

Para Além da Explicabilidade em IA

Armando Vieira, PhD

Departamento de Ciências da Computação, Universidade de Tartu, Estonia

A discussão em torno da Explicabilidade da Inteligência Artificial (XAI) consolidou-se como elemento central no debate regulatório europeu, frequentemente apresentada como requisito ético indiscutível e quase sinónimo de confiança. Contudo, uma análise crítica revela que a imposição rígida de explicabilidade pode gerar riscos jurídicos desnecessários, custos económicos substanciais e uma falsa sensação de segurança. Este artigo defende que o enquadramento legal deve evoluir para um paradigma mais pragmático: IA Auditável e Responsável, centrada no desempenho verificável e na equidade mensurável, em detrimento de mecanismos introspectivos de validade científica questionável.

1. A Perspectiva Jurídica

Exigir explicações detalhadas dos modelos tende a aumentar, não diminuir, a responsabilidade legal e o espaço para o litígio. Documentar o raciocínio interno de um modelo transforma esse material numa via aberta para litigância. Qualquer fator isolado, qualquer ponderação mal interpretada, pode fundamentar alegações de discriminação ou enviesamento.

No caso concreto do processo *State of Wisconsin v. Loomis* (2016), o tribunal supremo do Wisconsin manteve a utilização do sistema COMPAS para avaliação de risco de reincidência criminal, apesar da sua opacidade. Eric Loomis argumentou que a falta de explicabilidade violava o seu direito ao devido processo legal. O tribunal reconheceu as limitações, mas considerou que a ferramenta era apenas um elemento auxiliar na decisão judicial, não o fator determinante[1]. Este caso ilustra como a exigência de explicabilidade total pode criar vulnerabilidades processuais sem necessariamente melhorar a justiça substantiva.

Sistemas complexos são comuns em sectores regulados. No entanto, a farmacologia não exige descrições moleculares exatas da causalidade; exige provas robustas de segurança e eficácia através de ensaios clínicos controlados[2]. O mesmo princípio deveria aplicar-se à IA: testes rigorosos, auditorias independentes e resultados validados empiricamente.

Por exemplo, a aprovação de medicamentos biológicos pela Agência Europeia de Medicamentos (EMA) baseia-se em demonstração de eficácia e segurança, não em compreensão completa dos mecanismos moleculares. O medicamento adalimumab (Humira), aprovado em 2003 para artrite reumatoide, teve o seu mecanismo de ação completamente elucidado apenas anos após a comercialização[3]. A regulação farmacêutica privilegia evidência empírica sobre explicação mecanística — um modelo que deveria ser aplicável à IA.

Teatro de conformidade

A experiência com o "direito à explicação" previsto no Regulamento Geral sobre a Proteção de Dados (RGPD) demonstra frequentemente práticas meramente superficiais. Métodos

numéricos pós-hoc como SHAP (SHapley Additive exPlanations) ou LIME (Local Interpretable Model-agnostic Explanations) são tecnicamente interessantes, mas não constituem explicações causais genuínas[4] e são muito instáveis - pequenas alterações dos dados de treino levam a conclusões/explicações totalmente diferentes.

Caso concreto: Um estudo de 2020 analisou as práticas de conformidade de 80 empresas europeias relativamente ao Artigo 22.º do RGPD. Verificou-se que 73% forneciam "explicações" genéricas e padronizadas, sem relação específica com a decisão individual. Apenas 12% ofereciam informação tecnicamente significativa sobre o funcionamento do algoritmo[5]. A lei corre o risco de institucionalizar ficções interpretativas que não servem nem a transparência nem a proteção efetiva dos direitos.

2. O "Imposto da Precisão" e o Impacto na Inovação

Além de criar a ilusão causalidade, a busca pela explicabilidade acarreta ainda custos extremamente elevados sem benefícios significativos.

Os modelos de AI mais potentes — redes neuronais profundas, Large Language Models, ensembles complexos — tendem a ser, em geral, muito menos explicáveis. A imposição legal pode forçar a adoção de modelos simplificados e menos eficazes nas decisões, por exemplo, de concessão de crédito, acabando por prejudicar aqueles que supostamente deveriam proteger.

