DA TRANSPARÊNCIA À CONFIABILIDADE: UMA PERSPECTIVA DA ENGENHARIA DE SISTEMAS PARA A INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL NA PRÁTICA CLÍNICA

FROM TRANSPARENCY TO RELIABILITY: A SYSTEMS ENGINEERING PERSPECTIVE TOWARDS EXPLAINABLE ARTIFICIAL INTELLIGENCE IN CLINICAL PRACTICE

João Santos [1]*

[1]Universidae Nove de Julho, Brasil

RESUMO

A incorporação de modelos de Inteligência Artificial (IA) na saúde promete revolucionar o diagnóstico e o tratamento, mas sua natureza como "caixa-preta" representa uma barreira crítica para a adoção clínica e a responsabilidade médica. A Inteligência Artificial Explicável (XAI) surge como uma resposta a essa barreira, buscando prover transparência às decisões algorítmicas. No entanto, a literatura corrente, embora rica em descrever métodos de XAI, ainda carece de uma abordagem estruturada sob a ótica da engenharia de sistemas. Este artigo avança na discussão ao propor um *framework* para o ciclo de vida de sistemas de XAI, argumentando que a meta não deve ser apenas a "explicabilidade", mas a construção de "sistemas de IA confiáveis" (*Trustworthy AI*). Analisamos a taxonomia dos métodos de XAI (intrínsecos e pós-hoc) sob a perspectiva de suas implicações de engenharia e propomos um ciclo de vida em quatro fases: (1) Governança de Dados e Mitigação de Viés; (2) Verificação e Validação das Explicações; (3) Integração Clínica e Fatores Humanos; e (4) Monitoramento Pós-Implementação e Deriva da Explicação. Discutimos os desafios e as fronteiras da pesquisa, com ênfase na causalidade, escalabilidade e nos marcos regulatórios pertinentes ao contexto brasileiro, como a Lei Geral de Proteção de Dados (LGPD). Concluímos que a transição da XAI do campo acadêmico para a prática clínica sustentável depende de um paradigma rigoroso de engenharia, focado na robustez, validação e confiabilidade do sistema como um todo.

Recebido: 11 de Janeiro de 2025

Aceito: 25 de Abril de 2025

* joaoanderson@gmail.com

Palavras chave: Inteligência Artificial Explicável, IA na Saúde, Engenharia de Sistemas, IA Confiável, Validação de Modelos, Aprendizado de Máquina.

ABSTRACT

The incorporation of Artificial Intelligence (AI) models into healthcare promises to revolutionize diagnosis and treatment, but their "black box" nature represents a critical barrier to clinical adoption and medical accountability. Explainable Artificial Intelligence (XAI) emerges as a response to this barrier, seeking to provide transparency to algorithmic decisions. However, the current literature, although rich in describing XAI methods, still lacks a structured approach from the perspective of systems engineering. This article advances the discussion by proposing a framework for the life cycle of XAI systems, arguing that the goal should not be only "explainability", but the construction of "trustworthy AI systems". We analyze the taxonomy of XAI methods (intrinsic and post-hoc) from the perspective of their engineering implications and propose a life cycle in four phases: (1) Data Governance and Bias Mitigation; (2) Verification and Validation of Explanations; (3) Clinical Integration and Human Factors; and (4) Post-Implementation Monitoring and Explanation Drift. We discuss the challenges and frontiers of research, with emphasis on causality, scalability, and regulatory frameworks relevant to the Brazilian context, such as the General Data Protection Law (LGPD). We conclude that the transition of XAI from academia to sustainable clinical practice depends on a rigorous engineering paradigm focused on the robustness, validation, and reliability of the system as a whole.

Keywords: Explainable Artificial Intelligence, AI in Healthcare, Systems Engineering, Trustworthy AI, Model Validation, Machine Learning.

1. INTRODUÇÃO

A ascensão do aprendizado de máquina (*machine learning*, ML) e, em particular, do aprendizado profundo (*deep learning*, DL), inaugurou uma nova era de possibilidades na área da saúde. Modelos algorítmicos demonstram performance comparável ou superior à de especialistas humanos em tarefas como a análise de imagens médicas, a predição de risco e o auxílio ao diagnóstico (Esteva et al., 2019; Rajkomar et al., 2019). No entanto, a crescente complexidade desses modelos, frequentemente envolvendo milhões de parâmetros, gerou um paradoxo: quanto mais potente o modelo, mais opaco seu processo decisório (Burrell, 2016; Lipton, 2018). Essa opacidade, ou o "problema da caixa-preta", não é apenas uma curiosidade acadêmica; é uma barreira fundamental à sua adoção clínica por razões éticas, legais e de segurança do paciente (Mittelstadt et al., 2019).

