DANIEL DE ARAUJO DOURADO

Regulação da inteligência artificial na saúde

São Paulo
2023

DANIEL DE ARAUJO DOURADO

Regulação da inteligência artificial na saúde

Tese apresentada à Faculdade de Medicina da Universidade de São Paulo para a obtenção do título de Doutor em Ciências

Programa: Saúde Coletiva

Orientador: Prof. Dr. Fernando Mussa Abujamra Aith

São Paulo

2023

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

Dourado, Daniel de Araujo
Regulação da inteligência artificial na saúde / Daniel de Araujo Dourado. -- São Paulo, 2023.
Tese(doutorado)--Faculdade de Medicina da Universidade de São Paulo.
Programa de Saúde Coletiva.
Orientador: Fernando Mussa Abujamra Aith.

Descritores: 1.Inteligência artificial 2.Marcos regulatórios em saúde 3.Direito sanitário 4.Regulação e fiscalização em saúde 5.Governança em saúde 6.Saúde digital

USP/FM/DBD-494/23

Responsável: Erinalva da Conceição Batista, CRB-8 6755

Nome: DOURADO, Daniel de Araujo

Título: Regulação da inteligência artificial na saúde

Tese apresentada à Faculdade de Medicina da Universidade de São Paulo para a obtenção do título de Doutor em Ciências

Aprovado em:

Banca Examinadora

Prof(a). Dr(a). _____

Instituição: _____

Julgamento: _____

Prof(a). Dr(a). _____

Instituição: _____

Julgamento: _____

Prof(a). Dr(a). _____

Instituição: _____

Julgamento _____

RESUMO

DOURADO, Daniel de Araujo. **Regulação da inteligência artificial na saúde**. 2023. Tese (Doutorado em Ciências) – Faculdade de Medicina, Universidade de São Paulo, São Paulo, 2023.

A Inteligência Artificial (IA) é um ponto de virada na trajetória histórica da humanidade, despertando interesse e preocupação crescentes. Na área da saúde, a IA promete induzir transformações profundas, alterando práticas e estruturas dos sistemas de saúde. A integração efetiva da IA na saúde requer um marco regulatório específico. Esta tese tem o objetivo de estabelecer um paradigma geral para a regulação da IA na saúde, baseando-se em fundamentos éticos e jurídicos e em diretrizes regulatórias. Para isso, foi realizada uma pesquisa qualitativa, utilizando abordagem teórica e empírica, fundamentada em uma revisão não sistemática da literatura global e na análise de propostas políticas e normativas de países e organizações internacionais. Primeiramente, o estudo identifica os conceitos básicos de IA, com foco em aprendizado de máquina (*machine learning*), abordando sua evolução histórica e o panorama atual na área da saúde. Evidencia-se que as diversas aplicações da IA na saúde compartilham características elementares e desafios, demonstrando que existem pontos comuns e suficientemente delimitados que justificam o reconhecimento da IA como objeto específico da regulação em saúde. A partir desse reconhecimento, são identificados e analisados fundamentos éticos e jurídicos para estruturar a regulação da IA na saúde, contemplando elementos de: direito à saúde; ética em IA; direitos humanos; princípios éticos próprios da IA na saúde; e governança de dados de saúde. Com base nesses fundamentos, a tese propõe diretrizes regulatórias para a IA na saúde em três dimensões: 1) segurança e eficácia; 2) transparência; e 3) responsabilidade. Em termos de segurança e eficácia, enfatiza-se a necessidade de admitir os sistemas de IA na categoria regulatória "*software* como dispositivo médico" (SaMD). Portanto, esses sistemas devem ser avaliados pelos padrões estabelecidos por entidades internacionais, mas acrescidos de dois aspectos essenciais: monitoramento contínuo de mudanças e eliminação de vieses para promover equidade. A transparência exige acesso a informações que fundamentam as decisões de sistemas de IA, priorizando a clareza sobre os contextos de desenvolvimento e implementação dos modelos, e deve ser expressa por meio de documentação padronizada. No que diz respeito à transparência dos modelos, é necessário instituir tratamentos regulatórios distintos para sistemas de IA interpretável e de IA explicável. Na dimensão da responsabilidade, são exploradas limitações dos atuais sistemas de responsabilidade civil para lidar com a IA e possíveis abordagens para solucionar essa questão jurídica. Em seguida, analisam-se as atuais propostas de adaptação das regras de responsabilidade civil para contemplar produtos baseados em IA. Por fim, são abordadas tendências emergentes que devem influenciar a regulação da IA na saúde, incluindo: o papel das legislações horizontais e da abordagem regulatória setorial; a perspectiva da estratégia regulatória baseada em ambientes de testagem (*sandboxes*); a necessidade e a complexidade de regular os modelos de fundação (*foundation models*) para uso na saúde. Conclui-se que os fundamentos analisados e as diretrizes propostas constituem uma base sólida para a elaboração de um paradigma regulatório robusto e abrangente para a IA na saúde. Espera-se contribuir para o debate em busca do avanço seguro, transparente e responsável desse campo em rápida expansão.

Palavras-chave: Inteligência artificial. Marcos regulatórios em saúde. Direito sanitário. Regulação e fiscalização em saúde. Governança em saúde. Saúde digital.

ABSTRACT

DOURADO, Daniel de Araujo. **Regulation of artificial intelligence in healthcare**. 2023. Doctoral thesis (Doctorate in Science) – "Faculdade de Medicina, Universidade de São Paulo", São Paulo, 2023.

Artificial Intelligence (AI) is a turning point in the historical trajectory of humanity, provoking growing interest and concern. In the healthcare sector, AI promises to induce profound transformations, changing practices and structures of healthcare systems. The effective integration of AI in healthcare requires a specific regulatory framework. This thesis aims to establish a general paradigm for the regulation of AI in healthcare, grounded in ethical-legal foundations and regulatory guidelines. To this end, it was conducted qualitative research, using a theoretical and empirical approach, based on a non-systematic review of global literature and the analysis of political and normative proposals from countries and international organizations. Firstly, the study identifies the key concepts and applications of AI, with a focus on machine learning, addressing its historical evolution and the current panorama in the healthcare area. Despite myriad applications, AI in healthcare exhibits fundamental characteristics and challenges, demonstrating that there are common and sufficiently delimited points that justify the recognition of AI as a specific object of healthcare regulation. Based on this recognition, the study identifies and analyzes ethical and legal foundations to structure the regulation of AI in health, covering elements of: the right to health; AI ethics; human rights; ethical principles specific to AI in healthcare; and health data governance. From these foundations, the thesis proposes regulatory guidelines for AI in healthcare in three dimensions: 1) safety and efficacy; 2) transparency; and 3) responsibility. In terms of safety and efficacy, it emphasizes the need to classify AI systems within the existing "software as a medical device" (SaMD) regulatory framework. Therefore, these systems must be evaluated according to standards established by international entities, but with two essential aspects added: monitoring continuous changes and eliminating biases to promote equity. Transparency requires access to information that underlies AI decisions, prioritizing clarity about the contexts of model development and deployment, and it must be expressed through standardized documentation. Regarding the transparency of models, it is necessary to institute different regulatory treatments for interpretable AI and explainable AI systems. In the responsibility dimension, the thesis explores limitations of current civil liability and tort law systems for dealing with AI and possible approaches to solving this legal issue. Next, it analyzes current proposals for adapting civil liability rules to cover products based on AI. Finally, emerging trends that should influence the regulation of AI in healthcare are addressed, including: the role of horizontal legislation and the sectoral regulatory approach; the perspective of a regulatory strategy based on testing environments (regulatory sandboxes); the need and complexity of regulating foundation models for use in healthcare. The conclusion is that the analyzed fundamentals and proposed guidelines constitute a solid basis for developing a robust and comprehensive regulatory paradigm for AI in healthcare. The prospect is to contribute to the debate in search of the safe, transparent, and responsible advancement of this rapidly expanding field.

Keywords: Artificial intelligence. Regulatory frameworks for health. Health law. Health care coordination and monitoring. Health governance. Digital health.

SUMÁRIO

1 INTRODUÇÃO

A Inteligência Artificial (IA) pode ser o maior evento da história de nossa civilização. Todos os aspectos das nossas vidas serão transformados por ela. Com esses termos, Stephen Hawking resumiu o impacto da revolução tecnológica que estamos vivendo, no seu célebre discurso de inauguração de um dos vários centros dedicados ao estudo e desenvolvimento da IA estabelecidos recentemente. Segundo ele, o surgimento de uma IA poderosa "será a melhor ou a pior coisa que já aconteceu à humanidade" e acertar nessa empreitada "é crucial para o futuro da nossa civilização e da nossa espécie" (UNIVERSITY OF CAMBRIDGE, 2016).

Tanto a expectativa como a preocupação em relação à IA têm aumentado progressivamente nas últimas décadas. A **regulação da IA** tornou-se uma prioridade global a partir dos anos 2020, como reconheceu a Cúpula do Grupo dos Sete (G7) em maio de 2023 ao estabelecer o "Processo de IA de Hiroshima" (G7 HIROSHIMA SUMMIT, 2023). Em julho desse mesmo ano, o Secretário-Geral da Organização das Nações Unidas (ONU) também atestou a importância dessa regulação em pronunciamento na primeira reunião do Conselho de Segurança especificamente dedicada a debater IA (UN PRESS, 2023).

A **saúde** é um dos setores nos quais as transformações impulsionadas pela IA prometem ser mais significativas. A era digital tem introduzido tecnologias com o potencial de modificar de forma substancial as práticas e serviços de saúde, influenciando as estruturas dos sistemas de saúde em todo o mundo. A implementação da IA na saúde oferece oportunidades sem precedentes para melhorar a qualidade dos cuidados e das ações preventivas tanto no nível individual quanto populacional, ampliar o acesso, reduzir custos e explorar novas fronteiras em prevenção, diagnóstico e tratamento (MATHENY *et al.*, 2022). Até meados da década de 2020, embora a assimilação da IA na área da saúde tenha ocorrido de maneira mais gradual do que em outros setores, tem-se formado um consenso sobre sua inevitável integração em diversos domínios da medicina e da saúde pública (SAHNI; CARRUS, 2023).

A regulação tem sido vista como elemento central para a efetiva incorporação da IA no setor da saúde (LANCET, 2023). A estruturação de **marcos regulatórios específicos para a IA na saúde** é considerada medida essencial para que essa tecnologia seja usada de forma segura, transparente, responsável e justa. Há crescente compreensão de que essa regulação exige novas estruturas de governança e uma abordagem global (WHO, 2021), seguindo a tendência que vem sendo observada em relação à regulação da IA em geral (BREMMER; SULEYMAN, 2023).

1.1 Estágio Atual de Discussão

Regulamentos e políticas para desenvolvimento ético, implementação e uso de tecnologias de IA na saúde estão em fase de elaboração em diferentes níveis e por diversas entidades nacionais e internacionais. Mas ainda não há leis específicas em vigor. A evolução muito acelerada do panorama da IA tem deixado essa atividade cada vez mais complexa.

A Organização Mundial da Saúde (OMS) tem liderado o processo para definição de uma estratégia global de saúde digital a fim de apoiar os sistemas nacionais de saúde, conforme aprovado por unanimidade pelos Estados Membros na 71ª Assembleia Mundial da Saúde (WHO, 2018). Nessa estratégia, insere-se a preparação de uma estrutura de governança e regulação da IA na saúde.

Em outubro de 2023, a OMS publicou o primeiro documento resultante do seu "Grupo de Trabalho sobre Considerações Regulatórias sobre IA para a Saúde", grupo multidisciplinar composto por gestores públicos e representantes de autoridades reguladoras, da academia e da indústria. Trata-se de uma abordagem inicial ao tema em áreas temáticas abrangentes, que ainda não tem intenção de ser uma orientação ou diretriz política e regulatória. A proposta é colaborar em parceria com grupos nacionais e internacionais estabelecidos, tendo como escopo a potencial convergência e a harmonização regulatória para a IA aplicada à saúde (WHO, 2023).

A OMS pretende, entre outras medidas, trabalhar de maneira coordenada com entidades intergovernamentais para identificar e formular leis e políticas, considerando a iniciativa de elaborar uma legislação modelo para ser usada como referência por governos que pretendam criar regulações próprias para IA na saúde (WHO, 2021). Algumas organizações internacionais têm contribuído ativamente nesse campo, como o Fórum Internacional de Reguladores de Dispositivos Médicos (IMDRF – *International Medical Device Regulators Forum*), a Organização Internacional para Padronização (ISO – *International Organization for Standardization*), a Organização para a Cooperação e Desenvolvimento Econômico (OCDE), entre outras.

No atual momento desse debate (até a primeira metade da década de 2020), busca-se definir princípios gerais que possam ser aplicados para regulação da IA na saúde em diferentes contextos: nos países de baixa, média e alta renda, nos setores público e privado, por governos e organizações internacionais.

Esse é o desafio para os próximos anos.

1.2 Objetivos

Diante desse panorama, a presente tese tem dois objetivos principais:

a) identificar **fundamentos éticos e jurídicos** para a construção teórica de um paradigma geral de regulação da IA na saúde;
b) caracterizar e analisar **diretrizes para elaboração desse paradigma regulatório** para a IA na saúde.

Para atingir esse escopo, definem-se os seguintes objetivos específicos:

a) propor princípios gerais para a regulação da IA na saúde que possam ser adaptados para aplicação em diferentes contextos e países;
b) identificar tendências regulatórias e perspectivas para uso da IA na saúde nos países e organizações internacionais que estão à frente desse processo;
c) analisar possíveis mudanças na regulação de sistemas de saúde para possibilitar a incorporação de sistemas de IA.

1.3 Método e Estrutura do Trabalho

Esta tese é desenvolvida em cinco seções, seguidas por uma de síntese conclusiva.

Primeiramente, são apresentados conceitos básicos do campo da inteligência artificial (IA) e da sua aplicação na área da saúde, a partir de estudo teórico e revisão não sistemática da literatura. A **seção 2** identifica definições e tipologias de IA, com foco no tipo aprendizado de máquina (*machine learning*), o seu modo de funcionamento e os subtipos mais relevantes. A **seção 3** aborda a evolução histórica e o panorama atual de IA na saúde, analisando as especificidades para o desenvolvimento e a implementação de sistemas de IA na assistência à saúde e em outros cenários no setor da saúde.

Os conceitos e análises explorados nessas duas seções iniciais da tese evidenciam que há elementos bem definidos no campo da IA que estão ou estarão presentes em qualquer cenário em que a tecnologia for usada na área da saúde. Embora não exista uma definição consensual da IA, suas aplicações na saúde estão baseadas em conceitos essenciais de estatística e ciência

da computação que são válidos independentemente dos usos particulares. O aprofundamento nesses domínios foge do escopo deste trabalho. O que se pretende é demonstrar que existem pontos comuns e suficientemente delimitados que justificam o reconhecimento da IA como objeto específico da regulação em saúde.

A partir desse reconhecimento, a tese aborda diretamente a regulação da IA na saúde. É o núcleo principal do trabalho. A partir de pesquisa qualitativa, com abordagem teórica e empírica, são identificadas as propostas políticas e as tendências regulatórias para IA na saúde atualmente existentes e que estão em desenvolvimento por governos de países e por entidades internacionais. Então, são propostos princípios e estratégias regulatórias com amparo na literatura global sobre o tema. As duas seções dessa parte exploram respectivamente os fundamentos para a regulação da IA na saúde e as diretrizes para sua estruturação.

A **seção 4** é dedicada a explorar os fundamentos para a regulação da IA na saúde, entendidos como elementos éticos e jurídicos consolidados e que devem servir de base para construção do arcabouço regulatório: o direito à saúde; a ética em IA; os direitos humanos; os princípios éticos específicos para a IA na saúde; e a governança de dados pessoais de saúde.

A **seção 5** propõe diretrizes para a regulação da IA na saúde. Abordam-se questões em que a estrutura regulatória está em fase em elaboração e possíveis caminhos para definição de um paradigma geral. Esta tese sustenta que isso deve ser feito com base em princípios gerais definidos em três dimensões: 1) segurança e eficácia; 2) transparência; e 3) responsabilidade.

Na dimensão de segurança e eficácia, são abordados os aspectos pertinentes à admissão de sistemas de IA em dispositivos médicos e enfatizados dois aspectos particulares para a regulação dessas ferramentas na área da saúde: o monitoramento contínuo das mudanças e a eliminação de vieses para assegurar equidade. Um ponto central é o argumento de que elementos de justiça algorítmica precisam ser integrados como requisitos regulatórios de segurança e eficácia em IA na saúde.

A transparência é tratada como uma dimensão independente por ser necessária para a todos os outros elementos da regulação da IA na saúde, tanto da dimensão de segurança e eficácia como da dimensão da responsabilidade. Analisam-se os desafios para admitir a explicação de resultados algorítmicos como requisito regulatório e enfatiza-se a ideia de que a documentação padronizada deve ser o mecanismo essencial para expressão da transparência.

A dimensão da responsabilidade é analisada considerando a atual organização jurídica dos sistemas de responsabilidade civil e enfatizando as particularidades na área da saúde. Com base nos desafios enfrentados para atribuição de responsabilidades em IA na saúde, exploram-se possíveis mecanismos e propostas atualmente existentes nesse campo.

Cabe ressaltar que outras dimensões relevantes para a regulação da IA, tais como as relacionadas à propriedade intelectual, ao direito trabalhista e à cibersegurança, não foram incluídas como objeto de análise. Embora importantes na construção de marcos regulatórios para a IA em geral, nessas outras dimensões não foram identificadas particularidades específicas que influenciem diretamente a construção de um paradigma geral de regulação da IA na saúde.

A **seção 6** contempla uma breve exposição sobre tendências emergentes. Exploram-se tópicos que estão em desenvolvimento ou em discussão viva por ocasião da pesquisa que deu origem a esta tese e que têm potencial para influenciar a formulação de marcos regulatórios para a IA na saúde: o papel das legislações horizontais e da abordagem regulatória setorial; a perspectiva da estratégia regulatória baseada em ambientes de testagem (sandboxes); a necessidade e a complexidade de regular os modelos de fundação (foundation models).

Por fim, a **seção 7** apresenta uma síntese conclusiva, listando os principais pontos abordados na tese, perspectivas e encaminhamentos nessa linha de pesquisa.

O desenvolvimento desta tese mostrou-se particularmente desafiador em razão da natureza bastante dinâmica do campo da IA. A pesquisa enfrentou a tarefa de manter-se alinhada com mudanças quase diárias nos objetos de investigação, o que exigiu atualizações frequentes no conteúdo, revisões e adaptações para incorporar os avanços e discussões mais recentes. Este trabalho não busca estabelecer conclusões definitivas, ele tem o propósito de contribuir para o debate atual sobre a regulação da IA na saúde num momento crucial de discussão. Espera-se que esta tese sirva como um recurso informativo e uma base para novas investigações e reflexões nesta época em que a IA assume posição central na inovação tecnológica, na medicina, na saúde pública e no futuro da humanidade.

2 INTELIGÊNCIA ARTIFICIAL: CONCEITOS E TIPOLOGIAS

2.1 Definições de IA

2.1.1 Definição do campo da IA

O termo Inteligência Artificial (IA) foi cunhado em 1956, durante a célebre Conferência de Dartmouth, definido como “a ciência e a engenharia de produzir máquinas inteligentes” (MCCARTHY *et al.*, 1955). Mas o campo operacional já existia antes disso. O trabalho mais notório que pavimentou o caminho para a definição de IA foi o de Alan Turing (1950), que propôs substituir a pergunta “pode uma máquina pensar?” por “pode uma máquina ser linguisticamente indistinguível de um humano?”. Essa proposta marcou o início do chamado "Teste de Turing", uma medida da capacidade de uma máquina de exibir comportamento inteligente equivalente ao de um humano na capacidade de produção de linguagem (BRINGSJORD; GOVINDARAJULU, 2022).

Mais de cinco décadas após ter sido um dos criadores da IA como campo de conhecimento, John McCarthy (2007) continuou a conceituá-la usando a expressão “máquinas inteligentes”. Ele estava ciente de que este não era um entendimento consensual, principalmente por não existir uma definição universalmente aceita de inteligência, um conceito que desempenha um papel crítico neste contexto. A inteligência não é um atributo uniforme, mas se apresenta em formas e graus variados entre humanos, diversos animais e certas máquinas. Na perspectiva da IA, a inteligência pode ser entendida como um aspecto computacional que facilita a realização de metas no mundo.

Assim, a IA é geralmente definida como o campo dedicado à criação de entidades artificiais que, em contextos apropriados, podem demonstrar comportamentos semelhantes aos de animais ou pessoas. Mais especificamente, trata-se de uma área multidisciplinar vinculada ao objetivo de desenvolver máquinas capazes de simular a inteligência humana, aprender a partir de experiências, adaptar-se a novos estímulos e executar tarefas que geralmente requerem a capacidade intelectual humana (RUSSELL; NORVIG, 2010).

2.1.2 Sistemas de IA

De acordo com Stuart Russell e Peter Norvig (2010), a inteligência artificial pode ser compreendida sob quatro perspectivas: 1) sistemas que pensam como humanos; 2) sistemas que pensam racionalmente; 3) sistemas que agem como humanos; e 4) sistemas que agem racionalmente.

Os **sistemas que pensam como humanos** constituem a primeira perspectiva, que envolve o desenho de sistemas de IA que imitam os processos de pensamento humano (NEWELL; SIMON, 1976). Esta abordagem, conhecida como modelagem cognitiva, visa replicar os padrões complexos e multifacetados da inteligência humana em máquinas. Baseia-se na crença de que a mente humana serve como o melhor modelo para o comportamento inteligente (HASSABIS *et al.*, 2017).

Os **sistemas que pensam racionalmente** formam a segunda perspectiva, muitas vezes referida como "IA lógica". A ideia principal é implementar o raciocínio lógico formal em máquinas, permitindo-lhes inferir e deduzir conhecimento com base em regras predefinidas (GENESERETH; NILSSON, 1987). Esta abordagem enfatiza a importância das representações simbólicas e lógicas do conhecimento, bem como a manipulação algorítmica dessas representações para gerar novo conhecimento (MCCARTHY, 1959).

A terceira perspectiva da IA envolve a criação de **sistemas que agem como humanos**. Esta abordagem é exemplificada pelo Teste de Turing (TURING, 1950). Técnicas neste domínio abrangem as atuais técnicas de processamento de linguagem natural (PLN), visão computacional e robótica, facilitando interações semelhantes às humanas com máquinas (HOVY, 1993).

A quarta perspectiva se concentra na criação de **sistemas que agem racionalmente**, também conhecida como a abordagem do "agente racional". A ênfase está na capacidade da máquina de tomar decisões apropriadas, dada a sua percepção do ambiente e os seus objetivos (RUSSELL; NORVIG, 2010). Esta perspectiva engloba áreas como o planejamento automatizado, o aprendizado por reforço e a tomada de decisão sob incerteza.

2.1.3 Trajetória e desafio persistente para definição de IA

A trajetória da IA tem sido marcada por períodos de intensa atividade e avanço, conhecidos como “verões da IA”, intercalados por períodos de diminuição de financiamento e interesse, os chamados “invernos da IA”. Apesar dessas oscilações, a IA tem alcançado progressos significativos nas últimas décadas, com avanços impulsionados sobretudo pelo campo do aprendizado de máquina (*machine learning*). A presente onda ganhou força na década de 2010 a partir do desenvolvimento de um subtipo do aprendizado de máquina, o aprendizado profundo (*deep learning*) (HAENLEIN; KAPLAN, 2019). Esse crescimento exponencial está diretamente relacionado ao recente desenvolvimento das capacidades de armazenamento, processamento, análise e transmissão de grandes quantidades de dados: a área que se denomina *big data* e que é base para o desenvolvimento da IA atual (HILBERT; LÓPEZ, 2011).

O termo “Inteligência Artificial” é hoje usado para se referir a um campo extremamente amplo de conhecimento e prática, que engloba ciência da computação, matemática, psicologia, linguística, filosofia e neurociência, entre outras disciplinas. Mas não há uma definição universalmente aceita de IA, o que continua sendo tema de ampla discussão no cenário global (EUROPEAN COMMISSION, 2019a). Entidades internacionais e países têm procurado encontrar definições mais ou menos abrangentes para IA, mas esse elemento persiste como um dos principais desafios para a regulamentação desse campo (RENDA; ENGLER, 2023).

2.2 Tipos de Inteligência Artificial

A IA tem sido geralmente classificada de duas formas principais: pelo grau de similaridade com a inteligência humana e pela sua funcionalidade.

2.2.1 IA Estreita e IA Geral

Com base no grau de similaridade com a inteligência humana, a IA é dividida em duas categorias principais: IA Estreita ou Restrita e IA Geral.

A **IA Estreita** (*Narrow AI*) – ou IA Fraca (*Weak AI*) – refere-se a sistemas projetados para realizar uma tarefa específica, como reconhecimento de voz, sistemas de recomendação ou reconhecimento de imagens. Esses sistemas podem superar os humanos em tarefas específicas, mas não conseguem compreender ou aprender além do que foram programados (KURZWEIL, 2005). Até o presente, todas as formas de IA desenvolvidas se enquadram nesta categoria.

Por outro lado, a **IA Geral** (*General AI*) – ou IA Forte (*Strong AI*) – contempla a possibilidade de uma única inteligência geral capaz de demonstrar todos os aspectos da inteligência humana, incluindo compreensão, raciocínio, aprendizado e consciência (GOERTZEL, 2014; SEARLE, 1980). Ou seja, esses sistemas teriam capacidade de executar qualquer tarefa intelectual que um ser humano possa fazer. Essa forma de IA continua sendo uma possibilidade teórica, já que ainda não existem modelos ou sistemas que atinjam este nível de inteligência. Alguns estudiosos argumentam que provavelmente isso nunca seja alcançado (FJELLAND, 2020).

2.2.2 IA Simbólica e IA Conexionista

Outra classificação importante da IA é baseada em sua funcionalidade, dividindo-se em IA Simbólica (*Symbolic AI*) e IA Conexionista (*Connectionist AI*).

A **IA Simbólica**, também conhecida como IA baseada em regras, IA clássica ou pelo acrônimo GOFAI (*Good Old-Fashioned AI*), utiliza representações simbólicas de problemas, lógica e algoritmos de busca (NEWELL; SIMON, 1976). Ela opera no princípio de usar símbolos como blocos de construção da cognição, imitando como os humanos usam símbolos para atribuir significados. A IA simbólica se destaca em ambientes com regras e objetivos claros, como um jogo de xadrez. No entanto, ela tem dificuldades com aplicações que provavelmente encontrarão variações, pois é extremamente desafiador criar regras para todos os cenários possíveis devido à complexidade e variabilidade do mundo real (SANTORO *et al.*, 2021).

A **IA Conexionista**, também conhecida como "redes neurais" (*neural networks*) ou "processamento paralelo distribuído" (*parallel distributed processing*), tenta imitar a estrutura e função do cérebro humano. Esta forma de IA é caracterizada por nós interconectados – chamados de "neurônios" – que processam informações em paralelo, o que é bastante diferente dos modelos tradicionais de IA simbólica que usam processamento sequencial e regras explícitas

(RUMELHART; MCCLELLAND, 1986). Ao contrário da IA simbólica, que depende da codificação humana, a IA conexionista se torna mais inteligente por meio da exposição a dados e do aprendizado de padrões e relações associadas. Essa ideia se baseia na crença de que a inteligência emerge das conexões e interações de nós de processamento simples. Nos modelos conexionistas, o conhecimento não é armazenado em um local ou banco de dados centralizado. Em vez disso, é distribuído por muitos nós, e a força das conexões entre esses nós codifica a informação (SMOLENSKY, 1988). Os sistemas desse tipo baseiam-se em grandes quantidades de elementos de processamento (ou neurônios artificiais), cada um contendo unidades ponderadas (que determinam a relevância e direção da informação processada), uma função de transferência (que condensa várias entradas em uma única saída) e uma saída. A função de transferência é encarregada de avaliar múltiplas entradas e condensá-las em um valor de saída único, ao passo que cada peso determina a relevância e a direção da informação processada.

O debate sobre qual o melhor caminho entre IA simbólica e IA conexionista persiste por muitas décadas (MINSKY, 1991). Após anos de preponderância da abordagem simbólica, recentemente tem havido crescente predomínio da IA conexionista. As arquiteturas conexionistas têm demonstrado melhor desempenho em tarefas complexas, como reconhecimento de imagem, visão computacional, predição e aprendizado supervisionado. No entanto, o treinamento dessas redes requer um poder computacional significativamente maior, e as suposições feitas para sua construção podem simplificar muito os detalhes dos sistemas neurais subjacentes (BUCKNER; GARSON, 2019). Por isso, esse debate segue aceso em busca de caminhos para integrar IA simbólica e IA conexionista (MARCUS, 2001).

2.3 Aprendizado de Máquina (*Machine Learning*)

O principal propulsor da atual evolução da IA é o campo conhecido como aprendizado de máquina (***machine learning***). Os termos "inteligência artificial" (IA) e "*machine learning*" (ML), embora não sejam sinônimos, são frequentemente usados de maneira intercambiável, considerando que que ML é o subtipo de IA com maior aplicação atualmente e com maior potencial futuro. No entanto, o campo de ML tem definição mais precisa, já que não recorre necessariamente ao conceito inexato de "inteligência".

2.3.1 Definição de ML

Em linhas gerais, *machine learning* (ML) é uma família de técnicas de modelagem estatística e matemática que usa uma variedade de abordagens para aprender e melhorar automaticamente a previsão de um estado-alvo, sem programação explícita. O objetivo é permitir que as máquinas encontrem padrões, façam previsões ou executem tarefas quando encontram dados novos ou não vistos – o que é denominado generalização. As técnicas de ML essencialmente convertem dados e experiências em novos conhecimentos, geralmente na forma de modelos matemáticos. Uma vez criados, esses modelos podem então ser utilizados para realizar tarefas que, de outra forma, seriam muito complexas ou demoradas para serem desenvolvidas sem a assistência do computador (MITCHELL, 1997).

Na perspectiva da ciência da computação, *machine learning* se diferencia da programação tradicional de computadores – geralmente chamada de "sistema baseado em regras" (*rule-based system*) – em que se codificam manualmente funções que mapeiam entradas específicas para suas saídas correspondentes. Basicamente, todas as interações do computador consistem em uma entrada, uma função e uma saída. Isso pode ser representado de forma simplificada pela notação **y = f (x)**, onde y é a saída, f é a função e x é a entrada. O **processamento** é justamente a função que transforma a entrada na saída. Na abordagem tradicional de programação, as regras são deliberadamente definidas para processar as entradas de modo que produzam as saídas desejadas.

2.3.2 Como funciona o aprendizado de máquina

Em *machine learning*, o termo "aprendizado" refere-se à intenção de criar sistemas capazes de aprender com a experiência, assim como um ser humano, com intervenção externa mínima. Nesse caso, o programa procura (até encontrar) uma função que pode mapear com precisão os dados de entradas para saídas. Em seguida, essa função é empregada para processar novas entradas e produzir novas saídas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Ou seja, como na programação tradicional, um programa ainda precisa ser escrito, mas o objetivo é essencialmente diferente: aprender uma função de mapeamento precisa em vez de pré-estabelecer essa função. Portanto, apesar de muitas vezes ser antropomorfizado, o aprendizado em ML é

fundamentalmente matemático, usando algoritmos que extrapolam parâmetros de conjuntos de dados para informar a formação e os ajustes de um modelo.

No campo da IA, *machine learning* é reconhecido como a subcategoria que envolve algoritmos e modelos que permitem às máquinas aprenderem com dados e se aprimorarem automaticamente através da experiência. Essa área surgiu na década de 1970 e se beneficiou da ascensão de métodos estatísticos e probabilísticos em computação durante as décadas de 1980 e 1990. Houve também uma ênfase em soluções específicas desse domínio, em parte devido às limitações dos primeiros sistemas de IA (BISHOP, 2006). O advento de poderosas tecnologias de computação e a explosão de dados disponíveis impulsionou um grande crescimento da área de ML a partir dos anos 2000 (JORDAN; MITCHELL, 2015).

2.3.3 Paradigmas de ML

Atualmente, o campo de *machine learning* é formado por técnicas de IA que usam principalmente métodos estatísticos para criar modelos preditivos a partir do reconhecimento de padrões complexos nos dados. Identificam-se quatro paradigmas de ML:

a) aprendizado supervisionado;
b) aprendizado não-supervisionado;
c) aprendizado semissupervisionado;
d) aprendizado por reforço.

Aprendizado supervisionado (*supervised learning*) é a categoria de ML em que o modelo é treinado num conjunto de dados rotulado, no qual as respostas corretas são conhecidas, com o objetivo de fazer previsões para dados não vistos. O modelo aprende a prever a saída dos dados de entrada durante o processo de treinamento, que envolve aprender uma função que mapeia uma entrada para uma saída com base em pares de exemplos de entrada-saída. Depois que o modelo é treinado, ele pode ser usado para prever a saída de novos dados de entrada não vistos. As técnicas de aprendizado supervisionado são geralmente divididas em tarefas de regressão e classificação: regressão prevê saídas contínuas (variáveis numéricas) e classificação prevê saídas distintas ou não contínuas (variáveis categóricas) (MURPHY, 2012).

Aprendizado não-supervisionado (*unsupervised learning*), por outro lado, não depende de uma saída conhecida durante a fase de treinamento. Os dados de entrada não são

emparelhados com nenhum dado de saída correto (o que se denomina dados não-rotulados), de modo que o modelo aprende a identificar padrões e estruturas apenas com os dados de entrada. Esse tipo de aprendizado é usado quando o objetivo é descobrir padrões ocultos ou estruturas intrínsecas nos dados. As técnicas de aprendizado não-supervisionado são usadas para análise estatística de dados em tarefas como agrupamento de instâncias semelhantes e redução de dimensionalidade, em que o objetivo é simplificar as entradas sem perder muita informação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Aprendizado semissupervisionado (*semi-supervised learning*) é um intermédio entre os aprendizados supervisionado e não-supervisionado, técnica também denominada "supervisão fraca" (*weak supervision*). É um tipo de *machine learning* empregado para aproveitar condições em que há grandes quantidades de dados não-rotulados disponíveis, em combinação com conjuntos menores de dados rotulados. O algoritmo recebe algumas informações de supervisão, mas não necessariamente para todos os exemplos, e busca rotular os dados não-rotulados (ao resolver um problema de classificação) ou agrupá-los (ao resolver um problema de agrupamento). As técnicas de aprendizado semissupervisionado têm ganhado crescente relevância no treinamento de modelos de linguagem (CHAPELLE; SCHÖLKOPF; ZIEN, 2010).

Aprendizado por reforço (*reinforcement learning*) é um tipo de *machine learning* em que um agente aprende a tomar decisões interagindo com seu ambiente. O computador enfrenta uma situação semelhante a um jogo, realizando determinadas ações e observando os resultados com base em recompensas (positivas ou negativas) que recebe por suas ações. O agente aprende por tentativa e erro, seu o objetivo principal é aprender uma política – que significa uma estratégia que determina qual ação o agente deve tomar em cada circunstância para maximizar a recompensa total (SUTTON; BARTO, 2018).

As técnicas consideradas como *machine learning* tradicional são algoritmos tipicamente usados em contextos de aprendizado supervisionado. Há diversas modalidades de algoritmos nessa categoria, tais como regressão linear, regressão logística, máquinas de vetores de suporte (SVM – *support vector machines*), árvores de decisão e florestas aleatórias (*random forests*), entre outros. Essas abordagens são chamadas de tradicionais para distinguir das técnicas de aprendizado profundo (*deep learning*), o subtipo mais recente de ML.

2.3.4 Aprendizado contínuo (*continual learning*)

No campo de *machine learning*, merece destaque a abordagem conhecida como "aprendizado contínuo" (*continual learning*) ou "aprendizado incremental (*incremental learning*).

O **aprendizado contínuo** funciona a partir de um fluxo de dados não estacionário. O objetivo fundamental é resolver um problema conhecido como "esquecimento catastrófico", no qual modelos treinados em novas distribuições de dados tendem a esquecer capacidades previamente adquiridas. O processo de aprendizado ocorre continuamente, de modo que o modelo é atualizado à medida que novos dados são disponibilizados. Esse aprendizado incremental pode ocorrer: por tarefa (*task-incremental*) – aprender sequencialmente a resolver uma série tarefas distintas; por domínio (*domain-incremental*) – aprender a resolver a mesma tarefa em contextos diferentes; ou por classe (*class-incremental*) – aprender a discriminar entre um número crescente de classes (VAN DE VEN; TUYTELAARS; TOLIAS, 2022). É um campo que busca aproximar as habilidades da IA com a capacidade de aprendizado ao longo da vida que humanos e animais possuem. No processo de aprendizagem, os modelos precisam manter um equilíbrio entre a plasticidade, adaptando-se aos novos dados, e a estabilidade, retendo o conhecimento dos dados antigos (PARISI *et al.*, 2019).

Uma modalidade relevante de aprendizado contínuo é o denominado "aprendizado online" (***online learning***), empregado em situações em que os dados são gerados em função do tempo. No *online learning*, os dados ficam disponíveis sequencialmente e são usados para atualizar as previsões em cada etapa, diferentemente do que ocorre no tradicional "aprendizado em lote" (*batch learning*), no qual os dados de treinamento são fornecidos antecipadamente. Ao propiciar atualizações instantâneas do modelo para quaisquer novas instâncias de dados, é possível desenvolver modelos mais eficientes e escaláveis para tarefas de ML de grande dimensão e de alta velocidade (HOI *et al.*, 2021).

2.3.5 Aprendizado profundo (*deep learning*)

O termo **aprendizado profundo (*deep learning*)** distingue uma classe particular de sistemas de aprendizado de máquina (*machine learning*), caracterizados por sua estrutura organizacional única e capacidade de aprender funções complexas.

Em contraste com os métodos tradicionais de *machine learning*, onde o número de parâmetros é normalmente limitado pela quantidade de características de entrada, os modelos de *deep learning* (DL) superam essa restrição organizando os parâmetros em camadas hierárquicas. Essa estrutura em camadas permite uma interação complexa entre as características de entrada. Elas são repetidamente multiplicadas e combinadas, com as saídas de uma camada de parâmetros atuando como entrada para a próxima camada. Essas operações se repetem até que uma previsão seja feita. A maior interação entre as características e os parâmetros do modelo, possibilitada pela arquitetura de aprendizado profundo, aumenta significativamente a capacidade do sistema de aprender e representar funções complexas (GOODFELLOW; BENGIO; COURVILLE, 2016).

A concepção de *deep learning* origina-se do paradigma da IA conexionista. Os modelos de DL utilizam redes neurais artificiais para imitar a estrutura e função do cérebro humano. Eles empregam várias camadas ocultas, cada uma atuando como uma unidade de processamento que realiza transformações não-lineares dos dados de entrada. O número de camadas ocultas pode variar de acordo com o modelo e o problema a ser resolvido – redes neurais com mais de duas camadas ocultas são consideradas profundas, embora possa haver muitas mais. Assim, as transformações executadas em todas as camadas juntas permitem que a rede neural aprenda representações complexas dos dados (LECUN; BENGIO; HINTON, 2015).

As **redes neurais** aprendem com erros, assim como os demais sistemas de aprendizado supervisionado, ajustando os pesos de suas conexões para aprimorar gradualmente o desempenho ao longo do tempo. Nas redes neurais, isso é feito por um processo chamado **retropropagação** (*backpropagation*). Esse processo envolve a propagação de erros de volta através da rede, de modo que o peso de cada conexão pode ser ajustado para minimizar o erro. O aprendizado é um processo iterativo, no qual a rede melhora seu desempenho ao longo do tempo, ajustando continuamente os pesos de suas conexões com base nos erros que comete durante o treinamento (RUMELHART; HINTON; WILLIAMS, 1986).