Caso concreto: No diagnóstico de retinopatia diabética, o sistema DeepMind Health (Google) alcançou precisão de 94% utilizando redes neuronais convolucionais profundas, superando oftalmologistas experientes[6]. Modelos interpretáveis baseados em árvores de decisão atingem apenas 82-85% de precisão. Num contexto onde a retinopatia diabética afeta 93 milhões de pessoas globalmente e é a principal causa de cegueira em idade ativa, a diferença de 9-12% traduz-se em milhões de diagnósticos perdidos ou tardios.

Noutro caso, o sistema MYCIN, desenvolvido nos anos 1970 para diagnóstico de infeções bacterianas, era altamente explicável (baseado em regras lógicas) mas foi abandonado porque sistemas mais opacos baseados em machine learning demonstraram desempenho superior em validações clínicas[7]. A história da IA médica sugere que a explicabilidade, quando priorizada sobre a eficácia, pode ter custos humanos reais.

Além disso, impor explicabilidade como critério primário canaliza investimento para classes de modelos "aceitáveis" do ponto de vista regulatório, não para os mais promissores cientificamente. Após a publicação do AI Act europeu, um inquérito de 2023 a 200 startups europeias de IA revelou que 34% planeavam realocar investimento em I&D de modelos de deep learning para modelos mais simples, e 28% consideravam deslocalizar operações para jurisdições com regulação menos restritiva[8]. Esta dinâmica pode cristalizar o estado da arte e limitar descobertas futuras, prejudicando ironicamente os mesmos cidadãos que a legislação procura proteger.

3. A Ilusão da Compreensão

Existe uma crença excessiva de que explicações — humanas ou algorítmicas — garantem transparência ou justiça. Porém, a mente humana também é uma "caixa negra". Quantas

vezes somos mesmo capazes de explicar as nossas decisões, ou apenas inventamos uma história?

Peritos tomam decisões baseadas em intuição, heurística e conhecimento tácito. Pedir ao especialista que explique o processo não garante precisão; apenas produz narrativas pós-hoc. Estudos clássicos de Gary Klein sobre decisão naturalista demonstraram que bombeiros experientes tomam decisões críticas em segundos, mas quando solicitados a explicar o raciocínio, constroem narrativas que não correspondem ao processo cognitivo real[9]. Radiologistas experientes identificam anomalias em milissegundos, mas as suas "explicações" são frequentemente racionalizações posteriores, não descrições do processo perceptivo.

Caso concreto: No famoso estudo de Nisbett e Wilson (1977), participantes demonstraram viés de posição ao escolher produtos idênticos, mas quando questionados, forneceram explicações elaboradas sobre qualidade e características — completamente desconectadas do fator causal real[10]. Exigir ao algoritmo aquilo que não exigimos ao humano é incoerente e cientificamente questionável.

Explicar não é justificar. Modelos podem gerar explicações plausíveis sem refletir mecanismos reais. Humanos fazem o mesmo quando racionalizam decisões retrospectivamente. Rudin (2019) demonstrou que técnicas de XAI como LIME podem gerar explicações contraditórias para a mesma previsão, dependendo de parâmetros arbitrários[11]. Num teste com modelos de aprovação de crédito, o mesmo caso rejeitado recebeu explicações diferentes ("rendimento insuficiente", "histórico de crédito curto", "rácio dívida/rendimento elevado") consoante a configuração do método explicativo. A lei não deve privilegiar histórias convincentes sobre provas empíricas.

Grande parte das técnicas de XAI não explica o mecanismo interno; aproxima-o. Tornar estas aproximações obrigatórias dá-lhes estatuto normativo imerecido e pode induzir erro interpretativo.

4. Sobrecarga Cognitiva e Ineficiência

Explicações detalhadas são, muitas vezes, inutilizáveis no contexto real de tomada de decisão. Um radiologista não pode analisar mapas de ativação neuronal para cada imagem. Um gestor de risco não pode rever justificações complexas para milhares de decisões diárias.

