Revista Ibero-Americana de Humanidades, Ciências e Educação — REASE

OPEN ACCESS

doi.org/10.51891/rease.v12i1.23937

INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI) EM SAÚDE DIGITAL: TENSÕES ENTRE TRANSPARÊNCIA ALGORÍTMICA, PROPRIEDADE INTELECTUAL E ÉTICA DO CUIDADO

Maryane Francisca Araujo de Freitas Cavalcante[1]
João Gabriel Freitas Cavalcante[2]
Aryadynna Santos Feitosa[3]
Erimar Pereira da Rocha[4]
Maria Raimunda D'Jesus Neta[5]
Lucileide Aquino do Nascimento[6]
Leonilson Neri dos Reis[7]
Francisco das Chagas Batista Santos[8]
Maisa Barbosa Santos[9]
Francisco das Chagas dos Santos Aguiar[10]
Láila Raila Leal Dias[11]

RESUMO: A incorporação da inteligência artificial (IA) na saúde digital tem ampliado a eficiência e a capacidade analítica dos sistemas de cuidado, mas também intensificado desafios éticos, jurídicos e sociais associados à opacidade algorítmica. Este artigo tem como objetivo analisar criticamente as tensões entre transparência algorítmica, propriedade intelectual e ética do cuidado no contexto da Inteligência Artificial Explicável (XAI) aplicada à saúde digital. Trata-se de um ensaio teórico crítico, fundamentado em revisão integrativa da literatura científica e análise documental de marcos regulatórios nacionais e internacionais publicados entre 2021 e 2026. A abordagem adotada é interdisciplinar, articulando contribuições da saúde coletiva, da bioética, do direito e dos estudos em ciência e tecnologia. Os resultados indicam que a XAI é condição central para a confiança clínica, a segurança do paciente e a autonomia profissional, especialmente em aplicações de alto risco, ao mesmo tempo em que evidencia conflitos estruturais com regimes de proteção da propriedade intelectual, notadamente o uso de segredos comerciais. Conclui-se que a adoção de mecanismos intermediários de governança, como transparência em camadas, auditorias independentes e supervisão humana significativa, é essencial para conciliar inovação tecnológica, proteção de direitos fundamentais e integridade ética do cuidado em saúde.

Palavras-chave: Inteligência Artificial Explicável. Saúde Digital. Transparência Algorítmica. Propriedade Intelectual. Ética do Cuidado. Tomada de Decisão Automatizada. Regulação. Segurança do Paciente.

[1] Mestranda em Propriedade Intelectual – Instituto Federal do Piauí (IFPI).
[2] Graduando Bacharelado em Ciências da Computação – Universidade Federal do Piauí (UFPI).
[3] Mestranda em Propriedade Intelectual – Instituto Federal do Piauí (IFPI).
[4] Mestrando em Propriedade Intelectual – Instituto Federal do Piauí (IFPI).
[5] Mestranda em Propriedade Intelectual – Instituto Federal do Piauí (IFPI).
[6] Mestranda em Propriedade Intelectual – Instituto Federal do Piauí (IFPI).
[7] Mestrando em Propriedade Intelectual – Instituto Federal do Piauí (IFPI).
[8] Mestrando em Propriedade Intelectual – Instituto Federal do Piauí (IFPI).
[9] Mestranda em Propriedade Intelectual – Instituto Federal do Piauí (IFPI).
[10] Especialista em Metodologias de Ensino - Falculdade Evangélica do Meio Norte (FAEME).
[11] Mestre em Engenharia de Materiais – Instituto Federal do Piauí.

1. INTRODUÇÃO

A incorporação da inteligência artificial (IA) na saúde digital tem promovido transformações estruturais nos processos de diagnóstico, monitoramento, apoio à decisão clínica e organização do cuidado. Sistemas de apoio à decisão, aplicativos de saúde mental, algoritmos preditivos de risco e ferramentas automatizadas de triagem passaram a desempenhar papel estratégico nos sistemas de saúde, influenciando diretamente as condutas profissionais e as trajetórias terapêuticas dos pacientes. Esses avanços reforçam a promessa de maior eficiência, precisão e personalização do cuidado, ampliando a capacidade de resposta dos serviços de saúde e o potencial de qualificação das práticas assistenciais.

Entretanto, tais benefícios coexistem com desafios éticos, jurídicos e sociais relevantes, especialmente relacionados à opacidade dos algoritmos, à responsabilização por decisões automatizadas e à proteção de direitos fundamentais. Embora a IA ofereça oportunidades para otimizar planos terapêuticos, aprimorar a gestão dos sistemas de saúde e fortalecer o autocuidado dos pacientes, sua adoção exige marcos regulatórios e éticos robustos, capazes de assegurar transparência, segurança e centralidade do cuidado humano nos processos decisórios mediados por tecnologia (Organização Mundial da Saúde, 2021).