Em resposta, o campo da Inteligência Artificial Explicável (XAI) emergiu com o objetivo de tornar as decisões de modelos de IA compreensíveis para os usuários humanos (Arrieta et al., 2020). Revisões sistemáticas, como a de Bharati et al. (2022), categorizaram exaustivamente os métodos de XAI e suas aplicações, respondendo às

perguntas de "por que, como e quando" utilizá-los. Essa base é fundamental, mas, sob a ótica da engenharia, a explicabilidade é uma propriedade necessária, porém não suficiente. O verdadeiro desafio reside na engenharia de **sistemas de IA confiáveis** (*Trustworthy AI*), onde a explicação é apenas um dos componentes, ao lado de robustez, justiça (*fairness*), privacidade e governança (Falco & Shneiderman, 2023).

No contexto brasileiro, a implementação de tais sistemas em uma escala como a do Sistema Único de Saúde (SUS) amplifica esses desafios. A heterogeneidade dos dados, a necessidade de validação em populações diversas e a conformidade com a Lei Geral de Proteção de Dados (LGPD, Lei nº 13.709/2018) exigem uma abordagem que transcenda a simples aplicação de uma técnica de XAI a um modelo treinado.

Este artigo, portanto, propõe uma mudança de paradigma: da busca pela "explicação" para a engenharia de "confiabilidade". Argumentamos que a XAI deve ser tratada não como um passo final, mas como um processo integrado ao longo de todo o ciclo de vida do sistema de IA. Para tanto, apresentamos um *framework* conceitual de engenharia para o desenvolvimento, validação e monitoramento de sistemas de XAI na prática clínica, com o objetivo de fomentar uma discussão mais pragmática e alinhada às necessidades de implementação no mundo real.

2. Taxonomia de Engenharia dos Métodos de XAI

A literatura classifica os métodos de XAI de diversas formas. Do ponto de vista da engenharia de sistemas, a distinção mais funcional é entre abordagens *ante-hoc* (intrínsecas) e *pós-hoc*.

2.1 Modelos Intrinsecamente Interpretáveis (Ante-hoc)

Estes são modelos transparentes por design. Incluem algoritmos clássicos como regressão linear, árvores de decisão e sistemas baseados em regras (Rudin, 2019). A principal vantagem de engenharia é que a explicação é o próprio modelo; não há uma camada adicional de aproximação que precise ser validada. Por exemplo, uma árvore de decisão para predição de risco cardiovascular oferece um fluxo de regras explícito e auditável (Letham et al., 2015).

O *trade-off* fundamental, contudo, é entre interpretabilidade e performance. Para capturar as relações não-lineares complexas presentes em dados médicos de alta dimensão (genômica, imagens), esses modelos podem ser insuficientes, levando a uma perda de acurácia preditiva em comparação com abordagens de caixa-preta (Goodman &

Flaxman, 2017). A escolha por um modelo intrínseco é, portanto, uma decisão de projeto que deve ponderar os requisitos de performance e a criticidade da transparência absoluta.

2.2 Métodos de Explicação Pós-hoc

Estes métodos são aplicados a modelos de caixa-preta já treinados, funcionando como uma camada de análise para explicar predições individuais ou o comportamento global do modelo (Molnar, 2020). São a força motriz da XAI moderna, pois permitem o uso de modelos de alta performance (e.g., redes neurais profundas) sem sacrificar completamente a interpretabilidade. Do ponto de vista da engenharia, eles se dividem em duas classes principais.

2.2.1 Métodos Agnósticos ao Modelo. Estes métodos tratam o modelo de IA como uma caixa-preta, analisando apenas as relações entre entrada e saída. Sua grande vantagem de engenharia é a flexibilidade, pois podem ser aplicados a qualquer tipo de algoritmo. Os exemplos mais proeminentes são (i) **LIME (Local Interpretable Model-agnostic Explanations)**, que explica uma predição individual ao treinar um modelo interpretável mais simples (e.g., regressão linear) em uma vizinhança local da instância de interesse (Ribeiro et al., 2016), e que, apesar de intuitivo, apresenta desafios de engenharia relacionados à instabilidade das explicações e à definição do que constitui uma "vizinhança local" (Alvarez-Melis & Jaakkola, 2018); e (ii) **SHAP (SHapley Additive exPlanations)**, que, baseado na teoria dos jogos cooperativos, atribui a cada característica de entrada (*feature*) um valor que representa sua contribuição para a predição, garantindo consistência teórica (Lundberg & Lee, 2017). O SHAP tornou-se um padrão na indústria, mas seu custo computacional pode ser proibitivo para modelos muito complexos ou para cenários que exigem explicações em tempo real, um requisito comum em sistemas de apoio à decisão clínica (Strumbelj & Kononenko, 2014).