As tecnologias de *deep learning* têm aplicações práticas em vários domínios devido à sua capacidade de aprender com grandes quantidades de dados e fazer previsões precisas. Destacam-se especialmente nas áreas de reconhecimento de imagem, reconhecimento de fala e processamento de linguagem natural (NLP – *natural language processing*). As principais arquiteturas de redes neurais são as redes sem realimentação (*feedforward*), a perceptron multicamadas (MLP – *multilayer perceptrons*), as redes neurais convolucionais (CNNs – *convolutional neural networks*) – arquitetura projetada para processar imagens – e as redes neurais recorrentes

(RNNs – *recurrent neural networks*) – arquitetura projetada para processar dados sequenciais, usadas para tarefas como reconhecimento de fala e tradução automática (LECUN; BENGIO; HINTON, 2015). Mais recentemente, tem ganhado muita projeção a arquitetura *transformer*, principalmente no campo de NLP (VASWANI *et al.*, 2017).

Embora as redes neurais sejam uma tecnologia em constante desenvolvimento, já se mostraram promissoras em uma ampla variedade de aplicações. Conforme continuam a evoluir, é bem provável que se tornem ainda mais eficientes e amplamente adotadas no futuro.

3 A INTELIGÊNCIA ARTIFICIAL NA SAÚDE

3.1 Evolução Histórica e Panorama Atual da IA na Saúde

A saúde foi identificada desde cedo como uma das áreas mais propícias para uso da IA. Os primeiros sistemas de assistência à decisão clínica foram concebidos e desenvolvidos por pesquisadores na década de 1950, nos primórdios do campo da IA (MILLER, 1994). A partir da década de 1970, foram criados vários programas para simular o raciocínio humano especializado, com objetivo de auxiliar médicos na formulação de hipóteses diagnósticas em casos difíceis, interpretar exames clínicos e escolher tratamentos adequados (SZOLOVITS; PATIL; SCHWARTZ, 1988). Esses programas foram aplicações significativas da então predominante IA simbólica e se tornaram bastante prevalentes nas décadas de 1980 e 1990, um período de ascensão dos sistemas de suporte à decisão clínica (CDSS – *clinical decision support systems*).

3.1.1 Sistemas de suporte à decisão clínica

Sistema de suporte à decisão clínica (CDSS) é um *software* projetado para fornecer auxílio direto à tomada de decisões clínicas a fim de aprimorar as decisões médicas com conhecimento clínico direcionado, informações do paciente e outras informações de saúde. Características de cada paciente individual são combinadas com uma base de conhecimento clínico computadorizada e avaliações ou recomendações específicas para o paciente são então apresentadas ao clínico para uma decisão (SUTTON *et al.*, 2020).

A incorporação dos CDSSs da primeira geração (décadas de 1980 e 1990) na assistência à saúde ficou bem restrita. Esses sistemas baseados em regras demonstraram as limitações dessa abordagem tecnológica: principalmente, o alto custo de manutenção e a necessidade de atualizações regulares que exigiam revisões de autoria humana. Era muito desafiador codificar interações complexas entre diferentes partes do conhecimento fornecidas por diferentes especialistas. Além disso, o desempenho dos sistemas era circunscrito pela precisão do conhecimento médico pré-existente (YU; BEAM; KOHANE, 2018).

3.1.2 *Machine learning* e *deep learning* na saúde

Nesse contexto, o advento do campo de *machine learning* (ML) foi muito bem recebido na área da saúde e na medicina (OBERMEYER; EMANUEL, 2016). As técnicas de ML permitem o desenvolvimento de sistemas que podem detectar padrões até então desconhecidos nos dados, sem a necessidade de especificar regras de decisão para cada tarefa específica ou contabilizar interações complexas entre recursos de entrada. A disponibilidade de grandes quantidades de dados na assistência médica, aliada ao aumento exponencial na capacidade computacional, impulsionou uma crescente expectativa em torno da incorporação substancial da IA na saúde, sobretudo a partir da década de 2000 (RAJKOMAR; DEAN; KOHANE, 2019).

O atual grande aumento no interesse pela IA na área da saúde é atribuído à aplicação bem-sucedida de técnicas de *deep learning* (DL) em vários domínios (HINTON, 2018). Considera-se o ano de 2012 como marco para a melhora significativa dos sistemas de DL no desempenho em tarefas de classificação de imagens e para o consequente impulso no uso dessas ferramentas em diferentes setores (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). Por isso, áreas que se baseiam na identificação de padrões em imagens, como radiologia, patologia e dermatologia, foram pioneiras no uso de DL na saúde e na medicina (TOPOL, 2019b).

Desde então, o uso de IA na saúde e medicina tem crescido rapidamente, expandindo-se para além da abordagem de visão computacional em muitas outras áreas, como no uso do processamento de linguagem natural (NLP) para análise de dados de registros eletrônicos de saúde e de técnicas de aprendizado por reforço (*reinforcement learning*) na cirurgia assistida por robótica (ESTEVA *et al.*, 2019). Além disso, essas técnicas têm demonstrado um desempenho promissor na realização de previsões úteis e precisas em diferentes cenários clínicos (RAJKOMAR; OREN; *et al.*, 2018).

Mais recentemente, uma nova fronteira se abriu com a explosiva ascensão dos **modelos de fundação** (*foundation models*) a partir da década de 2020, dentre os quais se destacam os grandes modelos de linguagem (LLMs – *large language models*). O cenário está em constante mudança e evoluindo muito rapidamente (ACOSTA *et al.*, 2022; HAUG; DRAZEN, 2023; MOOR *et al.*, 2023).

A IA vem se mostrando cada vez mais promissora nos mais diversos âmbitos das áreas da saúde e da medicina. Merecem destaque as aplicações em diagnósticos baseados em imagens, medicina personalizada, análise preditiva, desenvolvimento de medicamentos e em atividades administrativas.

3.1.3 Diagnósticos baseados em imagens

As técnicas de *deep learning* (DL) têm demonstrado proficiência crescente em **análise e interpretação de imagens médicas**. Os modelos baseados em IA são capazes de detectar alterações em radiografias, mamografias, tomografias computadorizadas, ressonâncias magnéticas, dentre outras modalidades de exames de imagem, muitas vezes igualando ou superando a precisão de especialistas humanos (LITJENS *et al.*, 2017).

A radiologia é reconhecida como a área da medicina que testemunhou de forma pioneira o potencial das ferramentas de IA, embora o impacto real ainda seja esperado para o futuro (RAJPURKAR; LUNGREN, 2023).

A patologia é outra área em que também se tem visto avanços consideráveis, na leitura e identificação de padrões em imagens de anatomia patológica com uso de sistemas de DL (HARRISON *et al.*, 2021).

Das áreas de medicina clínica e cirúrgica, a capacidade aumentada de análise de imagens tem-se revelado muito útil em dermatologia e em oftalmologia. Por exemplo, já há modelos de DL bem desenvolvidos para identificação de lesões suspeitas de câncer de pele e para rastreamento e avaliação de fundo de olho e de diferentes tipos de lesões de retina (YU; BEAM; KOHANE, 2018).

3.1.4 Medicina personalizada

Técnicas de *machine learning* (ML) são utilizadas para analisar as informações genéticas, estilo de vida e outros dados de saúde dos pacientes para fornecer recomendações de tratamento personalizadas (JIANG *et al.*, 2017).

Essa abordagem é conhecida como **medicina de precisão** e tem o objetivo de usar a biologia individual – em vez da biologia populacional – em todos os estágios da jornada médica de um paciente. Assim, é possível adequar os tratamentos a pacientes individuais, melhorando sua eficácia e reduzindo os efeitos colaterais. Isso é feito com base no perfil da doença, nas informações diagnósticas ou prognósticas e na resposta ao tratamento. Os modelos de ML analisam variações genômicas e fatores contribuintes do tratamento médico, como idade, sexo, raça, histórico familiar, perfil imunológico, perfil metabólico, microbioma e vulnerabilidade ambiental (BOHR; MEMARZADEH, 2020).

Há recentes avanços no uso de IA em medicina molecular, tais como no sequenciamento genético de DNA e RNA e na medição de alta dimensão de proteínas e metabólitos, que começam a ser integrados nas chamadas aplicações **multiômicas** (GOMES; ASHLEY, 2023).

3.1.5 Análise preditiva

Algoritmos de ML tradicional e, mais recentemente, de DL são usados para prever tendências e desfechos nas mais diversas situações clínicas com base em dados de saúde em grande escala, tanto históricos como colhidos em tempo real.

Essas ferramentas têm inúmeras aplicações na previsão de progressão de doenças, taxas de readmissão de pacientes e os riscos potenciais à saúde, permitindo intervenções proativas (WANG; KUNG; BYRD, 2018). Como os modelos de ML são projetados para fazer previsões a partir de padrões estatísticos em vez de seguir o conhecimento médico especializado, os sistemas de apoio à decisão clínica (CDSS) vêm sendo desenvolvidos cada vez mais com técnicas "não baseadas em conhecimento" (SUTTON *et al.*, 2020).

Além disso, técnicas de ML podem ser muito úteis no âmbito coletivo, para auxiliar na tomada de decisões em saúde pública e no planejamento para alocação de recursos humanos e financeiros (BATES *et al.*, 2014). Um relevante exemplo disso é o uso de ferramentas de IA para identificar e rastrear surtos de doenças infecciosas e monitorar estratégias de mitigação (BROWNSTEIN *et al.*, 2023).

3.1.6 Descoberta e desenvolvimento de medicamentos

Modelos de ML, principalmente os baseados em aprendizado por reforço e em redes neurais, têm sido usados de maneira muito promissora para prever as propriedades de medicamentos em potencial e otimizar seu desenvolvimento (VAMATHEVAN *et al.*, 2019).

Essas ferramentas possibilitam obtenção de informações químicas de grandes bancos de dados de compostos e fazem previsões relevantes sobre muitos elementos do processo de descoberta de novas drogas com base na estrutura química e nas características dos seus possíveis alvos biológicos, tais como predições sobre eficácia, atividade biológica, seletividade, perfis farmacocinéticos e toxicológicos de novas drogas (CHEN, H. *et al.*, 2018). Nesse sentido, a IA tem potencial para reduzir significativamente o tempo e o custo associados aos métodos tradicionais de descoberta de medicamentos (NATURE, 2023a).

Nesse cenário, uma importante revolução vem da aceleração extraordinária do progresso para a **previsão da estrutura tridimensional de proteínas**. Grandes incrementos nesse campo têm sido obtidos desde 2021, a partir do desenvolvimento do modelo de IA capaz de prever com precisão atômica quase todas as estruturas proteicas conhecidas presentes em organismos (JUMPER *et al.*, 2021). A velocidade desse avanço tem sido tão grande que, cerca de um ano depois, esse mesmo modelo já se mostrou capaz de prever praticamente todas as proteínas existentes na natureza e até mesmo inventar novas (CALLAWAY, 2022). Abre-se uma nova era para as ciências naturais. A forma tridimensional das proteínas é o que determina sua função nas células e serve de base para o desenvolvimento de muitos medicamentos.

3.1.7 Administração em saúde e assistência médica

Técnicas de ML vêm sendo empregadas com sucesso para otimizar inúmeras tarefas administrativas, como agendamento de consultas, gerenciamento de registros eletrônicos, processamento de solicitações de seguros, faturamento, e muitas outras (BATES *et al.*, 2014).

Ao analisar grandes quantidades de dados e automatizar processos, essas ferramentas podem melhorar a eficiência e também liberar os profissionais de saúde para se concentrarem mais nas atividades assistenciais (TOPOL, 2019a). Espera-se que esse seja o campo no qual uso

da IA na saúde seja mais rapidamente incorporado e que tenha um grande impacto na redução de custos dos sistemas de saúde (SAHNI *et al.*, 2023).

3.1.8 Outras aplicações

Além dessas, há muitas outras aplicações da IA na saúde atualmente em desenvolvimento e com potencial de ampliação. Dois exemplos: 1) em **cirurgia robótica**, robôs com sistemas de ML podem realizar procedimentos complexos com alta precisão e técnicas minimamente invasivas, aprimorando continuamente seu desempenho com base no aprendizado a partir de dados de cirurgias anteriores; 2) **dispositivos vestíveis** (*wearables*) integrados a modelos de ML e de reconhecimento de fala vêm sendo associados a telemedicina para monitoramento remoto, suporte e reabilitação de pacientes (ROSKI *et al.*, 2022).

3.2 Desenvolvimento de Modelos de IA na Saúde

Apesar das especificidades inerentes à modalidade de aplicação e ao contexto de uso, existem elementos comuns no processo de criação dos diferentes tipos de sistemas de IA utilizados na área da saúde. Em particular, algumas etapas devem ser observadas no desenvolvimento e na validação de modelos de *machine learning* (ML) destinados a diagnosticar, prognosticar ou recomendar ações voltadas aos cuidados preventivos e assistência à saúde.

3.2.1 Seleção do problema e definição da tarefa

A etapa inicial no desenvolvimento de modelos de ML para assistência à saúde é a **seleção do problema**. É essencial que se defina e caracterize o problema a ser abordado, o que é feito identificando uma questão clínica específica que se pretenda resolver com uso de IA e em seguida avaliando se ela pode ser resolvida e se é útil resolvê-la usando IA. Dessa forma, antes de iniciar o desenvolvimento de um modelo, devem ser consideradas questões como “quem é o usuário-alvo?”, “quais serão as intervenções adotadas com base nos resultados?”, “quais são

os mecanismos de execução dessas intervenções?", "qual é o risco de falha e de eventos adversos?", "qual a capacidade de intervenção face aos recursos existentes?" e "qual é a mudança desejada após a intervenção?" (LIU, H. *et al.*, 2022). O objetivo é avaliar a utilidade, a viabilidade, os custos, os desafios de implementação, a possibilidade de aceitação clínica e a manutenção a longo prazo de uma solução baseada em IA para cuidados de saúde.

Outro ponto importante nessa fase preliminar é **definir a tarefa de previsão**. A máquina pode aprender com base em experiências humanas – assumindo que a experiência possa ser quantificada e expressa em termos de uma função matemática – ou pode ser autorizada a aprender extraindo percepções anteriormente desconhecidas (CHEN, P. C.; LIU; PENG, 2019). Esse é o momento de estabelecer a abordagem de aprendizado a ser empregada: supervisionada, não supervisionada ou por reforço.

3.2.2 Conceitos fundamentais em aprendizado supervisionado

O maior valor prático de ML atualmente advém do aprendizado supervisionado. Esse paradigma é amplamente empregado em aplicações nas quais dados históricos são usados para prever eventos futuros, que é o principal uso da IA na saúde hoje (BEAM; KOHANE, 2018). As principais definições no desenvolvimento desses modelos (LIU, H. *et al.*, 2022):

a) **conjunto de dados de treinamento** (*training dataset*): conjunto de dados usado para aprender os parâmetros de um modelo;

b) **conjunto de dados de validação** (*validation dataset*): conjunto de dados usado para ajustar os hiperparâmetros de um modelo;

c) **hiperparâmetro** (*hyperparameter*): configuração externa ao modelo cujo valor é usado como configuração para o algoritmo de aprendizado;

d) **conjunto de dados de teste** (*test dataset*): conjunto de dados independente do conjunto de dados de treinamento, mas que segue a mesma distribuição dele;

e) **sensibilidade** (*sensitivity*): proporção de positivos reais que são identificados corretamente em uma classificação binária – também chamada de taxa de verdadeiros positivos, *recall* ou probabilidade de detecção;

f) **especificidade** (*specificity*): proporção de negativos reais que são identificados corretamente em uma classificação binária – também chamada de taxa de verdadeiros negativos;

g) **precisão** (*precision*): proporção de positivos previstos que são verdadeiros positivos – também chamada de valor preditivo positivo;

h) **acurácia** (*accuracy*): proporção de previsões corretas em relação ao total de previsões realizadas;

i) **curva ROC** (*Receiver Operating Characteristic*): gráfico gerado pela plotagem da taxa de verdadeiros positivos (sensibilidade) em relação à taxa de falsos positivos (1 - especificidade) em diferentes limiares de discriminação entre positivo e negativo;

j) **curva PR** (*Precision-Recall*): gráfico gerado pela plotagem da precisão em relação à sensibilidade em diferentes limiares de discriminação entre positivo e negativo.

3.2.3 Seleção e pré-processamento de dados

A etapa subsequente após a seleção do problema é a **criação dos conjuntos de dados**. Primeiramente, são coletados os dados brutos, que é um conjunto total de dados que representam situações em que a tarefa pretendida já foi realizada. Em seguida é executado o **pré-processamento de dados**, que é processo de limpeza, transformação e padronização dos dados brutos para criar um conjunto de dados adequado para uso em algoritmos de *machine learning* (RAJKOMAR; DEAN; KOHANE, 2019).

No desenvolvimento de modelos na área da saúde, podem ser usados dados estruturados ou dados não-estruturados. **Dados estruturados** são informações organizadas de maneira predefinida, geralmente em formato tabular com linhas e colunas (por isso, também podem ser chamados de dados tabulares) como dados demográficos de pacientes, resultados de exames laboratoriais, códigos de diagnósticos e outros dados tipicamente armazenados em registros de prontuários eletrônicos. **Dados não-estruturados**, por outro lado, são informações que não estão em conformidade com um formato ou modelo específico, como imagens ou relatórios de texto escritos em linguagem natural (ESTEVA *et al.*, 2019).

3.2.4 Treinamento de modelos

Como já mencionado, no aprendizado supervisionado os modelos são treinados usando dados rotulados. Os dados são organizados em **instâncias**, que são **pares individuais de entrada e saída** (também chamados de exemplos), em que as entradas são denominadas características ou **recursos** (*features*) e as saídas, **rótulos** (*labels*). O conjunto total de dados é dividido nos subconjuntos de treinamento, validação e teste. O modelo é treinado apenas no conjunto de treinamento, para discernir a função que vincula as entradas às saídas. O conjunto de validação, mantido separado do processo de treinamento, é usado para avaliar o desempenho de generalização do modelo conforme ele evolui e auxilia os desenvolvedores na tomada de decisões críticas sobre variáveis que influenciam o processo de treinamento – os chamados hiperparâmetros. Por fim, o conjunto de teste, reservado até a conclusão do processo de desenvolvimento, é usado para verificação final do desempenho do modelo em exemplos completamente inéditos (CHEN, P. C.; LIU; PENG, 2019; RAJKOMAR; DEAN; KOHANE, 2019).

Em termos matemáticos, treinar um modelo significa minimizar a diferença entre a saída da função do algoritmo e o rótulo verdadeiro, para cada amostra no conjunto de treinamento. Essa diferença é chamada de "perda". Se o algoritmo for mal treinado, ele terá uma perda alta. Se for bem treinado, a diferença entre o rótulo verdadeiro e a saída do algoritmo será pequena em média e, portanto, a perda resultante também será pequena. O modelo consegue isso ajustando seus **parâmetros** (também chamados de pesos) que são coeficientes numéricos usados para realizar operações nos dados de entrada (por exemplo, soma ou multiplicação numa equação linear) de modo a minimizar a perda, comumente utilizando um algoritmo específico, como o "método do gradiente" (*gradient descent*). Quando um modelo prevê valores numéricos reais, está abordando um problema de regressão, concentrando-se na identificação de resultados contínuos. Por outro lado, se os rótulos utilizados forem essencialmente números que funcionam como identificadores de categoria, o modelo está resolvendo um problema de classificação. Em uma tarefa de classificação, a predição do modelo é apresentada como uma probabilidade de que um determinado conjunto de características pertença a cada categoria (MURPHY, 2012).

O processo de seleção de uma arquitetura de modelo e de treinamento do modelo envolve essencialmente o balanceamento entre o ajuste insuficiente ou "subajuste" (*underfitting*) e o ajuste excessivo ou "sobreajuste" (*overfitting*). Esse balanceamento é conhecido como "compensação entre viés e variância" (*bias-variance trade-off*), em que viés representa a

diferença entre as previsões médias e os valores corretos dos rótulos e variância significa a medida do quanto as previsões do modelo variam em torno de sua média (BISHOP, 2006). O ***underfitting*** ocorre quando um modelo de baixa capacidade é usado em relação à complexidade do problema e ao tamanho do conjunto de dados e, portanto, não é capaz de prever dados novos (generalização) por não ter conseguido capturar a relação entre as entradas e as saídas. O ***overfitting*** acontece quando o modelo se ajusta tanto aos dados de treinamento que age como se tivesse simplesmente "memorizado" esses dados, de modo que também é incapaz de fazer a generalização em dados não vistos anteriormente (CHEN, P. C.; LIU; PENG, 2019).

Esse processo de treinamento em aprendizado supervisionado adota uma estrutura cíclica. Inicia-se com a configuração do ambiente de treinamento, seleção de hiperparâmetros e inicialização do modelo com uma função aleatória. O modelo, então, aprende com os exemplos do conjunto de treinamento e o desempenho é avaliado no conjunto de validação. Caso o modelo supere desempenhos anteriores, essa versão é mantida. Esse ciclo se repete até que o desempenho no conjunto de validação atinja um platô. Esse processo é repetido com várias configurações de hiperparâmetros para desempenho ideal, pois diferentes hiperparâmetros podem resultar em modelos distintos. Finalmente, uma vez que o desempenho do modelo no conjunto de validação é satisfatório, ele é testado no conjunto de teste (GOODFELLOW; BENGIO; COURVILLE, 2016; RASCHKA, 2020).

3.2.5 Avaliação de modelos

Avaliar o desempenho do sistema de ML é uma etapa crítica no processo de desenvolvimento. A escolha da métrica de avaliação deve refletir o uso pretendido do modelo.

Métricas como **sensibilidade, especificidade e precisão** são dependentes de limiar – para que sejam calculadas, o limiar de discriminação entre positivo e negativo deve estar previamente definido. Essas métricas têm mais importância na área da saúde em comparação com estudos básicos de ML, devido à natureza binária de tomada de decisão das aplicações clínicas. Há um *trade-off* entre as medidas de sensibilidade, especificidade e precisão que precisa ser resolvido com base na situação clínica em questão. A opção entre sensibilidade e precisão representa bem esse ponto. Por exemplo, se o modelo for usado como uma ferramenta de triagem, em que previsões que sejam falsos negativos podem ter sérias consequências, a sensibilidade é

uma métrica apropriada – pois um classificador altamente sensível pode descartar com segurança uma doença quando seu resultado é negativo. Por outro lado, se o modelo for usado para diagnóstico, resultados falsos positivos podem levar a tratamentos desnecessários, de modo que a precisão é uma escolha melhor – já que maior precisão significa maior probabilidade de que os positivos previstos sejam verdadeiros positivos (WYNANTS *et al.*, 2019).

A **acurácia** é uma métrica amplamente usada em validação de modelos de ML, mas que precisa ser vista com muita cautela na assistência à saúde. Isso porque é frequente que sejam empregados conjuntos de dados altamente desequilibrados, nos quais os rótulos não são distribuídos igualmente (número de rótulos verdadeiros positivos é muito diferente do número de rótulos verdadeiros negativos). Ou seja, o que se pretende é fazer predição de um desfecho relativamente pouco comum – uma doença ou outra condição clínica, mesmo que prevalente, é um evento proporcionalmente raro considerando a população geral. Por isso, os sistemas de ML na saúde são particularmente suscetíveis ao chamado "paradoxo da acurácia": modelos com alta acurácia podem falhar em captar informações cruciais numa dada tarefa de classificação (VALVERDE-ALBACETE; PELÁEZ-MORENO, 2014). No contexto da saúde, um modelo com alta acurácia, que prevê corretamente todos os negativos, mas não é capaz de discriminar os positivos, não tem aplicação prática.

A **curva ROC** é o gráfico em que a taxa de verdadeiros positivos (sensibilidade) do modelo é colocada no eixo y e a taxa de falsos positivos (1 - especificidade), no eixo x. Ela representa visualmente a relação entre sensibilidade e especificidade, permitindo avaliar o desempenho do modelo em toda a sua faixa de operação (ou seja, em todos os limiares de discriminação entre positivo e negativo). Em geral há um *trade-off* entre sensibilidade e especificidade, pois conforme se aumenta uma normalmente se diminui a outra. Por isso, a análise da curva ROC possibilita a escolha do melhor ponto operacional do classificador, que é o limiar de discriminação entre positivos e negativos a ser usado numa aplicação específica.

Um dos principais indicadores para aferir o desempenho dos modelos preditivos de ML na saúde é a **área sob a curva ROC**, que é uma métrica independente de limiar. A área sob a curva ROC é um valor entre 0,5 (classificador totalmente aleatório, que acerta 50% das previsões) e 1,0 (classificador prefeito, que acerta 100% das previsões) que fornece um número único que resume a eficácia de um determinado modelo (LIU, H. *et al.*, 2022). Vale destacar que em conjuntos de dados altamente desequilibrados a curva PR (*Precision-Recall*) fornece uma base melhor para comparar classificadores, a partir de uma lógica de análise bem semelhante à utilizada para a curva ROC (SAITO; REHMSMEIER, 2015).

A avaliação dos desempenhos dos modelos orienta empiricamente a escolha daquele mais adequado para o problema específico que se pretende resolver. Mas não existe um modelo ou um tipo de algoritmo universalmente melhor, pois um conjunto de suposições que funciona bem em um domínio pode funcionar mal em outro. Por isso, na prática é preciso desenvolver muitos tipos diferentes de modelos, treinados em muitos tipos diferentes de algoritmos, para tentar abarcar a grande variedade de dados do mundo real. Essa propriedade é conhecida como teorema "não há almoço grátis" em ML (MURPHY, 2012).

3.3 Desafios para Implementação da IA na Saúde

Os avanços recentes no desenvolvimento de modelos de ML específicos para a área da saúde são considerados preliminares. A incorporação das aplicações de IA em ambientes clínicos de forma ampla é vista como promessa para o futuro (HAUG; DRAZEN, 2023).

Identificam-se quatro etapas essenciais no processo de implementação de soluções de IA na saúde (SENDAK *et al.*, 2020):

a) *design* e desenvolvimento (identificação do problema certo para resolver);
b) avaliação e validação do modelo;
c) difusão e dimensionamento do produto (aplicação em escala);
d) manutenção e monitoramento contínuos.

Até o começo da década de 2020, ainda há poucos exemplos de uso efetivo de IA na prestação de cuidados de saúde e esparsas evidências de que processos ou resultados tenham sido aprimorados quando ferramentas de IA são implantadas (HE *et al.*, 2019). Existem muitos desafios a serem enfrentados para alcançar implementação prática dessa tecnologia na assistência à saúde.

3.3.1 Escolha dos dados

Em geral, dados de saúde são bastante heterogêneos e provenientes de diversas fontes, como registros de prontuários eletrônicos, imagens e exames diagnósticos, genômica, entre outros, mesclando dados estruturados e não estruturados. A **escolha e o pré-processamento dos**

dados são tão ou mais relevantes do que a escolha do algoritmo do modelo – vale a máxima "entra lixo, sai lixo" (KILKENNY; ROBINSON, 2018). Um dos mais importantes fatores que dificultam o desenvolvimento de soluções de IA para a saúde é a escassez de grandes volumes de dados rotulados de boa qualidade, necessários para treinamento de modelos de ML, principalmente os baseados em algoritmos de DL (SHAH, P. *et al.*, 2019)

Nessa etapa, é fundamental observar que modelos preditivos baseados em dados de saúde são suscetíveis a vieses causados por erros de classificação de condições clínicas e por seleções não representativas da população-alvo, especialmente se os dados forem coletados de prontuários médicos (GIANFRANCESCO *et al.*, 2018).

Também é essencial reconhecer que os dados de saúde são gerados em um ambiente bastante dinâmico, com mudanças frequentes nas populações e nas práticas clínicas ao longo do tempo. A capacidade preditiva dos modelos exige vigilância constante do desempenho e sobretudo uso de dados recentes, pois a vida útil de dados clínicos pode ser estimada em meses (CHEN, J. H. *et al.*, 2017).

3.3.2 Utilidade clínica e viabilidade

Outro ponto crucial é admitir que as métricas utilizadas na validação de modelos de ML frequentemente falham em refletir relevância clínica (KELLY *et al.*, 2019). Avaliar a **utilidade clínica** é definir e caracterizar o problema a ser abordado pelo sistema de IA e determinar se esse problema pode ser resolvido (ou vale a pena resolver) usando IA. Não há indicador que possa por si só sintetizar os atributos necessários de um modelo para aplicação na assistência à saúde. Por isso, as métricas precisam ser apresentadas em composições variadas: área sob a curva ROC (ou área sob a curva PR), sensibilidade e especificidade em um ponto operacional, valores preditivos positivos e negativos (como já mencionado, acurácia muitas vezes tem menor aplicação prática), entre outras formas de apresentação.

Algumas métricas distintas são usadas visando avaliar utilidade clínica. O "número necessário para tratar" (NNT – *number needed to treat*) expressa o número de pacientes que precisam ser tratados para que um paciente se beneficie (ou o número de verdadeiros positivos em que se precisa tomar medidas ou tratar para prevenir mais um evento). O "número necessário para causar danos" (NNH – *number needed to harm*) representa o número de pessoas que

receberam determinada intervenção que levaria apenas uma pessoa a ser prejudicada (ou o número de pessoas que precisam ser tratadas para produzir mais um evento adverso) – compara o aumento absoluto do risco de resultados indesejáveis em vez de olhar para os resultados desejáveis (LAUPACIS; SACKETT; ROBERTS, 1988).

Mesmo assim, as métricas em si mesmas não capturam a utilidade clínica do sistema de IA. Ou seja, se (e como) o uso do modelo pode levar a alterações benéficas no atendimento de pacientes. Para isso, têm sido propostas técnicas como a "análise da curva de decisão", que busca quantificar o benefício líquido de utilizar um modelo para orientar ações subsequentes (VICKERS; ELKIN, 2006). Uma análise de curva de decisão considera a probabilidade limite de um evento em que os custos relativos de previsões falso-positivas e falso-negativas são levados em consideração para derivar o benefício líquido do modelo em diferentes limiares de discriminação. A ideia é que sejam assimiladas as características do ambiente assistencial na avaliação do desempenho de um modelo para compreender se ele será útil dadas as restrições existentes nesse ambiente. Nesse sentido, características como o número de ações que a equipe de atendimento pode realizar, o custo e a eficácia presumida dessas ações, e a chance de o paciente seguir a ação recomendada precisam ser consideradas durante o treinamento do modelo – e não depois que o modelo de melhor desempenho já tiver sido selecionado (SHAH, N. H.; MILSTEIN; BAGLEY, 2019).

Isso propicia ferramentas para que aplicações de IA sejam concebidas, desenvolvidas e validadas considerando o alinhamento com o estado clínico desejado, definindo **pares de predição-ação**. As predições feitas pelos modelos não devem ser apenas melhorias no conhecimento acerca do problema, mas devem estar associadas a intervenções específicas que demonstrem melhorar resultados relevantes. É possível que em determinadas situações, dado o custo, a complexidade logística e a eficácia da ação, não existam zonas operacionais viáveis para que o sistema de IA tenha utilidade clínica. Os pares de predição-ação precisam ser observados ao longo de todo o processo de implementação dos modelos, já que apenas demonstrar alta capacidade preditiva pode não garantir melhores resultados se não houver ações eficazes a serem tomadas a partir das previsões (FIHN *et al.*, 2022).

Em suma, um dos principais desafios na implementação de soluções de IA na assistência à saúde é encontrar mecanismos para avaliar utilidade, viabilidade e impacto clínico dos sistemas de ML. Considerar o que é necessário para implantação dos modelos no ambiente clínico e o efeito geral que eles podem ter nos cuidados clínicos e nos padrões de atendimento.

3.3.3 Generalização

Um bom desempenho de um sistema de ML em um pequeno conjunto de dados de validação não significa necessariamente que haverá generalização para outras populações em diferentes contextos clínicos. A discrepância entre o desenvolvimento cientificamente sólido de um modelo de IA e seu uso eficaz no mundo real é chamado de "abismo da IA" (KEANE; TOPOL, 2018).

Há várias razões possíveis para esse fenômeno. A causa mais direta é que **a distribuição de novos dados pode ser diferente daquela na qual o algoritmo foi treinado**. Mas uma mudança no conjunto de dados também pode acontecer quando não se consideram alterações nas ações e práticas clínicas ao longo do tempo – que podem ser provocadas pela própria introdução do modelo preditivo – ou por mudanças nas populações-alvo (FINLAYSON *et al.*, 2021). Outra possibilidade é modelo acidentalmente ajustar-se com base em **fatores de confusão** desconhecidos (chamados de "ruídos"), em vez encontrar os sinais verdadeiros.

Atualmente, grande parte dos sistemas de IA ainda não conseguem alcançar generalização confiável, o que pode comprometer a aplicabilidade clínica. A avaliação adequada do desempenho clínico e da generalização no mundo real requer validação externa usando conjuntos de dados de tamanho adequado coletados de instituições diferentes daquelas que forneceram os dados para o treinamento do modelo (KELLY *et al.*, 2019).

Por outro lado, isso pode indicar que a implementação de modelos de ML precise se concentrar naqueles ambientes e situações específicas em que tenham demonstrada utilidade clínica (FUTOMA *et al.*, 2020).

3.3.4 Vieses

Os modelos de ML estão sujeitos a vieses ao longo de todo o ciclo de vida. A possibilidade de causar consequências indesejáveis e injustiças é um importante desafio para a aplicação da IA na assistência à saúde. Podem-se identificar diferentes tipos de viés: viés histórico, viés de representação, viés de medição, viés de agregação, viés de avaliação e viés de implementação (SURESH; GUTTAG, 2021).

O **viés histórico** surge quando as interpretações de um modelo são consideradas desiguais devido à influência de seu estado passado ou atual, o que é contrário às normas e valores sociais. Refere-se à tendência de tomar decisões com base em preconceitos estabelecidos ou crenças preconcebidas. Por exemplo, é necessário observar que dados históricos de saúde são tipicamente dominados por dados de homens brancos.

O **viés de representação** (ou viés de amostragem), se manifesta quando o conjunto de dados utilizado para treinar um sistema de IA falha em espelhar a distribuição real da população para a qual o sistema foi projetado. Isso acontece quando certos segmentos da população-alvo não são adequadamente representados no conjunto de dados de treinamento.

O **viés de medição** aparece quando há ruídos nos dados distribuídos de forma desigual entre diferentes grupos, levando a disparidades no desempenho. É possível que em algumas situações os únicos recursos e rótulos acessíveis e mensuráveis sejam substitutos ruidosos para a variável real de interesse. Embora os dados não possam ser alterados e alguns vieses históricos possam estar inerentemente ligados aos dados, é fundamental ter consciência desse problema para buscar estratégias de mitigação.

O **viés de agregação** surge durante o desenvolvimento do modelo, quando são feitas tentativas de combinar populações com distribuições distintas da variável que está sendo estudada. É necessário reconhecer que uma abordagem “de tamanho único” não é aplicável ao desenvolvimento de modelos e que esse tipo de situação requer modelos separados para populações distintas ou a inclusão de variáveis demográficas no modelo para compensar as variações sistemáticas.

O **viés de avaliação** ocorre durante a validação e o ajuste do modelo quando os dados de teste não são representativos da população final que utilizará as soluções de IA. Uma solução para esse problema é a validação externa do modelo de IA em diferentes dados não vistos selecionados da população-alvo. Também pode ocorrer se as métricas de desempenho forem usadas incorretamente, por isso é aconselhável o uso de métricas de avaliação completas e detalhadas.

O **viés de implementação** surge quando o modelo é usado incorretamente ou seus resultados são mal interpretados. Ou seja, se o uso pretendido do modelo se desviar de seu uso real. Refere-se à interação entre a sociedade e a solução de IA.

3.3.5 Opacidade

Outro desafio significativo para implementação de modelos de ML na saúde está relacionado à transparência dos algoritmos: o **problema da opacidade**. Os algoritmos usados em ML (sobretudo os baseados em DL) são frequentemente chamados de "caixas pretas" devido às suas complexas estruturas internas e processos de tomada de decisão. Como os modelos usam inúmeras variáveis para chegar a um determinado resultado, as representações matemáticas são na maior parte das vezes ininteligíveis para os humanos.

A opacidade tem relação com essa dimensão de "caixa preta" (*black box*), mas não só. Há também a opacidade imposta por sigilo corporativo ou de estado (intencional) – pois o compartilhamento de códigos ou conjuntos de dados específicos pode revelar segredos comerciais ou divulgar dados confidenciais do usuário – e a opacidade decorrente do "analfabetismo técnico" dos usuários (BURRELL, 2016). Essa opacidade pode ser muito problemática na área da saúde, pois os padrões éticos e legais exigem uma justificativa clara para as decisões clínicas.

Por isso, a dimensão da transparência algorítmica e a questão da aplicação de modelos interpretáveis ou explicáveis são elementos centrais na ética e na regulação da IA na saúde (BABIC *et al.*, 2021a), como aprofundado adiante.

3.3.6 Interação humano-máquina

A interação entre humanos e máquinas desempenha um papel significativo na implementação de soluções IA na assistência à saúde. O uso de sistemas de suporte à decisão clínica (CDSS) é altamente afetado por fatores organizacionais, como recursos, pessoal, habilidades, treinamento, cultura, fluxo de trabalho e processos (SUTTON *et al.*, 2020). Por isso, o desempenho de modelos de ML nesse contexto é bastante influenciado por elementos muito complexos, como a mudança de comportamento das pessoas pela introdução do *software* no cenário de trabalho, a capacidade dos usuários interpretarem as saídas dos modelos e o nível de confiança nas recomendações dadas pela máquina (EMANUEL; WACHTER, 2019).

Os profissionais de saúde podem resistir à adoção de soluções baseadas em IA em ambientes clínicos por diferentes razões, como serem céticos em relação ao seu desempenho, não

terem o treinamento necessário para usar essas ferramentas com eficiência ou se sentirem ameaçados em sua autonomia profissional. Além de educação de profissionais e pacientes, o desenvolvimento de modelos de ML que possam ser interpretados – ou ao menos razoavelmente compreendidos – e a reavaliação de processos de trabalho são elementos a serem considerados para que haja confiança e viabilidade na implementação da IA na saúde (YU; KOHANE, 2019).

Por outro lado, os profissionais de saúde também podem estar sujeitos ao **viés da automação**, que é a tendência de favorecer decisões tomadas por sistemas automatizados em detrimento daquelas feitas por humanos, mesmo quando as decisões automatizadas estão erradas (GODDARD; ROUDSARI; WYATT, 2012). Esse fenômeno pode ser particularmente grave à medida que os sistemas de IA apresentarem desempenho progressivamente melhor, pois isso pode levar ao excesso de confiança e potencial negligência no julgamento humano, algo que evidentemente pode ter sérias implicações na assistência à saúde.

Mesmo que um sistema de IA seja projetado para apenas dar recomendações, em vez de realizar tarefas de diagnóstico e tratamento, pode haver consequências indesejadas na interação com os usuários, como viés de confirmação e fadiga de alerta (PHANSALKAR *et al.*, 2013).

Uma estratégia para mitigar esses riscos é usar o chamado “modo silencioso” na implementação de sistemas de IA na assistência à saúde. No **modo silencioso**, o produto AI é implantado e as previsões são feitas em tempo real, mas nenhuma ação é tomada a partir das previsões (SENDAK *et al.*, 2020). Essa modalidade de avaliação serve para observar se o usuário interpreta corretamente as saídas dos modelos e se a aplicação é feita de forma adequada e na população certa. Abordagem que pode ser muito útil para organizar fluxos de trabalho e configurações de produtos, testando a usabilidade da interface e o efeito de soluções de IA na tomada de decisões clínicas (WIENS *et al.*, 2019).