Grande parte dos sistemas contemporâneos de inteligência artificial, sobretudo aqueles baseados em aprendizado profundo, opera por meio de modelos de difícil interpretação, caracterizados como *black boxes*. Essa opacidade desafia princípios fundamentais da prática em saúde, como a autonomia do paciente, o consentimento informado e a segurança clínica. Além disso, compromete a possibilidade de questionamento e responsabilização por decisões automatizadas, uma vez que a lógica decisória subjacente frequentemente não é rastreável nem mesmo por seus desenvolvedores (Amann *et al.*, 2020).

Nesse contexto, emerge a Inteligência Artificial Explicável (*Explainable Artificial Intelligence – XAI*) como resposta ética, técnica e regulatória a tais limitações. A XAI propõe métodos e modelos capazes de tornar compreensíveis os processos decisórios algorítmicos, especialmente para profissionais de saúde e pacientes. Com isso, ela distingue-se dos sistemas baseados em regras do passado ao buscar explicabilidade em arquiteturas complexas, sem comprometer o desempenho analítico característico do aprendizado profundo (Amann *et al.*, 2020; Núcleo de Informação e Coordenação do Ponto BR, 2024).

Entretanto, a exigência de transparência e explicabilidade na saúde digital insere-se em um contexto normativo e econômico marcado pela proteção da propriedade intelectual,

especialmente pelo uso de segredos comerciais e industriais para resguardar algoritmos, modelos e bases de dados. Essa configuração produz uma tensão estrutural, na qual a demanda ética e regulatória por explicações adequadas para decisões que impactam a saúde e a vida dos indivíduos confronta-se com os mecanismos jurídicos destinados a incentivar a inovação tecnológica (Organização Mundial da Saúde, 2021).

Esse dilema expressa o conflito entre o interesse público na compreensão, contestação e responsabilização de decisões automatizadas e o interesse privado na preservação de ativos tecnológicos estratégicos, fundamentais para a competitividade e o retorno econômico. Assim, a saúde digital torna visível a necessidade de soluções regulatórias intermediárias, capazes de equilibrar transparência, proteção da inovação e compromisso ético com o cuidado.

Nessa perspectiva, este artigo tem como objetivo analisar criticamente as tensões entre transparência algorítmica, propriedade intelectual e ética do cuidado no contexto da inteligência artificial explicável em saúde digital. Busca-se compreender como distintos marcos teóricos e regulatórios enquadram e respondem a essa problemática. Ademais, discute-se a construção de caminhos de governança da IA capazes de conciliar a preservação da inovação tecnológica com a proteção dos direitos dos pacientes e a integridade do cuidado em saúde.

2. REFERENCIAL TEÓRICO

A XAI emerge como resposta às limitações éticas, técnicas e sociais dos sistemas de IA opacos, particularmente aqueles baseados em modelos complexos de aprendizado profundo. No campo da saúde, a explicabilidade ultrapassa a compreensão técnica do algoritmo, exigindo a produção de justificativas clinicamente relevantes, contextualizadas e passíveis de comunicação sobre decisões e recomendações automatizadas (Núcleo de Informação e Coordenação do Ponto BR, 2024).

Nesse contexto, a opacidade das chamadas *black boxes* configura não apenas um desafio tecnológico, mas também uma barreira ética ao exercício da autonomia do paciente, ao consentimento informado e à segurança clínica. Diante desse cenário, a XAI pressupõe a adoção de níveis diferenciados de explicação, capazes de tornar os processos decisórios algorítmicos compreensíveis. Esses níveis devem ser adequados a pessoas públicas como profissionais de saúde, pacientes, gestores e reguladores; como condição para decisões responsáveis e eticamente sustentáveis no cuidado em saúde (Amann *et al.*, 2020).

Para Organização Mundial da Saúde (2021), do ponto de vista ético, a XAI articula-se diretamente aos princípios da bioética ao exigir que decisões automatizadas em saúde sejam

compreensíveis e justificáveis. A autonomia pressupõe que os pacientes compreendam, ao menos em termos gerais, como são produzidas as decisões que afetam sua saúde. Por sua vez, os princípios da beneficência e da não maleficência demandam a identificação e a mitigação de riscos associados a vieses algorítmicos, erros de classificação e generalizações inadequadas.

Nesse sentido, a XAI não se configura apenas como um requisito técnico, mas como um pressuposto ético indispensável para a adoção responsável da IA na prática clínica. Ao contribuir para a prevenção de discriminações estruturais e para a proteção da dignidade humana e da segurança do paciente, a explicabilidade torna-se condição central para alinhar inovação tecnológica aos valores fundamentais do cuidado em saúde.

Diante disso, a ética do cuidado amplia o debate sobre o uso da inteligência artificial na saúde ao enfatizar as dimensões relacionais, contextuais e de vulnerabilidade que caracterizam a prática clínica. Diferentemente de abordagens estritamente normativas, essa perspectiva reconhece que decisões em saúde não se reduzem a atos técnicos, mas constituem práticas situadas, sustentadas por confiança, responsabilidade e atenção às singularidades dos sujeitos (Amann *et al*., 2020; Organização Mundial da Saúde, 2021).

