S., Gonçalves, T., Nonato, L. G. “Explainable Artificial Intelligence (XAI) – From Theory to Methods and Applications.” *IEEE Access*, 2024.

[6] Gunning, D., Aha, D. “DARPA’s Explainable Artificial Intelligence Program.” *AI Magazine*, 2019.

[7] Hulsen, T. “Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare.” *AI Journal*, 2023.

[8] Pinto, G. B. S., Mello, C. E. R., Garcia, A. C. B. “Explicações de Inteligência Artificial Conscientes de Contexto em Aplicações de Mercado de Trabalho.” *Revista Brasileira de Informática*, UNIRIO, 2024.

[9] Aïvodji, U., Gambs, S., Arai, H., and Hara, S., “Characterizing the Risk of Fairwashing,” Proc. 35th Conf. on Neural Information Processing Systems (NeurIPS), Sydney, Australia, 2021.