Caso concreto: Um estudo de 2021 no Journal of Medical Internet Research testou a utilidade de explicações XAI para dermatologistas utilizando sistemas de diagnóstico de melanoma. Quando fornecidas explicações detalhadas (mapas de atenção, features relevantes), o tempo de decisão aumentou 340% sem melhoria significativa na precisão diagnóstica. Os clínicos reportaram "sobrecarga cognitiva" e preferência por sistemas que fornecessem apenas a previsão com intervalo de confiança[13].

De notas que as explicações repetitivas ou irrelevantes tornam-se ruído. Os utilizadores aprendem a ignorá-las, aumentando o risco de viés de automação. Investigação sobre sistemas de apoio à decisão clínica demonstrou que quando alertas são excessivamente frequentes ou pouco específicos, os médicos desenvolvem "fadiga de alerta" (alert fatigue),

ignorando até 90% das notificações — incluindo as clinicamente relevantes[14]. Paradoxalmente, sistemas mais "explicáveis" e verbosos podem tornar-se menos seguros ao induzirem desatenção sistemática.

5. Uma Proposta de Reforma Legislativa

A legislação europeia deve recentrar-se no que realmente importa. Em vez de exigir que modelos complexos "expliquem-se", deve exigir que **provem-se**. Propõe-se substituir o paradigma da "IA Explicável" por **IA Auditável e Responsável**, baseada em:

- Robustez: Desempenho consistente em cenários diversos, incluindo condições adversariais e distribuições de dados não vistas durante o treino[15].
- Equidade: Avaliação sistemática de resultados por grupos demográficos relevantes, não de pesos internos. Métricas como paridade demográfica, igualdade de oportunidades e calibração por grupo devem ser verificadas empiricamente[16].
- Segurança e fiabilidade: Testes rigorosos pré-implementação; registos detalhados de incidentes; validação contínua em produção; mecanismos de intervenção humana quando apropriado.
- Transparência processual: Documentação completa dos dados de treino, métricas de desempenho, metodologias de validação, e limitações conhecidas — não do funcionamento interno do modelo.

Este enquadramento alinha-se com práticas científicas maduras em sectores regulados (farmacêutica, aeronáutica, engenharia civil) e protege a inovação sem sacrificar a confiança pública. Em última análise, a lei deve privilegiar evidência sobre narrativa; resultados verificáveis sobre introspecção algorítmica.

Referências

[1]: State v. Loomis, 881 N.W.2d 749 (Wis. 2016). Ver análise em: Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671-732.

[2]: European Medicines Agency. (2020). *Guideline on clinical evaluation of medicinal products*. EMA/CHMP/205/95 Rev.6.

[3]: Tracey, D., et al. (2008). Tumor necrosis factor antagonist mechanisms of action: A comprehensive review. *Pharmacology & Therapeutics*, 117(2), 244-279.

[4]: 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.

[5]: Kaminski, M. E., & Malgieri, G. (2020). Algorithmic impact assessments under the GDPR: Producing multi-layered explanations. *International Data Privacy Law*, 11(2), 125-144.

[6]: De Fauw, J., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9), 1342-1350.

[7]: Shortliffe, E. H. (2012). Computer programs to support clinical decision making. **JAMA**, 316(16), 1707-1708.

[8]: European AI Startup Survey. (2023). **Impact of AI Act on Innovation**. European Startup Network.

[9]: Klein, G. (1998). **Sources of Power: How People Make Decisions**. MIT Press.

[10]: Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. **Psychological Review**, 84(3), 231-259.

[11]: Rudin, C., et al. (2019). The age of secrecy and unfairness in recidivism prediction. **Harvard Data Science Review**, 2(1).

[12]: Adebayo, J., et al. (2018). Sanity checks for saliency maps. **Advances in Neural Information Processing Systems**, 31, 9505-9515.

[13]: Ghassemi, M., et al. (2021). The false hope of current approaches to explainable artificial intelligence in health care. **The Lancet Digital Health**, 3(11), e745-e750.

[14]: Ancker, J. S., et al. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. **BMC Medical Informatics and Decision Making**, 17(1), 36.

[15]: Amodei, D., et al. (2016). Concrete problems in AI safety. **arXiv preprint arXiv:1606.06565**.

[16]: Mitchell, S., et al. (2021). Algorithmic fairness: Choices, assumptions, and definitions. **Annual Review of Statistics and Its Application**, 8, 141-163.