Nesse sentido, a introdução de sistemas de IA opacos tende a fragilizar a relação de cuidado ao deslocar a autoridade decisória para processos automatizados de difícil compreensão, enfraquecendo a confiança e a corresponsabilidade entre profissionais e pacientes. Esse cenário reforça a explicabilidade como exigência ética fundamental para preservar a centralidade da relação humano-humano e garantir um cuidado responsável, sensível às necessidades e vulnerabilidades do paciente (Núcleo de Informação e Coordenação do Ponto BR, 2024).

Em contraposição, os regimes de propriedade intelectual, exercem um papel central e ambivalente no desenvolvimento da inteligência artificial aplicada à saúde digital. Diferentemente das patentes, que pressupõem a divulgação pública da invenção em troca de exclusividade temporária, os segredos comerciais permitem a proteção indefinida de algoritmos e bases de dados confidenciais, desde que mantidos em sigilo. Essa estratégia configura uma forma de opacidade intencional voltada à preservação de vantagens competitivas, mas que limita o acesso externo a informações essenciais para a transparência, a avaliação crítica e a responsabilização dos sistemas(Denis *et al*., 2021).

Para enfrentar os desafios éticos associados ao uso da inteligência artificial em saúde, a Organização Mundial da Saúde propõe seis princípios orientadores: proteção da autonomia humana, promoção do bem-estar e da segurança, garantia de transparência e explicabilidade, fortalecimento da responsabilidade, promoção da inclusão e da equidade e estímulo a uma IA

sustentável. Esses princípios consolidam a explicabilidade como eixo central da governança ética, especialmente em contextos nos quais decisões automatizadas impactam diretamente a vida e a saúde das pessoas (Organização Mundial da Saúde, 2021; Núcleo de Informação e Coordenação do Ponto BR, 2024).

No contexto brasileiro, a regulação da inteligência artificial encontra-se em fase de transição, evoluindo de diretrizes genéricas para um modelo de governança baseado em risco e transparência qualificada. Esse processo está ancorado na Lei Geral de Proteção de Dados e em iniciativas legislativas recentes, como o Projeto de Lei nº 2338/2023, que prevê o direito à explicação sobre os critérios que orientam o funcionamento dos sistemas, especialmente em aplicações de alto risco. Essas iniciativas buscam equilibrar a promoção da inovação tecnológica com a proteção de direitos fundamentais, sobretudo em aplicações de alto risco, como aquelas voltadas ao diagnóstico e ao cuidado em saúde (Núcleo de Informação e Coordenação do Ponto BR, 2024).

A Organização Mundial da Saúde (2021) destaca a necessidade de uma abordagem em que a implementação da inteligência artificial na saúde deve adotar uma supervisão humana significativa (*human-in-the-loop*), assegurando que os profissionais permaneçam no controle das decisões clínicas. Nessa abordagem, a IA é concebida como ferramenta de apoio ao julgamento profissional, destinada a ampliar a capacidade decisória e não a substituir a responsabilidade, a diligência e a autonomia do cuidado humano.

Além disso, as diretrizes internacionais e nacionais convergem ao afirmar que a transparência deve ser qualificada e proporcional ao risco, bem como adaptada à capacidade de compreensão dos diferentes destinatários das explicações. No Brasil, o direito à revisão de decisões automatizadas pressupõe o acesso a informações claras sobre os critérios utilizados. Da mesma forma, as propostas regulatórias mais recentes reforçam a exigência de explicações compatíveis com a dignidade humana, mesmo diante das tensões impostas por segredos comerciais e industriais (Núcleo de Informação e Coordenação do Ponto BR, 2024).

A tensão entre a exigência ética de explicabilidade e a proteção econômica conferida pelos regimes de propriedade intelectual, especialmente pelo segredo comercial, constitui um dos desafios mais complexos da saúde digital, frequentemente descrito na literatura como *opacidade intencional*. Para Denis *et al.* (2021), enquanto a explicabilidade é um requisito ético indispensável para assegurar a autonomia do paciente, o consentimento informado e a segurança clínica, empresas de tecnologia recorrem à proteção da propriedade intelectual para resguardar modelos algorítmicos e preservar sua competitividade.

Essa dinâmica produz um conflito estrutural entre transparência e inovação, no qual demandas éticas, clínicas e regulatórias confrontam-se com estratégias de sustentabilidade econômica. Partindo desse cenário, o referencial teórico adotado neste estudo reconhece que tal tensão não se resolve por soluções simplistas, como a abertura irrestrita de códigos-fonte, mas requer abordagens intermediárias de governança capazes de equilibrar interesses públicos e privados de forma ética e responsável.