2.2.2 Métodos Específicos ao Modelo. Diferentemente dos agnósticos, estes métodos são projetados para classes específicas de modelos (majoritariamente redes neurais), aproveitando sua arquitetura interna para gerar explicações. Sua vantagem de engenharia é a potencial maior fidelidade ao comportamento do modelo. Exemplos incluem (i) **Mapas de Ativação (Grad-CAM)**, que, utilizado em Redes Neurais Convolucionais (CNNs), usa os gradientes da última camada convolucional para produzir um mapa de calor que destaca as regiões da imagem de entrada mais importantes para a decisão (Selvaraju et al., 2017), sendo visualmente poderoso para validar se um modelo de radiologia está, de fato, "olhando" para a patologia correta; e (ii) **Propagação de**

Relevância (LRP), que decompõe a predição de uma rede neural, redistribuindo-a em cascata reversa até a camada de entrada para indicar a contribuição de cada pixel ou *feature* (Bach et al., 2015). A escolha entre métodos pós-hoc agnósticos e específicos é uma decisão de arquitetura. Métodos agnósticos oferecem flexibilidade para experimentação com diferentes modelos de IA, enquanto métodos específicos podem fornecer explicações mais fiéis, mas atrelam a solução de XAI a uma arquitetura de modelo particular.

3. O CICLO DE VIDA DA ENGENHARIA DE SISTEMAS XAI

A implementação de XAI na prática clínica não pode ser um *afterthought*. Propomos que ela seja integrada em um ciclo de vida de engenharia rigoroso, composto por quatro fases interconectadas.

Fase 1: Governança de Dados e Mitigação de Viés

A confiança em um sistema de IA começa com a confiança nos dados. Explicações geradas a partir de dados com viés (*bias*) não são apenas inúteis, mas perigosas, pois podem criar uma falsa sensação de segurança em uma decisão fundamentalmente falha (Obermeyer et al., 2019). A engenharia de um sistema XAI deve, portanto, começar com (i) uma análise de viés, auditando sistematicamente os dados de treinamento para identificar e mitigar vieses demográficos, socioeconômicos ou de subpopulações; (ii) a documentação rigorosa da proveniência, coleta, limpeza e limitações dos dados através de *Data Sheets for Datasets*, seguindo frameworks como o proposto por Gebru et al. (2021); e (iii) a garantia de que todo o processo de manipulação de dados esteja em conformidade com a LGPD, especialmente no que tange ao tratamento de dados sensíveis de saúde.

Fase 2: Verificação e Validação das Explicações (V&V)

Uma vez que um método de XAI é implementado, como podemos confiar na explicação que ele gera? Esta é uma questão central de engenharia. A fase de V&V deve ir além da simples inspeção visual e incorporar métricas quantitativas, como (i) **Fidelidade (*Fidelity*)**, que mede o quão bem a explicação se aproxima do comportamento do modelo de caixa-preta; (ii) **Robustez (*Robustness*)**, que avalia se a explicação se mantém estável para pequenas perturbações na entrada; e (iii) **Consistência (*Consistency*)**, que verifica se modelos funcionalmente equivalentes produzem

explicações semelhantes para a mesma entrada (Doshi-Velez & Kim, 2017). A validação de explicações é um campo ativo de pesquisa, mas é um passo de engenharia não negociável para sistemas de missão crítica.

Fase 3: Integração Clínica e Fatores Humanos

Uma explicação tecnicamente perfeita é inútil se não for compreensível e acionável pelo usuário final — o profissional de saúde. Esta fase envolve a colaboração estreita com médicos, enfermeiros e outros especialistas para projetar a interface homem-máquina (Tonekaboni et al., 2019). Os desafios de engenharia incluem (i) o design da interface, questionando como apresentar a explicação de forma a não sobrecarregar o usuário e se integrar ao fluxo de trabalho clínico; (ii) a mitigação do viés de automação, projetando o sistema para reduzir o risco de que os médicos confiem excessivamente na recomendação da IA (Goddard et al., 2012); e (iii) o desenvolvimento de programas de treinamento e alfabetização em IA para que os usuários finais compreendam as capacidades e as limitações do sistema.