3.3.7 Monitoramento contínuo

O monitoramento dos sistemas de IA tem como principal escopo garantir segurança e eficácia durante todo o seu ciclo de vida. Mesmo que modelos de *machine learning* adequadamente treinados e avaliados tenham integração bem-sucedida na prática clínica, eles permanecem sujeitos a deterioração de desempenho ao longo do tempo. Podem ocorrer variações previsíveis (embora inevitáveis), quando há variabilidade inerente aos dados usados pelo modelo,

e variações decorrentes de alguma mudança inesperada no sistema, quando ocorrem alterações na distribuição do conjunto de dados. Monitorar esses sistemas envolve atividades complexas de controle estatístico, de tal forma que as eventuais variações sejam identificadas e corrigidas corretamente (FENG *et al.*, 2022).

Essa tarefa torna-se ainda mais desafiante em sistemas de IA baseados em aprendizado contínuo (DAVIS; WALSH; MATHENY, 2022). Embora seja uma abordagem muito promissora para otimizar as decisões de gestão clínica em tempo real, a cautela em torno do uso de técnicas de aprendizado incremental na área da saúde é atribuída à dificuldade para monitorar continuamente esses modelos (LEE, Cecilia S.; LEE, 2020).

Mesmo nos modelos de aprendizado tradicional (não-contínuo), o monitoramento exige atualizações periódicas a fim de mitigar possíveis quedas de desempenho. Isso depende da disponibilidade de dados de treinamento de qualidade, o que demanda a adoção de estratégias permanentes para coleta e pré-processamento dos dados. Por isso, no uso de sistemas de IA na prática clínica, é fundamental que sejam definidas as responsabilidades pelo monitoramento e atualização. Além de atribuir essa tarefa desenvolvedores, pode ser necessário criar equipes de garantia de qualidade nos ambientes clínicos. O papel dos reguladores também deve ser crucial nesse contexto (FENG *et al.*, 2022).

4 FUNDAMENTOS PARA REGULAÇÃO DA IA NA SAÚDE

4.1 Direito à Saúde e Ética em IA

4.1.1 Efetivação da saúde como direito humano fundamental

O reconhecimento do direito à saúde como um direito humano fundamental é bem estabelecido no direito internacional (MANN *et al.*, 1994), consagrado na Declaração Universal dos Direitos Humanos (UNITED NATIONS, 1948) e desenvolvido no Pacto Internacional sobre Direitos Econômicos, Sociais e Culturais (UNITED NATIONS, 1966b). Atualmente, está consolidado o entendimento de que o direito à saúde não significa um vago "direito a ser saudável", mas sim o direito ao mais alto padrão de saúde atingível e a instalações, bens, serviços e condições necessárias para a realização desse direito. Isso envolve tanto liberdades quanto direitos, que incluem o direito de controlar o próprio corpo e o direito a um sistema de proteção à saúde que ofereça igualdade de oportunidades para que todos desfrutem do mais alto nível possível de saúde (UNITED NATIONS, 2000).

Esse fenômeno político e jurídico é associado à transição do Estado liberal para o Estado social: ou seja, a incorporação dos direitos econômicos e sociais feita pelos Estados contemporâneos para além dos direitos relacionados às liberdades individuais (BOBBIO, 1996). A admissão de direitos sociais, como o direito à saúde, modificou profundamente a atuação do Estado, de uma posição passiva para uma posição ativa no que diz respeito aos direitos fundamentais. Diferentemente da possibilidade de aplicação direta típica dos direitos fundamentais de defesa e liberdades públicas, a efetivação dos direitos sociais depende de conduta positiva do Estado no sentido de implementar as prestações que constituem o objeto de cada direito fundamental (BÖCKENFÖRDE, 2017). Isso é, as condições para o exercício dos direitos sociais precisam ser construídas pelos governos.

Nesse sentido, para assegurar o direito à saúde os Estados devem assumir uma série de obrigações em três categorias básicas: respeitar, proteger e cumprir (OHCHR; WHO, 2008). A **obrigação de respeitar** exige que os Estados se abstenham de interferir direta ou indiretamente no direito à saúde. A **obrigação de proteger** exige que os Estados impeçam que terceiros interfiram no direito à saúde, criando leis e adotando medidas para garantir que os atores privados

obedeçam aos padrões de direitos humanos ao fornecer cuidados de saúde ou serviços que afetem a saúde. E a **obrigação de cumprir** exige que os Estados adotem medidas legislativas, administrativas, orçamentárias, judiciais, executivas e outras apropriadas para realizar plenamente o direito à saúde, como organizar políticas de saúde e assegurar condições sanitárias para suas populações.

4.1.2 Regulação e direito à saúde

A atividade regulatória na saúde refere-se a um sistema de regras e procedimentos, estabelecidos e executados por uma autoridade reconhecida, destinados a orientar, dirigir e controlar as ações e decisões relacionadas à saúde. A regulação é, portanto, intervenção estatal no mercado de saúde e no sistema sanitário público, mediante normas, atos legislativos e administrativos, incentivos e sanções, com intuito de assegurar eficiência e equidade na assistência à saúde, melhorando os resultados gerais de saúde na sociedade (GOSTIN, 2008).

Na **regulação em saúde**, o Estado controla diretamente ou influencia diversas variáveis como quantidade, qualidade e preço de produtos, ações e serviços. Isso é feito nas diversas esferas da área da saúde e existem muitas formas possíveis de categorizar a regulação na saúde, de acordo com os objetos, as funções e os efeitos buscados pela atuação estatal (ROBERTS *et al.*, 2008). De uma perspectiva mais abrangente, há a **regulação econômica em saúde**, que envolve políticas e práticas destinadas a influenciar ou controlar o comportamento econômico dos provedores de saúde, pagadores e consumidores, buscando corrigir imperfeições no mercado (FOLLAND; GOODMAN; STANO, 2013). Como função dos sistemas de saúde, é exercida a **regulação da assistência à saúde**, que compreende as regras relativas à prestação de serviços de saúde e à disponibilidade recursos, incluindo a regulamentação de profissionais, instituições e sistemas de saúde. Outra dimensão essencial nesse contexto é a **regulação de produtos para saúde**, que diz respeito a medidas de controle de produtos que afetem a saúde (como alimentos, medicamentos e dispositivos médicos) a fim de reduzir riscos à saúde individual e coletiva (FIELD, 2006).

Importa ressaltar que a regulação na saúde pode ser feita de forma direta ou indireta. A **regulação direta** é aquela que envolve principalmente o controle da qualidade, segurança e eficácia dos serviços e produtos de saúde, como no licenciamento de profissionais de saúde, na

certificação de estabelecimentos de saúde e na aprovação de novos medicamentos e dispositivos médicos. A **regulação indireta** é representada principalmente pelo sistema de responsabilidade civil, que influencia os resultados da saúde por meio de mecanismos legais que tratam de lesões e danos pessoais, podendo impactar significativamente os resultados de saúde ao moldar comportamentos dos indivíduos e das organizações (GOSTIN, 2008).

A regulação é, portanto, elemento essencial para a efetivação do direito à saúde. A atividade regulatória integra o feixe de obrigações dos Estados – nos sentidos de proteger e cumprir – para operacionalizar a saúde como direito fundamental (OHCHR; WHO, 2008). O dever estatal de garantir o direito ao mais alto padrão de saúde atingível compreende o direito a uma regulação em saúde efetiva e guiada pelos princípios basilares dos direitos humanos (HUNT; BACKMAN, 2008; WHO, 2019).

4.1.3 Por que regular a IA na saúde

A incorporação de sistemas de IA na assistência à saúde desafia a regulação em saúde. Deve-se reconhecer que as ações e serviços de saúde, que sempre foram prestados principalmente por pessoas, começam a ser fortemente influenciados e até mesmo executados por sistemas automatizados (RICHMAN, 2018). Além disso, os atuais modelos regulatórios não estão preparados para lidar com a alta velocidade de desenvolvimento das ferramentas digitais, com grande variedade e variabilidade (IQBAL; BILLER-ANDORNO, 2022).

As ferramentas desenvolvidas para a saúde a partir de modelos de ML representam reconhecidamente os maiores desafios nesse cenário (PARIKH; OBERMEYER; NAVATHE, 2019), em função da capacidade de aprendizado com a experiência do mundo real e de adaptação para melhorar continuamente o desempenho (BATES, 2023; GOTTLIEB; SILVIS, 2023b). Diferentemente dos objetos da regulação em saúde tradicional, como medicamentos e dispositivos médicos, os sistemas de IA/ML podem mudar continuamente mesmo depois de implementados. Regulá-los é como tentar atingir um alvo em constante movimento.

Superar esses desafios é uma **necessidade tecnológica, política e jurídica**. Regular a IA na saúde é crucial para assegurar que essas novas tecnologias tenham qualidade e segurança comprovadas e sobretudo que o seu uso na assistência à saúde esteja alinhado com preceitos éticos e com a observância dos direitos humanos (WHO, 2021).

4.1.4 Ética em IA e Direitos Humanos

Ética e IA é um campo interdisciplinar que aborda as implicações morais, legais e sociais das tecnologias de IA. É um domínio em rápida evolução que explora os dilemas e desafios éticos impostos pela crescente integração da IA nos diversos aspectos da vida, a partir de fundamentos da filosofia, da ciência da computação, do direito e das ciências sociais. O escopo desse campo é orientar o desenvolvimento e o uso da IA em harmonia com os valores humanos e normas sociais (DUBBER; PASQUALE; DAS, 2020).

As implicações da IA para os direitos humanos têm sido cada vez mais debatidas. Admite-se que os sistemas de IA podem ajudar promover os direitos humanos, melhorando acesso à informação e o monitoramento de abusos ao redor do mundo, ajudando a identificar padrões de discriminação ou violência e mesmo auxiliando na administração da justiça em muitos países e jurisdições. No entanto, o uso da IA também apresenta riscos significativos aos direitos humanos, o que tem sido alvo de crescente preocupação no cenário internacional.

Entidades como o Conselho da Europa (COUNCIL OF EUROPE, 2018, 2019), o Alto Comissariado das Nações Unidas para os Direitos Humanos (HUMAN RIGHTS COUNCIL, 2020) e a Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO, 2019, 2022) têm adotado recomendações específicas sobre o impacto dos sistemas algorítmicos nos direitos humanos, especialmente a partir da segunda metade da década de 2010. Essas recomendações abordam a necessidade de uso ético da IA e demonstram preocupações quanto a pontos cruciais como o risco de desrespeito à privacidade, falta de transparência e a possibilidade de perpetuação de vieses e preconceitos pelos sistemas de IA.

O cerne do debate em torno de IA e direitos humanos conduzido pelas entidades internacionais está no reconhecimento de que os requisitos morais e legais básicos universais e inalienáveis devem restringir e guiar o uso das novas tecnologias na sociedade: **dignidade humana, igualdade, liberdade, não discriminação**. Como os direitos humanos admitidos em tratados assinados e ratificados pelos Estados são juridicamente vinculantes, existe no direito internacional a estrutura necessária para que governos e organizações tenham o dever de respeitá-los. Isso não supre a necessidade de regulação do setor, mas traça o caminho no qual ela deve ser construída a partir de balizas éticas fundamentais.

4.1.5 Princípios gerais de ética em IA

O primeiro padrão sobre ética e IA reconhecido pela comunidade internacional foi a Recomendação do Conselho de Inteligência Artificial da Organização para a Cooperação e Desenvolvimento Econômico – OCDE (OECD, 2019). Os princípios de IA da OCDE foram adotados em maio de 2019 pelos 36 países membros da OCDE e admitidos em diversos países, tendo sido endossados pelos governos do G20 em junho de 2019. A recomendação não é juridicamente vinculante, mas vem tendo importante influência política na definição de padrões internacionais e na elaboração de leis nacionais na busca de promover o desenvolvimento e uso de IA confiável e que respeite os direitos humanos e os valores democráticos. Os princípios gerais para ética em IA da OCDE são:

a) **crescimento inclusivo, desenvolvimento sustentável e bem-estar** – IA confiável em busca de resultados benéficos para as pessoas e para o planeta;
b) **valores centrados no ser humano e justiça** – respeito ao estado de direito, direitos humanos e valores democráticos durante todo o ciclo de vida dos sistemas de IA;
c) **transparência e explicabilidade** – compromisso com divulgação responsável de informações que permitam promover uma compreensão geral dos sistemas de IA e com a possiblidade de que afetados adversamente contestem os resultados desses sistemas;
d) **robustez e segurança** – os sistemas de IA devem ser seguros e protegidos durante todo o ciclo de vida, sendo garantida a rastreabilidade;
e) **responsabilidade** *(accountability)* – atores envolvidos na IA devem ser responsáveis pelo bom funcionamento dos sistemas e pelo respeito aos demais princípios.

Também em 2019 foram publicadas as Diretrizes Éticas para IA Confiável pela Comissão Europeia, resultado do trabalho de um grupo independente de especialistas de alto nível em IA, composto por representantes da sociedade civil, da indústria e da academia, formado em junho de 2018 (EUROPEAN COMMISSION, 2019b). Essas diretrizes estão fundamentadas em quatro princípios éticos: 1) respeito da autonomia humana; 2) prevenção de danos; 3) equidade; e 4) explicabilidade. Com base nesses princípios, são definidos sete requisitos principais que os sistemas de IA devem atender para serem considerados confiáveis:

a) **ação e supervisão humanas** – preservar autonomia e tomada de decisões por humanos, por estratégias como *human-in-the-loop* (ou *human-in-command*);

b) **robustez técnica e segurança** – serem resilientes e seguros, precisos, confiáveis e reprodutíveis;
c) **privacidade e governança de dados** – respeitar a privacidade, a qualidade e a integridade dos dados, garantindo o acesso legitimado aos dados;
d) **transparência** – serem rastreáveis, com decisões explicáveis e assegurando que os humanos estejam cientes quando interagirem ou forem afetados por sistemas de IA;
e) **diversidade, não discriminação e equidade** – evitar vieses injustos e serem acessíveis a todos;
f) **bem-estar social e ambiental** – beneficiar todos os seres humanos, incluindo gerações futuras, levando em consideração o meio ambiente e o impacto social;
g) **responsabilidade** *(accountability)* – serem auditáveis, assegurar comunicação dos impactos negativos e reparação adequada e acessível em caso de danos.

Embora haja bastante convergência quanto aos valores essenciais, existem diferentes propostas em relação aos preceitos e requisitos da ética em IA. Uma possível abordagem é a adaptação dos princípios clássicos da bioética: respeito à **autonomia**, **não-maleficência**, **beneficência** e **justiça**. Propõe-se que esses quatro princípios básicos acrescidos do princípio da **explicabilidade** – incorporando nele a noção de responsabilidade – podem fornecer uma estrutura ética necessária para compreender as oportunidades e riscos no sentido de orientar o uso da IA na sociedade (FLORIDI *et al.*, 2018).

Da segunda metade da década de 2010 até o início dos anos 2020, a abordagem regulatória para uso de sistemas de IA começou a se estruturar por meio de códigos de conduta não vinculantes (*soft law*), já que leis específicas sobre IA começaram a ser debatidas e elaboradas, mas não chegaram a entrar em vigor. Vale dizer que, nesse mesmo período, as legislações de privacidade de dados tornaram-se praticamente onipresentes no mundo (GREENLEAF, 2023).

Das centenas de documentos contendo princípios éticos ou diretrizes para IA produzidos por entidades governamentais, conselhos de especialistas para assessoramento de entes públicos, institutos de pesquisa e empresas privadas, tem se formado uma convergência global em torno de cinco princípios éticos:

a) transparência;

b) justiça e equidade;

c) não-maleficência;

d) responsabilidade;

e) privacidade.

No entanto, ainda existe significativa divergência em relação ao sentido desses princípios, como devem ser interpretados e qual o caminho para implementá-los (JOBIN; IENCA; VAYENA, 2019).

4.2 Princípios Éticos para IA na Saúde

A importância de se definir princípios éticos específicos para regulação da IA na saúde começou a ser debatida já na segunda metade da década de 2010, quando surgiram as primeiras evidências de que campo de *machine learning* (ML) seria altamente promissor e revolucionário para a área da saúde (INTERNATIONAL BIOETHICS COMMITTEE, 2017).

Em 2021, a Organização Mundial da Saúde (OMS) publicou o guia inaugural com diretrizes sobre ética e governança da IA na saúde com pretensão de alcance global, resultado de dois anos de trabalho desenvolvido por um amplo grupo de especialistas (WHO, 2021). Esse documento consolida os primeiros princípios básicos considerados consensuais no campo da ética em IA especificamente para a área saúde:

a) proteger a autonomia;

b) promover o bem-estar humano, a segurança humana e o interesse público;

c) garantir transparência, explicabilidade e inteligibilidade;

d) promover responsabilidade e prestação de contas;

e) assegurar inclusão e equidade;

f) promover inteligência artificial responsiva e sustentável.

Esses princípios da OMS para ética da IA na saúde têm o propósito de orientar desenvolvedores, usuários e reguladores no contexto do desenvolvimento, avaliação, implementação e avaliação contínua de tecnologias de IA na saúde. As diretrizes estão baseadas na conjugação dos princípios bioéticos básicos (autonomia, não-maleficência, beneficência e justiça) com os demais princípios gerais de ética em IA (privacidade, transparência, responsabilidade).

Abre-se uma nova perspectiva na ética em saúde. Há muito se reconhece o dever especial dos cuidadores de saúde no sentido de promover e resguardar esses valores éticos essenciais em relação aos pacientes (BEAUCHAMP; CHILDRESS, 1979). Agora, uma vez que sistemas de IA passam a ser usados em ambientes clínicos e por profissionais de saúde para realizar tarefas ou como suporte para tomada de decisões que antes eram reservadas a humanos, os responsáveis por projetar, programar e implantar essas tecnologias também devem estar sujeitos a obrigações éticas específicas (WHO, 2021).

Apesar de haver algumas diferenças de terminologia, os princípios da OMS condensam os preceitos essenciais da ética em IA e fornece uma base sólida para a construção do campo específico da ética em IA na saúde. Portanto, adotar os seis princípios conforme delimitados pela entidade internacional é um referencial adequado para compreender os fundamentos éticos para a regulação da IA na saúde.

4.2.1 Autonomia

O princípio da **autonomia** é pedra angular da bioética. Refere-se ao direito dos indivíduos de tomarem decisões informadas sobre suas próprias vidas, saúde e bem-estar (BEAUCHAMP; CHILDRESS, 1979). Na ética da IA na saúde, requer que o uso de sistemas computacionais não prejudique a autonomia humana. Significa assegurar que os seres humanos permaneçam no controle total dos sistemas de saúde e das decisões médicas (WHO, 2021).

Os sistemas de IA usados na área da saúde têm o potencial de influenciar significativamente os processos de tomada de decisão. Os modelos de ML têm capacidade para fazer diagnósticos, prever resultados e evoluções clínicas e para fornecer recomendações para opções de tratamento. Esses recursos podem melhorar muito a prestação de cuidados de saúde, mas representam risco para a autonomia das pessoas se as recomendações dos sistemas de IA forem seguidas sem considerar os valores, preferências e direitos dos pacientes de tomar decisões informadas sobre sua própria saúde (MITTELSTADT *et al.*, 2016). Além disso, em sistemas de IA desenvolvidos para autocuidado, como por exemplo em tecnologias vestíveis (*wearables*), as informações fornecidas diretamente para os pacientes podem dificultar a tomada de decisão livre (OWENS; CRIBB, 2019). O usuário pode se sentir compelido a seguir recomendações por perceber uma autoridade do sistema de IA ou por não compreender como o sistema funciona,

risco que tem se tornado particularmente relevante com a crescente aplicação de *chatbots* baseados grandes modelos de linguagem (HAUPT; MARKS, 2023).

A autonomia está diretamente relacionada à transparência, pois envolve garantir que os prestadores de cuidado tenham as informações necessárias para fazer uso seguro e eficaz dos sistemas de IA e que as pessoas entendam o papel que tais sistemas desempenham em seus cuidados (CHAR; ABRÀMOFF; FEUDTNER, 2020). Portanto, para proteger a autonomia é importante garantir que os pacientes sejam adequadamente informados sobre o papel da IA nas ações de saúde que os afetam e que sejam capazes de entender e de contestar decisões baseadas em IA (GROTE; BERENS, 2020).

Também se vincula à proteção da privacidade e da confidencialidade, bem como à garantia do consentimento. É essencial que sejam observadas as obrigações impostas pelas leis de proteção de dados e que o uso de modelos de ML no diagnóstico, prognóstico e planos de tratamento seja incorporado ao processo de consentimento informado e válido – e não apenas aos genéricos termos de acordo de usuário (GERKE; MINSSEN; COHEN, 2020).

4.2.2 Bem-estar humano, segurança e interesse público

O princípio ético de promover o bem-estar humano, a segurança e o interesse público condensa a essência dos princípios clássicos de **não-maleficência** e **beneficência**. Não-maleficência é a exigência de que o uso de IA não prejudique as pessoas – o conhecido aforismo latino *primum non nocere* (“primeiro, não faça nenhum mal”). A beneficência diz respeito à necessidade de que, além de terem seus riscos minimizados, as tecnologias de IA tenham maximizados os seus benefícios para as pessoas humanas e para o mundo natural (FLORIDI *et al.*, 2020).

Para isso, devem ser estabelecidos requisitos regulatórios de segurança e eficácia para usos e indicações bem definidos, além de medidas de controle de qualidade para monitoramento do desempenho das aplicações de IA ao longo do tempo (MINSSEN *et al.*, 2020). Os sistemas de IA devem funcionar de acordo com regras de desenvolvimento, validação e implementação de modo a evitar perigos e lesões às pessoas e ao mesmo tempo fornecer informações preditivas corretas, precisas e confiáveis para a maioria das situações em que forem aplicados (CHAR; ABRÀMOFF; FEUDTNER, 2020). Considerando o caráter dinâmico e a natureza adaptativa

dos modelos de ML usados na assistência à saúde, essa avaliação precisa ser contínua e feita em tempo real (GILBERT *et al.*, 2021).

A **prevenção de danos** exige que o uso de tecnologias de IA não resulte em nenhum dano físico ou mental. Os modelos devem ser projetados para equilibrar os "alertas" fornecidos pelos seus resultados com salvaguardas apropriadas para proteger indivíduos de estigmatização ou discriminação devido ao estado de saúde (WHO, 2021). Isso inclui, mais uma vez, a necessidade incontornável de observar as obrigações de privacidade de dados pessoais (VAYENA; BLASIMME; COHEN, 2018).

A **promoção do interesse público** envolve a obrigação de desenvolvedores e usuários dos sistemas de IA de maximizar os benefícios do uso dessa tecnologia não apenas para os indivíduos, mas para toda a sociedade (FLORIDI *et al.*, 2018). Deve haver a constante preocupação de tornar esses benefícios amplamente distribuídos e acessíveis, considerando as perspectivas de saúde pública e saúde global. As tecnologias de IA não devem ser reservadas a grupos privilegiados, mas devem ser disponibilizadas a todos que possam ter bom proveito delas, garantindo o acesso aos ambientes com menos recursos, como os países de renda média e baixa e a suas populações (SCHWALBE; WAHL, 2020).

4.2.3 Transparência, explicabilidade e inteligibilidade

A **transparência** é o princípio ético geral mais frequente nos códigos de conduta que estabelecem diretrizes para o uso da IA (JOBIN; IENCA; VAYENA, 2019). Na ética da IA aplicada à saúde, é preconizado que os sistemas sejam inteligíveis e compreensíveis para desenvolvedores, reguladores e usuários (incluindo profissionais de saúde e pacientes). As informações relevantes sobre os sistemas de IA precisam ser documentadas antes da implementação e continuar a ser divulgadas regularmente após aprovação para uso. Nesse sentido, é imprescindível que se facilite a consulta pública e a compreensão sobre o funcionamento dos modelos de IA no mundo real (WHO, 2021).

Exige-se o fornecimento de informações precisas e adequadas acerca dos pressupostos e limitações dos sistemas de IA, de modo a abarcar os protocolos de operação e os métodos de seleção, processamento e rotulagem dos dados, bem como as condições de desenvolvimento e de validação dos modelos. Isso é essencial para possibilitar auditorias nos sistemas de IA,

permitindo mapear corretamente o seu funcionamento a fim de identificar erros e antecipar possíveis consequências (LIU, X. *et al.*, 2022).

Além disso, espera-se que as tecnologias de IA sejam explicáveis de acordo com a capacidade de entendimento daqueles a quem a explicação for direcionada, fornecendo esclarecimentos a respeito do funcionamento e das condições das decisões dos algoritmos para profissionais de saúde, pacientes e demais usuários dos sistemas (WATSON *et al.*, 2019). Operacionalizar a explicabilidade dos modelos de ML é um dos principais desafios no cenário atual e representa um dos elementos cruciais para a regulação da IA na saúde (BABIC *et al.*, 2021a).

O princípio ético de garantir transparência, explicabilidade e inteligibilidade é considerado essencial para possibilitar a efetivação dos demais princípios no campo da ética em IA. Compreendida em sentido amplo, a transparência diz respeito a como o público recebe informações para justificar ou explicar as decisões baseadas em sistemas de IA, bem como os detalhes sobre quem pode ser responsabilizado por essas decisões (FLORIDI *et al.*, 2018). Por isso, o exercício da autonomia, da equidade, da responsabilização e prestação de contas demandam transparência e compreensão acerca da lógica de funcionamento da IA.

4.2.4 Responsabilidade e prestação de contas

A responsabilidade e a prestação de contas (*accountability*) são elementos interligados num único princípio ético para o uso de tecnologias de IA. Na aplicação na área da saúde, esse princípio significa que os desenvolvedores e provedores são responsáveis por garantir que os sistemas de IA executem as tarefas específicas para as quais foram desenvolvidos, que a IA seja usada em condições apropriadas e por pessoas devidamente treinadas. Isso frequentemente requer o estabelecimento de pontos de supervisão humana para assegurar que o modelo algorítmico permanece em um caminho de desenvolvimento eficaz e eticamente adequado (WHO, 2021). Além disso, deve haver meios para atribuir responsabilidade e propiciar reparação para indivíduos e grupos que sofrerem danos decorrentes de decisões baseadas em sistemas de IA.

Embora os termos sejam frequentemente usados como sinônimos, há diferenças essenciais entre os sentidos de responsabilidade e prestação de contas (*accountability*).

A **responsabilidade** refere-se ao papel das pessoas em sua relação com os sistemas de IA. É o dever de responder por suas ações, que existe antes que qualquer tarefa ou ação seja

realizada. Por isso, são necessários mecanismos para vincular as decisões dos modelos da IA aos dados de entrada e às ações tomadas pelos humanos envolvidos no desenvolvimento e implementação desses sistemas à medida que a cadeia de responsabilidade aumenta devido à participação de mais atores nesse processo (DIGNUM, 2020). Responsabilidade diz respeito a todo o sistema social e técnico no qual as ferramentas de IA estão inseridas e que engloba pessoas, máquinas e instituições.

A **prestação de contas** (*accountability*) está relacionada à expectativa de que todos aqueles responsáveis pelo desenvolvimento e implementação cumpram os padrões definidos para garantir o funcionamento adequado dos sistemas de IA durante todo o ciclo de vida (NOVELLI; TADDEO; FLORIDI, 2023). Refere-se à exigência de que os atores envolvidos sejam capazes de explicar e justificar as decisões aos usuários e ao público – no que se articula diretamente com a transparência –, em conformidade com os valores e normas socialmente reconhecidos e com as leis aplicáveis.

O princípio ético de responsabilidade e prestação de contas é o fundamento da responsabilização civil (*liability*) para reparação de danos que sejam causados pelo uso de sistemas de IA na assistência à saúde (PRICE; GERKE; COHEN, 2022).

4.2.5 Inclusão e equidade

Os preceitos éticos inclusão e equidade são elementos do princípio bioético **justiça**, que se refere à distribuição justa e equitativa de recursos e oportunidades na saúde, de modo a garantir que os cuidados sejam distribuídos de forma equânime entre os diferentes grupos de uma sociedade (BEAUCHAMP; CHILDRESS, 1979). Esse princípio está fundamentado nos conceitos éticos de justiça e equidade em sentido amplo, segundo os quais os benefícios e ônus devem ser distribuídos equitativamente entre todos os membros da sociedade (RAWLS, 1999).

Na ética da IA na saúde, significa que as tecnologias de IA devem ser projetadas para ter acesso mais abrangente possível, independentemente de idade, sexo, gênero, renda, raça, etnia, orientação sexual, capacidade ou outras características (WHO, 2021). Além disso, os sistemas de IA devem estar disponíveis para uso nos países de baixa e média renda. Os algoritmos precisam ser desenvolvidos para não reproduzirem preconceitos e desequilíbrios presentes nas

sociedades que discriminem determinados grupos sociais e devem minimizar essas disparidades existentes (SCHWALBE; WAHL, 2020).

Portanto, é necessário que haja monitoramento e avaliação contínuos que assegurem que os sistemas de IA não mantenham ou piorem qualquer tipo de preconceito e discriminação (RAJKOMAR; HARDT; *et al.*, 2018). Os atores envolvidos em todo o processo de desenvolvimento e implementação de sistemas de IA na saúde devem estar ativamente engajados na identificação de vieses e na mitigação do dano potencial que pode ser causado por eles a indivíduos e populações (NTOUTSI *et al.*, 2020).

O que se pretende alcançar a partir da efetivação do princípio de inclusão e equidade é a igualdade justa de oportunidades associada ao tratamento não discriminatório – também chamado "princípio da diferença" (RAWLS, 1999). Além disso, é preciso operacionalizar transparência ao longo de todo o processo para que seja assegurado o direito igual à justificação sobre as condições e decisões relacionadas à implementação de sistemas de IA nos cuidados de saúde individuais e coletivos (GIOVANOLA; TIRIBELLI, 2023).

4.2.6 Responsividade e sustentabilidade

Responsividade e sustentabilidade são elementos considerados interligados num único princípio na ética da IA aplicada à saúde (WHO, 2021). Os sistemas de IA precisam responder de forma adequada às expectativas relacionadas ao seu uso, sobretudo para atender às necessidades de saúde individuais e coletivas das populações dos locais em que sejam implementadas. Também precisam ser sustentáveis, desenvolvidos e usados de maneira viável nas perspectivas econômica, social e ambiental (UNESCO, 2022).

Para promover a **responsividade**, a capacidade de resposta dos sistemas de IA na saúde deve ser avaliada de forma contínua, sistemática e transparente. Governos locais, instituições acadêmicas, centros de pesquisa, agências internacionais, organizações não governamentais, indústria e sociedade civil devem estar envolvidos no desenvolvimento e na implementação de tecnologias de IA para a saúde (ALAMI *et al.*, 2020).

As tecnologias de IA devem promover ampla **sustentabilidade** dos sistemas de saúde, ambientes e locais de trabalho, minimizar suas consequências ambientais e aumentar a eficiência energética (JIA *et al.*, 2023). Além disso, o uso de IA na saúde também requer que governos

e empresas promovam treinamentos e adaptações específicas para os profissionais de saúde e atuem para reduzir o impacto da perda de postos de trabalho como consequência do uso de sistemas automatizados (HOSNY; AERTS, 2019).

4.3 Governança de Dados de Saúde

Nos últimos anos, tem-se afirmado frequentemente que os dados são o petróleo do século XXI. Por essa analogia, pode-se dizer que a IA desempenhará neste século um papel semelhante ao que o motor de combustão cumpriu na transformação da sociedade a partir da primeira metade do século XX (WOOD, 2019). Há quem prefira comparar os dados ao oxigênio, pois representam um recurso indispensável para a infraestrutura contemporânea e que podem beneficiar toda a sociedade ao serem compartilhados (IE UNIVERSITY, 2022). Independentemente da metáfora adotada, é incontestável que a revolução digital impulsionada pela IA depende de grandes quantidades de dados de alta qualidade e que é preciso haver estruturas robustas para a governança desses dados.

4.3.1 Conceitos básicos

A **governança de dados** refere-se ao processo pelo qual se atribui autoridade e controle sobre dados e do consequente exercício dessa autoridade por meio de tomada de decisões que assegurem a gestão e o uso eficaz e ético tanto dos dados como das tecnologias relacionadas a eles (PLOTKIN, 2021). As organizações e pessoas envolvidas no tratamento de dados devem aplicar e monitorar a observância de padrões e regras estabelecidos para garantir a responsabilidade e a prestação de contas (*accountability*) ao longo de todo o ciclo de vida dos dados e dos sistemas desenvolvidos a partir deles. Esse processo compreende o gerenciamento de disponibilidade, usabilidade, qualidade, integridade, privacidade e segurança dos dados. Portanto, a governança de dados é a base para uma IA confiável (JANSSEN *et al.*, 2020).

Na área da saúde, os dados são historicamente admitidos como fundamentais para a tomada de decisões e têm assumido uma importância crescente em todo o planejamento, assistência e demais atividades no contexto sanitário. Nas últimas duas décadas, a definição de dados

de saúde expandiu-se de maneira muito significativa. Além dos tradicionais registros clínicos e eletrônicos oriundos de sistemas de saúde e seguradoras, hoje são incorporados dados provenientes de diversas outras fontes, como redes sociais, aplicativos, dispositivos vestíveis e *sites* de busca, entre outros. Esse ecossistema de dados de saúde, também conhecido como o conjunto dos "megadados biomédicos" (*biomedical big data*), inclui os dados de fontes padrão de serviços de saúde e pesquisa, associados a dados sociais, econômicos, comportamentais, ambientais e outros tipos dados não diretamente relacionados com a saúde, mas que são convertidos em dados de saúde. Tem se formado consenso no sentido de reconhecer a necessidade de mecanismos de controle de acesso e utilização dos dados de saúde, bem como de estruturas de governança específicas para os megadados biomédicos (VAYENA; BLASIMME, 2017).

4.3.2 Proteção de dados como fundamento para regulação da IA na saúde

O uso de dados de saúde para desenvolvimento e implementação de sistemas de IA levanta várias preocupações. A qualidade dos dados – especialmente os provenientes dos países de baixa e média renda – e os preconceitos sistêmicos decorrentes da sub-representação de diversos grupos, podem comprometer a eficácia dos modelos de IA. A privacidade individual é uma preocupação constante, dado o risco de discriminação e ameaças a dignidade das pessoas, sobretudo em populações estigmatizadas e vulneráveis (WHO, 2023). O excesso na coleta de dados e sua reutilização potencialmente antiética também suscitam dilemas relacionados a direitos humanos. Além desses, há ainda o risco de ampliação da disparidade entre quem controla os dados e quem os fornece. Esse "colonialismo de dados" que pode exacerbar desequilíbrios de poder entre países, principalmente se elementos como consentimento, privacidade e autonomia não forem adequadamente observados (WHO, 2021).

A formação do campo político e normativo de governança de dados nas mais diversas jurisdições tem-se baseado em elementos comuns da área de privacidade e proteção de dados pessoais e construído a partir dos paradigmas norte-americano e europeu, com recente preponderância do último sobre o primeiro. Embora não haja normas de direito internacional sobre essa matéria até o começo da década de 2020, ubiquidade das leis de privacidade de dados em escala global consolidou-se nesse período (GREENLEAF, 2023), de tal forma que se constituiu um campo autônomo, com interfaces significativas com institutos de direito público e direito privado nas diversas jurisdições (BERNIER; MOLNÁR-GÁBOR; KNOPPERS, 2022).

Na regulação da IA aplicada à saúde, a dimensão da proteção e privacidade de dados pessoais constitui um elemento já bem estabelecido (MCNAIR; PRICE, 2022). Portanto, a proteção de dados pode ser considerada um fundamento para a regulação da IA na saúde, ao lado dos princípios éticos e de direitos humanos. Um dos pontos centrais na governança de dados de saúde é o direito à privacidade.

4.3.3 Privacidade de dados pessoais

A **privacidade de dados pessoais** é um instituto bem assentado no direito contemporâneo. Por isso, todas as iniciativas para regulação da IA lidam com a proteção da privacidade de dados pessoais como fundamento ético – e, mais recentemente, também jurídico – a partir do qual o arcabouço regulatório deve ser construído.

A noção atual do direito à privacidade começou a ser desenvolvida no final do século XIX. Naquela época, estava mais relacionada aos aspectos da privacidade física e do respeito ao direito à propriedade, mas passou a incorporar o direito mais geral do indivíduo de "ser deixado em paz" (WARREN; BRANDEIS, 1890). Essa concepção é próxima da ideia moderna de privacidade no contexto da governança de dados.

No momento histórico de transição para o Estado social, a privacidade foi assimilada pela legislação em matéria de direitos humanos, a partir da Declaração Universal dos Direitos Humanos (UNITED NATIONS, 1948). Desde então, houve desenvolvimento doutrinário e jurisprudência formada pelas cortes de direitos humanos. A privacidade é hoje admitida como direito humano fundamental cujo núcleo tem origem no reconhecimento de que todas as pessoas devem ser respeitadas como indivíduos com direito a um espaço para existir sem interferências. Como nos demais direitos humanos, os direitos são mantidos em relação aos direitos concorrentes de outros, em equilíbrio entre os membros da sociedade. Logo, a privacidade não é entendida como um direito absoluto, pois há sempre exceções previstas nos documentos normativos e admitidas pelos tribunais (TOWNEND, 2021).

Proteger a privacidade de dados tornou-se um elemento essencial para a governança democrática na era digital. Em 1988, o Comitê de Direitos Humanos da ONU (UNITED NATIONS, Human Rights Committee, 1988) postulou a necessidade de legislações de proteção de dados pessoais para salvaguardar o direito fundamental à privacidade reconhecido pelo Pacto

Internacional dos Direitos Civis e Políticos (UNITED NATIONS, 1966a). Embora a maioria dos países tenham criado novas leis de proteção de dados nas últimas duas décadas, ainda não existem tratados internacionais vinculantes que tratem especificamente de governança de dados no nível global (BERNIER; MOLNÁR-GÁBOR; KNOPPERS, 2022).

Da perspectiva dos direitos humanos, os indivíduos devem sempre ter controle sobre seus dados. Para isso, deve ser assegurado acesso a mecanismos de consentimento individual e de proteção dos dados pessoais. Esses mecanismos estão atualmente incorporados nas diversas leis de proteção de dados adotadas em todo o mundo. De forma geral, essas leis incluem padrões para realização de atividades de processamento de dados e estabelecem obrigações para controladores e processadores, tanto públicos como privados, prevendo sanções em caso de violações (PRIVACY INTERNATIONAL, 2018).

As violações de privacidade relacionam-se com regras contextuais sobre como as informações podem circular em relação aos atores envolvidos, ao processo pelo qual as informações são acessadas, à frequência e à finalidade do acesso. Ocorre violação quando há quebra dessas regras: se algum ator errado tem acesso às informações, se o processo para acesso não é observado, ou se a finalidade do acesso é inadequada. Essas regras buscam resguardar os indivíduos por razões basicamente de duas categorias: consequencialistas e deontológicas. Razões consequencialistas dizem respeito a consequências negativas que violações de privacidade possam causar aos titulares dos dados – como um aumento no valor do prêmio do seguro ou uma discriminação no emprego, por exemplo. As razões deontológicas precedem a existência qualquer dano ao indivíduo, pois se considera que violações de privacidade são eticamente preocupantes mesmo se não houver consequências negativas diretas para o titular (PRICE; COHEN, 2019). Os dados de saúde são geralmente considerados uma categoria especial de dados pessoais, por serem identificados como informações sensíveis nas quais as violações de privacidade são particularmente propensas a causar danos aos titulares.