3. METODOLOGIA

O estudo caracteriza-se como um ensaio teórico crítico, de natureza qualitativa e abordagem interdisciplinar, voltado à análise das tensões entre transparência algorítmica, propriedade intelectual e ética do cuidado no contexto da XAI aplicada à saúde digital. A opção metodológica privilegia a problematização conceitual e normativa dos fenômenos analisados. Essa abordagem permite articular contribuições da saúde coletiva, da bioética, do direito e dos estudos em ciência, tecnologia e sociedade.

A pesquisa bibliográfica foi conduzida por meio de uma revisão integrativa da literatura, seguindo etapas sistemáticas de identificação, seleção, análise e síntese dos estudos. Foram consideradas publicações científicas indexadas nas bases PubMed, Scopus, Web of Science, IEEE Xplore e SciELO, publicadas no período de 2021 a 2026, de modo a assegurar a atualidade e relevância dos achados frente às rápidas transformações tecnológicas e regulatórias no campo da saúde digital.

A estratégia de busca utilizou descritores controlados e não controlados, combinados por operadores booleanos, incluindo os termos: *explainable artificial intelligence*, *digital health*, *clinical decision support*, *algorithmic transparency*, *intellectual property*, *trade secrets* e *ethics of care*. Os critérios de inclusão abrangeram estudos teóricos, revisões, análises regulatórias e pesquisas aplicadas relacionadas ao uso de IA em saúde. Assim, foram excluídos trabalhos sem aderência temática direta ou com foco exclusivamente técnico desvinculado de implicações éticas, clínicas ou jurídicas.

Paralelamente à revisão da literatura científica, realizou-se análise documental de marcos regulatórios e diretrizes nacionais e internacionais pertinentes ao tema. Foram examinados documentos de organismos multilaterais, legislações sobre proteção de dados pessoais, propostas normativas emergentes sobre inteligência artificial e relatórios institucionais. A análise concentrou-se em abordagens de governança baseada em risco, explicabilidade algorítmica e supervisão humana significativa.

O material selecionado foi organizado em eixos temáticos analíticos, possibilitando a construção de uma síntese crítica orientada pelas tensões centrais do estudo. A análise buscou identificar convergências, lacunas e conflitos entre os diferentes referenciais teóricos e normativos examinados. Dessa forma, produziu-se uma interpretação integrada que fundamenta as discussões e conclusões do artigo, sem pretensão de generalização empírica, priorizando a consistência conceitual, ética e regulatória da XAI em saúde digital.

4. RESULTADOS E DISCUSSÕES

4.1 Explicabilidade algorítmica, confiança clínica e segurança do paciente

A XAI é amplamente reconhecida como um requisito central para a adoção responsável da IA na saúde. Em contextos de saúde digital, a confiança clínica é particularmente sensível à opacidade dos modelos complexos, tornando a XAI fundamental para a transição de sistemas de "caixa-preta" para abordagens transparentes e auditáveis. Essa capacidade de compreensão das decisões automatizadas é essencial para assegurar a segurança do paciente e a equidade, especialmente em cenários diagnósticos nos quais erros podem resultar em consequências críticas (Marques *et al.*, 2025; Ali *et al.*, 2023).

Segundo Yang *et al.* (2023), a XAI surge como uma resposta crítica às limitações dos modelos de *caixa-preta*, ao promover maior transparência, interpretabilidade e alinhamento ético nos processos decisórios automatizados. Nessa direção, a literatura destaca que a explicabilidade constitui um elemento estruturante para a confiança e a governança dos sistemas de inteligência artificial. Tal necessidade torna-se ainda mais relevante em domínios sensíveis como saúde, finanças e direito, nos quais a crescente complexidade dos modelos de aprendizado profundo intensifica os riscos associados à opacidade algorítmica.

Nos estudos de Hulsen (2023), a XAI tem como propósito central tornar os modelos de inteligência artificial compreensíveis aos seres humanos. Essa finalidade torna-se especialmente relevante em contextos críticos como o cuidado em saúde, nos quais decisões automatizadas podem produzir impactos diretos e significativos sobre a vida dos pacientes. Isso fortalece a confiança de profissionais e pacientes, qualificando os processos decisórios mediados por algoritmos.

Dessa forma, a análise dos estudos selecionados evidencia consenso de que a XAI constitui um elemento estruturante para a confiança no atendimento e para a segurança do paciente. Os modelos de aprendizado profundo apresentam elevado desempenho preditivo ao identificar padrões complexos e não lineares em grandes volumes de dados. Contudo, sua

complexidade estrutural e o elevado número de parâmetros dificultam a compreensão das bases que sustentam diagnósticos e recomendações terapêuticas. Essa opacidade limita a capacidade dos profissionais de saúde de avaliar criticamente os resultados produzidos pelos sistemas automatizados (Kiseleva; Kotzinos; De Hert, 2022).