Fase 4: Monitoramento Pós-Implementação e Deriva da Explicação

O lançamento de um sistema de XAI não é o fim do ciclo de vida. Modelos de IA degradam com o tempo devido à "deriva de conceito" (*concept drift*), quando a distribuição estatística dos dados do mundo real muda (Widmer & Kubat, 1996). Isso implica que não apenas a acurácia do modelo pode cair, mas também a validade de suas explicações. A engenharia de monitoramento deve prever (i) o monitoramento contínuo da performance do modelo; (ii) a detecção de mudanças nas distribuições dos dados de entrada; e (iii) o monitoramento da "deriva da explicação" (*explanation drift*), um conceito que propomos para descrever a monitorização da estabilidade e fidelidade das explicações ao longo do tempo como um alerta precoce de que o modelo não está mais se comportando como esperado.

4. DESAFIOS E FRONTEIRAS DA PESQUISA

A operacionalização deste ciclo de vida enfrenta desafios significativos que definem as fronteiras da pesquisa em XAI aplicada. Tais desafios incluem (i) a distinção entre causalidade e correlação, visto que a maioria dos métodos atuais de XAI identifica correlações, sendo a integração com inferência causal a próxima fronteira para explicações mais robustas (Pearl, 2019); (ii) a escalabilidade e o custo computacional,

pois a implementação em larga escala, como no SUS, exige métodos de XAI computacionalmente eficientes; (iii) a regulamentação, já que agências como a ANVISA estão desenvolvendo diretrizes para *software as a medical device* (SaMD), e a validação rigorosa das explicações será um requisito central (Benjamens et al., 2020); e (iv) a necessidade de equipes interdisciplinares, pois o sucesso da engenharia de sistemas XAI depende da colaboração entre engenheiros, cientistas de dados, médicos, eticistas e reguladores.

5. CONCLUSÃO

A Inteligência Artificial Explicável (XAI) é indispensável para destravar o potencial da IA na saúde. Contudo, para mover a XAI da teoria para a prática clínica diária, especialmente em contextos complexos e de larga escala como o brasileiro, é preciso adotar uma rigorosa perspectiva de engenharia de sistemas. O foco deve ser ampliado da "transparência" de uma única predição para a "confiabilidade" de todo o sistema ao longo de seu ciclo de vida.

Propusemos um *framework* de engenharia em quatro fases — governança de dados, V&V das explicações, integração clínica e monitoramento pós-implementação — como um roteiro para o desenvolvimento de sistemas de IA confiáveis. Acreditamos que a adoção de tal abordagem estruturada é o caminho para garantir que as soluções de IA na saúde sejam não apenas inteligentes e precisas, mas também seguras, justas e verdadeiramente úteis para médicos e pacientes. O desafio para nós, engenheiros e acadêmicos, é construir pontes entre o potencial algorítmico e a realidade clínica, transformando a promessa da XAI em um benefício tangível e confiável para a sociedade.

REFERÊNCIAS

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahuja, S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ, 7*, e7702. https://doi.org/10.7717/peerj.7702

Alvarez-Melis, D., & Jaakkola, T. S. (2018). *On the robustness of interpretability methods*. ArXiv. https://arxiv.org/abs/1806.08049

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 1-9. https://doi.org/10.1186/s12911-020-01332-6

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE, 10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digital Medicine, 3*(1), 118. https://doi.org/10.1038/s41746-020-00324-0

Bharati, S., Mondal, M. R. H., & Podder, P. (2022). A review on explainable artificial intelligence for healthcare: Why, how, and when? *IEEE Reviews in Biomedical Engineering*.

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721-1730). Association for Computing Machinery. https://doi.org/10.1145/2783258.2788613

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. ArXiv. https://arxiv.org/abs/1702.08608

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24-29. https://doi.org/10.1038/s41591-018-0316-z

Falco, G., & Shneiderman, B. (2023). A framework for trustworthy AI systems. *Communications of the ACM, 66*(3), 22-25. https://doi.org/10.1145/3582426

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92. https://doi.org/10.1145/3458723

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121-127. https://doi.org/10.1136/amiajnl-2011-000271

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, 38*(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics, 9*(3), 1350-1371. https://doi.org/10.1214/15-AOAS848

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31-57. https://doi.org/10.1145/3236386.3241340

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems 30* (pp. 4765-4774).

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287574

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447-453. https://doi.org/10.1126/science.aax2342

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62*(3), 54-60. https://doi.org/10.1145/3241036

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).

Strumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems, 41*(3), 647-665. https://doi.org/10.1007/s10115-013-0679-x

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference* (pp. 359-380). PMLR.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine learning, 23*(1), 69-101. https://doi.org/10.1007/BF00116900