Há atualmente dois paradigmas regulatórios bem definidos em matéria de privacidade e proteção de dados: o norte-americano e o europeu. O paradigma norte-americano é baseado na abordagem setorial, com legislações específicas, tais como para a saúde e para o setor financeiro. O paradigma europeu assenta-se na abordagem holística fundada numa lei geral de proteção de dados. Muitos países têm adotado características legislativas híbridas entre esses dois paradigmas, mas o padrão europeu tem exercido influência crescente e determinante como referência global (BERNIER; MOLNÁR-GÁBOR; KNOPPERS, 2022). Para a governança de dados de saúde, é necessário contemplar esses dois paradigmas.

4.3.4 Paradigma norte-americano

O primeiro padrão específico de governança de dados de saúde com importância global foi o adotado pelos Estados Unidos (EUA) com base na "Lei de Responsabilidade e Portabilidade de Seguros de Saúde" (**HIPAA** – *Health Insurance Portability and Accountability Act*), de 1996. Os EUA são dos poucos países que ainda não instituíram uma legislação nacional abrangente de privacidade e proteção de dados até o começo da década de 2020, mantendo essa matéria regulamentada por leis estaduais (GREENLEAF, 2023). A regulação setorial baseada na HIPAA permanece em vigor e com relevante influência, sobretudo em razão da concentração de empresas de tecnologia nos EUA.

O regime regulatório instituído pela HIPAA tem o propósito de definir padrões federais para garantir a segurança das informações de saúde dos cidadãos dos EUA (OFFICE FOR CIVIL RIGHTS, 2021). A lei busca assegurar confidencialidade, integridade e disponibilidade dessas informações, enquanto concede acesso a prestadores, câmaras de compensação e planos e seguradoras de saúde. Para isso, cria diversas regras dentre as quais destacam-se as de privacidade (*Privacy Rule*), segurança (*Security Rule*) e aplicação (*Enforcement Rule*).

A **Regra de Privacidade** da HIPAA regula o uso e a divulgação de "informações de saúde protegidas" (*protected health information* – PHI) pelas chamadas "entidades abrangidas" (*covered entities*). Essas entidades são os prestadores de serviços de saúde, as câmaras de compensação, as seguradoras e os planos de saúde. Para que seja permitida divulgação de qualquer dado de saúde identificável, exige-se autorização por escrito do titular, a menos que se aplique alguma das exceções previstas na lei:

a) divulgação para o próprio indivíduo;

b) tratamento, pagamento ou operações de cuidados de saúde;

c) divulgação com oportunidade de concordar ou contestar;

d) uso e divulgação incidental (permitidos de outra forma);

e) atividade de interesse público (razões de saúde pública, casos de violência, processos judiciais ou administrativos, interesse para pesquisa, ameaça grave a saúde ou a segurança);

f) conjunto de dados limitado para fins de pesquisa, saúde pública ou operações de cuidados de saúde.

Não há restrições ao uso ou divulgação de informações de saúde não identificadas. Os conjuntos de dados dos quais forem removidos 18 tipos de identificadores (como nome, endereço, dados pessoais, números de documentos, datas de atendimento etc.) podem ser compartilhados livremente tanto para fins de pesquisa como com interesse comercial.

A **Regra de Segurança** da HIPAA, publicada em 2003, complementa a regra de privacidade a partir da especificação de padrões para proteção de um subconjunto de informações, denominadas "informações eletrônicas de saúde protegidas" (*electronic protected health information* – e-PHI). Essa regra especifica uma série de procedimentos administrativos, técnicos e de segurança física que as entidades abrangidas devem usar para garantir confidencialidade, integridade e disponibilidade dos e-PHI (que não se aplicam a informações transmitidas oralmente ou por escrito). Em 2009, a Lei HITECH (*Health Information Technology for Economic and Clinical Health Act*) ampliou o alcance da regra de segurança da HIPAA para incluir os parceiros comerciais das entidades abrangidas pela lei, embora limitado às relações e ambientes de cuidados de saúde.

A **Regra de Aplicação** da HIPAA contém disposições relativas à conformidade, investigações e à imposição de penalidades financeiras civis por violações das demais regras administrativas criadas pela própria lei (OFFICE FOR CIVIL RIGHTS, 2021).

O paradigma norte-americano de governança de dados de saúde está essencialmente baseado na HIPAA, alterada pela Lei HITECH de 2009, mas essa abordagem tem importantes lacunas. A estratégia principal para proteger privacidade de dados pessoais, ao mesmo tempo em que permite o compartilhamento de informações, é a desidentificação (ou anonimização) dos dados. Contudo, a remoção dos 18 identificadores especificados pela HIPAA não impede que os dados possam se tornar reidentificáveis por meio de triangulações feitas com outros conjuntos de dados (PRICE; COHEN, 2019).

Além disso, a maioria dos atuais dados de saúde não se enquadram no regime da HIPAA. A legislação não cobre dados de saúde gerados por empresas que não sejam classificadas como "entidades abrangidas" e seus parceiros comerciais diretos, nem abarca informações sobre saúde criadas pelos próprios titulares – por exemplo, em mecanismos de busca e redes sociais (COHEN; MELLO, 2018). Tecnologias emergentes baseadas em dispositivos vestíveis, internet das coisas, dados genômicos, bem como ferramentas de atendimento remoto – especialmente impulsionadas durante a pandemia de Covid-19 – também carecem de proteção no

regime HIPAA. Esses estão entre os principais motivos alegados para defender a importância de uma atualização desse marco regulatório (THEODOS; SITTIG, 2020).

Portanto, parte considerável dos dados pessoais que servem de combustível – ou oxigênio – para o desenvolvimento de sistemas de IA na saúde não se encaixam no paradigma norte-americano de governança.

4.3.5 Paradigma europeu

A governança de dados de saúde na União Europeia (UE) está baseada no Regulamento Geral sobre a Proteção de Dados (**GDPR** – *General Data Protection Regulation*), aprovado em 2016, que entrou em vigor em maio de 2018 (EUROPEAN UNION, 2016). O GDPR é uma lei abrangente concebida para substituir a Diretiva de Proteção de Dados de 1995, com intuito de harmonizar as leis de privacidade e proteção de dados nos países membros da UE e remodelar a forma como as organizações europeias públicas e privadas abordam a governança de dados.

O GDPR adota seis princípios fundamentais para tratamento de dados pessoais (art. 5º):

a) licitude, lealdade e transparência;
b) limitação das finalidades – que devem ser determinadas, explícitas e legítimas;
c) minimização dos dados – limitados ao necessário para as finalidades;
d) exatidão – dados inexatos devem ser apagados ou retificados;
e) limitação da conservação – dados armazenados apenas durante o período necessário para as finalidades;
f) integridade e confidencialidade – garantia de segurança.

Por definição, dados pessoais são informações relacionadas a uma pessoa singular identificada ou identificável, direta ou indiretamente, tendo em conta a tecnologia disponível à data do tratamento dos dados e a evolução tecnológica (GDPR, art. 4º, 1). Assim, os princípios da proteção de dados não se aplicam a informações consideradas anônimas, que são aquelas em que o titular não possa ser identificado (GDPR, considerando 26).

O **consentimento** é considerado a pedra angular do marco regulatório de proteção de dados europeu. Ele deve ser livre, específico, informado e inequívoco. O ônus de provar o consentimento recai sobre os responsáveis pelo tratamento de dados e o titular pode retirar seu

consentimento a qualquer momento. Essa concepção busca conceder aos indivíduos maior controle sobre os seus dados pessoais (KUNER; BYGRAVE; DOCKSEY, 2020).

Os **dados de saúde** são qualificados como uma das "categorias especiais de dados pessoais", nos termos da lei europeia. Vale ressaltar que o GDPR define dados de saúde como "todos os dados relativos ao estado de saúde de um titular de dados que revelem informações sobre a sua saúde física ou mental no passado, no presente ou no futuro" (considerando 35). Ou seja, todo e qualquer dado que revele informação de saúde ou que possa ser convertido em dado de saúde, independentemente da sua fonte. Essa definição ampliada reflete a intenção do legislador de incorporar a jurisprudência do Tribunal de Justiça da União Europeia e do Tribunal Europeu de Direitos Humanos (BYGRAVE; TOSONI, 2020).

Assim enquadrados, os dados de saúde submetem-se ao regime de tratamento de **categorias especiais de dados pessoais** estabelecido pelo GDPR (art. 9º), ao lado dos dados genéticos e biomédicos, dos dados que revelem a origem racial ou étnica, as opiniões políticas, as convicções religiosas ou filosóficas, ou a filiação sindical, e dos dados relativos à vida sexual ou orientação sexual de uma pessoa. Há considerável sobreposição entre os "dados relativos à saúde", os "dados genéticos" e os "dados biométricos". As três modalidades são subcategorias de dados pessoais e, portanto, devem satisfazer os critérios definidos pelo GDPR: conter informação relativa a uma pessoa singular identificada ou identificável. O tratamento desses dados de categorias especiais é proibido, exceto em situações específicas:

a) houver consentimento explícito do titular dos dados;
b) for necessário para cumprimento de obrigações e exercício de direitos específicos (direito do trabalho e da seguridade social);
c) for necessário para proteção de interesses vitais do titular ou de outra pessoa, quando o titular estiver incapacitado de dar seu consentimento;
d) for feito por uma fundação, associação ou outro organismo sem fins lucrativos com fins políticos, filosóficos, religiosos ou sindicais;
e) tiverem sido manifestamente tornados públicos pelo titular;
f) for necessário em ações legais e processos judiciais;
g) for necessário por motivos de interesse público;
h) for necessário para efeitos de medicina preventiva ou do trabalho ou para prestação de cuidados de saúde;
i) houver interesse público na área da saúde pública;

j) for necessário para fins de arquivo de interesse público, de investigação científica ou histórica ou para fins estatísticos.

O regime jurídico do GDPR atribui uma série de direitos aos titulares dos dados que conformam o paradigma geral de governança de dados que se aplica aos dados de saúde (arts. 12º a 22º). A começar pelo **direito à transparência** das informações e das comunicações, acompanhado de regras bem definidas para o exercício dos direitos, que asseguram que os indivíduos recebam informações claras, concisas e acessíveis sobre o seu processamento de dados. Os direitos dos indivíduos são:

a) **direito de acesso** – saber se seus dados estão sendo processados, a finalidade do tratamento, e poder acessar os próprios dados;
b) **direito de retificação** – corrigir dados errados ou imprecisos;
c) **direito ao apagamento de dados** – "direito a ser esquecido";
d) **direito à limitação do tratamento** – suspender o processamento de seus dados em algumas situações específicas;
e) **direito de portabilidade** – receber os seus dados pessoais e transmitir a outro responsável pelo tratamento;
f) **direito de oposição** – opor-se ao tratamento dos seus dados para fins específicos;
g) **direito de não ficar sujeito a decisões exclusivamente automatizadas**.

Esse último direito listado diz respeito mais diretamente à regulação da IA e está previsto no art. 22º do GDPR. O dispositivo legal autoriza decisões automatizadas se forem necessárias para um contrato entre o titular e o controlador dos dados, se for autorizado por lei ou se estiver baseado em consentimento explícito do titular. Pelo fato de aplicar-se apenas a decisões "*exclusivamente* com base no tratamento automatizado", e os sistemas de IA em estarem em larga medida relacionados a julgamentos humanos, acredita-se que essa regra terá pouca aplicabilidade prática (MENDOZA; BYGRAVE, 2017).

De uma forma geral, a aplicação é considerada um ponto chave do regime instituído pelo GDPR. Ele estabelece sanções com multas administrativas significativas (2% ou 4% do volume de negócios anual total, com base na gravidade da infração) e dá amplos poderes para as Autoridades de Proteção de Dados dos países para investigar, intervir e até interromper o processamento de dados, além de instaurar processos judiciais.

Portanto, o paradigma europeu de governança de dados de saúde integra o marco regulatório geral de proteção e privacidade de dados da EU. Ele segue a tradição europeia nessa

matéria, aprimorando e aprofundando institutos jurídicos presentes na Diretiva de Proteção de Dados de 1995 (HOOFNAGLE; VAN DER SLOOT; BORGESIUS, 2019). Ao determinar que os dados – e não os responsáveis pelo seu tratamento – são o objeto central da regulação, o paradigma europeu escapa daquele que é apontado como um dos principais problemas do paradigma norte-americano, já que o marco regulatório do GDPR abrange, em tese, todos os dados de saúde.

Cabe mencionar que quando o uso de dados de saúde estiver amparado por razões de interesse público no **domínio da saúde pública**, o GDPR cria exceções às regras estabelecidas, prevendo derrogações à proibição de tratamento dos dados. Assim sendo, em situações em que o desenvolvimento de sistemas de IA for comprovadamente baseado em motivação de relevância sanitária, os dados pessoais podem ser empregados sem consentimento (GDPR, art. 9º, 2) e conservados por tempo indeterminado (GDPR, art. 17º, 3).

O GDPR representa um marco importante na governança de dados e tornou-se a referência legislativa global nessa matéria (KUNER; BYGRAVE; DOCKSEY, 2020). Por isso, o paradigma europeu é hoje o principal padrão para governança de dados de saúde e tem sido replicado por muitos países.

4.3.6 Leis de proteção de dados no Brasil e em outros países

Até os primeiros anos da década de 2020, algumas das maiores jurisdições do mundo, além dos EUA e da UE, criaram leis específicas de proteção e privacidade de dados, que contemplam dados de saúde.

O **Brasil** aprovou a **Lei Geral de Proteção de Dados Pessoais** (LGPD) em 2018 (BRASIL, 2018), que entrou em vigor em agosto de 2020 (as normas sobre sanções administrativas entraram em vigor em agosto de 2021). A LGPD é explicitamente inspirada no GDPR e espelha muitos dos institutos criados pelo padrão europeu. Os dados de saúde são classificados na categoria "dado pessoal sensível" (art. 5º, II) que reproduz a "categoria especial" do GDPR. Pelo regime da LGPD, o tratamento de dados pessoais sensíveis, só pode ocorrer em situações especificadas em lei (art. 11):

a) com consentimento do titular, para finalidades específicas;
b) para cumprimento de obrigação legal ou regulatória;

c) para realização de estudos por órgão de pesquisa;
d) para exercício regular de direitos;
e) para proteção da vida ou da incolumidade física do titular ou de terceiros;
f) para tutela da saúde, em procedimento realizado por profissionais de saúde, serviços de saúde ou autoridade sanitária;
g) para garantia da prevenção à fraude e à segurança do titular.

A lei brasileira dispõe de dois dispositivos específicos para os dados de saúde (LGPD, art. 11, § 4º e § 5º). O primeiro veda comunicação ou uso compartilhado entre controladores de dados de saúde com objetivo de obter vantagem econômica, exceto para portabilidade de dados solicitada pelo titular ou para prestação de serviços de saúde em benefício do titular e para as transações financeiras e administrativas relacionadas. O segundo dispositivo proíbe operadoras de planos privados de saúde de utilizar dados de saúde para praticar seleção de riscos, contratação e exclusão de beneficiários.

Assim, a LGPD cria no Brasil um marco regulatório para proteção de dados de saúde em consonância com o paradigma europeu e alinhado aos padrões internacionais. Estabelece diretrizes rigorosas para a coleta, armazenamento, tratamento e compartilhamento de dados pessoais, que incluem os dados de saúde. Reconhece dados de saúde como uma categoria especial de dados pessoais, dada a sua sensibilidade, e exige que as organizações obtenham consentimento explícito dos titulares para seu uso, salvo em circunstâncias específicas previstas em lei. Concede aos titulares dos dados o direito de acessar, corrigir e solicitar a eliminação de seus dados, bem como de revogar o consentimento a qualquer momento. Impõe obrigações às entidades que tratam esses dados, como garantir sua segurança, confidencialidade e transparência no tratamento, estabelecendo sanções administrativas para violações. E cria a Autoridade Nacional de Proteção de Dados (ANPD), entidade responsável por fiscalizar e garantir a conformidade com a lei. Em suma, a LGPD estabelece um quadro robusto para governança de dados de saúde no Brasil (AITH; DALLARI, 2022).

Na **China**, o marco regulatório de governança de dados de saúde é definido principalmente pela Lei de Proteção de Informações Pessoais (mais conhecida pela sigla da tradução para o inglês PIPL – *Personal Information Protection Law*), que entrou em vigor em novembro de 2021. Na lei chinesa, há muitos elementos em comum com as legislações norte-americana e europeia. Os dados de saúde estão contemplados em institutos semelhantes aos do paradigma europeu, como a consideração de dados de saúde em uma categoria mais sensível de dados pessoais, a definição do consentimento individual como regra para o processamento de

informações, o direito de retirada do consentimento e o direito ao apagamento. Há também explícita menção a "emergência de saúde pública, ou necessária para proteger a vida, a saúde e a propriedade de pessoas físicas em emergência" como circunstâncias para a não exigência de consentimento no processamento de dados pessoais. Logo, apesar de todas as diferenças político-institucionais, a China tem também um marco regulatório de governança de dados de saúde bastante semelhante aos padrões internacionais (WANG, C. *et al.*, 2022).

Na **Índia**, até os anos 2020, privacidade e proteção de dados dependiam de análises casuísticas em decisões judiciais a respeito da Lei de Tecnologia da Informação (*Information Technology Act*) de 2000, diante da ausência de definição específica de dados pessoais no ordenamento jurídico e da falta de delimitação explícita do direito à privacidade da Constituição indiana. Isso foi em parte solucionado pela jurisprudência formada pela Suprema Corte da Índia em 2017, que deu suporte ao entendimento do direito à privacidade como um direito fundamental (CHATTERJEE, 2019). Em agosto de 2023, os legisladores indianos aprovaram a Lei de Proteção de Dados Pessoais Digitais (*Digital Personal Data Protection Bill*). Essa nova lei define dados pessoais como informações relacionadas a uma pessoa identificada ou identificável e estabelece as condições para tratamento de dados mediante consentimento individual. Além disso, concede direitos aos titulares, incluindo o direito de obter informações, buscar correção e apagamento (mas sem o chamado "direito ao esquecimento). Também cria o Conselho de Proteção de Dados da Índia e define penalidades administrativas para violações (PRS LEGISLATIVE RESEARCH, 2023). Embora a nova lei não tenha dispositivos específicos sobre dados de saúde, o país mais populoso do mundo também passa a ter um marco regulatório para governança de dados pessoais.

Além desses, outros países populosos recentemente aprovaram leis de proteção de dados. A **Indonésia** promulgou a lei *Personal Data Protection* ("*PDP Law"*) em outubro de 2022 (REPUBLIC OF INDONESIA, 2023). Na **Nigéria**, entrou em vigor em junho de 2023 o *Nigeria Data Protection Act* (NDPC, 2023).

Até o início de 2023, **162 países** já haviam aprovado leis de privacidade de dados e pelo menos outros 20 países tinham projetos de lei em análise. Espera-se que até o final da década de 2020 haja leis de proteção de dados em todos os países (GREENLEAF, 2023).

4.3.7 Mecanismos para proteção de dados de saúde

A construção de uma IA confiável na área da saúde precisa empregar mecanismos técnicos, regulatórios e de gestão que busquem assegurar a efetivação dos princípios de privacidade e proteção de dados pessoais. Os objetivos básicos da regulação de dados de saúde são dar aos titulares o controle sobre seus dados e responsabilizar os controladores pelos eventuais desvios ou violações, porém evitando encargos desnecessários para o uso de dados pessoais para fins de pesquisa e na saúde pública (COHEN; MELLO, 2018).

O mecanismo central nos atuais paradigmas de proteção de dados é a **desidentificação**. Por definição, dados pessoais são informações que se referem a pessoa natural identificada ou identificável. Logo, qualquer informação que perca a possibilidade ser atribuído a um titular específico perde a qualificação de dado pessoal e passa a não estar sob abrigo das normas protetivas. O GDPR usa o conceito de "pseudonimização", que diz respeito ao tratamento que impossibilita os dados de serem associados a um titular sem recorrer a informações suplementares, que devem ser mantidas separadamente (EUROPEAN UNION, 2016). Mas essa referência não é consensual em leis inspiradas no paradigma europeu. A LGPD brasileira usa preferencialmente a ideia de anonimização, definida como "utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo" (art. 5º), reservando o termo pseudonimização (art. 13) para o uso de dados em estudos de saúde pública (BRASIL, 2018). Sob o paradigma norte-americano, é mais comum o uso do termo *de-identification* em sentido amplo (que provém da HIPAA), em referência a todos os métodos destinados a impedir que dados sejam relacionados a indivíduos específicos. Na literatura biomédica em geral, ainda é preciso construir uniformidade em relação ao uso desses termos (CHEVRIER *et al.*, 2019).

Em sistemas de IA para aplicação na área da saúde, a função da desidentificação é bastante importante, uma vez que os modelos são treinados em grandes bancos de dados sensíveis. Estratégias de desidentificação são utilizadas há algum tempo em estudos de pesquisa multicêntricos que compartilham dados de registros eletrônicos de saúde de diferentes ambientes ou instituições (KUSHIDA *et al.*, 2012). Recentemente, têm sido proposta a utilização de modelos de *deep learning* para desidentificar esse tipo de dado, facilitando o uso no desenvolvimento de outros modelos de IA (AHMED; AZIZ; MOHAMMED, 2020). A anonimização ou pseudonimização são meios para garantir que os modelos sejam desenvolvidos salvaguardando a

privacidade das pessoas e que os seus resultados não sejam usados para rastrear até os pacientes individuais.

Entretanto, os métodos de desidentificação nem sempre são bem-sucedidos. Os dados muitas vezes podem ficar suscetíveis a reidentificação a partir da reconstrução do conjunto de dados (ROCHER; HENDRICKX; DE MONTJOYE, 2019). Além disso, pode não ser possível desidentificar completamente alguns tipos de dados de saúde, como sequências genômicas, pois as relações de identidade podem ser inferidas com base nas informações presentes nos próprios dados que se pretende anonimizar (MAY, 2018). A grande diversidade de fontes de dados de saúde, incluindo os que são fornecidos pelos próprios titulares, tem acrescentado mais uma camada de dificuldade na tarefa de desidentificação. É o caso dos dados de saúde gerados por dispositivos vestíveis, por exemplo (CHIKWETU *et al.*, 2023). Há ainda circunstâncias em que o anonimato pode reduzir os benefícios positivos dos dados para os próprios titulares ou prejudicar o seu direito de exercer controle sobre os dados (VAYENA; BLASIMME, 2017). Por isso, é essencial que os marcos regulatórios exijam dos controladores de dados o emprego de métodos cada vez mais seguros e sofisticados de desidentificação (MURDOCH, 2021).

O **consentimento** é o principal mecanismo normativo presente nas legislações de proteção de dados para garantir o direito à privacidade. A concepção do consentimento individual é diretamente relacionada ao intuito de conceder autonomia e controle aos titulares dos dados. Para isso, é necessário que o consentimento seja livre e informado, ou seja, fornecido apenas após explicações simples e acessíveis sobre quais dados serão utilizados e sobre as possíveis consequências do uso, incluindo a compreensão dos riscos envolvidos (GERKE; MINSSEN; COHEN, 2020). Vêm sendo consideradas abordagens alternativas, como o "consentimento dinâmico" e o "amplo consentimento". Consentimento dinâmico significa possibilitar modificações periódicas para diferentes usos, de modo que os titulares possam autorizar alguns e negar outros, alterando suas escolhas em relação à utilização e reutilização de seus dados ao longo do tempo (VAYENA; BLASIMME, 2017). Amplo consentimento é uma forma de facilitar o uso secundário de dados de saúde sem comprometer privacidade e autonomia, como tem sido proposto por instituições que depositam dados em biorrepositórios e recolhem consentimento para uso futuro em pesquisas (SPECTOR-BAGDADY *et al.*, 2020).

Da perspectiva da regulação a IA, é também relevante considerar as circunstâncias em que o consentimento pode ser dispensado. Os marcos regulatórios de proteção de dados preveem que isso deve ocorrer nas situações em que deve prevalecer o interesse público sobre o direito à privacidade. A **dispensa do consentimento** decorre da ponderação que resulta na

admissão de que o direito fundamental do indivíduo seja relativizado diante de um objetivo legítimo em benefício público, desde que seja necessário, proporcional e que não haja outros meios menos restritivos para consecução desse objetivo (TOWNEND, 2021). Exemplo claro: numa situação em que o uso de dados de saúde de muitos indivíduos for essencial para o desenvolvimento de um sistema de IA potencialmente capaz de salvar vidas. Nesse caso, partilhar os dados de saúde para o bem público poderia até mesmo ser considerado um dever da autoridade sanitária (COHEN, 2018). Para isso, deve haver uma condição de saúde pública bem estabelecida e que seja identificada e informada de forma transparente, além de não existir interesse comercial ou monetário. A questão jurídica de base é a busca do equilíbrio entre situações em que os dados de saúde devem ser considerados um bem público – em que se autoriza a dispensa de consentimento – e situações em que deve prevalecer o seu estatuto privado. A proteção da privacidade é a regra geral, que só pode ser relativizada em contextos fáticos nos quais a partilha de dados gerar benefícios claros para a saúde pública que de outra forma não estariam disponíveis (PRICE; COHEN, 2019).

No contexto de **proteção de dados**, o "aprendizado federado" (*federated learning*) é um mecanismo diretamente relacionado às tecnologias de IA que merece destaque. Trata-se de uma técnica de *machine learning* que consiste no treinamento de modelos em vários dispositivos ou servidores que contêm amostras de dados locais, sem a necessidade de centralizar os dados. A principal vantagem do **aprendizado federado** é possibilitar o desenvolvimento de modelos com maior volume e diversidade de dados, preservando privacidade e confidencialidade, além de garantir mais segurança, pois os dados brutos não saem do seu local original (YANG *et al.*, 2019). Cria-se um ambiente colaborativo para o desenvolvimento de modelos de IA que facilita a conformidade com as normas de proteção de dados pessoais e ao mesmo tempo melhora a robustez da gestão de dados, pois minimiza a possibilidade de haver um único ponto de falha (por empregar dados de diversas fontes). Assim, além de assegurar privacidade e segurança dos dados, pode melhorar o desempenho e a generalização dos modelos (LI *et al.*, 2020). Na IA aplicada à saúde, embora se reconheçam limitações como a heterogeneidade dos dados de saúde e dificuldades em relação a rastreabilidade e responsabilização, o aprendizado federado é considerado bastante promissor (RIEKE *et al.*, 2020).

Além desses, há também mecanismos institucionais para proteção de dados pessoais que são relevantes para a regulação da IA aplicada à saúde. A estruturação de **centros de dados** (*data hubs*) que agregam dados de saúde é um dos principais desses mecanismos. Um centro de dados de saúde é uma infraestrutura técnica digital com a missão principal de permitir a

partilha de dados de saúde, assegurando acessibilidade em conformidade com a regulação de governança de dados (ALVAREZ-ROMERO *et al.*, 2023). Um bom exemplo é o Espaço Europeu de Dados de Saúde (*European Health Data Space* – EHDS), apresentado como uma proposta regulatória pela Comissão Europeia em maio de 2022 (EUROPEAN COMMISSION, 2022d). O EHDS tem o objetivo de criar um "mercado único dos produtos e serviços de saúde digitais, por intermédio da harmonização das regras", principalmente dispositivos médicos e sistemas de IA para uso na saúde. A concepção visa apoiar a utilização de dados de saúde para inovação, pesquisa e elaboração de políticas, resguardando o controle dos dados pessoais pelos titulares.

Em suma, a governança de dados é fundamento para a regulação da IA na saúde e são necessárias medidas para que seja exercida de maneira sólida e eficaz (WHO, 2021). Os países devem ter legislações em matéria de proteção de dados pessoais que contemplem requisitos para utilização de dados de saúde, especialmente o direito ao consentimento informado. Devem ser estabelecidas autoridades independentes de proteção de dados, com estruturas adequadas para monitorar o cumprimento das regulações de proteção de dados. Precisa haver transparência sobre as condições de uso dos dados de saúde pelas entidades que controlarem esses dados, sejam públicas ou privadas. O objetivo principal do compartilhamento de dados de saúde sempre deve ser a busca do maior benefício público possível.

5 DIRETRIZES PARA REGULAÇÃO DA IA NA SAÚDE

5.1 Segurança e Eficácia

As primeiras diretrizes a serem estabelecidas para regulação da IA na saúde pertencem à dimensão de segurança e eficácia. São requisitos cruciais que decorrem diretamente da natureza do uso desses sistemas de IA. Os modelos de IA/ML desenvolvidos para aplicação na área da saúde ficam submetidos à atividade estatal de regulação de produtos para saúde quando passam a funcionar como dispositivos médicos – especificamente, ao serem qualificados na categoria de "*softwares* como dispositivos médicos". Segurança e eficácia são os objetivos primordiais da regulação de dispositivos médicos.

5.1.1 *Softwares* como dispositivos médicos (SaMD)

A definição do termo "dispositivo médico" está uniformizada desde 2005, a partir do trabalho da Força-Tarefa de Harmonização Global (GHTF – *Global Harmonization Task Force*), constituída por representantes de autoridades reguladoras sanitárias de diversos países. Pela definição harmonizada, **dispositivo médico** é qualquer instrumento, aparelho, máquina, implante, reagente *in vitro*, *software*, ou qualquer outro material destinado a ser utilizado isoladamente ou em combinação, por seres humanos, para uma ou mais das seguintes finalidades:

a) diagnóstico, prevenção, monitoramento, tratamento ou alívio de doenças;
b) diagnóstico, monitoramento, tratamento, alívio ou compensação de uma lesão;
c) investigação, substituição, modificação ou suporte da anatomia ou de um processo fisiológico;
d) apoio ou sustentação da vida;
e) controle da concepção;
f) desinfecção de dispositivos médicos;
g) fornecimento de informações para fins médicos ou de diagnóstico por meio de exame *in vitro* de amostras derivadas do corpo humano.

Além disso, é requisito para ser enquadrado como dispositivo médico que tal objeto não atinja a ação principal pretendida no corpo humano por meios farmacológicos, imunológicos ou metabólicos, mas que possa ser auxiliado na função pretendida por tais meios (GHTF, 2005). Essa definição tornou-se referência global e tem sido adotada como base por muitas agências reguladoras nos últimos anos. Embora seja considerada uma definição complexa, por ter sido concebida principalmente para reguladores (ARONSON; HENEGHAN; FERNER, 2020), continua sendo utilizada em várias jurisdições como definição essencial da regulação sanitária aplicada a dispositivos médicos.

Desde 2011, as autoridades sanitárias de Austrália, Brasil, Canadá, China, União Europeia, Japão e Estados Unidos estabeleceram o Fórum Internacional de Reguladores de Dispositivos Médicos (**IMDRF – *International Medical Device Regulators Forum***). O IMDRF funciona como um grupo permanente que dá continuidade ao trabalho da GHTF, com o escopo de acelerar a harmonização e a convergência regulatória internacional de dispositivos médicos. Além dos fundadores, atualmente compõem o IMDRF as autoridades reguladoras de Rússia, Singapura, Coreia do Sul e Reino Unido, bem como representantes da OMS e das agências reguladoras da Suíça e da Argentina, como observadores oficiais (IMDRF, 2023).

O IMDRF tem definido diretrizes para estruturação de marcos regulatórios de *software* como dispositivo médico (**SaMD – *Software as a Medical Device***) desde 2013. O conceito geral é o de que todo *software* destinado a ser usado para uma ou mais finalidades médicas, desde que execute essas finalidades sem fazer parte de um dispositivo médico de hardware, deve ser classificado como um SaMD (IMDRF, 2013). A regulação de SaMD vem vendo um componente essencial do cenário regulatório de dispositivos médicos, especialmente porque aplicativos e sistemas computacionais desempenham um papel cada vez mais importante em todas as dimensões da área da saúde.

O paradigma geral para a regulação de SaMD é baseado na **classificação de risco**, assim como ocorre na regulação dos demais dispositivos médicos. Portanto, todo *software* qualificado como dispositivo médico deve necessariamente ser avaliado para categorização feita com base na combinação dois elementos: 1) **a importância da informação fornecida** pelo SaMD para a decisão na assistência à saúde; e 2) **a situação ou condição de saúde** em que o SaMD é aplicado. As quatro categorias (I, II, III e IV) representam níveis de impacto crescente no paciente ou na saúde pública, de modo que a categoria I significa baixo impacto e a categoria IV o nível de impacto mais elevado (Quadro 5.1). A categorização de um SaMD conforme o risco é o que deve orientar quais requisitos regulatórios se aplicam a ele (IMDRF, 2014).

Quadro 5.1 – Categorias de SaMD

Situação ou condição do estado de saúde	Importância das informações fornecidas pelo SaMD para a decisão de saúde		
	Tratar ou diagnosticar	Conduzir a gestão clínica	Informar a gestão clínica
Crítico	IV	III	II
Grave	III	II	I
Não-grave	II	I	I

Fonte: Adaptado de IMDRF (2014, p.14).

A principal referência internacional para gerenciamento de risco em dispositivos médicos é o padrão **ISO 14971** (atualmente na 3ª Edição, publicada em 2019), da Organização Internacional para Padronização (ISO – *International Organization for Standardization*), que especifica a terminologia, os princípios e o processo para gestão de riscos de dispositivos médicos, incluindo os SaMD. Esse padrão parte da premissa de que a utilização de um dispositivo médico envolve um grau inerente de risco, conceituado a partir dois componentes principais: 1) a probabilidade da ocorrência de danos; e 2) quão graves podem ser as consequências desses danos. Os riscos podem estar relacionados com lesões para o paciente, para o utilizador ou para terceiros, bem como com danos a propriedades ou o meio ambiente (ISO, 2019).

O processo de **gestão de riscos**, tal como definido pela ISO 14971:2019, começa pela determinação de potenciais fontes de danos (perigos) que podem estar associadas a um determinado dispositivo. Uma vez identificados os perigos, exploram-se as contingências de sequências de eventos que podem levar a situações em que há exposição a tais perigos. Em seguida, determina-se a probabilidade e a gravidade do dano, chamada estimativa de risco. O risco estimado é então confrontado com critérios de aceitabilidade de risco – que são previamente determinados pelo fabricante do dispositivo como parte do plano de gestão de risco. Essa etapa define se são necessárias medidas de controle de risco, que são executadas até que a estimativa de risco seja compatível com os critérios de aceitabilidade, de modo a equilibrar os riscos residuais em relação aos benefícios esperados do uso do dispositivo – se o risco não for redutível a níveis aceitáveis, o dispositivo deve retornar à fase de projeto. Esse processo de identificação de perigos e estimativa dos riscos associados é chamado de **análise de risco**. Todo esse procedimento para avaliação e revisão contempla as fases de projeto, produção e pós-produção. Esses

parâmetros para gestão de riscos que se aplicam a SaMD também devem ser empregados na regulação de dispositivos médicos baseados em sistemas de IA (ODAIBO, 2021).

Ainda na gestão de riscos, as diretrizes para regulação de SaMD contemplam o monitoramento durante o **ciclo de vida** dos *softwares*, compatível com a gestão da qualidade e a aplicação de critérios de boas práticas industriais, conforme padronizado pela IEC 62304, documento publicado pela Comissão Eletrotécnica Internacional (IEC – *International Electrotechnical Commission*) em 2006 (IEC, 2006). Os fabricantes devem seguir requisitos para desenvolvimento, implementação e documentação de *software* robusto e confiável, proporcional ao risco e conforme o uso pretendido. Devem ser conduzidos procedimentos de vigilância pós-comercialização, tanto por avaliação da percepção dos usuários como por mecanismos de detecção automática de falhas. Além disso, os fabricantes de SaMD devem ter um grau de controle adequado para gerenciamento de modificações nos seus produtos, uma vez que as alterações nos *softwares* podem ter efeitos significativos na situação dos cuidados de saúde e no ambiente sociotécnico de utilização do SaMD (IMDRF, 2014).

Há também diretrizes estabelecidas pelo IMDRF específicas para aplicação de um **sistema de gestão da qualidade** (QMS – *Quality Management System*). Enquanto os padrões internacionalmente reconhecidos para muitos setores industriais estão na família de normas ISO 9000, os específicos para os dispositivos médicos estão atualmente estabelecidos pelo padrão **ISO 13485** (atualmente na 3ª Edição, publicada em 2016), que especifica requisitos para um sistema de gestão da qualidade para todas as etapas produtivas – projeto e desenvolvimento, produção, armazenamento e distribuição, instalação e manutenção de dispositivos médicos (ISO, 2016). Conforme esse padrão, um sistema de gestão de qualidade para SaMD deve se basear em:

> a) uma estrutura organizacional que proporcione governança, responsabilização e prestação de contas (*accountability*);
>
> b) um conjunto de processos de suporte ao ciclo de vida do SaMD – planejamento de produto, gestão de riscos, controle de documentos e registros, monitoramento e análise para melhoria de processos e produtos (incluindo terceirizados); e
>
> c) um conjunto de processos de realização e uso do SaMD – desenho do produto, desenvolvimento, avaliação e validação, implementação, manutenção, desativação.

Esses elementos devem ser escaláveis de acordo com o tamanho da organização e devem manter o foco em assegurar segurança, eficácia e desempenho do SaMD (IMDRF, 2015).

A **avaliação clínica** é uma dimensão crucial para a regulação de SaMD. O IMDRF desenvolveu uma estrutura para avaliação clínica que tem sido adotada pelas autoridades sanitárias dos países que o compõem e por agências reguladoras globalmente. Essa proposta integra os conceitos de gestão de riscos e os princípios do sistema de gestão da qualidade com intuito de aferir o impacto de um SaMD a fim de torná-lo clinicamente significativo para os usuários. A abordagem parte da concepção de que um SaMD é um *software* que utiliza um algoritmo (que pode ser baseado em regras ou um modelo de IA) e que funciona pelo processamento de dados de entrada para produzir uma saída. A saída representa um resultado destinado a fins médicos ou de saúde, conforme a intenção de uso definida pelo fabricante (IMDRF, 2017).

Todos os dispositivos médicos requerem avaliação clínica, independentemente da sua classificação de risco. Os riscos e benefícios associados ao uso de um SaMD estão relacionados a possíveis saídas imprecisas ou erradas que possam impactar o manejo clínico ou alguma situação de saúde. Esses riscos devem ser considerados ao determinar o tipo e a quantidade de dados necessários para apoiar a finalidade pretendida e as alegações clínicas esperadas. Assim, os fabricantes devem justificar o nível de evidência clínica fornecida e demonstrar que o SaMD foi testado adequadamente no ambiente clínico pretendido e que as intenções de uso têm a relação risco-benefício resultante claramente aceitável à luz do estado da arte (KWADE, 2021).

O **Processo de Avaliação Clínica** estabelecido pelo IMDRF é concebido como um conjunto de atividades contínuas conduzidas para avaliação e análise da segurança, da eficácia e do desempenho clínico de um SaMD (IMDRF, 2017). Esse processo é integrado por três componentes:

a) associação clínica válida;
b) validação analítica;
c) validação clínica.

Associação clínica válida refere-se à medida em que o resultado obtido por um SaMD é clinicamente aceito ou bem fundamentado com base num quadro científico estabelecido ou conjunto de evidências, de forma a corresponder a situações do mundo real e condições de saúde identificadas na definição do SaMD. Também chamada de “validade científica”, é um indicador da aceitação de quanto significado e confiança podem ser atribuídos aos resultados do SaMD. Evidências existentes podem ser apresentadas a partir de literatura científica,

pesquisas clínicas originais, diretrizes de sociedades profissionais ou análises de dados secundários. É também possível gerar evidências conduzindo novos ensaios clínicos.

Validação analítica (também conhecida como validação técnica) mede a capacidade de um SaMD gerar o resultado técnico pretendido com acurácia, precisão e confiabilidade – ou seja, de processar corretamente os dados sempre da mesma maneira. Esse componente mostra objetivamente se o *software* foi construído corretamente e atende às especificações em conformidade com as necessidades do usuário e com os usos pretendidos. Os desenvolvedores devem gerar evidências para demonstrar que a saída do SaMD é tecnicamente o que se espera de acordo com o uso pretendido. Essas evidências podem ser produzidas durante as atividades de validação (como parte do sistema de gestão da qualidade ou de boas práticas de engenharia de *software*) ou podem ser geradas por meio de bancos de dados coletados previamente.