Com isso, limitações tornam-se particularmente críticas em aplicações nas quais decisões automatizadas impactam diretamente as condutas terapêuticas, como sistemas de apoio à decisão clínica, diagnóstico por imagem e monitoramento contínuo por sensores. Nesses contextos, a ausência de XAI amplia de forma significativa os riscos clínicos e éticos, sendo reconhecida como um dos principais entraves à segurança do paciente e à integridade da medicina digital (Alam; Kaur; Kabir, 2023; Maharajpet; Abhilash; Bedre, 2024).

Nos sistemas de apoio à decisão clínica (CDSS), a adoção de modelos do tipo *caixa-preta* tende a enfraquecer o papel ativo do profissional de saúde. Essa configuração favorece o viés de automação e a aceitação acrítica de recomendações algorítmicas. Nesse contexto, a XAI desempenha função estratégica ao oferecer racionalidades interpretáveis que permitem ao médico compreender, validar, contextualizar ou refutar as sugestões do sistema, preservando a autonomia profissional e a responsabilidade clínica no processo decisório (Ueda *et al.*, 2024, Kaur; Shukla, 2025).

Segundo Hettikankanamage *et al.* (2025), a explicabilidade algorítmica desempenha papel central na mitigação dos riscos associados à opacidade de modelos preditivos complexos em aplicações biomédicas. Nessa perspectiva, a literatura indica que a adoção de técnicas de inteligência artificial nesses domínios está diretamente condicionada à capacidade dos modelos de fornecer explicações compreensíveis. Isso ocorre porque transparência, responsabilização e aceitação clínica constituem requisitos essenciais em contextos sensíveis à segurança do paciente.

Assim, no diagnóstico por imagem, embora a inteligência artificial apresente elevado desempenho em áreas como radiologia e histopatologia, a ausência de explicabilidade compromete a verificação da relevância clínica dos biomarcadores utilizados. Técnicas visuais, como Grad-CAM e mapas de saliência, contribuem para indicar regiões de interesse nos exames. Contudo, a literatura aponta que a explicação visual isolada é frequentemente insuficiente para fundamentar decisões diagnósticas em contextos de alta complexidade clínica (Ali *et al.*, 2023).

No entanto, para superar essas limitações, a literatura recomenda a adoção de abordagens híbridas ou simbólicas que integrem evidências visuais, explicações em linguagem natural e

bases de conhecimento médico. Essa integração permite contextualizar as predições algorítmicas de modo que o profissional de saúde possa validá-las ou refutá-las com base em seu julgamento clínico. Ademais, enfatiza-se que a utilidade clínica dessas explicações deve ser validada sistematicamente com usuários finais, e não apenas por métricas computacionais, a fim de assegurar apoio efetivo à tomada de decisão segura (Hettikankanamage *et al.*, 2025, Yang *et al.*, 2023).

Em ambientes de monitoramento contínuo por sensores e aplicações de Internet das Coisas (IoT) em saúde, o uso de modelos opacos pode comprometer a interpretação clínica ao negligenciar a heterogeneidade dos pacientes, artefatos de sinal e interações multimodais entre dados fisiológicos. Técnicas de explicabilidade, como LIME, SHAP e abordagens híbridas, contribuem para ampliar a compreensão dos modelos e mitigar vieses e erros de generalização. Contudo, a literatura ressalta que essas abordagens ainda são avaliadas predominantemente por métricas computacionais, sem validação sistemática com usuários finais, o que limita sua efetiva aplicabilidade em fluxos reais de trabalho clínico (Maharajpet; Abhilash; Bedre, 2024).

A opacidade algorítmica está associada a riscos éticos e clínicos significativos, sobretudo pela dificuldade de identificar vieses que podem levar a diagnósticos incorretos ou a tratamentos desiguais para grupos sub-representados. Essa limitação compromete a equidade e a segurança do cuidado em saúde. Além disso, a falta de transparência fragiliza os mecanismos de responsabilização, ao dificultar a atribuição de responsabilidades legais e profissionais em casos de erro induzido por sistemas automatizados (Kiseleva; Kotzinos; De Hert, 2022).

Para superar esses entraves, Kiseleva, Kotzinos e De Hert (2022) defendem a adoção de uma abordagem de transparência em múltiplas camadas, adequada aos diferentes atores envolvidos no uso e na regulação da inteligência artificial em saúde. Nessa perspectiva, a XAI deixa de ser compreendida apenas como uma métrica técnica de desempenho. Ela passa a operar como um sistema de prestação de contas, capaz de alinhar o desenvolvimento algorítmico às exigências éticas, regulatórias e clínicas da assistência à saúde.

Diante dessas evidências, os estudos defendem uma abordagem centrada no ser humano, na qual a XAI ultrapasse sua função técnica e passe a empoderar o julgamento clínico. Para isso, as explicações devem ser contextualizadas, personalizadas ao nível de literacia em saúde dos usuários e integradas aos fluxos de trabalho, garantindo que a IA atue como instrumento transparente, confiável e eticamente alinhado ao cuidado em saúde..