Validação clínica mede a capacidade de um SaMD gerar um resultado clinicamente significativo para a situação de saúde pretendida. Refere-se ao impacto positivo que pode ter na saúde de indivíduos ou da população, especificado como benefício clínico, mensurável e relevante, para saúde individual ou saúde pública. Essa avaliação deve ser feita pelo fabricante durante o desenvolvimento de um SaMD antes e depois da comercialização. A relação entre os resultados do SaMD e as situações clínicas de interesse podem ser demonstradas a partir de dados existentes de estudos realizados para o mesmo uso pretendido ou para um uso diferente, desde que a extrapolação seja devidamente justificada. Também é possível gerar novos dados clínicos para um uso específico pretendido. A validação clínica pode ser aferida e apresentada por diversas métricas, tais como sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo, número necessário para tratar (NNT), número necessário para causar dano (NNH) e outras medidas usadas em epidemiologia e estatística (v. seções 3.3.5 e 3.4.2).

As diretrizes do IMDRF para regulação da SaMD enfocam ainda o **aprendizado contínuo usando dados do mundo real**. A concepção central é baseada na coleta de dados e análise contínua num ambiente pós-comercialização, pois esse monitoramento propicia relevante contribuição para evolução da funcionalidade e do uso pretendido do SaMD. Os dados de desempenho no mundo real devem incluir informações sobre segurança e eficácia a partir de geração contínua de evidências clínicas, que podem fornecer provas de que a validação analítica ou a validação clínica de um SaMD é superior ou inferior às medidas inicialmente avaliadas pelo fabricante. Essas evidências clínicas adicionais devem ser utilizadas pelo fabricante para atualizar a avaliação clínica, o que eventualmente pode resultar na habilitação de novas

funcionalidades do SaMD ou na desativação de funcionalidades até então existentes. Em decorrência da avaliação clínica contínua, a categorização do risco e o uso pretendido do SaMD podem mudar.

Essas diretrizes não são específicas para dispositivos médicos baseados em *machine learning* (ML). Aplicam-se a outras situações em que haja coleta de informações pós-comercialização em um SaMD (IMDRF, 2017). Mas toda a abordagem é plenamente convergente com elementos essenciais reconhecidos para monitoramento contínuo e atualização de modelos para a IA/ML na área da saúde (FENG *et al.*, 2022).

5.1.2 Dispositivos médicos baseados em *machine learning* (MLMD)

Até o segundo semestre de 2023, ainda não havia diretrizes definidas pelo IMDRF particularmente para dispositivos médicos baseados em *machine learning*. Um grupo de trabalho constituído por representantes de diversos países e entidades internacionais de todos os continentes foi formado com o objetivo de construir um consenso regulatório nesse campo, a partir da definição de critérios de boas práticas (GMLP – *Good Machine Learning Practice*) e princípios harmonizados globalmente.

O primeiro documento técnico produzido por esse grupo foi publicado em 2022 e propõe estabelecer termos e definições para promover consistência a fim de fornecer uma base para o desenvolvimento de futuras diretrizes para uma nova categoria de dispositivos médicos: **MLMD – *Machine Learning-enabled Medical Devices*** (IMDRF, 2022). O documento apresenta definições atualmente harmonizadas para termos como “viés”, “treinamento”, “aprendizado contínuo”, “aprendizado supervisionado”, “aprendizado semissupervisionado”, “aprendizado não-supervisionado” “conjunto de dados de teste” e “conjunto de dados de treinamento” (v. seções 2.3.3 e 3.2.2). Esses e outros termos estão atualmente estabelecidos no principal padrão sobre conceitos e terminologia de IA: o **ISO/IEC 22989**, cuja primeira edição também foi publicada em 2022 (ISO/IEC, 2022).

Um dos principais elementos das diretrizes regulatórias a serem criadas especificamente para a nova categoria MLMD será abordar a **capacidade de mudança** inerente aos modelos de *machine learning*. Propõe-se observar tanto alterações nos próprios modelos de ML como alterações no ambiente de uso relativamente aos dados de treinamento dos modelos de IA/ML

(IMDRF, 2022). Assim, começam a ser delineados os aspectos que deverão ser objeto da regulação da IA na saúde na dimensão de segurança e eficácia.

Em relação às **mudanças nos MLMD**, a recomendação é que sejam observados os atributos que descrevem todos os elementos envolvidos:

a) **causa** – origem da mudança (ex.: ajustes dos modelos, novos treinamentos, novos métodos, novos dados, novos algoritmos etc.)
b) **efeito** – alteração resultante (ex.: mudanças nas indicações de uso, no desempenho, nas entradas ou nas saídas dos modelos);
c) **desencadeante** – evento que provoca a mudança (ex.: limitações de desempenho ou limitações de dados, mudanças no ambiente etc.);
d) **domínio** – mudança homogênea (em todo o âmbito de aplicação do modelo) ou heterogênea (adaptação local específica, ex.: a uma região ou a uma população);
e) **efetivação** – mecanismo para implementação da mudança, externo (feita por desenvolvedor ou usuário) ou interno (feita por *software* dentro do dispositivo).

No que se refere às **mudanças no ambiente dos MLMD**, busca-se apreciar os aspectos associados a alterações na configuração dos dispositivos em relação aos dados de desenvolvimento dos modelos de ML:

a) **causa** – origem da mudança em relação ao ambiente de desenvolvimento (ex.: alterações na forma ou na qualidade dos dados de entrada, alterações na população, alterações na prática clínica etc.);
b) **efeito** – piora ou melhora no desempenho, segurança ou eficácia;
c) **domínio** – mudança homogênea (em todo o âmbito de aplicação do modelo) ou heterogênea (adaptação local específica, ex.: a uma região ou a uma população).

A abordagem regulatória específica para dispositivos médicos baseados em ML com perspectiva de harmonização internacional está ainda em fase de definição (até o final do ano de 2023) e, muito provavelmente, deve levar alguns anos para ser construída. Enquanto isso, os primeiros produtos que buscam aprovação têm sido submetidos à regulação existente para SaMD em geral. Os critérios para regular os MLMD começaram a ser desenhados a partir do final da década de 2010, principalmente nos Estados Unidos e na União Europeia. Esses dois paradigmas guardam muitas semelhanças (GILBERT *et al.*, 2021), mas recentemente têm apresentado divergências relevantes, sobretudo em razão da adoção de estratégias regulatórias mais flexíveis no cenário norte-americano (GILBERT; ANDERSON; *et al.*, 2023).

5.1.3 Paradigma norte-americano

A agência reguladora norte-americana FDA (*Food and Drug Administration*) adota o paradigma do IMDRF para regulação de SaMD, aplicando as suas diretrizes (FDA, 2020), exceto por utilizar três classes de risco: classe I – risco mais baixo; classe II – risco intermediário; classe III – risco mais elevado; em vez das quatro categorias do IMDRF.

Na atual regulação dos EUA, há três caminhos possíveis para aprovação de um SaMD:

a) **notificação pré-comercialização** – mais conhecida por "autorização 510(k)", é concedida quando se demonstra que um SaMD é "substancialmente equivalente" (ou seja, tão seguro e eficaz quanto) outro dispositivo comercializado legalmente;
b) **classificação *De Novo*** – usada para novos SaMD da classe I ou classe II de risco para os quais não existem equivalentes comercializados e que demonstrem segurança e eficácia pela avaliação geral de risco;
c) **aprovação pré-comercialização** – usada para os SaMD da classe III de risco, tem um processo regulatório mais rigoroso, que deve ser apoiado por evidências científicas completas.

Os primeiros dispositivos médicos baseados em IA/ML vêm sendo aprovados por meio desse marco regulatório existente para SaMD, sendo a absoluta maioria pelo caminho da notificação simples, a "autorização 510(k)" – de modo que a avaliação regulatória completa para esses dispositivos tem sido residual (BENJAMENS; DHUNNOO; MESKÓ, 2020). Na ausência de uma estrutura regulatória própria, a agência norte-americana criou a regra do "**algoritmo travado**", limitando as autorizações para comercialização apenas da versão do SaMD que foi submetida, sem permitir alterações após a aprovação regulatória. A FDA definiu que os modelos deveriam permanecer "travados" no sentido de fornecer o mesmo resultado cada vez que a mesma entrada fosse aplicada e não mudar com o uso. Essa estratégia foi assumida pela incapacidade dos reguladores de lidar com o chamado "problema da atualização", uma vez que não havia requisitos estabelecidos para avaliar sistemas de IA adaptativos. Entretanto, trata-se de uma estratégia ineficaz e que pode até oferecer mais riscos, pois modelos de IA que não mudam com o uso têm menor utilidade clínica e ficam mais suscetíveis a deterioração de desempenho ao longo do tempo (BABIC *et al.*, 2019; PARIKH; OBERMEYER; NAVATHE, 2019).

A FDA reconheceu que o marco regulatório de SaMD não é adequado para dispositivos baseados em IA/ML, por não ter sido concebido para uso em tecnologias que se adaptam continuamente e mudam ao longo do tempo e porque os procedimentos previstos para alterações pós-comercialização não têm a dinâmica necessária para regular essa modalidade de produto de saúde. Nesse contexto, a agência norte-americana publicou, em abril de 2019, uma **proposta de estrutura regulatória para mudanças em SaMD baseados em *machine learning***. Nessa proposta, a FDA delineou uma abordagem regulatória do ciclo de vida total dos dispositivos, que prevê um "plano de controle de mudanças predeterminado" para submissões pré-comercialização. Esse plano englobaria as modificações antecipadas, referidas como "pré-especificações de SaMD", e um método empregado para efetuar tais alterações de forma controlada, gerenciando os riscos aos pacientes, intitulado "protocolo de mudança de algoritmo". As mudanças podem ser de três categorias:

a) desempenho clínico ou técnico-analítico;
b) entrada usada pelo algoritmo e associação clínica com a saída do SaMD;
c) uso pretendido do SaMD conforme a classificação de risco.

Nesse enfoque, a FDA busca incentivar um comprometimento por parte dos fabricantes quanto à transparência e o monitoramento do desempenho no mundo real dos SaMD baseados em ML. Tem o escopo de possibilitar que tanto a FDA quanto os fabricantes monitorem os dispositivos desde a fase pré-comercial até seu desempenho pós-comercialização (FDA, 2019).

Essa proposta regulatória, apresentada em um documento de discussão aberto para receber contribuições das partes interessadas, deu origem ao **Plano de Ação para SaMD baseados em IA/ML**, publicado em janeiro de 2021. Nesse plano, a FDA ressalta a intenção de atualizar a estrutura regulatória, desenvolvendo uma regulação específica – assim como o IMDRF propõe para a nova categoria MLMD. Apresenta o conceito de "boas práticas" específicas para os dispositivos de IA/ML (GMLP – *Good Machine Learning Practice*), envolvendo gestão de dados, treinamento e avaliação de modelos, interpretabilidade e documentação. Enuncia o compromisso em apoiar uma abordagem regulatória "centrada no paciente", enfocando a transparência dos fabricantes para com os usuários dos SaMD baseados em IA/ML, e que ao mesmo tempo procure identificar e eliminar vieses algorítmicos, especialmente étnico-raciais e socioeconômicos. Por fim, o plano prevê a criação de pilotos de avaliação de "desempenho no mundo real" para avançar na abordagem de ciclo de vida total do produto (FDA, 2021).

Outro aspecto essencial do paradigma regulatório norte-americano é a abordagem aos **sistemas de suporte à decisão clínica** (CDSS – *clinical decision support systems*). A "Lei de Curas" (*21st Century Cures Act*), promulgada em 2016, modificou o processo de aprovação de medicamentos e dispositivos médicos pela FDA com intuito de acelerar a entrada de novos produtos no mercado do país. Essa lei foi um dos principais marcos do movimento político que teve como objetivo a mudança do papel da FDA de "fiscalizador de segurança" para "facilitador de inovação" (LIEVEVROUW; MARELLI; VAN HOYWEGHEN, 2022).

Dentre várias medidas nesse sentido, a Lei de Curas redefiniu requisitos regulatórios para *softwares* de suporte a decisões cínicas, gerando incerteza em relação à aplicabilidade dos procedimentos até então adotados pela agência. Após alguns anos de debates, em documento de orientação publicado em setembro de 2022, a FDA definiu os critérios para distinguir *softwares* de suporte à decisão clínica que são considerados dispositivos médicos – e, portanto, qualificados como SaMD da perspectiva regulatória – de *softwares* considerados "não-dispositivos" – que não ficam submetidos ao processo regulatório e passam a ser denominados "*Non-Device CDS*". Em síntese, um *software* que cumpra todos os quatro critérios, cumulativamente, passa a ser reconhecido pela FDA como **não-dispositivo médico** (FDA, 2022):

a) (1) não adquire, processa ou analisa imagens, sinais ou padrões médicos;
b) (2) exibe, analisa ou imprime informações médicas geralmente comunicadas entre profissionais (ex.: estudos clínicos revisados por pares, diretrizes de prática clínica);
c) (3) fornece recomendações (informações ou alternativas) ao profissional de saúde sobre prevenção, diagnóstico ou tratamento, sem fornecer uma diretiva específica;
d) (4) permite que o profissional de saúde revise de forma independente a base dessas recomendações para não depender principalmente delas para tomar uma decisão.

Embora nessas diretivas a FDA não faça distinção de *softwares* baseados em IA/ML, essa nova orientação pode colocar uma parcela dos sistemas de IA desenvolvidos para aplicação na área da saúde fora do alcance regulatório, já que atualmente a grande maioria dos sistemas de suporte à decisão clínica utiliza *machine learning*. Mas necessariamente deve haver restrição ao enquadramento como não-dispositivo médico dos sistemas que empreguem modelos opacos. O preenchimento do critério 4 é inviável em caso de uso de algoritmos complexos que funcionem como caixa preta, pois não permitem que profissionais de saúde revisem de forma independente a base para as recomendações que resultam deles (GERKE, 2023).

Além disso, o atual paradigma norte-americano permite que *softwares* sejam excluídos da regulação de dispositivos médicos pelo uso pretendido. Se um fabricante declara que um sistema de IA é destinado a ser um "produto de bem-estar geral", ele deixa de estar sujeito ao quadro regulatório de SaMD. A Lei de Curas expressamente dispõe que *softwares* desenvolvidos "para manter ou encorajar um estilo de vida saudável" e que não sejam diretamente relacionados com diagnóstico, cura, mitigação, prevenção ou tratamento de condições clínicas não podem ser enquadrados como dispositivos médicos (MINSSEN *et al.*, 2020).

Embora tenha havido relevante flexibilização da abordagem regulatória norte-americana desde a entrada em vigor da Lei de Curas, essas estratégias flexíveis alinham-se também com a **discricionariedade na aplicação** das regras pela FDA. Essa prática frequentemente adotada pela agência consiste no uso de seu poder discricionário conferido por lei para dispensar determinados requisitos de registro a produtos que são considerados de menor risco (BABIC *et al.*, 2021b; GILBERT; ANDERSON; *et al.*, 2023). Contudo, até o momento ainda não está claro como essas critérios serão usados para qualificar um sistema baseado em *machine learning* como não-dispositivo médico, de modo que o impacto nesse mercado ainda está por ser observado nos próximos anos (VAN LAERE; MUYLLE; CORNU, 2021).

Por fim, cabe observar um argumento relevante que começa a surgir no contexto norte-americano, mas que pode influenciar abordagens regulatórias globalmente: a necessidade de mudança de perspectiva dos reguladores, da avaliação de produtos para a **avaliação de sistemas**. A ideia central é que a estratégia tradicional da regulação em saúde é insuficiente para regular os SaMD baseados em IA/ML. Em vez de apreciar a adequação de dispositivos como produtos individuais, a regulação deveria considerar a interação humano-máquina (v. seção 3.3.6) e envolver a apreciação não apenas do produto em si, como também da sua interação profissionais de saúde, pacientes, outros dispositivos e com os dados coletados e utilizados. Essa "visão de sistema" pode ser o caminho para garantir segurança e eficácia dos dispositivos desenvolvidos com sistemas de IA, sobretudo diante do seu caráter adaptativo. Isso implicaria em alterações significativas nas atuais normas e mesmo na estrutura dos órgãos reguladores, algo que deve ser considerado no futuro (GERKE *et al.*, 2020).

5.1.4 Paradigma europeu

O marco regulatório de *softwares* como dispositivos médicos na UE está atualmente estabelecido pelo Regulamento de Dispositivos Médicos (**MDR – *Medical Device Regulation***), promulgado em 2017 e que começou a ser aplicável a partir de maio de 2021 (EUROPEAN UNION, 2017). O MDR é uma norma geral para todos os tipos de dispositivos médicos, que substitui a antiga Diretriz de Dispositivos Médicos (MDD – *Medical Device Directive*), no intuito de atualizar a regulação diante do progresso tecnológico ocorrido nos últimos anos. Nesse sentido, o novo regulamento mantém boa parte dos institutos criados pela diretriz anterior e atualiza alguns deles (NIEMIEC, 2022).

O MDR estabelece "requisitos gerais de segurança e desempenho", com normas relativas a aspectos a fabricação, rotulagem, documentação e princípios gerais de classificação de risco aplicáveis a todo e qualquer dispositivo médico. A verificação quanto à adequação a esses requisitos é feita num processo chamado "avaliação da conformidade", que atribui a **marca CE** (*Conformité Européenne*) aos produtos que os cumprem, representando a autorização para distribuição ou colocação no mercado europeu. Merece destaque o mecanismo de rastreabilidade definido por meio de um sistema de identificação única dos dispositivos, denominado "sistema UDI" (sigla inglesa de *Unique Device Identification*).

A **classificação de risco** no paradigma europeu utiliza as mencionadas diretrizes do IMDRF e adota o ISO 14971 como padrão harmonizado para gestão de riscos de dispositivos médicos (EUROPEAN UNION, 2022). Portanto, é obrigatória a demonstração pelos fabricantes de que o desempenho dos dispositivos atinge um patamar a partir do qual os riscos de segurança sejam aceitáveis em relação aos benefícios esperados. Mas o MDR emprega uma nomenclatura distinta para categorização de riscos em escala crescente: **classes I**, **IIa**, **IIb** e **III**.

Uma importante particularidade do marco regulatório estabelecido pelo MDR diz respeito à forma como são realizados os processos de avaliação da conformidade. A conformidade dos dispositivos da classe I de risco pode ser autodeclarada pelo fabricante. No caso de dispositivos de maior risco (classes IIa, IIb e III), a conformidade com os requisitos deve ser verificada por empresas privadas designadas pelos Estados-Membros, que recebem a qualificação de "organismos notificados". Assim, a principal agência reguladora europeia, a Agência Europeia de Medicamentos (EMA – *European Medicines Agency*), não é o principal regulador de dispositivos médicos no espaço europeu (MINSSEN *et al.*, 2020).

No que se refere à regulação de SaMD, há regras específicas para a **classificação de risco de *softwares*** (MDR, Anexo VIII, seção 6.3, Regra nº 11):

a) *softwares* destinados a fornecer informações a serem utilizadas para a tomada de decisões com finalidade diagnóstica ou terapêutica são, como regra geral, classificados na classe IIa;
b) se tais decisões tiverem um impacto que possa causar uma deterioração grave do estado de saúde de uma pessoa ou uma intervenção cirúrgica, são classificados na classe IIb;
c) caso as decisões puderem provocar morte ou deterioração irreversível do estado de saúde de uma pessoa, passam a ser enquadrados na classe III;
d) *softwares* destinados à monitorização de processos fisiológicos são, em regra, classificados na classe IIa;
e) *softwares* destinados à monitorização de parâmetros fisiológicos vitais, quando a natureza das variações desses parâmetros for tal que possa resultar em perigo imediato para o paciente, são enquadrados na classe IIb;
f) todos os outros *softwares* são classificados como classe I.

Com base nessas regras, muitos dispositivos que antes da entrada em vigor do MDR se enquadravam na classe de risco mais baixa passaram a uma classe de risco mais elevada. Acredita-se que a maioria dos dispositivos médicos baseados em IA/ML serão classificados nas classes IIa ou IIb e consequentemente submetidos obrigatoriamente a avaliações por "organismos notificados" (MINSSEN *et al.*, 2020). Esse aspecto é visto como uma fragilidade importante da abordagem regulatória europeia, pois atualmente se tem verificado uma escassez da capacidade regulatória em razão da indisponibilidade da infraestrutura necessária para sua efetivação. O número insuficiente de organismos notificados designados pelas autoridades reguladoras dos países europeus nos primeiros meses de vigência do MDR pode atrasar a entrada de produtos no mercado. Assim, uma estratégia pensada para garantir segurança e eficácia pode comprometer o acesso a inovações na área da saúde (GILBERT; ANDERSON; *et al.*, 2023).

A **avaliação clínica** recebe um capítulo detalhado no MDR. O processo de avaliação clínica está alinhado às diretrizes gerais definidas pelo IMDRF, embora não sejam empregados os mesmos termos. As etapas delineadas são (MDR, Anexo XIV):

a) preparação (e atualização) de um plano de avaliação clínica;

b) identificação de dados clínicos relevantes (gerados pelo fabricante e disponíveis na literatura científica) e de lacunas nas evidências disponíveis;

c) avaliação dos dados disponíveis, incluindo a qualidade metodológica e validade científica, bem como sua relevância para a avaliação clínica do dispositivo;

d) pesquisa clínica (envolvendo seres humanos) – obrigatória para dispositivos de classe III e implantáveis, necessária para outros dispositivos se houver lacunas nas evidências;

e) análise de todos os dados clínicos relevantes para avaliar se demonstram conformidade com os requisitos gerais de segurança e desempenho.

Os fabricantes podem reivindicar a equivalência do dispositivo em avaliação com um outro dispositivo que já possua marcação CE, utilizando dados clínicos desse dispositivo já registrado (NIEMIEC, 2022).

Além disso, o Grupo de Coordenação dos Dispositivos Médicos (MDCG –*Medical Device Coordination Group*), criado pelo MDR para elaborar orientações e aconselhar as autoridades europeias, definiu **diretrizes específicas para avaliação clínica e de desempenho de *softwares* para dispositivos médicos** (MDCG, 2020). Essas diretrizes são diretamente baseadas no processo de avaliação clínica conforme estabelecido pelo IMDRF (composto pelos três elementos: associação clínica válida, validação analítica e validação clínica) e propõem métodos para realização dos estudos clínicos necessários.

O **monitoramento pós-comercialização** previsto pelo MDR detalha aspectos de vigilância e fiscalização do mercado para todos os tipos de dispositivos médicos, como a obrigatoriedade de um plano de acompanhamento clínico que deve descrever como os dados clínicos são coletados e avaliados ao longo da vida útil do produto. No entanto, o atual marco regulatório europeu ainda não contempla nenhuma norma específica para sistemas baseados em IA/ML. Embora a abordagem regulatória de SaMD possa ser aplicada a sistemas de IA, restou uma lacuna importante no paradigma europeu ao não tratar de protocolos de mudanças de algoritmos e de desempenho no mundo real de modelos adaptativos na sua legislação atualizada sobre dispositivos médicos (GILBERT *et al.*, 2021).

A previsão é que essa lacuna seja fechada em 2024. A EMA pretende consolidar normas para avaliação do ciclo de vida total de dispositivos médicos baseados em IA/ML, com base em consulta pública para avaliar especificamente essa dimensão regulatória, orientando sobretudo o monitoramento contínuo pós-comercialização desses produtos (EMA, 2023).

5.1.5 Equidade como elemento indispensável

A **equidade**, também entendida como **justiça algorítmica** (*algorithmic fairness*), deve ser incorporada como um componente essencial na dimensão de segurança e eficácia na regulação da IA na saúde. Com fundamento nos preceitos dos direitos humanos, assegurar equidade é uma imposição ética inerente ao uso dessa tecnologia (v. seção 4.2.5). Um sistema de IA só pode ser qualificado como seguro e eficaz para ser usado na saúde se ele for justo, o que significa operar de maneira imparcial e sem discriminação.

Atualmente, ainda não há normas vinculantes aplicáveis que sejam específicas para assegurar justiça algorítmica. Alguns reguladores têm recorrido às leis gerais antidiscriminação existentes até que haja legislações próprias (MATTHEWS; MURPHY, 2023). É o caso da *Federal Trade Commission* (FTC), agência reguladora de proteção ao consumidor dos EUA, que afirma o entendimento de que o uso de sistemas de IA discriminatórios em qualquer setor econômico deve ser tratado como violação à lei federal regente (JILLSON, 2021).

Na regulação da IA na saúde, a apreciação da justiça algorítmica deve ser vista como elemento do **processo de avaliação clínica** de dispositivos médicos baseados em IA/ML. Nas etapas de validação analítica (técnica) e de validação clínica, os desenvolvedores e fabricantes devem fornecer explicações científicas e clínicas adequadas sobre o desempenho dos sistemas de IA, o que necessariamente precisa incluir as populações e contextos em que se pretende usar os SaMD. Relatar informações como gênero, raça e etnia das pessoas cujos dados foram usados no treinamento dos modelos de ML pode ajudar a evitar potenciais vieses e preconceitos e a identificar situações ou populações nas quais os sistemas de IA podem não funcionar conforme esperado. Portanto, a apresentação dessas informações tem que ser considerada como requisito regulatório no monitoramento contínuo dos dispositivos de IA (WHO, 2023).

Esse aspecto é particularmente importante na regulação de sistemas de IA na área da saúde, tendo em vista que grande parte desses modelos são desenvolvidos utilizando dados de registros médicos, que são geralmente desequilibrados e tendenciosos. Os conjuntos de dados de treinamento dos modelos de ML devem ser **representativos de diferentes categorias** (por exemplo, considerando prevalências distintas de determinada doença em subgrupos populacionais). Além disso, devem **incorporar dados de determinantes sociais da saúde**: condições como ocupação, renda e local de moradia podem desequilibrar modelos de IA para apresentarem viés racial ou de gênero (OBERMEYER *et al.*, 2019; RAJKOMAR; HARDT; *et al.*, 2018).

Nesse sentido, é crucial que a regulação da IA na saúde avalie criteriosamente os cenários clínicos – e, tanto quanto possível, também geográficos, políticos, sociais e econômicos – em que os modelos tiverem sido desenvolvidos e validados. Reguladores devem considerar que há uma evidente disparidade global no desenvolvimento da IA na saúde, já que atualmente as bases de dados dos EUA e da China são desproporcionalmente representadas, pois esses países concentram a produção na área (CELI *et al.*, 2022). E mesmo que haja desenvolvimento e validação no próprio país em que se pretende implementar uma ferramenta baseada em IA/ML, é essencial que se observe que, em regra, sistemas de IA funcionam principalmente nos contextos em que foram treinados. Eles podem desempenhar completamente bem num cenário e falhar totalmente em outro (FUTOMA *et al.*, 2020). Assim como um dispositivo de IA treinado e validado num determinado contexto só pode ser considerado seguro e eficaz em outro contexto mediante a devida comprovação no processo de avaliação clínica e no monitoramento contínuo, a dimensão da equidade também deve ser constantemente acompanhada.

Mesmo que os vieses sejam identificados e evitados no treinamento e validação de modelos, os sistemas de IA ainda são propensos a alterações a partir da implementação. Pode haver mudanças na relação entre entradas e saídas em razão de alterações no ambiente ("desvio de conceito"), mudanças por haver dados de entrada diferentes dos dados de treinamento e validação ("desvio de covariável") ou instabilidades do sistema de IA. Por isso, dispositivos médicos baseados em ML precisam ter **capacidade de adaptação e aprendizado com dados do mundo real**. A regulação da IA na saúde tem que permitir melhorias nos modelos com dados novos, porque elas podem ser necessárias para obter resultados benéficos para subgrupos e diminuição de vieses – a estratégia regulatória do "algoritmo travado" é inadequada (BABIC *et al.*, 2019).

A necessidade de identificar e mitigar os vieses a que os sistemas de IA são suscetíveis (v. seção 3.3.4) é um aspecto amplamente reconhecido na ciência de *machine learning*. Em todos os setores em que a tecnologia é empregada, os resultados de modelos de ML necessariamente estão atrelados à qualidade dos dados de treinamento e às intenções de quem os desenvolve (FRA, 2022; O'NEIL, 2016). Embora inexista uma noção universal de justiça, os preceitos fundamentais da ética em IA exigem a busca constante por mecanismos para assegurar equidade nas técnicas de ML. Muitas abordagens têm sido propostas, que podem ser basicamente agrupadas em três categorias (CORBETT-DAVIES *et al.*, 2023):

a) anticlassificação;
b) paridade de classificação;
c) calibração.

A **anticlassificação** ou "justiça por desconhecimento" (*fairness through unawareness*) requer que todos os atributos protegidos – como raça, etnia, gênero, idade, orientação sexual, religião, nacionalidade, presença de alguma deficiência – sejam excluídos dos modelos, para que não sejam usados no processo de tomada de decisão. Isso também inclui a omissão de características desprotegidas que sejam *proxies* de atributos protegidos (por exemplo: profissão, endereço residencial). A justificativa é evitar a discriminação direta que pode ocorrer se forem considerados esses atributos que provém de dados sensíveis (VEALE; BINNS, 2017). A deficiência dessa abordagem é que há muitas situações em que a distribuição de variáveis de interesse (como as relacionadas com risco à saúde) diferem entre subpopulações, fenômeno conhecido como "problema da inframarginalidade" (CORBETT-DAVIES *et al.*, 2023). Portanto, essa abordagem é pouco útil nos contextos em que os modelos de ML precisam incluir atributos protegidos para melhorar a precisão, quando é necessário permitir que os sistemas de IA usem essas informações sensíveis aprendendo a evitar preconceitos que levem a resultados injustos.

A **paridade de classificação** envolve métodos para garantir que os sistemas de IA tenham desempenho preditivo consistente em diferentes subpopulações, especialmente aquelas definidas por atributos protegidos, de forma que os modelos sejam igualmente precisos para pessoas de grupos protegidos e não protegidos. Isso pode ser aferido em diferentes métricas. Pode-se exigir que a sensibilidade (taxa de verdadeiros positivos) seja igual entre os grupos, o que é chamado de "igualdade de oportunidades". Ou que tanto a sensibilidade como a especificidade (ou seja, as taxas de verdadeiros positivos e de falsos positivos) sejam iguais, conhecida como "probabilidades equalizadas". Ou ainda que o valor preditivo positivo seja idêntico, geralmente referido como "paridade preditiva". Usam-se técnicas para forçar os modelos a terem alguma dessas propriedades. No entanto, é importante ressaltar que métricas iguais não equivalem necessariamente a resultados equitativos e que essas definições também são problemáticas quando distribuições de riscos são diferentes entre grupos ("inframarginalidade"), pois podem piorar a precisão geral dos modelos (RAJKOMAR; HARDT; *et al.*, 2018).

A **calibração** diz respeito à correspondência entre as previsões feitas pelos modelos de ML e os resultados observados em diferentes grupos populacionais. Em geral, parte da comparação entre o resultado real e o resultado esperado fornecido por um sistema de IA. Por exemplo, se um modelo prevê que os pacientes têm um risco de 20% de uma determinada doença em 10 anos, então aproximadamente 20% dos pacientes em cada subgrupo (categorizados por raça, gênero ou outra variável) aos quais é atribuída esta pontuação de risco devem de fato desenvolver a doença nos próximos 10 anos. Técnicas de calibração são particularmente relevantes em

contextos clínicos nos quais pontuações de risco sejam usadas para informar decisões sobre assistência à saúde. A calibração tem sido considerada estratégia mais adequada para melhorar a equidade dos sistemas de IA na saúde, pois modelos que prevejam resultados precisos para todos os subgrupos populacionais podem apresentar maior benefício líquido (v. seção 3.3.2) quando há seleção de limiares apropriados (PFOHL *et al.*, 2022).

Uma abordagem regulatória completa deve contemplar aspectos de justiça algorítmica em todas as etapas de avaliação de dispositivos de IA para aplicação na saúde. Modelos de ML desenvolvidos em populações não representativas muito dificilmente terão utilidade clínica e aplicabilidade, exceto em contextos muito específicos. É preciso que a regulação da IA na saúde integre a exigência reportar informações demográficas dos dados de treinamento e validação dos modelos. Isso está diretamente relacionado a questões importantes sobre transparência e comunicação de informações.

5.2 Transparência

A admissão de que transparência deve ser um requisito regulatório para a IA aplicada à saúde decorre de seu reconhecimento praticamente consensual como preceito ético fundamental nesse campo (v. seção 4.2.3).

Importa evocar que o sentido abrangente de **transparência** significa que todas as informações que justifiquem as decisões tomadas a partir de resultados provenientes de sistemas de IA precisam ser disponibilizadas às pessoas envolvidas. Esse é um aspecto considerado crucial para construir a confiança da sociedade nas tecnologias baseadas em IA (EUROPEAN COMMISSION, 2019b; FLORIDI *et al.*, 2018). Portanto, transparência em sentido amplo diz respeito à explicação do ambiente institucional em que os sistemas de IA são desenvolvidos, implantados e geridos. O foco é a compreensão geral das pessoas e organizações responsáveis pelo desenvolvimento, utilização e regulação da IA. Assim, a transparência pode ser instrumentalizada por mecanismos como a adoção de documentação padronizada sobre como os sistemas e modelos de IA são criados, treinados e implementados, bem como em processos de avaliação de impacto dos sistemas de IA nos diferentes contextos de utilização (MITTELSTADT, 2022).

Mas o conceito de transparência é também empregado em referência a aspectos particulares do campo da IA, especialmente em *machine learning*. Nesse contexto, a ideia de

transparência poder estar relacionada à compreensão de como um algoritmo funciona, sem necessariamente ingressar nos dados de treinamento ou nas previsões individuais de um modelo (OECD, 2019), situação em que é usualmente denominada **transparência do algoritmo** (USACM, 2017). Ou pode significar o acesso aos elementos necessários para compreender como um modelo toma decisões, o que demanda conhecer não só o algoritmo, como também o modelo treinado e os dados usados no treinamento. Para isso, é necessário ter uma visão holística dos recursos do modelo e de cada um dos seus componentes aprendidos, como pesos e parâmetros (LIPTON, 2018; MOLNAR, 2022). Esse sentido de transparência relaciona-se com os conceitos de **interpretabilidade e explicabilidade** dos sistemas de IA.

5.2.1 IA interpretável e IA explicável

Na dimensão da transparência na IA, interpretabilidade e explicabilidade são termos estreitamente interligados e frequentemente sobrepostos, pois ainda não há definições e limites de sentido amplamente acordados nesse campo (MITTELSTADT, 2022).

Uma definição sólida é que **interpretabilidade** representa o grau em que a causa de uma decisão tomada por um modelo de IA é compreensível para um observador humano (MILLER, 2019). Portanto, um modelo é totalmente interpretável quando um ser humano pode compreender o conjunto completo de causas de um determinado resultado. Contrapõe-se a um modelo opaco ou de caixa preta, em que raramente se tem qualquer noção concreta de como ou por que um resultado específico foi obtido a partir das entradas (BURRELL, 2016). Nesse sentido, a transparência pode ser considerada no nível de todo o modelo ou no nível de seus componentes. Assim, um modelo interpretável é ou um modelo simples – no qual um humano seja capaz de juntar os dados de entrada com os parâmetros de forma a fazer os cálculos e produzir uma previsão – ou um modelo que pode ser decomposto para que cada parte (entrada, parâmetro, cálculo) admita um entendimento mais ou menos intuitivo (LIPTON, 2018).

A ideia de **explicação** está geralmente relacionada ao modo pelo qual um observador humano pode obter a compreensão (MILLER, 2019). Assim, explicação é uma interpretação *post hoc*, em que informações dos resultados algorítmicos são extraídas após as decisões ou previsões terem sido feitas, o que também pode ser feito em modelos opacos (LIPTON, 2018). Por essa perspectiva, explicabilidade pode ser vista como um atributo que independe da

interpretabilidade inerente ao modelo, porque nos modelos que não sejam em si mesmos interpretáveis a explicação deve ser buscada por outros meios. Por isso, o termo **explicabilidade** tem sido usualmente empregado tanto para se referir aos modelos inerentemente interpretáveis como para as explicações obtidas para os modelos caixa preta (AMANN *et al.*, 2022).

Em que pese a frequente sobreposição de sentidos, a partir desse referencial é possível adotar definições de IA interpretável e IA explicável especialmente úteis para aplicação na área da saúde (AMANN *et al.*, 2022; BABIC *et al.*, 2021a).

A **IA interpretável** faz referência aos sistemas de IA/ML baseados em funções que sejam inerentemente transparentes, no sentido de serem modelos passíveis de interpretação por seres humanos – não significa serem imediatamente compreensíveis por todos, mas que exista a possibilidade de serem interpretados por pessoas com conhecimento na área. Deve-se reconhecer que interpretabilidade global de modelos é difícil de ser obtida na prática, porque a partir de um certo número de características um modelo passa a ser inconcebível para humanos – como, por exemplo, imaginar um espaço com mais de três dimensões. Ou seja, uma abordagem de IA interpretável deve necessariamente utilizar modelos com uma quantidade limitada de parâmetros e pesos (para que possam ser inteligíveis) e ser baseada em algoritmos de "caixa-branca", tais como funções lineares (cujos parâmetros correspondam a pesos que relacionem entradas e saídas) ou árvores de decisão que criem mapas intuitivos a partir de regras claramente estabelecidas e compreensíveis (MOLNAR, 2022).

A **IA explicável** diz respeito a uma abordagem essencialmente diferente, pois parte de modelos de caixa preta, que são aqueles nos quais a compreensão é humanamente impossível. Dado um modelo opaco usado para fazer previsões, um segundo algoritmo é treinado para se ajustar às previsões feitas pelo primeiro. De forma geral, desenvolve-se um modelo explicativo que encontra uma função interpretável que se aproxime ao máximo das saídas obtidas pelo modelo de caixa preta. Esse modelo substituto é usado para fornecer explicações *post hoc* para os resultados do modelo original opaco, mas não é suficientemente preciso para fazer previsões reais (já que é necessário reduzir o número de características até torná-lo inteligível). A explicação pode ser dada em termos de quais atributos dos dados de entrada no modelo opaco são mais importantes para uma previsão específica ou criar um modelo linear facilmente compreensível cujos resultados se assemelham aos do modelo original. Assim, a IA explicável basicamente consiste na tarefa de encontrar um modelo de "caixa branca" para explicar as previsões feitas por um modelo de caixa preta (BABIC *et al.*, 2021a).

Em suma: a IA interpretável é também explicável, em razão de sua "explicabilidade inerente"; a IA explicável não é originalmente interpretável, ela se torna passível de interpretação a partir da "explicabilidade *post hoc*" (LIPTON, 2018).

5.2.2 Direito à explicação

O **direito à explicação** (*right to explanation*) sobre decisões automatizadas diz respeito ao reconhecimento de que deve ser assegurado a todas as pessoas o direito de saber como são tomadas as decisões baseadas em IA que afetam suas vidas. Trata-se de um conceito relativamente recente, que vem se consolidando principalmente desde as discussões na elaboração do GDPR (v. seção 4.3.5).

O **GDPR** garante direitos a todas as pessoas que sejam influenciadas por decisões automatizadas. Pelo regulamento europeu, os titulares dos dados têm direito de saber quando há decisões automatizadas que os afetem significativamente ou produzam efeitos na sua esfera jurídica. Também têm direito de receber "informações úteis relativas à lógica subjacente, bem como a importância e as consequências previstas de tal tratamento" (arts. 13.º, 14.º e 15.º). Além disso, todo titular de dados tem "direito de não ficar sujeito a nenhuma decisão tomada exclusivamente com base no tratamento automatizado, incluindo a definição de perfis", a não ser que a decisão seja necessária para celebração ou execução de um contrato ou for baseada no consentimento explícito do titular – nesses casos, o GDPR dá ao titular dos dados o direito de solicitar intervenção humana e de contestar a decisão automatizada (art. 22.º).