4.2 Tensões ético-jurídicas, governança e propriedade intelectual na XAI em saúde

O segundo eixo dos resultados evidencia que a implementação da XAI na saúde digital é atravessada por tensões entre a necessidade ética de transparência e os interesses econômicos de proteção da inovação. Embora a literatura reconheça a explicabilidade como como um pilar para a segurança do paciente, ela enfrenta barreiras jurídicas ligadas à propriedade intelectual e desafios técnicos de governança. Essa estratégia visa proteger algoritmos e bases de dados proprietárias, mas contribui para a manutenção de formas de opacidade intencional no uso da IA em contextos clínicos (Marques *et al.*, 2025; Kaur; Shukla, 2025).

Kaur e Shukla (2025) examinam de forma integrada a articulação entre direito, ética e governança da inteligência artificial aplicada à saúde. Os autores ressaltam que a explicabilidade e a responsabilização algorítmica constituem pilares indispensáveis para a adoção responsável dessas tecnologias em contextos clínicos.

Para Alam, Kaur e Kabir (2023), o uso de algoritmos na saúde suscita tensões ético-jurídicas que extrapolam a dimensão técnica e impactam diretamente direitos fundamentais. A opacidade dos modelos do tipo *caixa-preta* compromete a autonomia e o consentimento informado, uma vez que pacientes e profissionais têm dificuldade em compreender os riscos e a lógica das decisões automatizadas. Além disso, ha desafios relevantes quanto à responsabilização por erros induzidos por IA, reforçando que, embora o médico permaneça como decisor final, ele necessita de explicações adequadas para exercer julgamento clínico crítico e evitar a prática defensiva ou acrítica.

Outra tensão central refere-se ao equilíbrio entre privacidade, explicabilidade e justiça. A oferta de explicações detalhadas pode, em certos casos, expor dados sensíveis ou permitir engenharia reversa dos modelos, criando riscos adicionais à segurança da informação. Paralelamente, modelos treinados em bases de dados não representativas tendem a reproduzir ou ampliar vieses estruturais, tornando a XAI um instrumento essencial para auditoria, identificação e mitigação de discriminações contra grupos sub-representados no cuidado em saúde (Hulsen, 2023).

Em relação a propriedade intelectual, segundo Tschider e Ho (2024), a proteção por segredo comercial tem sido amplamente adotada em aplicações de inteligência artificial na saúde. Essa opção decorre, sobretudo, da elevada complexidade dos algoritmos e das dificuldades de atender aos requisitos de divulgação exigidos pelo sistema patentário. Nesse contexto, a literatura ressalta que tecnologias de IA em saúde podem ser protegidas por diferentes regimes de propriedade intelectual, cada um com impactos distintos sobre a transparência, a auditabilidade e a governança dos sistemas.

Nesse campo, a proteção da inovação frequentemente entra em conflito com o dever de transparência clínica, configurando o que a literatura denomina *opacidade intencional*. Segundo Kiseleva, Kotzinos e De Hert (2022), desenvolvedores recorrem a regimes de segredo comercial para resguardar algoritmos e bases de dados proprietárias, o que dificulta auditorias independentes e a prestação de contas, especialmente porque, diferentemente das patentes, essa forma de proteção pode ser indefinida. Assim, as limitações do sistema patentário para algoritmos complexos reforçam o tensionamento com o interesse público, uma vez que alegações de segredo comercial não devem prevalecer quando estão em jogo direitos fundamentais, como a vida, a integridade física e a segurança do paciente.

Essa estratégia de proteção econômica, orientada à preservação de vantagens competitivas e à garantia de retorno financeiro, tensiona diretamente as exigências éticas e regulatórias de transparência e *accountability*. Nesse cenário, a implementação da XAI evidencia um conflito estrutural entre o imperativo ético da transparência e as estratégias de proteção de mercado adotadas pelos agentes tecnológicos. Essa prática demanda soluções regulatórias e institucionais capazes de equilibrar esses interesses concorrentes de forma proporcional e socialmente responsável (Tschider; Ho, 2024).

Do ponto de vista regulatório, os resultados indicam uma convergência internacional em direção a modelos de governança baseados em risco para a inteligência artificial em saúde. Nesses modelos, aplicações consideradas de alto risco são submetidas a exigências reforçadas de transparência, documentação, supervisão humana e monitoramento pós-implantação. Com isso, diretrizes e debates em contextos nacionais, indicam que a explicabilidade não deve ser entendida como abertura irrestrita de códigos-fonte, mas como obrigação regulatória proporcional ao risco e ao impacto clínico da aplicação (Kaur; Shukla, 2025).

Nesse cenário, a regulação tende a adotar um modelo de transparência em camadas, no qual o nível de detalhamento técnico é elevado para auditores e autoridades reguladoras, enquanto, para médicos e pacientes, a explicação prioriza a utilidade clínica e a compreensibilidade. Essa abordagem busca assegurar que a IA atue como um parceiro seguro, auditável e alinhado às exigências éticas e assistenciais da prática em saúde (Kiseleva; Kotzinos; De Hert, 2022).