O sentido dessas garantias é definido no considerando 71 do GDPR, que dá a orientação interpretativa desses dispositivos (embora um considerando não tenha força normativa). O titular dos dados deve ter "o direito de obter a intervenção humana, de manifestar o seu ponto de vista, de obter uma explicação sobre a decisão tomada na sequência dessa avaliação e de contestar a decisão". Ou seja, além de receber uma explicação inteligível, cria-se o direito à oportunidade de ser ouvido, de questionar e pedir revisão da decisão automatizada. É o que vem sendo chamado de "**devido processo algorítmico**" (KAMINSKI, 2019).

Desde a publicação do GDPR, mesmo antes de sua entrada em vigor, há intensas discussões sobre a existência e o alcance do direito à explicação nas decisões automatizadas (BYGRAVE, 2020). A questão essencialmente tem origem na aceitação do fato de que, como

um algoritmo de IA pode usar inúmeras variáveis para chegar a um determinado resultado (sobretudo com o uso crescente de algoritmos que geram modelos caixa preta), a forma complexa de representação matemática é na maior parte das vezes humanamente ininteligível.

Em linhas gerais, o debate atualmente se divide em duas concepções. Por um lado, há os que defendem a viabilidade e o escopo do direito à explicação apenas no que diz respeito à **funcionalidade geral do sistema**, em vez de sobre decisões específicas e circunstâncias individuais (WACHTER; MITTELSTADT; FLORIDI, 2017). Por outro lado, há o entendimento de que a explicação também deve **incluir decisões específicas**, com transparência limitada apenas pela dimensão intrinsecamente opaca dos algoritmos, de modo a permitir que o titular dos dados exerça seus direitos de acordo com o GDPR e em conformidade com os princípios e leis de direitos humanos (SELBST; POWLES, 2017).

Um entendimento alternativo é o de que o GDPR estabelece um sistema de "**transparência qualificada**" sobre a tomada de decisões algorítmicas. Essa leitura sustenta que a lei define regras direcionadas de diferentes graus de profundidade e escopo voltadas a diferentes destinatários, de modo que se deve fornecer um tipo de informação aos indivíduos e outro tipo aos especialistas e reguladores. Assume-se que existe o direito individual à explicação, mas a transparência não se limita às revelações ao público e a comunicações aos indivíduos afetados. Deve incluir também revelações de outro nível de profundidade em situações como a avaliação regulatória ou supervisão interna de uma empresa. O regulador pode obter acesso ao código-fonte, enquanto o indivíduo deve receber uma comunicação resumida e em linguagem acessível ao público em geral (KAMINSKI, 2019).

O debate iniciado a partir do paradigma europeu vem influenciando outras jurisdições, como é o caso do Brasil (v. seção 4.3.6). No **Brasil**, a Lei Geral de Proteção de Dados Pessoais (LGPD) não trata especificamente da regulação de IA (os termos "inteligência artificial" e "algoritmo" sequer aparecem no texto da lei), mas como foi abertamente inspirada no GDPR e incorporou muito da sua racionalidade, introduz no ordenamento jurídico brasileiro o direito à explicação e à revisão de decisões automatizadas (DOURADO; AITH, 2022).

O **direito à revisão de decisões automatizadas** está definido explicitamente no art. 20 da LGPD, que concede ao titular o direito a "solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses". Diferentemente do GDPR, a lei brasileira não prevê o direito de não se sujeitar a decisão

exclusivamente automatizada e nem de obter intervenção humana em caso de revisão.[1] Contudo, é possível que a obrigação de supervisão humana venha ser adotada por regulamentação infralegal, já que essa é atualmente a forma mais adequada para instrumentalizar esse direito.

O **direito à explicação** não aparece textualmente na lei brasileira (assim como no GDPR), mas decorre da interpretação sistemática da própria LGPD em conjunto com dispositivos constitucionais e da legislação de proteção ao consumidor (MONTEIRO; MACHADO; SILVA, 2021). A lei brasileira garante a todo aquele afetado por decisões automatizadas o direito a obter informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados. Essa expressão da transparência só pode ser garantida por meio de explicação.

5.2.3 Mecanismos para explicação em IA na saúde

A **explicabilidade** em sentido abrangente – contemplando sistemas de IA inerentemente interpretáveis e explicações *post hoc* em sistemas opacos – tem sido reconhecida como aspecto essencial tanto da transparência na IA em geral (v. seção 4.1.5) como da transparência da IA aplicada à saúde (v. seção 4.2.3).

A saúde é um setor em que a busca pela explicabilidade é vista como fundamental, principalmente para o uso de sistemas de IA para realização de atividades de assistência (HERZOG, 2022; HOLZINGER *et al.*, 2017). Admite-se que IA na saúde demanda a maior transparência possível sobre o funcionamento e as condições das decisões dos algoritmos, tanto da perspectiva dos profissionais de saúde como da perspectiva dos pacientes e usuários (WATSON *et al.*, 2019). Um dos principais gargalos para o uso clínico de sistemas de IA diz respeito à necessidade de construir uma relação de confiança para que os profissionais de saúde usem sistemas automatizados na prática assistencial. Se os profissionais de saúde puderem entender como as decisões são tomadas, os resultados dos algoritmos podem se assemelhar a outras ferramentas de diagnóstico e tratamento (MORLEY *et al.*, 2020). Igualmente importante é dar aos pacientes a possibilidade de compreender a lógica de decisões automatizadas que impactem condutas

[1] Na redação original aprovada pelo Congresso Nacional (de agosto de 2018), o art. 20 da LGPD previa o direito do titular de solicitar revisão de decisões automatizadas "por pessoa natural". Mas esse dispositivo foi alterado pela Medida Provisória nº 869/2018 (de dezembro de 2018), retirando a possibilidade de obter intervenção humana. Essa alteração foi mantida na lei de conversão da MP (Lei nº 13.853/2019), que define a redação da LGPD até o momento atual.

clínicas nos seus cuidados de saúde. Essa necessidade decorre diretamente do princípio da autonomia no uso de IA na saúde e da sua articulação com o princípio da transparência, que está na essência do direito à explicação (AMANN *et al.*, 2020).

Tal preocupação deve estar cada vez mais presente em diversas situações clínicas. Por exemplo, atualmente já há modelos de IA/ML capazes de definir critérios para transplantes de órgãos, como alocação, correspondência entre doador e receptor e de previsão de sobrevida dos pacientes transplantados (KHORSANDI *et al.*, 2021). É possível que esses sistemas de IA sejam futuramente usados na prática e que haja diferenças de ordem nas filas de transplantes em comparação com as que são hoje definidas por critérios clínicos feitos apenas por pessoas. A ética da IA ancorada nos direitos humanos não autoriza que decisões dessa natureza sejam tomadas unicamente com base em sistemas de caixa preta. Por essas razões, mecanismos para fornecer explicações algorítmicas são necessários para a IA na saúde.

Explicações sobre um sistema de IA podem ser buscadas para justificar decisões, para melhorar o controle, para aprimorar modelos ou para adquirir novos conhecimentos. Em todas essas situações, o objetivo do usuário (seja profissional de saúde, paciente ou regulador) é muito relevante para a explicabilidade. Por isso, projetar sistemas para fornecer explicações é bastante complexo, principalmente para obter explicabilidade *post hoc* (ROSCHER *et al.*, 2020). De todo modo, o campo da IA explicável (usualmente designado pelo termo XAI – *eXplainable Artificial Intelligence*) está em expansão e atualmente há muitas pesquisas desenvolvidas por empresas, órgãos de normalização, organizações sem fins lucrativos e instituições públicas com intuito de criar sistemas de IA que possam explicar suas previsões (GUNNING *et al.*, 2019).

Usualmente não é viável nem necessário que uma explicação forneça todo o processo de tomada de decisão de um modelo de ML. A explicabilidade é considerada essencial para situações em que alguma falha precisa ser determinada em uma instância específica do sistema de IA, principalmente quando resultados algorítmicos são usados para fazer recomendações ou tomar decisões que estariam habitualmente sujeitas à discrição humana. Nesse sentido, uma explicação precisa ser apta a responder a pelo menos um dos seguintes pontos (OECD, 2019):

a) **principais fatores de decisão** – indicar os fatores importantes para uma previsão feita por IA, preferencialmente ordenados por significância;
b) **fatores determinantes de decisão** – esclarecer fatores que afetam decisivamente o resultado;

c) **resultados divergentes** – esclarecer por que dois casos de aparência semelhante podem apresentar resultados diferentes.

Portanto, as explicações precisam trazer **informações interpretáveis** por humanos sobre os fatores utilizados pelos modelos de IA para chegar a um resultado e o peso relativo de cada fator. E devem ser capazes de oferecer respostas a **questões contrafactuais**, para que se saiba se um fator levado em conta numa decisão algorítmica foi determinante para um resultado específico (DOSHI-VELEZ *et al.*, 2019).

A maneira mais acessível de obter explicações é utilizar apenas algoritmos que criam modelos interpretáveis, tais como regressão linear, regressão logística e árvores de decisão – ou seja, usar IA interpretável. **Regressão linear** é um algoritmo que resolve problemas de regressão, definindo uma função que faz a previsão pretendida como uma soma ponderada dos recursos de entrada. **Regressão logística** é uma extensão do algoritmo de regressão linear para resolver problemas de classificação, usando uma função que limita as saídas em resultados entre 0 e 1 (probabilidades). **Árvores de decisão** são algoritmos que dividem os dados várias vezes de acordo com determinados valores de corte nos recursos (entradas), criando diferentes subconjuntos (nós) do conjunto de dados (cada instância em um subconjunto), prevendo resultados a partir do resultado de cada nó. Árvores de decisão podem ser usadas tanto para problemas de regressão como para problemas de classificação (v. seção 3.2.4). De uma forma geral, os modelos de ML desenvolvidos por esses algoritmos – e alguns outros que também são baseados em abordagens estatísticas – podem ser diretamente interpretados por meio de técnicas e cálculos compreensíveis por humanos (MOLNAR, 2022).

Entretanto, o campo da IA na saúde vem cada vez mais sendo impulsionado pelo uso de modelos desenvolvidos por algoritmos mais complexos de *machine learning* e *deep learning* que funcionam na lógica de caixa preta (v. seção 3.1). Para esses modelos, só é possível obter explicações por meio de técnicas *post hoc* – empregar a IA explicável. Muitas técnicas vêm sendo utilizadas nesse campo, geralmente divididas em métodos globais e métodos locais: os métodos globais procuram descrever o comportamento médio dos modelos e são usados para compreender os mecanismos gerais dos dados; os métodos locais têm o objetivo de explicar previsões individuais dos modelos.

Atualmente, as técnicas mais utilizadas na IA explicável em geral e na área da saúde são métodos locais, dentre os quais merecem destaque as técnicas LIME e SHAP. Ambas consistem no uso de um modelo substituto interpretável que explica um modelo complexo (opaco),

conforme o paradigma da IA explicável. A técnica **LIME** cria o modelo substituto local gerando um novo conjunto de dados que busca aproximar as previsões correspondentes às do modelo de caixa preta (RIBEIRO; SINGH; GUESTRIN, 2016). A técnica **SHAP** faz isso calculando a contribuição de cada recurso para a previsão, com base nos chamados "valores de Shapley" – método da teoria dos jogos de coalizão que calcula a contribuição média de cada integrante em todas coalizões possíveis (LUNDBERG; LEE, 2017). Há ainda métodos específicos para explicar modelos de redes neurais, como a técnica de **mapas de saliência** (também chamada de "atribuição de *pixels*"), que destaca *pixels* relevantes para uma determinada classificação de imagem. E dezenas de outras técnicas para os diferentes tipos de algoritmos (MOLNAR, 2022).

No entanto, os limites para IA explicável são bastante relevantes.

5.2.4 Limites para explicação em IA na saúde

A ideia de que a explicação deve ser um requisito obrigatório para a transparência da IA tem sido amplamente debatida. Inicialmente vista como quase consensual, a explicabilidade algorítmica tem sido alvo de controvérsia que vem adquirindo relevância crescente a partir da década de 2020. Isso está diretamente relacionado ao reconhecimento das limitações da IA explicável e tem impactos importantes na regulação da IA na saúde (AMANN *et al.*, 2022; BABIC; COHEN, 2023).

Primeiramente, é preciso considerar que existe um ***trade-off* entre explicabilidade e desempenho** de modelos de ML. Por definição, para que um sistema de IA seja explicável é necessária uma redução das variáveis da solução a um conjunto pequeno o suficiente para que fique acessível ao entendimento humano. Isso tende a inviabilizar o uso de alguns sistemas em problemas complexos. Como já explicitado, modelos de *deep learning* e de outras técnicas de ML capazes de processar muitas variáveis podem prever probabilidades de diagnósticos clínicos com precisão, mas são humanamente incompreensíveis. Nesse sentido, a ideia de explicação mais ampla, baseada na máxima transparência, torna-se incompatível com o uso de sistemas de IA que busquem alta acurácia preditiva (LONDON, 2019).

Vale ressaltar que esse *trade-off* não necessariamente existe em qualquer situação. Em sistemas de IA desenvolvidos para resolver problemas que possuem dados estruturados com uma boa representação de características significativas (como pode ser o caso no contexto da

saúde), muitas vezes têm sido observadas diferenças mínimas de desempenho entre os algoritmos complexos, como redes neurais, e os classificadores mais simples de IA interpretável, após adequado pré-processamento (v. seção 3.2.3). Essa constatação tem conduzido a argumentação de que o uso de algoritmos interpretáveis deve ser sempre priorizado, principalmente no desenvolvimento de modelos para serem usados em situações em que a explicação seja essencial, como pode acontecer em condições sensíveis ou críticas na área da saúde (RUDIN, 2019).

Além disso, a explicação de sistemas de IA é limitada pelas reais possibilidades oferecidas pelos mecanismos existentes. As técnicas atualmente disponíveis para explicabilidade são capazes de oferecer descrições amplas de como um sistema de IA funciona em sentido geral, mas são muito superficiais ou não confiáveis para decisões individuais. Na prática, as explicações podem ser muito úteis em processos globais de IA, como desenvolvimento de modelos e auditorias, mas raramente são informativas com relação a resultados específicos dados pelos algoritmos que funcionam como caixa preta.

Outro aspecto a ser levado em conta é a constatação de que usuários tendem a confiar excessivamente em explicações dadas pelas ferramentas de IA explicável, muitas vezes sem observar que essas explicações não têm garantidas de desempenho. Isso é problemático porque, como as explicações são apenas aproximações do processo de decisão do modelo opaco, cria uma fonte adicional de erro – tanto o modelo original como o modelo explicativo podem estar errados (GHASSEMI; OAKDEN-RAYNER; BEAM, 2021). Esse ponto é particularmente preocupante na área da saúde diante da perspectiva de adoção em escala crescente de soluções de IA em ambientes clínicos.

Também começam a surgir evidências de que as atuais técnicas de IA explicável trazem efeitos negativos para a equidade nos sistemas de IA. Os métodos de explicabilidade *post hoc* podem apresentar lacunas significativas de desempenho entre subgrupos, indicando melhores explicações para alguns do que para outros (BALAGOPALAN *et al.*, 2022). Ou seja, ao lançar mão de uma explicação, o usuário provavelmente está servido de um modelo de desempenho inferior e que pode apresentar falhas potencialmente prejudiciais em perspectivas muito relevantes. Por isso, o uso da IA explicável precisa ser feito com cautela, principalmente na área da saúde (GHASSEMI; OAKDEN-RAYNER; BEAM, 2021).

Diante dessas limitações, tem ganhado consistência o entendimento de que as abordagens de explicabilidade atualmente existentes não oferecem condições para que a explicação seja um requisito regulatório para a IA na saúde. Por permitirem apenas interpretações *post hoc*

aproximadas dos modelos opacos, as ferramentas disponíveis de IA explicável não são capazes de orientar ações e planejamento, tampouco de tornar transparentes as verdadeiras razões subjacentes das decisões algorítmicas. Pelos seus próprios limites técnicos, a atual IA explicável é naturalmente "insincera" e, portanto, inadequada para servir de base para julgamentos morais e legais (BABIC; COHEN, 2023).

Nesse sentido, uma possível abordagem para a regulação da IA na saúde poderia ser exigir que desenvolvedores de modelos caixa preta relatassem o desempenho de modelos interpretáveis testados e validados para os mesmos usos pretendidos. Isso permitiria uma avaliação direta da existência ou não do *trade-off* entre explicabilidade e desempenho, provavelmente incentivando o uso de algoritmos de IA interpretável sempre que possível. Uma proposta mais forte seria não permitir o uso de sistemas opacos para usos de alto risco se existisse um sistema interpretável com o mesmo nível de desempenho (RUDIN, 2019). Estratégias semelhantes devem ser consideradas pelos reguladores.

5.2.5 Documentação como mecanismo de transparência

A **documentação precisa e abrangente** é o principal mecanismo para assegurar a transparência na regulação da IA na saúde. O registro adequado de informações claras e detalhadas sobre os métodos, recursos e decisões tomadas ao longo de todo o ciclo de vida dos sistemas de IA, além de necessário para garantir confiança nessas ferramentas, deve ser um requisito regulatório essencial. Reguladores devem ter acesso a documentação adequada que contemple desde a concepção, desenvolvimento, treinamento e validação de modelos até a implementação e o período pós-implementação (WHO, 2023). O caminho regulatório apresentado para garantir segurança e eficácia de dispositivos baseados em IA é vinculado à transparência, pois essas dimensões estão intrinsecamente associadas.

Algumas propostas podem servir de referência para definir quais informações devem ser fornecidas e a forma como essas informações precisam estar organizadas de modo a melhorar a transparência da IA na saúde. Uma das iniciativas mais relevantes atualmente é a **TRIPOD-AI**, uma extensão específica para IA da declaração TRIPOD (*Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis*), protocolo de notificação padronizado que inclui uma lista de verificação de 22 itens projetada para melhorar a

qualidade dos relatórios de estudos que desenvolvem, validam ou atualizam modelos preditivos com finalidade clínica (diagnóstica ou prognóstica). Está associada à **PROBAST-AI**, que está sendo desenvolvida como extensão da PROBAST (*Prediction model Risk Of Bias ASsessment Tool*), ferramenta que avalia risco de viés e aplicabilidade de modelos preditivos a partir de 20 questões (em quatro domínios: participantes, preditores, resultado e análise). Essas ferramentas têm sido concebidas para orientar pesquisadores e revisores a avaliarem a qualidade de estudos e interpretar achados cientificamente relevantes. E podem ser também muito úteis para auxiliar reguladores (COLLINS *et al.*, 2021, 2015; WOLFF *et al.*, 2019).

Há outras iniciativas semelhantes e é bastante provável que novos protocolos e listas de verificação padronizadas para avaliação de sistemas de IA na saúde sejam desenvolvidas nos próximos anos. Um bom exemplo é a **MINIMAR** (*MINimum Information for Medical AI Reporting*), que propõe padrões para relatórios com informações mínimas necessárias para compreender o uso pretendido, as populações-alvo, os possíveis vieses ocultos e a capacidade de generalização dos modelos de ML. A proposta é exigir que os sistemas de IA incluam informações sobre as populações usadas nos dados de treinamento (fontes de dados e seleção de coorte) e sobre a demografia desses dados, de forma a permitir comparar com a população na qual os modelos serão implementados. Além disso, propõe como requisitos que sejam fornecidas informações detalhadas sobre arquitetura e desenvolvimento dos modelos, propiciando a interpretação da intenção de uso em comparação com sistemas de IA semelhantes. O objetivo é dar aos usuários – ou aos reguladores – um melhor entendimento sobre o funcionamento dos sistemas de IA considerando a transparência como abordagem para identificar melhores práticas de ML tendo em vista a justiça algorítmica (HERNANDEZ-BOUSSARD *et al.*, 2020).

5.3 Responsabilidade

A responsabilização efetiva é uma determinação ética essencial para a IA aplicada à saúde (v. seção 4.2.4). A responsabilidade é a dimensão que representa a **regulação indireta da IA na saúde** (v. seção 4.1.2), pois os mecanismos legais influenciam as ações e cuidados realizados por pessoas físicas e pessoas jurídicas, de modo que a incidência sobre os resultados de saúde acontece indiretamente (GOSTIN, 2008). Portanto, é um componente regulatório a ser exercido no âmbito dos sistemas de responsabilidade civil.

5.3.1 Sistemas de responsabilidade civil: elementos fundamentais

Um **sistema de reponsabilidade civil** expressa um conjunto de respostas culturais a desafios amplos de abordagem de riscos e atribuição de responsabilidades, que compreende mecanismos de compensação por danos e obrigações relacionadas a lesões, que podem ser compartilhadas ou direcionadas a pessoas ou organizações (ENGEL; MCCANN, 2009). É definido por regras, instituições e procedimentos geralmente entendidos como um ramo específico do direito privado, área que regula a relação jurídica entre particulares.

O termo **responsabilidade** é utilizado no direito para se referir a situações em que alguma pessoa natural ou pessoa jurídica deve arcar com consequências de um ato, fato ou negócio que tenha causado danos a outra pessoa. Baseia-se no princípio jurídico de que toda atividade que acarreta prejuízo gera para quem a exerceu o dever de indenizar aquele que foi prejudicado. O fundamento do instituto da responsabilidade civil é o de que a obrigação de indenizar busca restaurar um equilíbrio patrimonial e/ou moral que tenha sido violado.

Há significativas diferenças entre os sistemas de responsabilidade civil entre países e jurisdições, particularmente se comparados os ordenamentos jurídicos filiados ao direito anglo-saxão (*common law*) – nos quais a responsabilidade civil é denominada *tort law* – e ao direito romano-germânico (*civil law*). Mas existem elementos comuns nesses sistemas: são compostos por regimes de responsabilidade subjetiva e regimes de responsabilidade objetiva. Em linhas gerais, a **responsabilidade subjetiva** (ou responsabilidade por culpa) é baseada na premissa de que o comportamento de uma pessoa deve estar em conformidade com o padrão de cuidado esperado de um "cidadão médio". A responsabilização deve incidir quando esse padrão não é observado, o que é comumente referido como negligência (também as possíveis variações: imprudência e imperícia). A **responsabilidade objetiva** (ou responsabilidade sem culpa) é baseada na lógica de risco de uma atividade. A responsabilização ocorre independentemente de culpa (ou negligência), bastando que se comprove nexo causal entre a atividade realizada e o dano sofrido por outrem. Essas dimensões têm sido verificadas nos diferentes sistemas de responsabilidade civil globalmente (BUSSANI; SEBOK; INFANTINO, 2022).

A partir dessas dimensões essenciais da responsabilidade civil, é possível traçar uma abordagem para delimitação de um paradigma geral da responsabilidade no âmbito da regulação da IA na saúde.

5.3.2 Desafios para atribuição de responsabilidade em IA na saúde

A tendência geral em relação aos sistemas de IA na saúde é de que sejam admitidos como dispositivos médicos, equiparando-os a produtos de interesse à saúde, o que tem implicações diretas na aplicação da reponsabilidade civil (MINSSEN; MIMLER; MAK, 2020).

Os diferentes sistemas jurídicos tradicionalmente responsabilizam moral e juridicamente os fabricantes por defeitos de produtos ou consequências de sua operação. Essa concepção, denominada **responsabilidade dos produtos**, é considerada a essência da noção de responsabilidade objetiva. Foi construída com base nos argumentos de que a imposição de custos aos produtores serviria de incentivo para melhorar a segurança de produtos e ao mesmo tempo distribuir os custos de forma justa entre produtores e consumidores. A ideia nuclear é que os fabricantes são responsáveis por produtos defeituosos ou perigosos para os usuários, devendo reparar os danos por eles causados (BUSSANI; SEBOK; INFANTINO, 2022).

Os sistemas de IA baseados em IA/ML representam uma situação nova nesse cenário. Os fabricantes e operadores, em princípio, podem não ser capazes de prever o comportamento futuro dos modelos de ML e das decisões automatizadas, o que retira substrato para que sejam responsabilizados moralmente pelo seu funcionamento. Trata-se de uma **lacuna de responsabilidade**, que tem sido pensada desde que as atuais técnicas de IA começaram a ser desenvolvidas (MATTHIAS, 2004). Um dos caminhos possíveis é uma abordagem de responsabilidade moral distribuída, que desvincule a responsabilização da intencionalidade das ações, considerando que danos podem ser causados por uma rede de agentes, alguns humanos, outros puramente baseados em IA (FLORIDI, 2016).

A responsabilidade dos produtos para a IA na área da saúde enfrenta desafios adicionais. Há uma compreensão em algumas jurisdições de que na assistência à saúde aplica-se a "doutrina do intermediário instruído" (*learned intermediary doctrine*), segundo a qual o fabricante pode se eximir da responsabilização ao fornecer todas as informações necessárias para o uso correto de um produto a um intermediário (no caso, um profissional de saúde), que deve transmiti-las ao usuário final (o paciente). Essa doutrina é originária dos EUA e vem sendo aplicada em incidentes de responsabilidade em medicamentos e alguns dispositivos médicos que necessitam de prescrição para serem utilizados, impedindo que os demandantes processem diretamente os fabricantes para exigir indenizações de reparação de danos, uma vez que o médico (e não o paciente) é considerado consumidor. O raciocínio é que a decisão final sobre os cuidados

de saúde cabe ao prestador ou profissional de saúde. O problema é que quanto mais autonomia os sistemas de IA adquirirem e maior sua opacidade, mais difícil fica atribuir responsabilidade legal pelo seu comportamento aos seres humanos (SULLIVAN; SCHWEIKART, 2019). Além do mais, as estruturas regulatórias precisam se adaptar à crescente presença no mercado de consumo de sistemas de IA incorporados em aplicativos de uso direto pelo paciente, sem intermediação de nenhum profissional (BABIC *et al.*, 2021b).

O contexto da IA na saúde tem ainda outras particularidades. Antes de considerar a responsabilidade dos produtos, os conceitos de responsabilidade civil empregados na assistência à saúde são delimitados em relação aos agentes do cuidado. Nesse sentido, são geralmente vistos em duas perspectivas: dos profissionais de saúde e dos prestadores de assistência à saúde (como hospitais e demais equipamentos dos sistemas de saúde).

Na perspectiva dos **profissionais de saúde**, que tem como paradigma a responsabilidade médica, é comumente aceito em diversas jurisdições que existe uma obrigação de meio no exercício profissional. Entende-se que médicos e demais profissionais de saúde em geral não podem assegurar os resultados (como a cura de uma doença). Eles devem aplicar nas suas atividades o que é chamado de **padrão de cuidado**, que é a melhor técnica disponível e toda diligência possível, empregando os meios propícios e o compromisso ético para atingir o fim visado, mesmo que ele não venha a ser alcançado – com algumas exceções, como procedimentos estéticos e exames clínicos e radiológicos, em que se considera haver obrigação de resultado. Dessa forma, a responsabilidade civil é, em regra, **subjetiva**: há responsabilização se houver dano provocado por negligência (ou imprudência, ou imperícia).

Com o ingresso de **sistemas de IA em ambientes clínicos**, é preciso encontrar um equilíbrio na definição das responsabilidades morais e jurídicas no novo cenário em que a atividade se baseia na decisão conjunta entre humanos e sistemas automatizados. Um complicador nesse contexto decorre do uso crescente de modelos da ML desenvolvidos em algoritmos cada vez mais complexos e que, portanto, funcionam na lógica de caixa preta (PRICE, 2018). Mas essa definição de responsabilidades é considerada crucial para que haja adoção consistente de ferramentas de IA pela comunidade clínica (VAYENA; BLASIMME; COHEN, 2018). Médicos podem ser responsabilizados se seguirem recomendações de sistemas de IA que resultem em erros ou se ignorarem recomendações que teriam evitado desfechos negativos aos pacientes? É possível que decisões de sistemas IA sejam reconhecidas como fundamentos para diretrizes e protocolos clínicos, de modo que seu não cumprimento seja admitido como negligência? (SCHÖNBERGER, 2019). Essas respostas dependem de escolhas, como o tipo de

comportamento que se pretende encorajar ou desencorajar por um sistema legal, e do padrão de cuidados a serem estabelecidos à medida que o uso da IA na prática clínica se consolide (WHO, 2021).

Na perspectiva dos **prestadores de serviços de saúde**, a compreensão corrente é de que há responsabilidade civil **objetiva**. Entende-se que empresas de saúde ou de assistência médica, como hospitais e clínicas, ocupam a posição de fornecedores e, portanto, respondem objetivamente (ou seja, independentemente de culpa) pelos danos decorrentes dos serviços prestados. Por exemplo, hospitais podem ser responsabilizados indiretamente por erros cometidos pelos profissionais em atividade nas suas dependências. Também podem ser responsabilizados por não exercerem o devido cuidado na contratação, formação ou supervisão de funcionários ou por não manterem instalações e equipamentos adequados ("seleção ou supervisão negligente").

Assim, no contexto da incorporação de ferramentas de IA na assistência à saúde, a responsabilização pode servir como incentivo para que os prestadores de saúde adotem bastante cuidado na seleção de tecnologias e que procurem assegurar que médicos e demais profissionais de saúde recebam treinamentos e orientações adequados para evitar erros que possam advir da interação humano-máquina. No entanto, é razoável supor que dificilmente os prestadores de serviços de saúde terão as condições técnicas necessárias para avaliar por conta própria os sistemas de IA. Em regra, eles depositarão confiança nos desenvolvedores e, principalmente, nos reguladores (PRICE; GERKE; COHEN, 2022).

Isso remete o desafio de volta para a questão responsabilidade dos produtos e as já mencionadas dificuldades para atribuição de responsabilidade aos desenvolvedores ou fabricantes dos dispositivos médicos baseados em sistemas de IA (SULLIVAN; SCHWEIKART, 2019). Além disso, abre uma outra perspectiva, que é o **papel dos reguladores** na distribuição de responsabilidades. Ou seja, assumindo a obrigação da administração pública de garantir a segurança e a eficácia dos dispositivos médicos baseados em IA/ML – exercida por meio de agências reguladoras ou órgãos equivalentes –, em que medida deve-se considerar a responsabilização do Estado na distribuição de responsabilidades. Esse é um ponto ainda pouco debatido, que começa a ser pensado nos EUA, onde se cogita que fabricantes de dispositivos que tenham passado pelo procedimento completo de aprovação pré-comercialização (v. seção 5.1.3) estariam provavelmente protegidos contra ações judiciais estaduais que desafiem segurança ou eficácia dos SaMD. Por outro lado, fabricantes que tenham usado o caminho simplificado de notificação pré-comercialização poderiam ser demandados em ações de responsabilidade civil (PRICE; GERKE; COHEN, 2022).

Até meados da década de 2020, é bastante provável que esses desafios sobre atribuição de responsabilidade da IA em geral e especificamente da IA na saúde permaneçam em aberto. Os sistemas de responsabilidade civil constituem um campo no qual o papel das decisões judiciais é definidor, de forma que muitas incertezas persistirão até que casos concretos comecem a ser decididos pelos tribunais. E certamente haverá soluções muito diversas nos vários contextos culturais e jurisdições. Até que isso aconteça, pode-se analisar algumas propostas para abordar a dimensão da responsabilidade em IA na saúde.

5.3.3 Possíveis mecanismos para responsabilização

Algumas propostas vêm sendo apresentadas para buscar soluções jurídicas diante dos desafios na atribuição de responsabilidade em IA e que terão implicações na regulação da IA na saúde (SULLIVAN; SCHWEIKART, 2019).

Uma possibilidade seria empregar a **responsabilidade empresarial comum**. A atribuição de responsabilidades por resultados danosos ou defeitos nos sistemas de IA não deveria ser feita com foco em pessoas ou entidades específicas, e sim distribuída entre todos os grupos envolvidos no desenvolvimento e implementação desses sistemas de IA (VLADECK, 2014). Em vez de tentar apurar a origem e a natureza das falhas – tarefa que pode ser totalmente inviável em modelos opacos – a saída seria usar a lógica da responsabilidade civil solidária, que dá a quem for lesado a possibilidade de demandar a indenização por reparação de danos de todos ou de quem considerar mais apto a quitá-la, uma vez que todos podem ser considerados responsáveis. Mas essa lógica poderia inibir a entrada de potenciais atores na cadeia de produção.

Há ainda a proposta de conferir **personalidade "quase-jurídica"** aos sistemas de IA, dando-lhes alguns direitos e deveres (ALLAIN, 2013). A personalidade jurídica não poderia ser integral, pois os sistemas de IA não teriam capacidade para serem processados diretamente como sujeitos independentes. As empresas fabricantes teriam que responder às ações judiciais em nome dessas "pessoas quase-jurídicas", pagar as eventuais indenizações delas decorrentes ou contratar seguros para isso. Portanto, essa proposta criaria uma camada adicional de ficção jurídica, mas não resolveria propriamente a questão de atribuição de responsabilidade.

Uma alternativa seria replicar a lógica já existente em outros contextos e deslocar a responsabilidade para os **usuários dos sistemas de IA na assistência à saúde**: ou seja, para

médicos e profissionais de saúde ou para prestadores de saúde. Nessa perspectiva, a definição sobre a atribuição de responsabilidades decorreria da incorporação dos sistemas de IA nos padrões de cuidado – como pode acontecer se ferramentas de ML forem incluídas em protocolos clínicos e diretrizes terapêuticas. Mas isso exige que tanto os prestadores como os profissionais de saúde sejam capazes de avaliar os dispositivos baseados em IA/ML quanto à segurança e eficácia (incluindo possíveis vieses), observando critérios como os dos processos de avaliação clínica, principalmente em sistemas de caixa preta (PRICE, 2018). Portanto, uma solução de aplicação bastante complexa.

Na adoção de dispositivos médicos baseados em IA/ML como **sistemas de suporte à decisão clínica** (CDSS), as diferentes possibilidades de interação humano-máquina trarão as definições de atribuição de responsabilidade para a análise do padrão de cuidado. As decisões clínicas cabem aos profissionais de saúde, cuja responsabilização é sempre subjetiva: só há reconhecimento jurídico de responsabilidade civil nos casos em que existe resultado danoso e o profissional não segue o padrão de cuidado, configurando negligência. Em cada interação entre um profissional de saúde e um CDSS baseado em IA/ML, a recomendação do sistema de IA pode ser de acordo o padrão de cuidado ou diferente dele (considerando que tal sistema de IA ainda não esteja incorporado ao padrão de cuidado); o modelo de ML pode ter um desempenho ótimo ou ruim; o profissional pode seguir ou rejeitar a recomendação; e o resultado pode ser benéfico para o paciente ou causar algum dano (Quadro 5.2).

Quadro 5.2 – Possíveis resultados jurídicos relacionados ao uso de IA na prática clínica

Cenário	Saída da IA	Desempenho da IA	Ação do profissional	Resultado real	Resultado jurídico
1	Conforme o padrão de cuidado	Ótimo	Segue	Benéfico	Sem lesão e sem responsabilização
2			Rejeita	**Danoso**	**Lesão e responsabilização**
3		Ruim	Segue	**Danoso**	**Lesão, mas sem responsabilização**
4			Rejeita	Benéfico	Sem lesão e sem responsabilização
5	Fora do padrão de cuidado	Ótimo	Segue	Benéfico	Sem lesão e sem responsabilização
6			Rejeita	**Danoso**	**Lesão, mas sem responsabilização**
7		Ruim	Segue	**Danoso**	**Lesão e responsabilização**
8			Rejeita	Benéfico	Sem lesão e sem responsabilização

Fonte: Adaptado de Price, Gerke e Cohen (2019, p. 1766).

A partir dessas possibilidades de interação entre o profissional de saúde e o sistema de IA/ML, há oito possíveis cenários, em dois dos quais pode haver responsabilidade jurídica. A responsabilização, em tese, poderia existir nos quatro cenários em que há dano, que é requisito necessário para existir a admissão jurídica da lesão que dá origem à responsabilidade (se há resultado benéfico, não há que se falar em responsabilização). Mas nos cenários em que o profissional de saúde tiver observado o padrão de cuidado, não caberia responsabilização pelo dano provocado, porque não há culpa em sentido jurídico. Assim, haveria a possibilidade de responsabilização do profissional de saúde em dois cenários de resultado danoso: 1) se o profissional rejeitar uma recomendação correta do sistema de IA conforme o padrão de cuidado; 2) se o profissional seguir uma recomendação do sistema de IA fora do padrão de cuidado. Ambas as situações já ensejariam responsabilização, mesmo que não houvesse a recomendação da IA, em razão da não observância do padrão de cuidado (PRICE; GERKE; COHEN, 2019).

Portanto, essa solução jurídica desencorajaria fortemente a adoção de ferramentas de IA na prática clínica enquanto elas não forem formalmente incorporadas em protocolos clínicos, de modo a integrar o que é reconhecido como padrão de cuidado. É muito improvável que profissionais de saúde assumam o risco de tomar condutas clínicas fora do padrão de cuidado estabelecido, tanto do ponto de vista de ética profissional como da possibilidade de serem responsabilizados judicialmente por erros ou quaisquer resultados danosos. Embora isso possa resguardar os profissionais, seria desfavorável para a sociedade, por inibir o uso da IA na saúde em todo o seu potencial. Por isso, a efetiva introdução de sistemas de suporte à decisão clínica baseados em IA/ML depende primariamente da ação de entidades profissionais capazes de avaliar e incorporar essas ferramentas em protocolos clínicos e diretrizes terapêuticas que sejam amplamente reconhecidas e, mais uma vez, na atuação de reguladores na validação e no monitoramento contínuo desses dispositivos (WHO, 2021).

Por essas razões, existe a ponderação de que os regimes de responsabilidade civil podem não ser adequados para regular o uso da IA na saúde. A solução jurídica seria a adoção do chamado sistema de **compensação sem culpa** (*no-fault compensation*). Trata-se de uma abordagem para lidar com riscos e compensações por danos distinta dos sistemas de responsabilidade civil, que é aplicada em algumas jurisdições para erros médicos em geral (a Nova Zelândia foi pioneira, atualmente Dinamarca e Suécia têm sistemas semelhantes) e em situações específicas (como as indenizações por eventos adversos de vacinas nos EUA). Constituem-se fundos de compensação que podem ser financiados pelo Estado e complementados por seguros obrigatórios assumidos pelas empresas do setor. Num sistema de compensação sem culpa, o direito

à indenização depende da resposta positiva a duas questões: 1) se o resultado dos cuidados de saúde foi inesperado; e 2) se esse resultado foi prejudicial. Ou seja, não se confunde com a responsabilidade civil sem culpa (responsabilidade objetiva), porque não existe a fase de responsabilização (comprovação do nexo causal) e, principalmente, não há litígio. Em vez de ingressar com ação judicial, o sujeito que sofreu algum dano acessa uma instância administrativa e faz jus à indenização se for constatado ter ocorrido um resultado inesperado e prejudicial.

Logo, um sistema de compensação sem culpa poderia solucionar juridicamente a questão da responsabilização na IA aplicada à saúde sem precisar considerar os elementos de um sistema de responsabilidade civil, tais como: 1) exatamente por que a IA funcionou mal numa situação em que tenha provocado dano; 2) por que o profissional de saúde seguiu ou rejeitou a recomendação da saída do sistema de IA; 3) se outros profissionais em situação semelhante teriam seguido ou rejeitado tal recomendação; 4) se outro sistema de IA teria desempenho melhor. Ou seja, seria mais simples aferir o direito a indenização sem necessidade de comprovar negligência ou de encontrar um nexo causal específico. Também eliminaria potenciais injustiças que podem decorrer da impossibilidade de ultrapassar as camadas de opacidade (seja de natureza técnica ou segredo de negócio) de um modelo de caixa preta. Além disso, alguém que sofresse dano não precisaria decidir quem acionar num processo judicial (profissional de saúde, prestador, fabricante ou regulador), bastaria reclamar diretamente ao fundo específico em procedimento administrativo (HOLM; STANTON; BARTLETT, 2021).