Portanto, Kaur e Shukla (2025) destacam a adoção de mecanismos intermediários de governança configura uma estratégia viável e necessária para equilibrar a transparência algorítmica e a proteção da inovação no desenvolvimento de tecnologias em saúde. Esses

instrumentos ampliam a prestação de contas e a supervisão regulatória sem demandar a abertura irrestrita de códigos-fonte ou a exposição integral de ativos tecnológicos sensíveis.

Além disso, tais mecanismos são essenciais para mediar a tensão entre a proteção da propriedade intelectual e o imperativo ético de transparência na saúde. Embora o uso de segredos comerciais seja frequentemente mobilizado para preservar vantagens competitivas, essa forma de opacidade intencional não deve inviabilizar a inspeção técnica e ética necessária para assegurar a segurança, a equidade e a confiabilidade dos sistemas de inteligência artificial.

Diante disso, os achados deste estudo reforçam que a XAI deve ser compreendida como um instrumento de mediação entre interesses públicos e privados. Essa mediação permite compatibilizar a inovação tecnológica com a proteção de direitos fundamentais e a integridade do cuidado em saúde. Nesse sentido, a governança da XAI em saúde digital não se configura como uma solução técnica isolada, mas como um arranjo institucional complexo que exige articulação entre ética, regulação, prática clínica e regimes de propriedade intelectual.

4 CONCLUSÃO

A análise desenvolvida neste artigo evidencia que a XAI ocupa posição estratégica na consolidação de uma saúde digital eticamente responsável. Ao enfrentar os desafios da opacidade algorítmica, a XAI torna-se particularmente relevante em contextos de alto risco clínico. Os resultados demonstram que a explicabilidade é condição fundamental para a confiança clínica, a segurança do paciente e a preservação da autonomia profissional, especialmente em sistemas de apoio à decisão, diagnóstico por imagem e monitoramento contínuo, nos quais decisões automatizadas impactam diretamente as condutas terapêuticas.

Constata-se, entretanto, que a implementação da XAI não ocorre em um vazio normativo, mas em um cenário marcado por tensões estruturais entre o imperativo ético da transparência e os regimes de proteção da propriedade intelectual. A utilização recorrente de segredos comerciais como estratégia de proteção econômica configura formas de opacidade intencional, que tensionam a responsabilização, a auditabilidade e o interesse público. Nesse sentido, a literatura analisada reforça que a explicabilidade não deve ser entendida como abertura irrestrita de códigos-fonte, mas como obrigação regulatória e ética proporcional ao risco e ao impacto da aplicação em saúde.

Com isso, os achados indicam que a conciliação entre inovação tecnológica e proteção de direitos fundamentais pode ser viabilizada por meio da adoção de mecanismos intermediários de governança. Entre esses mecanismos destacam-se a transparência em

camadas, as auditorias independentes, a documentação técnica qualificada e a supervisão humana significativa. Tais arranjos permitem compatibilizar a proteção da propriedade intelectual com a necessidade de prestação de contas, fortalecendo modelos de governança baseados em risco e alinhados às diretrizes internacionais emergentes para a inteligência artificial em saúde.

Por fim, o estudo destaca que a ética do cuidado deve ocupar lugar central na governança da XAI em saúde digital, orientando tanto o desenvolvimento tecnológico quanto os marcos regulatórios. Mais do que tornar algoritmos compreensíveis, trata-se de assegurar que a IA fortaleça as relações de confiança, responsabilidade e sensibilidade às vulnerabilidades humanas que sustentam a prática em saúde. Para as pesquisas futuras, especialmente de natureza empírica, poderão aprofundar essas reflexões, contribuindo para políticas públicas e regulações mais responsivas às complexidades éticas, clínicas e sociais da saúde digital.

REFERÊNCIAS

ALAM, Mohammad Nazmul; KAUR, Mandeep; KABIR, Md. Shahin. **Explainable AI in healthcare: enhancing transparency and trust upon legal and ethical consideration**. *International Research Journal of Engineering and Technology (IRJET)*, v. 10, n. 6, p. 828–835, jun. 2023. Disponível em: https://www.irjet.net.

ALI, Sajid; ABUHMED, Tamer; EL-SAPPAGH, Shaker; MUHAMMAD, Khan; ALONSO-MORAL, Jose M.; CONFALONIERI, Roberto; GUIDOTTI, Riccardo; DEL SER, Javier; DÍAZ-RODRÍGUEZ, Natalia; HERRERA, Francisco. *Explainable Artificial Intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence*. **Information Fusion**, Amsterdam, v. 99, p. 101805, 2023. DOI: 10.1016/j.inffus.2023.101805. DOI: https://doi.org/10.1016/j.inffus.2023.101805.

AMANN, Julia; BLASIMME, Alessandro; VAYENA, Effy; FREY, Dietmar; MADAI, Vince I. *Explainability for artificial intelligence in healthcare: a multidisciplinary perspective*. **BMC Medical Informatics and Decision Making**, London, v. 20, n. 1, p. 310, 2020. DOI: https://doi.org/10.1186/s12911-020-01332-6.

DENIS, Gabriela; HERMOSILLA, María Paz; ARACENA, Claudio; SÁNCHEZ ÁVALOS, Roberto; GONZÁLEZ ALARCÓN, Natalia; POMBO, Cristina. *Uso responsável da inteligência artificial para políticas públicas: manual de formulação de projetos*. Washington, D.C.: Banco Interamericano de Desenvolvimento, 2021. Disponível em: https://www.iadb.org/.

HETTIKANKANAMAGE, Nadeesha; SHAFIABADY, Niusha; CHATTEUR, Fiona; WU, Robert M. X.; UD DIN, Fareed; ZHOU, Jianlong. **eXplainable artificial intelligence (XAI): a systematic review for unveiling the black box models and their relevance to biomedical imaging and sensing**. *Sensors*, Basel, v. 25, n. 21, p. 1–31, 2025. DOI: https://doi.org/10.3390/s25216649.

HULSEN, Tim. **Explainable Artificial Intelligence (XAI): concepts and challenges in healthcare**. *AI*, Basel, v. 4, n. 3, p. 652–666, 2023. DOI: 10.3390/ai4030034. Disponível em: https://doi.org/10.3390/ai4030034.

KISELEVA, Anastasiya; KOTZINOS, Dimitris; DE HERT, Paul. **Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations**. *Frontiers in Artificial Intelligence*, Lausanne, v. 5, e879603, 2022. DOI: https://doi.org/10.3389/frai.2022.879603.

MAHARAJPET, Sheela S.; ABHILASH, H. P.; BEDRE, Shrihari R. *A LIME-based explainable AI for healthcare IoT: building trust in clinical decision-making*. In: PANDIKUMAR, S.; THAKUR, Manish Kumar (org.). **Innovations and trends in modern computer science technology: overview, challenges and applications**. [S.l.]: QTanalytics, 2024. cap. 3, p. 22–29. DOI: https://doi.org/10.48001/978-81-980647-5-2-3.

MARQUES, Caroline Parra; SOUZA, Larissa Vitória Costa Carrazzoni de; COLTRI, Marcos Vinicius; FRANCO, Ademir. *Inteligência artificial a serviço da saúde: desafios éticos e legais na gestão de dados de pacientes com Alzheimer*. **Cadernos Ibero-Americanos de Direito Sanitário**, Brasília, v. 14, n. 3, p. 84–95, jul./set. 2025. DOI: 10.17566/ciads.v14i3.1357. DOI: https://doi.org/10.17566/ciads.v14i3.1357.

NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR (NIC.br). *Inteligência artificial na saúde: potencialidades, riscos e perspectivas para o Brasil*. São Paulo: Comitê Gestor da Internet no Brasil, 2024. (Cadernos NIC.br: Estudos Setoriais). Disponível em: https://www.nic.br/publicacao/inteligencia-artificial-na-saude-potencialidades-riscos-e-perspectivas-para-o-brasil/.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization, 2021. Disponível em: https://www.who.int/publications/i/item/9789240029200.

TSCHIDER, Charlotte A.; HO, Cynthia M. **Artificial intelligence and intellectual property in healthcare technologies**. In: HOFFMAN, Sharona; PODGURSKI, Andy (org.). *Research handbook on health, artificial intelligence and the law*. Cheltenham: Edward Elgar Publishing, 2024. cap. 11, p. 183–201.

UEDA, Daiju; KAKINUMA, Taichi; FUJITA, Shohei; KAMAGATA, Koji; FUSHIMI, Yasutaka; ITO, Rintaro; MATSUI, Yusuke; NOZAKI, Taiki; NAKAURA, Takeshi; FUJIMA, Noriyuki; TATSUGAMI, Fuminari; YANAGAWA, Masahiro; HIRATA, Kenji; YAMADA, Akira; TSUBOYAMA, Takahiro; KAWAMURA, Mariko; FUJIOKA, Tomoyuki; NAGANAWA, Shinji. **Fairness of artificial intelligence in healthcare: review and recommendations**. *Japanese Journal of Radiology*, Tóquio, v. 42, n. 1, 2024. DOI: https://doi.org/10.1007/s11604-023-01474-3.

YANG, Wenli; WEI, Yuchen; WEI, Hanyu; CHEN, Yanyu; HUANG, Guan; LI, Xiang; LI, Renjie; YAO, Naimeng; WANG, Xinyi; GU, Xiaotong; AMIN, Muhammad Bilal; KANG, Byeong. **Survey on explainable AI: from approaches, limitations and applications aspects**. *Human-Centric Intelligent Systems*, v. 3, p. 161–188, 2023. DOI: https://doi.org/10.1007/s44230-023-00038-y.