A criação de sistemas de compensação sem culpa demandaria modificações relevantes nos ordenamentos jurídicos na maioria das jurisdições. Além disso, há incerteza sobre o custo e a viabilidade para constituição de fundos de compensação. Por isso, embora se apresente como alternativa jurídica promissora é pouco provável que seja adotada num primeiro momento. A tendência atual é a promoção de ajustes nos sistemas de responsabilidade civil para adequar às peculiaridades dos produtos baseados em IA.

5.3.4 Propostas para adaptação da responsabilidade civil à IA

Atualmente, as duas principais propostas que visam adequar sistemas de responsabilidade civil para comportar produtos baseados em IA/ML estão em tramitação no processo legislativo da UE, após terem sido apresentadas pela Comissão Europeia em setembro de 2022. Uma

trata diretamente da adaptação de regras de responsabilidade civil (EUROPEAN COMMISSION, 2022c), a outra trata de alterações na diretiva relativa à responsabilidade decorrente de produtos defeituosos (EUROPEAN COMMISSION, 2022b). Ambas fazem parte de um pacote de medidas destinadas a apoiar a implantação da IA na Europa juntamente com a proposta legislativa que estabelece regras horizontais para sistemas de IA: o "Regulamento Inteligência Artificial" (mais comumente referido como *AI Act*), abordado adiante (v. seção 6.1).

A proposta relativa à **adaptação de regras de responsabilidade civil extracontratual à IA** trata de regras de responsabilidade subjetiva – ou seja, quando o demandante precisa provar a culpa do responsável pelo dano. Denominada **Diretiva Responsabilidade da IA** (AILD – *AI Liability Directive*), foi feita com base em uma avaliação de impacto, que levou à escolha de uma abordagem faseada. Numa primeira fase, propõe-se a adoção de três medidas para reduzir o ônus dos demandantes para produzir provas em ações de pedido de indenização por responsabilidade civil:

a) harmonizar o modo de documentação e divulgação de informações pelos fabricantes, para facilitar a identificação da ação ou omissão que tenha causado algum dano;
b) se a vítima demonstrar que o responsável descumpriu regras de segurança (definidas pelo *AI Act*), os tribunais podem presumir que o descumprimento causou o dano, mas admitindo prova em contrário (a chamada presunção de causalidade refutável);
c) se a única forma de provar o direito à indenização for demonstrar algo que aconteceu no interior da IA, o ônus da vítima é reduzido, mas admitindo prova em contrário (o responsável tem oportunidade de provar que não foi negligente).

Numa segunda fase, prevê uma reavaliação (após cinco anos) sobre a necessidade de harmonizar regras de responsabilidade civil objetiva para casos de uso de IA com um perfil de risco específico e a possibilidade de exigência de seguro obrigatório. O objetivo expresso da proposta é evitar as lacunas de responsabilidade em IA e preservar os direitos fundamentais das pessoas afetadas pelo uso das tecnologias baseadas em IA (EUROPEAN COMMISSION, 2022a).

A proposta relativa à **responsabilidade decorrente de produtos defeituosos** visa revisar o regime de responsabilidade dos produtos na UE, que está atualmente estabelecido pela Diretiva Responsabilidade dos Produtos (de 1985). Portanto, trata das regras para pedidos de indenização por responsabilidade objetiva, em que não há necessidade de comprovar culpa dos fabricantes. Essa revisão pretende alargar o âmbito das reclamações que podem ser apresentadas e o leque de danos que podem ser indenizados, facilitando as reparações aos consumidores.

Para isso, oferece uma interpretação mais ampla do conceito de "produto" para incluir *softwares*, sistemas de IA e serviços digitais. Amplia o conceito de "defeito" para abarcar riscos de segurança cibernética, de conectividade e falhas de atualizações de *softwares*. Também expande o escopo dos danos para incluir dados perdidos ou corrompidos e danos psicológicos. Além disso, facilita o ônus da prova para os requerentes caso o produto não cumpra requisitos de segurança de produtos da UE ou haja complexidade técnica ou científica que dificulte a comprovação de responsabilidade, como sistemas de IA de caixa preta (DE LUCA, 2023).

Em linhas gerais, essas duas propostas tendem a alterar significativamente os riscos de responsabilização para todos os fabricantes e fornecedores de produtos que incorporam sistemas de IA/ML no mercado europeu (como foram propostas como diretivas, precisam ser transpostas para os ordenamentos jurídicos dos países membros depois de aprovadas). Haverá maior exposição à litigância para as empresas que desenvolvem e implementam IA e maior proteção aos consumidores contra danos causados por esses produtos.

Entretanto, especificamente na IA aplicada na assistência à saúde, existe a preocupação de que persistam lacunas de responsabilidade nos casos em que haja danos causados pelo funcionamento interno de modelos de caixa preta quando os fabricantes tenham cumprido todos os requisitos exigidos – quer dizer, se não houver nenhuma violação de um dever de cuidado. A origem dessa preocupação é a possibilidade de haver erros decorrentes exclusivamente de decisões autônomas de sistemas de IA opacos, que não possam ser identificados nem pelos desenvolvedores nem pelo usuários (DUFFOURC; GERKE, 2023).

A definição de mecanismos para responsabilização no uso da IA ainda está numa fase incipiente. Uma delimitação dessa dimensão, tanto no âmbito geral como no específico da IA na saúde, depende da forma como serão formadas jurisprudências e, principalmente, de opções políticas que serão feitas nos próximos anos.

6 TENDÊNCIAS EMERGENTES

A regulação da IA é um dos temas mais efervescentes e debatidos na sociedade contemporânea. Mudanças rápidas e significativas são esperadas nos próximos anos e que terão impacto direto no campo da ética e regulação da IA específica para a saúde. Algumas tendências começam a se apresentar como promissoras nesse cenário e merecem um acompanhamento atento no futuro próximo.

6.1 Leis Horizontais ou Abordagem Setorial

É provável que o paralelo entre as abordagens europeia e norte-americana analisado acima, na dimensão de proteção de dados (v. seções 4.3.4 e 4.3.5), seja replicado no âmbito da regulação da IA de forma ampla. A UE discute a criação de uma lei horizontal, o AI Act (Regulamento Inteligência Artificial). Os EUA, sem perspectiva atual de ter legislação abrangente, caminham para reproduzir a estratégia da regulação setorial, conduzida pela atuação das agências federais e órgãos do Poder Executivo (THE WHITE HOUSE, 2023). Mas essa distinção não deve se manter na regulação da IA na saúde, porque será preciso definir uma regulamentação específica para o setor, em qualquer desses cenários (VOKINGER; GASSER, 2021).

A estratégia europeia começou a ser desenhada em fevereiro de 2020, com a publicação do "Livro Branco sobre a Inteligência Artificial" pela Comissão Europeia, documento que apresentou opções políticas para buscar o desenvolvimento seguro e confiável da IA. Ficou clara a intenção da UE de tomar à frente no caminho para o estabelecimento de um marco regulatório para a IA e "exportar os seus valores para todo o mundo" (EUROPEAN COMMISSION, 2020). Embora reconhecendo que a saúde é uma das áreas de aplicação mais importantes, esse documento não chegou a avançar em diretrizes regulatórias para o setor, limitando-se à enunciação de intenções políticas e princípios gerais (COHEN *et al.*, 2020).

A proposta de um marco regulatório para IA na Europa foi apresentada em abril de 2021 no projeto de lei do ***AI Act*** (EUROPEAN COMMISSION, 2021). Essa iniciativa legislativa é reconhecida como a primeira do mundo com intenção de estabelecer uma regulamentação abrangente para a IA e vem inspirando debates legislativos em países como Brasil e Canadá. O quadro jurídico proposto é focado nos usos específicos de sistemas de IA e dos riscos associados

a eles, com o objetivo de definir uma regulação geral comum aplicável a todos os setores (exceto o militar). O projeto busca uma definição "tecnologicamente neutra" dos sistemas de IA e propõe regras de conformidade de acordo com o risco que representam para os usuários:

a) **risco inaceitável** – sistemas considerados ameaças graves para pessoas serão banidos (como identificação biométrica em tempo real, pontuação social e manipulação cognitivo-comportamental de pessoas ou grupos vulneráveis específicos);
b) **alto risco** – sistemas que representem ameaças à saúde, à segurança ou aos direitos fundamentais exigem uma avaliação de conformidade obrigatória;
c) **risco limitado** – os sistemas que interagem com humanos (como *chatbots*) e que criam conteúdo (IA generativa) devem cumprir requisitos mínimos de transparência que permitam aos usuários tomar decisões informadas (incluindo a exigência de que usuários sejam avisados quando estiverem interagindo com IA).
d) **risco mínimo** – todos os outros sistemas não terão obrigações legais, mas serão incentivados a aderirem voluntariamente a códigos de conduta.

Os sistemas considerados de alto risco serão divididos em duas categorias. Uma categoria corresponde aos produtos já abrangidos por leis existentes em matéria de segurança dos produtos, como é o caso de dispositivos médicos (também máquinas, brinquedos, elevadores etc.). Outra categoria contempla domínios específicos (tais como identificação biométrica, educação e formação profissional, gestão de trabalhadores, assistência jurídica) em que os sistemas de IA terão que ser registrados em uma base de dados específica (MADIEGA, 2023). A proposta prevê a criação um sistema de governança a partir da criação do Comitê Europeu para a Inteligência Artificial (*European Artificial Intelligence Board*) e de autoridades nacionais para controle, supervisão e execução do regulamento.

Até o final de 2023, o projeto de lei ainda está em negociação no processo legislativo da UE. Mas independentemente de possíveis alterações na versão da lei a ser aprovada, a lei geral deve reforçar as regras de regulação de dispositivos médicos já discutidas (v. seção 5.1.4), pois a expectativa é que o escopo do *AI Act* não alcance normas mais específicas para o setor da saúde. Portanto, mesmo na estratégia regulatória baseada numa lei geral, haverá necessidade da construção de um marco regulatório próprio para a IA na saúde. Conforme analisado acima, o paradigma europeu precisará atualizar a regulação de dispositivos médicos para contemplar as particularidades dos sistemas baseados em IA/ML, como vem sinalizando a Agência Europeia de Medicamentos (EMA, 2023). Isso é o que deve acontecer nos países que adotarem a abordagem regulatória para IA baseada em legislações horizontais.

6.2 Ambientes de Testagem Regulatória (*Regulatory Sandboxes*)

Os ambientes de testagem regulatória (mais conhecidos pelo termo inglês ***sandboxes***) representam uma abordagem inovadora concebida originalmente no Reino Unido para setor de tecnologias financeiras (FCA, 2015). Um *sandbox* é um espaço legal criado para que empresas possam operar sob regras mais flexíveis (ou com menos regulamentação), supervisionados por uma autoridade reguladora, por um período limitado. A essência é permitir a experimentação de ideias e produtos inovadores num ambiente controlado, sem todas as exigências regulatórias, para potencialmente acelerar a entrada no mercado A estratégia de *sandboxes* regulatórios tem sido vista como muito promissora para buscar o equilíbrio entre o estímulo à inovação e a garantia de segurança em muitos setores, com destaque para a regulação da IA, tendo sido incorporada ao projeto europeu do *AI Act* (BUOCZ; PFOTENHAUER; EISENBERGER, 2023). Além da UE, há dezenas de países, como Japão, Noruega e Reino Unido, usando *sandboxes* regulatórios no setor financeiro (WORLD BANK, 2020).

Na regulação da IA, a proposta europeia prevê a criação de *sandboxes* com o papel duplo de promover a aprendizagem empresarial e de apoiar a aprendizagem regulatória, por propiciar a formulação de regimes jurídicos experimentais numa estrutura de riscos controlados (MADIEGA; VAN DE POL, 2022). O projeto do *AI Act* abre espaço para autorizações excepcionais no uso de dados pessoais para o desenvolvimento de sistemas de IA em ambientes de testagem nos domínios da saúde pública, proteção ambiental, segurança pública e na área de investigação ou repressão criminal. As regras e condições específicas para operação dos *sandboxes* devem ser estabelecidas pelas autoridades competentes dos Estados-Membros ou pela Autoridade Europeia para a Proteção de Dados (EUROPEAN COMMISSION, 2021).

A abordagem *sandbox* vem ganhando força como alternativa também no setor da saúde. A experiência das autorizações emergenciais de tratamentos e vacinas durante a pandemia de Covid-19 é vista como uma importante demonstração de como o uso de uma estratégia regulatória adaptável e dinâmica pode ser uma ferramenta valiosa para facilitar a incorporação tecnológica na saúde pública de forma ágil e segura (SHERKOW, 2022). A utilização cada vez mais frequente de *sandboxes* regulatórios é esperada para os próximos anos na avaliação de tecnologias em saúde (ATS), pois técnicas e processos inovadores e disruptivos vêm sendo introduzidos e se mostrando necessários para aprimorar os cuidados de saúde, especialmente no âmbito da saúde digital (LECKENBY *et al.*, 2021).

Nesse cenário, a aplicação de *sandboxes* na regulação da IA na saúde tem grande potencial. Ambientes de testagem podem permitir uma melhor compreensão dos sistemas de IA durante o desenvolvimento dos modelos, facilitar a formulação de políticas adequadas para reduzir riscos e encontrar mecanismos para assegurar segurança e eficácia antes de implementação definitiva para uso da população em geral (WHO, 2023). Entretanto, há ainda muitas incertezas em relação à definição de regras específicas para os *sandboxes* para a IA na saúde. Há também questões jurídicas a serem solucionadas, como determinar as jurisdições competentes para aplicá-los e encaminhar os problemas da responsabilização no uso da IA e da desigualdade de tratamento das empresas que não estejam participando dos ambientes experimentais (BUOCZ; PFOTENHAUER; EISENBERGER, 2023). A abordagem *sandbox* é bastante auspiciosa, porém ainda necessita de amadurecimento jurídico-institucional.

6.3 Regulação de Modelos de Fundação na Saúde (*Foundation Models*)

Os modelos de fundação (***foundation models***) representam o avanço mais recente no campo da IA. O termo *foundation model* foi cunhado em agosto de 2021 para designar qualquer modelo baseado em redes neurais profundas, treinado em vastos conjuntos de dados (muitas vezes não rotulados), que possa ser adaptado para realização de uma ampla gama de tarefas (desde tradução e criação de textos e imagens até a análise de exames diagnósticos). Essa versatilidade os diferencia dos demais modelos de ML, que se concentram em tarefas específicas, e se deve ao treinamento em dados extensos e diversos e à incorporação de técnicas de aprendizado de contexto. Esses modelos ultrapassaram as limitações das arquiteturas de *deep learning* até então existentes e atingiram a escala de centenas de bilhões de parâmetros.

São considerados o estágio mais evoluído já obtido na forma como uma tarefa "emerge" (é inferida automaticamente) no campo de IA/ML e, ao mesmo tempo, na "homogeneização" de técnicas. A "emergência", que resulta da escala dos modelos, permite que um modelo de linguagem possa ser adaptado para uma outra modalidade de tarefa simplesmente recebendo uma descrição da tarefa em linguagem natural, fazendo surgir uma resposta para a qual não houve treinamento específico. A "homogeneização" indica a consolidação dos métodos e arquiteturas para desenvolvimento de sistemas de IA de tal modo que praticamente qualquer tarefa possa ser executada a partir dos mesmos modelos de fundação (BOMMASANI *et al.*, 2022).

As aplicações dos *foundation models* na área da saúde são inúmeras e potencialmente transformadoras. Os modelos desenvolvidos nesse paradigma são capazes de interpretar vários tipos de informações médicas, como imagens, registros de prontuários eletrônicos, resultados de exames laboratoriais, genômica e textos médicos. Além disso, podem oferecer resultados como explicações em linguagem natural, recomendações, anotações e até mesmo conversar com o interlocutor humano (MOOR *et al.*, 2023). Embora inicialmente ainda não tenham sido desenvolvidos modelos de fundação específicos para a saúde, a extensão desses modelos para se tornarem **multimodais** – ou seja, capazes de interpretar não só texto, como também imagem, áudio e vídeo – amplifica exponencialmente as possibilidades de aplicação na saúde individual e coletiva (ACOSTA *et al.*, 2022; TOPOL, 2023).

Esse novo cenário traz muitos desafios adicionais para a regulação da IA na saúde. Como já começa a ser percebido com a introdução dos *chatbots* baseados em grandes modelos de linguagem (LLMs) em ambientes clínicos: há muitos potenciais benefícios, mas também riscos que não podem ser desconsiderados (LEE; BUBECK; PETRO, 2023). Passar esses *chatbots* pelo processo de avaliação de dispositivos médicos (que está sendo adaptado para SaMD com IA/ML) seria o caminho mais adequado, mas ainda não existem meios disponíveis para auditar nem validar segurança e eficácia desses sistemas (GILBERT; HARVEY; *et al.*, 2023; GOTTLIEB; SILVIS, 2023a). Além disso, os modelos de linguagem por trás dessas aplicações permanecem opacos. Eles são caixas pretas em termos de funcionamento e os detalhes de seu desenvolvimento são mantidos em segredo pelas grandes empresas de tecnologia que os controlam (NATURE, 2023b). Ao passo que não há regulação e tampouco possibilidade de incorporação no padrão de cuidado, profissionais de saúde que optem por utilizar *chatbots* de IA na prática clínica e seguir suas recomendações assumem o risco de serem responsabilizados pessoalmente por eventuais resultados desfavoráveis dessa escolha (MELLO; GUHA, 2023). Portanto, a implementação prática de *foundation models* na atividade clínica depende da estruturação de mecanismos para abordar questões éticas e regulatórias cruciais.

7 CONSIDERAÇÕES FINAIS

Esta tese buscou demonstrar como a evolução do campo da IA, desde o seu surgimento nos anos 1950 até a aceleração exponencial notável nas décadas de 2010 e 2020, consolidou-a como uma das principais tecnologias na área da saúde. Ao se tornar uma ferramenta diretamente relacionada à saúde, a **IA passou a ser objeto da regulação sanitária**. Portanto, estabelecer um paradigma para o marco regulatório da IA na saúde assume importância central na sociedade contemporânea.

O cenário atual da IA na saúde reflete os avanços em ***machine learning*** **(ML)**. Logo, representa a aplicação de técnicas de modelagem estatística aliadas à crescente disponibilidade de dados e a grandes capacidades computacionais para processá-los. Os modelos da ML são tão prevalentes que atualmente esse campo é frequentemente tratado como sinônimo de IA. Diversos subtipos e algoritmos de ML têm uso potencial na área da saúde, muitos dos quais usam as técnicas de *deep learning* (DL) – o estágio mais avançado da IA conexionista, que utiliza redes neurais. As aplicações são inúmeras, como os sistemas de suporte à decisão clínica, reconhecimento diagnóstico de imagens, medicina de precisão, análise preditiva para gestão em saúde, desenvolvimento de medicamentos, entre outras.

Independentemente da sua aplicação específica, os modelos de IA compartilham características elementares. O **desenvolvimento dos modelos** começa pela adequada seleção do problema e definição da tarefa a ser predita, seguidas pela seleção e pré-processamento dos dados. O treinamento dos modelos adere às práticas consolidadas na ciência de ML, principalmente do paradigma de aprendizado supervisionado, pois na área da saúde normalmente existe uma resposta certa que o algoritmo precisa aprender. Na **validação dos modelos**, utiliza-se um conjunto de métricas reconhecidas na epidemiologia e na ciência de dados, como sensibilidade, especificidade, precisão e área sob a curva ROC (*Receiver Operating Characteristic*).

Até meados dos anos 2020, a integração de sistemas de IA em contextos clínicos ainda é considerada incipiente. Isso está em parte relacionado aos **desafios para implementação** de ferramentas baseadas em ML na assistência à saúde. Um desses desafios é a escolha dos dados, que na área da saúde são caracteristicamente heterogêneos e suscetíveis a mudanças frequentes. Desenvolver modelos que sejam ao mesmo tempo clinicamente úteis e viáveis para uso prático é uma tarefa bastante complexa. Mesmo alcançando esses objetivos, os sistemas de IA na saúde enfrentam dificuldades devido a diversos tipos de vieses e a limitações na sua capacidade de

generalização. Frequentemente, esses sistemas precisam ser utilizados nos contextos em que foram criados, uma vez que seu desempenho tende a ser insatisfatório em contextos distintos. Por isso, há a necessidade de submetê-los a um monitoramento contínuo de seu desempenho no mundo real. Outra questão relevante é a opacidade, considerando que muitas ferramentas de IA são baseadas em modelos de "caixa preta", cujos processos internos não são diretamente compreensíveis. A interação humano-máquina também é um desafio significativo: mesmo assegurando um desempenho ótimo dos modelos de ML, a eficácia prática depende essencialmente de como os profissionais de saúde os interpretam e utilizam.

Considerando os atributos essenciais comuns a todos os modelos de IA/ML, bem como os desafios próprios para sua a implementação no setor da saúde, tem-se um substrato para a elaboração de um **paradigma regulatório** no contexto sanitário. Essa regulação deve possuir um escopo abrangente, que independa das aplicações específicas, e capaz de avaliar sistemas de IA para quaisquer usos pretendidos – desde sistemas de suporte a decisões clínicas em medicina personalizada até ferramentas destinadas a reorganizar fluxos de referência entre serviços de saúde, por exemplo.

Os primeiros **fundamentos para a regulação da IA na saúde** provém basicamente dos domínios dos direitos humanos e da ética em IA. Da admissão do **direito à saúde** como um direito humano fundamental decorre a obrigação dos Estados de regular todas as ações, serviços e produtos de interesse à saúde. Consequentemente, a regulação da IA transcende uma escolha política, tornando-se um imperativo jurídico na maioria dos países e jurisdições. O campo da **ética em IA** é também diretamente relacionado aos direitos humanos e visa assegurar valores essenciais: dignidade, igualdade, liberdade e não discriminação. Princípios éticos gerais em torno desses valores vêm sendo estabelecidos por entidades internacionais para consolidar parâmetros em escala global.

A partir dos preceitos da ética em IA e dos fundamentos da bioética, recentemente têm sido definidos **princípios éticos específicos para a IA na saúde**. Esses princípios são essenciais na construção de um marco regulatório para a IA na saúde:

a) garantia da autonomia humana em qualquer situação;
b) busca pelo benefício das pessoas e pelo interesse público, assegurando a prevenção de danos aos indivíduos e à sociedade;
c) transparência, com divulgação adequada das informações necessárias sobre o funcionamento dos sistemas de IA em todo o seu ciclo de vida;

d) responsabilização e prestação de contas dos encarregados pelo desenvolvimento e implementação dos sistemas de IA sobre seus resultados e efeitos na sociedade;
e) eliminação de preconceitos e discriminações, assegurando uma distribuição equitativa dos benefícios da IA;
f) sustentabilidade e responsividade às necessidades de saúde das populações.

A **governança de dados de saúde** é também fundamento para a regulação sanitária da IA, pois os dados representam o alicerce do IA. A dimensão de proteção e privacidade de dados pessoais é atualmente um elemento consolidado em quase todas as jurisdições do mundo, constituindo um campo juridicamente autônomo. O paradigma norte-americano de governança de dados de saúde foi pioneiro, mas é hoje reconhecidamente insuficiente diante das mudanças nos contextos social, político e tecnológico. O paradigma europeu emergiu como referência global nessa matéria após o Regulamento Geral sobre a Proteção de Dados (GDPR). Os dados de saúde são admitidos como uma categoria especial de dados pessoais, por isso sujeitos a um tratamento mais rigoroso. O uso de dados de saúde é limitado na finalidade e na duração, só permitido em circunstâncias previstas em lei. Até o início dos anos 2020, a grande maioria dos países já possuíam leis de proteção e privacidade de dados, em sua maioria inspirados no paradigma europeu. As questões centrais na dimensão da **proteção de dados pessoais** para fundamentar a regulação da IA na saúde compreendem:

a) a desidentificação dos dados, seja por anonimização ou pseudonimização, é o principal requisito das legislações de proteção de dados, mas há uma crescente complexidade técnica para assegurar sua eficácia;
b) em algumas situações, o anonimato pode restringir os benefícios positivos potenciais de sistemas de IA, aspecto que deve ser ponderado frente a essa exigência legal;
c) o consentimento individual, livre e informado, é mecanismo essencial na proteção de dados, mas é preciso observar situações que justifiquem sua dispensa, especialmente quando sistemas de IA na saúde tiverem claro interesse público;
d) técnicas de ML, como o "aprendizado federado" (*federated learning*), e estruturas institucionais, como os centros de dados (*data hubs*), são importantes para desenvolver a IA na saúde observando os preceitos legais de governança de dados;
e) autoridades independentes de proteção de dados são essenciais nesse contexto e devem apoiar a construção de marcos regulatórios para a IA na saúde.

Reconhecendo a IA como objeto da atividade regulatória e com base nos fundamentos éticos e jurídicos apresentados, esta tese argumenta que a regulação da IA na saúde deve ser estruturada em três dimensões: segurança e eficácia, transparência e responsabilidade.

Segurança e eficácia são os pilares da regulação de dispositivos médicos. Ao serem desenvolvidos para aplicação em tarefas que envolvam diagnóstico, prevenção, monitoramento, tratamento ou finalidades relacionadas, os sistemas de IA devem ser admitidos na categoria regulatória "software como dispositivo médico" (SaMD). Portanto, ferramentas baseadas em IA devem ser avaliadas sob a perspectiva regulatória de acordo com os padrões internacionais definidos pelo Fórum Internacional de Reguladores de Dispositivos Médicos (IMDRF) e pela Organização Internacional para Padronização (ISO), já adotados por autoridades reguladoras de vários países. As normativas específicas para dispositivos médicos baseados em *machine learning* (MLMD) ainda estão em fase de elaboração. Além das diretrizes das entidades internacionais, as abordagens e estratégias regulatórias traçadas nos Estados Unidos e na Europa vêm sendo referências globais. Os principais aspectos a serem considerados para a regulação da IA na saúde na dimensão de segurança e eficácia são:

a) sistemas de IA/ML na área da saúde devem ser categorizados de acordo com a classificação de risco aplicável aos SaMD (padrão IMDRF) e seguir os critérios de gestão de riscos e gestão da qualidade durante o ciclo de vida (padrões ISO e IMDRF);
b) sistemas de IA/ML devem passar por um processo de avaliação clínica (associação clínica válida, validação analítica e validação clínica), tanto antes implementação quanto no monitoramento contínuo (aprendizado com dados do mundo real);
c) a supervisão regulatória deve contemplar a natureza intrínseca dos sistemas baseados em IA/ML: a capacidade de mudança (padrão IMDRF em desenvolvimento);
d) elementos do novo paradigma norte-americano (proposta FDA em elaboração) devem ser incorporados nos marcos regulatórios: "protocolo de mudança de algoritmo", "boas práticas" em ML, pilotos de avaliação de desempenho no mundo real;
e) é necessário integrar mecanismos para avaliação sistêmica na regulação de dispositivos baseados em IA/ML, principalmente observando a interação humano-máquina;
f) a estrutura institucional deve se adaptar à dinâmica do setor, pois processos extensos podem atrasar a inovação (dificuldades atuais observadas no contexto europeu);
g) avaliação da equidade deve ser parte da avaliação clínica de sistemas de IA/ML, preservando a capacidade de adaptação dos modelos e usando técnicas de calibração para mitigar vieses que possam reproduzir preconceitos ou discriminação.

Transparência é considerada uma dimensão autônoma porque é essencial para todos os demais aspectos da regulação da IA na saúde. A compreensão ampla do conceito de transparência implica na necessidade de tornar acessíveis todas as informações que fundamentam as decisões originadas de modelos de IA/ML. Refere-se à clareza sobre os contextos técnico e institucional nos quais os sistemas de IA são concebidos, implementados e administrados, que é expressa por meio de documentação padronizada. Da perspectiva da ciência de ML, a transparência diz respeito a interpretabilidade e explicabilidade dos sistemas de IA. A compreensão de mecanismos e limites para a explicação em IA na saúde é crucial para definir a abrangência do direito à explicação e os requisitos regulatórios a serem estabelecidos nessa dimensão. São elementos-chave na regulação da IA na saúde no que diz respeito à transparência:

a) os marcos regulatórios devem diferenciar sistemas de IA interpretável (inerentemente explicáveis) de sistemas de IA explicável (explicáveis *post hoc*);
b) o direito à explicação (padrão GDPR) deve ser compreendido como expressão de uma "transparência qualificada": fornecer informações distintas conforme os destinatários;
c) os atuais mecanismos de explicação são capazes de fornecer fatores determinantes de decisão e esclarecer resultados divergentes em modelos interpretáveis (IA interpretável), mas representam apenas aproximações em modelos opacos (IA explicável);
d) pelas técnicas atuais, as explicações obtidas de modelos opacos são pouco confiáveis e, consequentemente, de pouca utilidade para guiar entendimentos morais e jurídicos;
e) os requisitos regulatórios de explicabilidade podem exigir explicações sobre decisões específicas em modelos de IA interpretável, mas devem se restringir a esclarecimentos sobre o funcionamento geral em sistemas de IA explicável;
f) deve ser considerada a exigência de comparação de desempenho com modelos interpretáveis como um requisito regulatório para sistemas baseados em modelos de IA opacos;
g) o mecanismo central na estratégia regulatória deve ser definir critérios para a padronização da documentação detalhada e completa de todas as etapas de desenvolvimento e implementação dos sistemas de IA, considerando listas já existentes como referências.

Responsabilidade constitui a dimensão da regulação indireta da IA na saúde, inserida no contexto dos sistemas de responsabilidade civil. Embora existam diferenças relevantes entre as jurisdições, a responsabilização jurídica geralmente se divide em dois regimes: a responsabilidade subjetiva (por culpa), aplicável aos profissionais de saúde, e a responsabilidade objetiva (independente de culpa), aplicável a provedores e fabricantes (responsabilidade dos produtos). Os sistemas de IA introduzem desafios inéditos a essas estruturas, especialmente ao adquirirem

autonomia e tomarem decisões que não dependem diretamente de operadores humanos. Isso gera uma lacuna na atribuição de responsabilidades, tornando o contexto da IA na saúde particularmente complexo. No âmbito da assistência à saúde, surgem dificuldades adicionais: é necessário delimitar claramente como as responsabilidades são atribuídas entre os diferentes atores envolvidos, como profissionais de saúde, fornecedores, fabricantes e reguladores, em situações em que a IA pode influenciar decisões clínicas, mesmo sem estar integrada ao padrão de cuidado estabelecido. Atualmente, há propostas em discussão para adaptar os sistemas de responsabilidade civil a essa nova realidade, tendo como referência principal a legislação europeia. Pontos pertinentes para a dimensão da responsabilidade na regulação da IA na saúde incluem:

a) algumas inovações jurídicas, como a responsabilidade empresarial comum ou conferir personalidade "quase-jurídica" aos sistemas de IA, são pouco promissoras;
b) a manutenção da lógica atual dos regimes de responsabilidade civil pode desestimular significativamente o uso de IA em ambientes clínicos até que essas ferramentas sejam oficialmente incluídas nos padrões de cuidado;
c) pode ser apropriado considerar a adoção de sistemas de compensação sem culpa, o que exigiria mudanças jurídicas e institucionais significativas na maioria dos países;
d) a tendência atual é a adaptação de regras de responsabilidade civil extracontratual e de responsabilidade dos produtos para facilitar o acesso aos demandantes e estimular cuidados por parte de fabricantes e provedores, mas os resultados ainda são incertos;
e) é possível que futuramente sejam criados regimes especiais de responsabilidade objetiva associados a seguros obrigatórios.

Os fundamentos analisados e as diretrizes propostas nesta tese constituem uma base sólida para a elaboração de um paradigma regulatório robusto e abrangente para a IA na saúde. Os objetivos da pesquisa foram atingidos, demonstrando a importância do tema para as áreas da medicina, da saúde pública e dos direitos humanos. Contudo, este estudo representa apenas o início de uma ampla linha de pesquisa, que se desdobra em diversas ramificações e tem aplicabilidade prática significativa nos contextos político, jurídico e sanitário. O campo da IA na saúde é vasto e está em rápida expansão, como ilustram as tendências emergentes discutidas. Desafios futuros incluem compreender o papel das leis horizontais e da abordagem regulatória setorial, estratégias para desenvolver *sandboxes* regulatórios para dispositivos baseados em IA/ML, e mecanismos de regulação para garantir que o uso de *foundation models* na prática clínica seja viável e esteja mais próximo de se tornar uma realidade.

Por enquanto.

REFERÊNCIAS

ACOSTA, Julián N.; FALCONE, Guido J.; RAJPURKAR, Pranav; TOPOL, Eric J. Multimodal biomedical AI. **Nature Medicine**, v. 28, n. 9, p. 1773–1784, set. 2022. https://doi.org/10.1038/s41591-022-01981-2.

AHMED, Tanbir; AZIZ, Md Momin Al; MOHAMMED, Noman. De-identification of electronic health record using neural network. **Scientific Reports**, v. 10, n. 1, p. 18600, 29 out. 2020. https://doi.org/10.1038/s41598-020-75544-1.

AITH, Fernando; DALLARI, Analluza Bolivar (Orgs.). **LGPD na saúde digital**. São Paulo, SP: Thomson Reuters Brasil, 2022.

ALAMI, Hassane; RIVARD, Lysanne; LEHOUX, Pascale; HOFFMAN, Steven J.; CADEDDU, Stéphanie Bernadette Mafalda; SAVOLDELLI, Mathilde; SAMRI, Mamane Abdoulaye; AG AHMED, Mohamed Ali; FLEET, Richard; FORTIN, Jean-Paul. Artificial intelligence in health care: laying the Foundation for Responsible, sustainable, and inclusive innovation in low- and middle-income countries. **Globalization and Health**, v. 16, n. 1, p. 52, 24 jun. 2020. https://doi.org/10.1186/s12992-020-00584-1.

ALLAIN, Jessica. From Jeopardy! to Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems. **Louisiana Law Review**, v. 73, n. 4, 1 ago. 2013. Disponível em: https://digitalcommons.law.lsu.edu/lalrev/vol73/iss4/7.

ALVAREZ-ROMERO, Celia; MARTÍNEZ-GARCÍA, Alicia; BERNABEU-WITTEL, Máximo; PARRA-CALDERÓN, Carlos Luis. Health data hubs: an analysis of existing data governance features for research. **Health Research Policy and Systems**, v. 21, n. 1, p. 70, 10 jul. 2023. https://doi.org/10.1186/s12961-023-01026-1.

AMANN, Julia; BLASIMME, Alessandro; VAYENA, Effy; FREY, Dietmar; MADAI, Vince I.; THE PRECISE4Q CONSORTIUM. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. **BMC Medical Informatics and Decision Making**, v. 20, n. 1, p. 310, 30 nov. 2020. https://doi.org/10.1186/s12911-020-01332-6.

AMANN, Julia; VETTER, Dennis; BLOMBERG, Stig Nikolaj; CHRISTENSEN, Helle Collatz; COFFEE, Megan; GERKE, Sara; GILBERT, Thomas K.; HAGENDORFF, Thilo; HOLM, Sune; LIVNE, Michelle; SPEZZATTI, Andy; STRÜMKE, Inga; ZICARI, Roberto V.; MADAI, Vince Istvan; INITIATIVE, on behalf of the Z.-Inspection. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. **PLOS Digital Health**, v. 1, n. 2, p. e0000016, 17 fev. 2022. https://doi.org/10.1371/journal.pdig.0000016.

ARONSON, Jeffrey K.; HENEGHAN, Carl; FERNER, Robin E. Medical Devices: Definition, Classification, and Regulatory Implications. **Drug Safety**, v. 43, n. 2, p. 83–93, 1 fev. 2020. https://doi.org/10.1007/s40264-019-00878-3.

BABIC, Boris; COHEN, I. Glenn. The Algorithmic Explainability “Bait and Switch”. Rochester, NY, 15 ago. 2023. Disponível em: https://papers.ssrn.com/abstract=4541487.

BABIC, Boris; GERKE, Sara; EVGENIOU, Theodoros; COHEN, I. Glenn. Algorithms on regulatory lockdown in medicine. **Science**, v. 366, n. 6470, p. 1202–1204, 6 dez. 2019. https://doi.org/10.1126/science.aay9547.

BABIC, Boris; GERKE, Sara; EVGENIOU, Theodoros; COHEN, I. Glenn. Beware explanations from AI in health care. **Science**, v. 373, n. 6552, p. 284–286, 16 jul. 2021a. https://doi.org/10.1126/science.abg1834.

BABIC, Boris; GERKE, Sara; EVGENIOU, Theodoros; COHEN, I. Glenn. Direct-to-consumer medical machine learning and artificial intelligence applications. **Nature Machine Intelligence**, v. 3, n. 4, p. 283-87, 20 abr. 2021b. https://doi.org/10.1038/s42256-021-00331-0.

BALAGOPALAN, Aparna; ZHANG, Haoran; HAMIDIEH, Kimia; HARTVIGSEN, Thomas; RUDZICZ, Frank; GHASSEMI, Marzyeh. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. 20 jun. 2022. **Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 20 jun. 2022. p. 1194–1206. Disponível em: https://dl.acm.org/doi/10.1145/3531146.3533179.

BATES, David W. How to regulate evolving AI health algorithms. **Nature Medicine**, v. 29, n. 1, p. 26–26, jan. 2023. https://doi.org/10.1038/s41591-022-02165-8.

BATES, David W.; SARIA, Suchi; OHNO-MACHADO, Lucila; SHAH, Anand; ESCOBAR, Gabriel. Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. **Health Affairs**, v. 33, n. 7, p. 1123–1131, jul. 2014. https://doi.org/10.1377/hlthaff.2014.0041.

BEAM, Andrew L.; KOHANE, Isaac S. Big Data and Machine Learning in Health Care. **JAMA**, v. 319, n. 13, p. 1317–1318, 3 abr. 2018. https://doi.org/10.1001/jama.2017.18391.

BEAUCHAMP, Tom L.; CHILDRESS, James F. **Principles of biomedical ethics**. Oxford: Oxford University Press, 1979.

BENJAMENS, Stan; DHUNNOO, Pranavsingh; MESKÓ, Bertalan. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. **npj Digital Medicine**, v. 3, n. 1, p. 1–8, set. 2020. https://doi.org/10.1038/s41746-020-00324-0.

BERNIER, Alexander; MOLNÁR-GÁBOR, Fruzsina; KNOPPERS, Bartha Maria. The international data governance landscape. **Journal of Law and the Biosciences**, v. 9, n. 1, lsac005, 2022. https://doi.org/10.1093/jlb/lsac005.

BISHOP, Christopher M. **Pattern recognition and machine learning**. New York: Springer, 2006(Information science and statistics).

BOBBIO, Norberto. **The age of rights**. Cambridge, UK : Oxford, OX, UK ; Cambridge, MA, USA: Polity Press ; Blackwell Publishers, 1996.

BÖCKENFÖRDE, Ernst-Wolfgang. Fundamental Rights Theory and Interpretation [1974]. *In*: KÜNKLER, Mirjam; STEIN, Tine (orgs.). **Constitutional and Political Theory**. Oxford, United Kingdom: Oxford University Press, 2017. p. 266–289. Disponível em: https://academic.oup.com/book/5358/chapter/148154297.

BOHR, Adam; MEMARZADEH, Kaveh. The rise of artificial intelligence in healthcare applications. *In*: BOHR, Adam; MEMARZADEH, Kaveh (orgs.). **Artificial Intelligence in Healthcare**. London ; San Diego, CA: Academic Press, 2020. p. 25–60. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780128184387000022.

BOMMASANI, Rishi; HUDSON, Drew A.; ADELI, Ehsan; ALTMAN, Russ; ARORA, Simran; VON ARX, Sydney; BERNSTEIN, Michael S.; BOHG, Jeannette; BOSSELUT, Antoine; BRUNSKILL, Emma; BRYNJOLFSSON, Erik; BUCH, Shyamal; CARD, Dallas; CASTELLON, Rodrigo; CHATTERJI, Niladri; CHEN, Annie; CREEL, Kathleen; DAVIS, Jared Quincy; DEMSZKY, Dora; … LIANG, Percy. On the Opportunities and Risks of Foundation Models. 12 jul. 2022. Disponível em: http://arxiv.org/abs/2108.07258.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). DOU, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

BREMMER, Ian; SULEYMAN, Mustafa. The AI Power Paradox. **Foreign Affairs**, 16 ago. 2023. Disponível em: https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox. Acesso em: 17 ago. 2023.

BRINGSJORD, Selmer; GOVINDARAJULU, Naveen Sundar. Artificial Intelligence. *In*: ZALTA, Edward N.; NODELMAN, Uri (orgs.). **The Stanford Encyclopedia of Philosophy**. Fall 2022 Edition. Stanford, CA: Metaphysics Research Lab, Stanford University, 2022. Disponível em: https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/.

BROWNSTEIN, John S.; RADER, Benjamin; ASTLEY, Christina M.; TIAN, Huaiyu. Advances in Artificial Intelligence for Infectious-Disease Surveillance. **New England Journal of Medicine**, v. 388, n. 17, p. 1597–1607, 27 abr. 2023. https://doi.org/10.1056/NEJMra2119215.

BUCKNER, Cameron; GARSON, James. Connectionism. *In*: ZALTA, Edward N. (org.). **The Stanford Encyclopedia of Philosophy**. Fall 2019. Stanford, CA: Metaphysics Research Lab, Stanford University, 2019. Disponível em: https://plato.stanford.edu/archives/fall2019/entries/connectionism/.

BUOCZ, Thomas; PFOTENHAUER, Sebastian; EISENBERGER, Iris. Regulatory sandboxes in the AI Act: reconciling innovation and safety? **Law, Innovation and Technology**, v. 15, n. 2, p. 357–389, 3 jul. 2023. https://doi.org/10.1080/17579961.2023.2245678.

BURRELL, Jenna. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. **Big Data & Society**, v. 3, n. 1, p. 1–12, jun. 2016. https://doi.org/10.1177/2053951715622512.

BUSSANI, Mauro; SEBOK, Anthony; INFANTINO, Marta. **Common Law and Civil Law Perspectives on Tort Law**. New York: Oxford University Press, 2022. DOI 10.1093/oso/9780195368383.001.0001. Disponível em: https://academic.oup.com/book/41844.

BYGRAVE, Lee A. Article 22 Automated individual decision-making, including profiling. *In*: BYGRAVE, Lee A.**The EU General Data Protection Regulation (GDPR)**. Oxford University Press, 2020. DOI 10.1093/oso/9780198826491.003.0055. Disponível em: https://academic.oup.com/book/41324/chapter/352298561.

BYGRAVE, Lee A.; TOSONI, Luca. Article 4(15). Data concerning health. *In*: KUNER, Christopher; BYGRAVE, Lee A.; DOCKSEY, Christopher. **The EU General Data Protection Regulation (GDPR)**. Oxford, United Kingdom: Oxford University Press, 2020. p. 217–224. Disponível em: https://doi.org/10.1093/oso/9780198826491.003.0021.

CALLAWAY, Ewen. 'The entire protein universe': AI predicts shape of nearly every known protein. **Nature**, v. 608, n. 7921, p. 15–16, 28 jul. 2022. https://doi.org/10.1038/d41586-022-02083-2.

CELI, Leo Anthony; CELLINI, Jacqueline; CHARPIGNON, Marie-Laure; DEE, Edward Christopher; DERNONCOURT, Franck; EBER, Rene; MITCHELL, William Greig; MOUKHEIBER, Lama; SCHIRMER, Julian; SITU, Julia; PAGUIO, Joseph; PARK, Joel; WAWIRA, Judy Gichoya; YAO, Seth; FOR MIT CRITICAL DATA. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. **PLOS Digital Health**, v. 1, n. 3, e0000022, 31 mar. 2022. https://doi.org/10.1371/journal.pdig.0000022.

CHAPELLE, Olivier; SCHÖLKOPF, Bernhard; ZIEN, Alexander (Orgs.). **Semi-supervised learning**. Cambridge, Massachusetts: MIT Press, 2010 (Adaptive computation and machine learning series).

CHAR, Danton S.; ABRÀMOFF, Michael D.; FEUDTNER, Chris. Identifying Ethical Considerations for Machine Learning Healthcare Applications. **The American Journal of Bioethics**, v. 20, n. 11, p. 7–17, 1 nov. 2020. https://doi.org/10.1080/15265161.2020.1819469.

CHATTERJEE, Sheshadri. Is data privacy a fundamental right in India? An analysis and recommendations from policy and legal perspective. **International Journal of Law and Management**, v. 61, n. 1, p. 170-90, fev. 2019. https://doi.org/10.1108/IJLMA-01-2018-0013.

CHEN, Hongming; ENGKVIST, Ola; WANG, Yinhai; OLIVECRONA, Marcus; BLASCHKE, Thomas. The rise of deep learning in drug discovery. **Drug Discovery Today**, v. 23, n. 6, p. 1241–1250, jun. 2018. https://doi.org/10.1016/j.drudis.2018.01.039.

CHEN, Jonathan H.; ALAGAPPAN, Muthuraman; GOLDSTEIN, Mary K.; ASCH, Steven M.; ALTMAN, Russ B. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. **International Journal of Medical Informatics**, v. 102, p. 71–79, 1 jun. 2017. https://doi.org/10.1016/j.ijmedinf.2017.03.006.

CHEN, Po-Hsuan Cameron; LIU, Yun; PENG, Lily. How to develop machine learning models for healthcare. **Nature Materials**, v. 18, n. 5, p. 410–414, maio 2019. https://doi.org/10.1038/s41563-019-0345-0.

CHEVRIER, Raphaël; FOUFI, Vasiliki; GAUDET-BLAVIGNAC, Christophe; ROBERT, Arnaud; LOVIS, Christian. Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. **Journal of Medical Internet Research**, v. 21, n. 5, p. e13484, 31 maio 2019. https://doi.org/10.2196/13484.

CHIKWETU, Lucy; MIAO, Yu; WOLDETENSAE, Melat K; BELL, Diarra; GOLDENHOLZ, Daniel M; DUNN, Jessilyn. Does deidentification of data from wearable devices give us a false sense of security? A systematic review. **The Lancet Digital Health**, v. 5, n. 4, p. e239–e247, abr. 2023. https://doi.org/10.1016/S2589-7500(22)00234-5.

COHEN, I. Glenn. Is There a Duty to Share Healthcare Data? *In*: VAYENA, Effy; LYNCH, Holly Fernandez; COHEN, I. Glenn; GASSER, Urs (orgs.). **Big Data, Health Law, and Bioethics**. Cambridge: Cambridge University Press, 2018. p. 209–222. Disponível em: https://doi.org/10.1017/9781108147972.020.

COHEN, I Glenn; EVGENIOU, Theodoros; GERKE, Sara; MINSSEN, Timo. The European artificial intelligence strategy: implications and challenges for digital health. **The Lancet Digital Health**, v. 2, n. 7, p. e376–e379, jul. 2020. https://doi.org/10.1016/S2589-7500(20)30112-6.

COHEN, I. Glenn; MELLO, Michelle M. HIPAA and Protecting Health Information in the 21st Century. **JAMA**, v. 320, n. 3, p. 231–232, 17 jul. 2018. https://doi.org/10.1001/jama.2018.5630.

COLLINS, Gary S.; DHIMAN, Paula; NAVARRO, Constanza L. Andaur; MA, Jie; HOOFT, Lotty; REITSMA, Johannes B.; LOGULLO, Patricia; BEAM, Andrew L.; PENG, Lily; CALSTER, Ben Van; SMEDEN, Maarten van; RILEY, Richard D.; MOONS, Karel GM. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. **BMJ Open**, v. 11, n. 7, p. e048008, 1 jul. 2021. https://doi.org/10.1136/bmjopen-2020-048008.

COLLINS, Gary S.; REITSMA, Johannes B.; ALTMAN, Douglas G.; MOONS, Karel G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. **BMJ**, v. 350, p. g7594, 7 jan. 2015. https://doi.org/10.1136/bmj.g7594.

CORBETT-DAVIES, Sam; GAEBLER, Johann D.; NILFOROSHAN, Hamed; SHROFF, Ravi; GOEL, Sharad. The Measure and Mismeasure of Fairness. 14 ago. 2023. Disponível em: http://arxiv.org/abs/1808.00023.

COUNCIL OF EUROPE. **Addressing the impact of algorithms on human rights**. Strasbourg, 2018. Disponível em: https://rm.coe.int/draft-recommendation-of-the-committee-of-ministers-to-states-on-the-hu/168095eecf.

COUNCIL OF EUROPE. **Unboxing Artificial Intelligence: 10 steps to protect Human Rights**. Strasbourg, maio 2019. Disponível em: https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64.

DAVIS, Sharon E.; WALSH, Colin G.; MATHENY, Michael E. Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings. **Frontiers in Digital Health**, v. 4, 2022. Disponível em: https://doi.org/10.3389/fdgth.2022.958284.

DE LUCA, Stefano. **New Product Liability Directive. Briefing: EU Legislation in Progress**. European Parliamentary Research Service, maio 2023. Disponível em: https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739341/EPRS_BRI(2023)739341_EN.pdf. Acesso em: 1 nov. 2023.

DIGNUM, Virginia. Responsibility and Artificial Intelligence. *In*: DUBBER, Markus D.; PASQUALE, Frank; DAS, Sunit (orgs.). **The Oxford Handbook of Ethics of AI**. New York: Oxford University Press, 2020. p. 213–231. DOI 10.1093/oxfordhb/9780190067397.013.12. Disponível em: https://academic.oup.com/edited-volume/34287/chapter/290662041.

DOSHI-VELEZ, Finale; KORTZ, Mason; BUDISH, Ryan; BAVITZ, Chris; GERSHMAN, Sam; O'BRIEN, David; SCOTT, Kate; SCHIEBER, Stuart; WALDO, James; WEINBERGER, David; WELLER, Adrian; WOOD, Alexandra. Accountability of AI Under the Law: The Role of Explanation. 20 dez. 2019.
Disponível em: http://arxiv.org/abs/1711.01134.

DOURADO, Daniel De Araujo; AITH, Fernando Mussa Abujamra. A regulação da inteligência artificial na saúde no Brasil começa com a Lei Geral de Proteção de Dados Pessoais. **Revista de Saúde Pública**, v. 56, p. 80, 9 set. 2022.
https://doi.org/10.11606/s1518-8787.2022056004461.

DUBBER, Markus Dirk; PASQUALE, Frank; DAS, Sunit (Orgs.). **The Oxford handbook of ethics of AI**. New York: Oxford University Press, 2020(Oxford handbooks series).

DUFFOURC, Mindy Nunez; GERKE, Sara. The proposed EU Directives for AI liability leave worrying gaps likely to impact medical AI. **npj Digital Medicine**, v. 6, n. 1, p. 77, 26 abr. 2023. https://doi.org/10.1038/s41746-023-00823-w.

EMA. The use of Artificial Intelligence (AI) in the medicinal product lifecycle. 17 jul. 2023. **European Medicines Agency**. Disponível em: https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle. Acesso em: 10 nov. 2023.

EMANUEL, Ezekiel J.; WACHTER, Robert M. Artificial Intelligence in Health Care: Will the Value Match the Hype? **JAMA**, v. 321, n. 23, p. 2281–2282, 18 jun. 2019.
https://doi.org/10.1001/jama.2019.4914.

ENGEL, David M.; MCCANN, Michael W. (Orgs.). **Fault lines: tort law as cultural practice**. Stanford, Calif: Stanford Law Books, 2009 (The cultural lives of law).

ESTEVA, Andre; ROBICQUET, Alexandre; RAMSUNDAR, Bharath; KULESHOV, Volodymyr; DEPRISTO, Mark; CHOU, Katherine; CUI, Claire; CORRADO, Greg; THRUN, Sebastian; DEAN, Jeff. A guide to deep learning in healthcare. **Nature Medicine**, v. 25, n. 1, p. 24–29, jan. 2019. https://doi.org/10.1038/s41591-018-0316-z.

EUROPEAN COMMISSION. **A definition of AI: main capabilities and scientific disciplines**. Brussels, 2019a. Disponível em: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines.

EUROPEAN COMMISSION. **Commission Staff Working Document.** Executive Summary of the Impact Assessment Report accompanying the document Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). Brussels, 28 set. 2022a.
Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022SC0320&qid=1676375584939.
Acesso em: 1 nov. 2023.

EUROPEAN COMMISSION. **Ethics guidelines for trustworthy AI**. Brussels, 2019b.
Disponível em: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

EUROPEAN COMMISSION. **Proposal for a Directive of the European Parliament and of the Council on liability for defective products**. Brussels, 28 set. 2022b. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0495&qid=1676374048444. Acesso em: 1 nov. 2023.

EUROPEAN COMMISSION. **Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)**. Brussels, 28 set. 2022c. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496&qid=1676374228766. Acesso em: 1 nov. 2023.

EUROPEAN COMMISSION. **Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts**. Brussels, 21 abr. 2021. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206. Acesso em: 18 out. 2023.

EUROPEAN COMMISSION. **Proposal for a Regulation of the European Parliament and of the Council on the European Health Data Space**. Strasbourg, 3 maio 2022d. Disponível em: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197. Acesso em: 4 out. 2023.

EUROPEAN COMMISSION. **White Paper on Artificial Intelligence: a European approach to excellence and trust**. Brussels, 19 fev. 2020. Disponível em: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

EUROPEAN UNION. **Commission Implementing Decision (EU) 2022/757 of 11 May 2022 amending Implementing Decision (EU) 2021/1182 as regards harmonised standards for quality management systems, sterilisation and application of risk management to medical devices**. European Commission, 17 maio 2022. Disponível em: https://eur-lex.europa.eu/eli/dec_impl/2022/757/oj. Acesso em: 1 out. 2023.

EUROPEAN UNION. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ, n. L 119, p. 1–88, 4 maio 2016. Disponível em: https://eur-lex.europa.eu/eli/reg/2016/679/oj. Acesso em: 20 set. 2023.

EUROPEAN UNION. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. OJ, n. L 117, p. 1–175, 5 maio 2017. Disponível em: http://data.europa.eu/eli/reg/2017/745/oj. Acesso em: 20 set. 2023.

FCA. **Regulatory sandbox**. Financial Conduct Authority, nov. 2015. Disponível em: https://www.fca.org.uk/publication/research/regulatory-sandbox.pdf.

FDA. **Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan**. U.S. Food and Drug Administration – FDA, jan. 2021. Disponível em: https://www.fda.gov/media/145022/download.

FDA. **Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff**. U.S. Food and Drug Administration – FDA, 28 set. 2022. Disponível em: https://www.fda.gov/media/109618/download.

FDA. **Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback**. U.S. Food and Drug Administration – FDA, 2 abr. 2019. Disponível em: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf.

FDA. Software as a Medical Device (SaMD). 9 set. 2020. FDA. Disponível em: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd. Acesso em: 13 out. 2023.

FENG, Jean; PHILLIPS, Rachael V.; MALENICA, Ivana; BISHARA, Andrew; HUBBARD, Alan E.; CELI, Leo A.; PIRRACCHIO, Romain. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. **npj Digital Medicine**, v. 5, n. 1, p. 1–9, maio 2022. https://doi.org/10.1038/s41746-022-00611-y.

FIELD, Robert I. **Health care regulation in America: complexity, confrontation, and compromise**. New York, NY: Oxford University Press, 2006.

FIHN, Stephan; SARIA, Suchi; MENDONÇA, Eneida; HAIN, Seth; MATHENY, Michael E.; SHAH, Nigam; LIU, Hongfang; AUERBACH, Andrew. Deploying Artificial Intelligence in Clinical Settings. *In*: MATHENY, Michael E.; ISRANI, Sonoo Thadaney; AHMED, Mahnoor; WHICHER, Danielle (orgs.). **Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril**. Washington, DC: National Academy of Medicine, 2022. p. 159–196.

FINLAYSON, Samuel G.; SUBBASWAMY, Adarsh; SINGH, Karandeep; BOWERS, John; KUPKE, Annabel; ZITTRAIN, Jonathan; KOHANE, Isaac S.; SARIA, Suchi. The Clinician and Dataset Shift in Artificial Intelligence. **New England Journal of Medicine**, v. 385, n. 3, p. 283–286, jul. 2021. https://doi.org/10.1056/NEJMc2104626.

FJELLAND, Ragnar. Why general artificial intelligence will not be realized. **Humanities and Social Sciences Communications**, v. 7, n. 1, p. 1–9, 17 jun. 2020. https://doi.org/10.1057/s41599-020-0494-4.

FLORIDI, Luciano. Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2083, p. 20160112, dez. 2016. https://doi.org/10.1098/rsta.2016.0112.

FLORIDI, Luciano; COWLS, Josh; BELTRAMETTI, Monica; CHATILA, Raja; CHAZERAND, Patrice; DIGNUM, Virginia; LUETGE, Christoph; MADELIN, Robert; PAGALLO, Ugo; ROSSI, Francesca; SCHAFER, Burkhard; VALCKE, Peggy; VAYENA, Effy. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. **Minds and Machines**, v. 28, n. 4, p. 689–707, dez. 2018. https://doi.org/10.1007/s11023-018-9482-5.

FLORIDI, Luciano; COWLS, Josh; KING, Thomas C.; TADDEO, Mariarosaria. How to Design AI for Social Good: Seven Essential Factors. **Science and Engineering Ethics**, v. 26, n. 3, p. 1771–1796, 3 abr. 2020. https://doi.org/10.1007/s11948-020-00213-5.

FOLLAND, Sherman; GOODMAN, Allen C.; STANO, Miron. **The economics of health and health care**. 7th ed. Upper Saddle River, N.J: Pearson, 2013.

FRA, European Union Agency for Fundamental Rights. **Bias in algorithms: artificial intelligence and discrimination**. Luxembourg: Publications Office of the European Union, 2022. Disponível em: https://data.europa.eu/doi/10.2811/25847.

FUTOMA, Joseph; SIMONS, Morgan; PANCH, Trishan; DOSHI-VELEZ, Finale; CELI, Leo Anthony. The myth of generalisability in clinical research and machine learning in health care. **The Lancet Digital Health**, v. 2, n. 9, p. e489–e492, 1 set. 2020. https://doi.org/10.1016/S2589-7500(20)30186-2.

G7 HIROSHIMA SUMMIT. **G7 Hiroshima Leaders' Communiqué**. 20 maio 2023. Disponível em: https://www.g7hiroshima.go.jp/documents/pdf/Leaders_Communique_01_en.pdf.

GENESERETH, Michael R.; NILSSON, Nils J. **Logical foundations of artificial intelligence**. Los Altos: Morgan Kaufmann, 1987.

GERKE, Sara. "Nutrition Facts Labels" for Artificial Intelligence/Machine Learning-Based Medical Devices—The Urgent Need for Labeling Standards. Rochester, NY, 1 fev. 2023. Disponível em: https://papers.ssrn.com/abstract=4404252.

GERKE, Sara; BABIC, Boris; EVGENIOU, Theodoros; COHEN, I. Glenn. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. **npj Digital Medicine**, v. 3, n. 1, p. 1–4, 7 abr. 2020. https://doi.org/10.1038/s41746-020-0262-2.

GERKE, Sara; MINSSEN, Timo; COHEN, I. Glenn. Ethical and legal challenges of artificial intelligence-driven healthcare. *In*: BOHR, Adam; MEMARZADEH, Kaveh (orgs.). **Artificial intelligence in healthcare**. London ; San Diego, CA: Academic Press, 2020. p. 295–336. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780128184387000125.

GHASSEMI, Marzyeh; OAKDEN-RAYNER, Luke; BEAM, Andrew L. The false hope of current approaches to explainable artificial intelligence in health care. **The Lancet Digital Health**, v. 3, n. 11, e745–e750, nov. 2021. https://doi.org/10.1016/S2589-7500(21)00208-9.

GHTF, Study Group 1. **Information Document Concerning the Definition of the Term "Medical Device"**. The Global Harmonization Task Force, 20 maio 2005. Disponível em: https://www.imdrf.org/sites/default/files/docs/ghtf/final/sg1/technical-docs/ghtf-sg1-n29r16-2005-definition-medical-device-050520.pdf.

GIANFRANCESCO, Milena A.; TAMANG, Suzanne; YAZDANY, Jinoos; SCHMAJUK, Gabriela. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. **JAMA Internal Medicine**, v. 178, n. 11, p. 1544–1547, nov. 2018. https://doi.org/10.1001/jamainternmed.2018.3763.

GILBERT, Stephen; ANDERSON, Stuart; DAUMER, Martin; LI, Phoebe; MELVIN, Tom; WILLIAMS, Robin. Learning From Experience and Finding the Right Balance in the Governance of Artificial Intelligence and Digital Health Technologies. **Journal of Medical Internet Research**, v. 25, n. 1, p. e43682, 14 abr. 2023. https://doi.org/10.2196/43682.

GILBERT, Stephen; FENECH, Matthew; HIRSCH, Martin; UPADHYAY, Shubhanan; BIASIUCCI, Andrea; STARLINGER, Johannes. Algorithm Change Protocols in the Regulation of Adaptive Machine Learning–Based Medical Devices. **Journal of Medical Internet Research**, v. 23, n. 10, e30545, 26 out. 2021. https://doi.org/10.2196/30545.

GILBERT, Stephen; HARVEY, Hugh; MELVIN, Tom; VOLLEBREGT, Erik; WICKS, Paul. Large language model AI chatbots require approval as medical devices. **Nature Medicine**, p. 1–3, 30 jun. 2023. https://doi.org/10.1038/s41591-023-02412-6.

GIOVANOLA, Benedetta; TIRIBELLI, Simona. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. **AI & SOCIETY**, v. 38, n. 2, p. 549–563, abr. 2023. https://doi.org/10.1007/s00146-022-01455-6.

GODDARD, Kate; ROUDSARI, Abdul; WYATT, Jeremy C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. **Journal of the American Medical Informatics Association**, v. 19, n. 1, p. 121–127, jan. 2012. https://doi.org/10.1136/amiajnl-2011-000089.

GOERTZEL, Ben. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. **Journal of Artificial General Intelligence**, v. 5, n. 1, p. 1–48, 1 dez. 2014. https://doi.org/10.2478/jagi-2014-0001.

GOMES, Bruna; ASHLEY, Euan A. Artificial Intelligence in Molecular Medicine. **New England Journal of Medicine**, v. 388, n. 26, p. 2456–2465, 29 jun. 2023. https://doi.org/10.1056/NEJMra2204787.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge, Massachusetts: The MIT Press, 2016 (Adaptive computation and machine learning). Disponível em: www.deeplearningbook.org.

GOSTIN, Lawrence O. **Public health law: power, duty, restraint**. Rev. and expanded 2nd ed. Berkeley : New York : Milbank Memorial Fund: University of California Press, 2008 (California/Milbank books on health and the public, 3).

GOTTLIEB, Scott; SILVIS, Lauren. How to Safely Integrate Large Language Models Into Health Care. **JAMA Health Forum**, v. 4, n. 9, e233909, 21 set. 2023a. https://doi.org/10.1001/jamahealthforum.2023.3909.

GOTTLIEB, Scott; SILVIS, Lauren. Regulators Face Novel Challenges as Artificial Intelligence Tools Enter Medical Practice. **JAMA Health Forum**, v. 4, n. 6, e232300, 8 jun. 2023b. https://doi.org/10.1001/jamahealthforum.2023.2300.

GREENLEAF, Graham. Global Data Privacy Laws 2023: 162 National Laws and 20 Bills. **SSRN Electronic Journal**, 2023. Disponível em: https://dx.doi.org/10.2139/ssrn.4426146.

GROTE, Thomas; BERENS, Philipp. On the ethics of algorithmic decision-making in healthcare. **Journal of Medical Ethics**, v. 46, n. 3, p. 205–211, mar. 2020. https://doi.org/10.1136/medethics-2019-105586.

GUNNING, David; STEFIK, Mark; CHOI, Jaesik; MILLER, Timothy; STUMPF, Simone; YANG, Guang-Zhong. XAI—Explainable artificial intelligence. **Science Robotics**, v. 4, n. 37, eaay7120, 18 dez. 2019. https://doi.org/10.1126/scirobotics.aay7120.

HAENLEIN, Michael; KAPLAN, Andreas. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. **California Management Review**, v. 61, n. 4, p. 5–14, ago. 2019. https://doi.org/10.1177/0008125619864925.

HARRISON, James H., Jr; GILBERTSON, John R.; HANNA, Matthew G.; OLSON, Niels H.; SEHEULT, Jansen N.; SORACE, James M.; STRAM, Michelle N. Introduction to Artificial Intelligence and Machine Learning for Pathology. **Archives of Pathology & Laboratory Medicine**, v. 145, n. 10, p. 1228–1254, 25 jan. 2021. https://doi.org/10.5858/arpa.2020-0541-CP.

HASSABIS, Demis; KUMARAN, Dharshan; SUMMERFIELD, Christopher; BOTVINICK, Matthew. Neuroscience-Inspired Artificial Intelligence. **Neuron**, v. 95, n. 2, p. 245–258, jul. 2017. https://doi.org/10.1016/j.neuron.2017.06.011.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. 2nd ed. New York, NY: Springer, 2009 (Springer series in statistics).

HAUG, Charlotte J.; DRAZEN, Jeffrey M. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. **New England Journal of Medicine**, v. 388, n. 13, p. 1201–1208, 30 mar. 2023. https://doi.org/10.1056/NEJMra2302038.

HAUPT, Claudia E.; MARKS, Mason. AI-Generated Medical Advice—GPT and Beyond. **JAMA**, v. 329, n. 16, p. 1349–1350, 25 abr. 2023. https://doi.org/10.1001/jama.2023.5321.

HE, Jianxing; BAXTER, Sally L.; XU, Jie; XU, Jiming; ZHOU, Xingtao; ZHANG, Kang. The practical implementation of artificial intelligence technologies in medicine. **Nature Medicine**, v. 25, n. 1, p. 30–36, jan. 2019. https://doi.org/10.1038/s41591-018-0307-0.

HERNANDEZ-BOUSSARD, Tina; BOZKURT, Selen; IOANNIDIS, John P A; SHAH, Nigam H. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. **Journal of the American Medical Informatics Association**, v. 27, n. 12, p. 2011–2015, 9 dez. 2020. https://doi.org/10.1093/jamia/ocaa088.

HERZOG, Christian. On the Ethical and Epistemological Utility of Explicable AI in Medicine. **Philosophy & Technology**, v. 35, n. 2, 30 maio 2022. https://doi.org/10.1007/s13347-022-00546-y.

HILBERT, Martin; LÓPEZ, Priscila. The World's Technological Capacity to Store, Communicate, and Compute Information. **Science**, v. 332, n. 6025, p. 60–65, abr. 2011. https://doi.org/10.1126/science.1200970.

HINTON, Geoffrey E. Deep Learning-A Technology With the Potential to Transform Health Care. **JAMA**, v. 320, n. 11, p. 1101–1102, 18 set. 2018. https://doi.org/10.1001/jama.2018.11100.

HOI, Steven C.H.; SAHOO, Doyen; LU, Jing; ZHAO, Peilin. Online learning: A comprehensive survey. **Neurocomputing**, v. 459, p. 249–289, out. 2021.

HOLM, Søren; STANTON, Catherine; BARTLETT, Benjamin. A New Argument for No-Fault Compensation in Health Care: The Introduction of Artificial Intelligence Systems. **Health Care Analysis**, v. 29, n. 3, p. 171–188, set. 2021. https://doi.org/10.1007/s10728-021-00430-4.

HOLZINGER, Andreas; BIEMANN, Chris; PATTICHIS, Constantinos S.; KELL, Douglas B. What do we need to build explainable AI systems for the medical domain? 28 dez. 2017. Disponível em: http://arxiv.org/abs/1712.09923.

HOOFNAGLE, Chris Jay; VAN DER SLOOT, Bart; BORGESIUS, Frederik Zuiderveen. The European Union general data protection regulation: what it is and what it means. **Information & Communications Technology Law**, v. 28, n. 1, p. 65–98, fev. 2019. https://doi.org/10.1080/13600834.2019.1573501.

HOSNY, Ahmed; AERTS, Hugo J. W. L. Artificial intelligence for global health. **Science**, v. 366, n. 6468, p. 955–956, 22 nov. 2019. https://doi.org/10.1126/science.aay5189.

HOVY, Eduard H. Automated discourse generation using discourse structure relations. **Artificial Intelligence**, v. 63, n. 1–2, p. 341–385, out. 1993. https://doi.org/10.1016/0004-3702(93)90021-3.

HUMAN RIGHTS COUNCIL. **Question of the realization of economic, social and cultural rights in all countries: the role of new technologies for the realization of economic, social and cultural rights**. Geneva: Office of the High Commissioner for Human Rights, 2020. Disponível em: https://www.ohchr.org/sites/default/files/HRBodies/HRC/RegularSessions/Session43/Documents/A_HRC_43_29.pdf.

HUNT, Paul; BACKMAN, Gunilla. Health systems and the right to the highest attainable standard of health. **Health and human rights**, United States, v. 10, n. 1, p. 81–92, 2008.

IE UNIVERSITY. Data is more than the new oil — it's oxygen. 5 abr. 2022. ie edu. Disponível em: https://www.ie.edu/blue-talks/data-is-more-than-the-new-oil-its-oxygen/. Acesso em: 19 set. 2023.

IEC. **IEC 62304:2006**. International Electrotechnical Commission – IEC, maio 2006. Disponível em: https://www.iso.org/standard/38421.html.

IMDRF. International Medical Device Regulators Forum (IMDRF). 14 set. 2023. **International Medical Device Regulators Forum**. Disponível em: https://www.imdrf.org. Acesso em: 7 out. 2023.

IMDRF, Artificial Intelligence Medical Devices Working Group. **Machine Learning-enabled Medical Devices: Key Terms and Definitions**. International Medical Device Regulators Forum – IMDRF, 9 maio 2022. Disponível em: https://www.imdrf.org/documents/machine-learning-enabled-medical-devices-key-terms-and-definitions.

IMDRF, SaMD Working Group. **Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations**. International Medical Device Regulators Forum – IMDRF, 18 set. 2014. Disponível em: https://www.imdrf.org/documents/software-medical-device-possible-framework-risk-categorization-and-corresponding-considerations.

IMDRF, SaMD Working Group. **Software as a Medical Device (SaMD): Application of Quality Management System**. International Medical Device Regulators Forum – IMDRF, 2 out. 2015. Disponível em: https://www.imdrf.org/documents/software-medical-device-samd-application-quality-management-system.

IMDRF, SaMD Working Group. **Software as a Medical Device (SaMD): Clinical Evaluation**. International Medical Device Regulators Forum – IMDRF, 21 set. 2017. Disponível em: https://www.imdrf.org/documents/software-medical-device-samd-clinical-evaluation.

IMDRF, SaMD Working Group. **Software as a Medical Device (SaMD): Key Definitions**. International Medical Device Regulators Forum – IMDRF, 9 dez. 2013. Disponível em: https://www.imdrf.org/documents/software-medical-device-samd-key-definitions.

INTERNATIONAL BIOETHICS COMMITTEE. **Report of the IBC on big data and health**. Paris: United Nations Educational, Cultural and Scientific Organization, 2017(SHS/YES/IBC-24/17/3 REV.2).
Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000248724.

IQBAL, Jeffrey David; BILLER-ANDORNO, Nikola. The regulatory gap in digital health and alternative pathways to bridge it. **Health policy and technology**, v. 11, n. 3, p. 100663–100663, 1 set. 2022. https://doi.org/10.1016/j.hlpt.2022.100663.

ISO. **ISO 13485:2016**. International Organization for Standardization – ISO, mar. 2016. Disponível em: https://www.iso.org/standard/59752.html.

ISO. **ISO 14971:2019**. International Organization for Standardization – ISO, dez. 2019. Disponível em: https://www.iso.org/standard/72704.html.

ISO/IEC. **ISO/IEC 22989:2022**. International Organization for Standardization – ISO / International Electrotechnical Commission – IEC, jul. 2022. Disponível em: https://www.iso.org/standard/74296.html.

JANSSEN, Marijn; BROUS, Paul; ESTEVEZ, Elsa; BARBOSA, Luis S.; JANOWSKI, Tomasz. Data governance: Organizing data for trustworthy Artificial Intelligence. **Government Information Quarterly**, v. 37, n. 3, p. 101493, 1 jul. 2020. https://doi.org/10.1016/j.giq.2020.101493.

JIA, Zhenge; CHEN, Jianxu; XU, Xiaowei; KHEIR, John; HU, Jingtong; XIAO, Han; PENG, Sui; HU, Xiaobo Sharon; CHEN, Danny; SHI, Yiyu. The importance of resource awareness in artificial intelligence for healthcare. **Nature Machine Intelligence**, v. 5, n. 7, p. 687–698, jul. 2023. https://doi.org/10.1038/s42256-023-00670-0.

JIANG, Fei; JIANG, Yong; ZHI, Hui; DONG, Yi; LI, Hao; MA, Sufeng; WANG, Yilong; DONG, Qiang; SHEN, Haipeng; WANG, Yongjun. Artificial intelligence in healthcare: past, present and future. **Stroke and Vascular Neurology**, v. 2, n. 4, dez. 2017. https://doi.org/10.1136/svn-2017-000101.

JILLSON, Elisa. Aiming for truth, fairness, and equity in your company's use of AI. 19 abr. 2021. **Federal Trade Commission**. Disponível em: https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai. Acesso em: 1 out. 2023.

JOBIN, Anna; IENCA, Marcello; VAYENA, Effy. The global landscape of AI ethics guidelines. **Nature Machine Intelligence**, v. 1, n. 9, p. 389–399, set. 2019. https://doi.org/10.1038/s42256-019-0088-2.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255–260, 17 jul. 2015. https://doi.org/10.1126/science.aaa8415.

JUMPER, John; EVANS, Richard; PRITZEL, Alexander; GREEN, Tim; FIGURNOV, Michael; RONNEBERGER, Olaf; TUNYASUVUNAKOOL, Kathryn; BATES, Russ; ŽÍDEK, Augustin; POTAPENKO, Anna; BRIDGLAND, Alex; MEYER, Clemens; KOHL, Simon A. A.; BALLARD, Andrew J.; COWIE, Andrew; ROMERA-PAREDES, Bernardino; NIKOLOV, Stanislav; JAIN, Rishub; ADLER, Jonas; … HASSABIS, Demis. Highly accurate protein structure prediction with AlphaFold. **Nature**, v. 596, n. 7873, p. 583–589, ago. 2021. https://doi.org/10.1038/s41586-021-03819-2.

KAMINSKI, Margot E. The Right to Explanation, Explained. **Berkeley Technology Law Journal**, v. 34, n. 1, p. 189–218, maio 2019. https://doi.org/10.15779/Z38TD9N83H.

KEANE, Pearse A.; TOPOL, Eric J. With an eye to AI and autonomous diagnosis. **npj Digital Medicine**, v. 1, n. 1, p. 1–3, 28 ago. 2018. https://doi.org/10.1038/s41746-018-0048-y.

KELLY, Christopher J.; KARTHIKESALINGAM, Alan; SULEYMAN, Mustafa; CORRADO, Greg; KING, Dominic. Key challenges for delivering clinical impact with artificial intelligence. **BMC Medicine**, v. 17, n. 1, p. 195, dez. 2019. https://doi.org/10.1186/s12916-019-1426-2.

KHORSANDI, Shirin Elizabeth; HARDGRAVE, Hailey J.; OSBORN, Tamara; KLUTTS, Garrett; NIGH, Joe; SPENCER-COLE, Richard T.; KAKOS, Christos D.; ANASTASIOU, Ioannis; MAVROS, Michail N.; GIORGAKIS, Emmanouil. Artificial Intelligence in Liver Transplantation. **Transplantation Proceedings**, v. 53, n. 10, p. 2939–2944, dez. 2021. https://doi.org/10.1016/j.transproceed.2021.09.045.

KILKENNY, Monique F; ROBINSON, Kerin M. Data quality: "Garbage in – garbage out". **Health Information Management Journal**, v. 47, n. 3, p. 103–105, 1 set. 2018. https://doi.org/10.1177/1833358318774357.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. ImageNet classification with deep convolutional neural networks. 3 dez. 2012. **Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1**. Red Hook, NY, USA: Curran Associates Inc., 3 dez. 2012. p. 1097–1105. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

KUNER, Christopher; BYGRAVE, Lee A.; DOCKSEY, Christopher. Background and Evolution of the EU General Data Protection Regulation (GDPR). *In*: KUNER, Christopher; BYGRAVE, Lee A.; DOCKSEY, Christopher. **The EU General Data Protection Regulation (GDPR)**. Oxford, United Kingdom: Oxford University Press, 2020. p. 1–47. Disponível em: https://doi.org/10.1093/oso/9780198826491.003.0001.

KURZWEIL, Ray. **The singularity is near: when humans transcend biology**. New York: Viking, 2005.

KUSHIDA, Clete A.; NICHOLS, Deborah A.; JADRNICEK, Rik; MILLER, Ric; WALSH, James K.; GRIFFIN, Kara. Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies. **Medical Care**, v. 50, p. S82, jul. 2012. https://doi.org/10.1097/MLR.0b013e3182585355.

KWADE, Zuzanna. Clinical evaluation of software. *In*: COBBAERT, Koen; BOS, Gert (orgs.). **Software as a medical device: regulatory and market access implications**. Rockville, Maryland: Regulatory Affairs Professionals Society, 2021. p. 63–74.

LANCET, The. AI in medicine: creating a safe and equitable future. **The Lancet**, v. 402, n. 10401, p. 503, 12 ago. 2023. https://doi.org/10.1016/S0140-6736(23)01668-9.

LAUPACIS, Andreas; SACKETT, David L.; ROBERTS, Robin S. An Assessment of Clinically Useful Measures of the Consequences of Treatment. **New England Journal of Medicine**, v. 318, n. 26, p. 1728–1733, 30 jun. 1988. https://doi.org/10.1056/NEJM198806303182605.

LECKENBY, Emily; DAWOUD, Dalia; BOUVY, Jacoline; JÓNSSON, Páll. The Sandbox Approach and its Potential for Use in Health Technology Assessment: A Literature Review. **Applied Health Economics and Health Policy**, v. 19, n. 6, p. 857–869, nov. 2021. https://doi.org/10.1007/s40258-021-00665-1.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, maio 2015. https://doi.org/10.1038/nature14539.

LEE, Cecilia S.; LEE, Aaron Y. Clinical applications of continual learning machine learning. **The Lancet Digital Health**, v. 2, n. 6, e279–e281, jun. 2020. https://doi.org/10.1016/S2589-7500(20)30102-3.

LEE, Peter; BUBECK, Sebastien; PETRO, Joseph. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. **New England Journal of Medicine**, v. 388, n. 13, p. 1233–1239, 30 mar. 2023. https://doi.org/10.1056/NEJMsr2214184.

LI, Tian; SAHU, Anit Kumar; TALWALKAR, Ameet; SMITH, Virginia. Federated Learning: Challenges, Methods, and Future Directions. **IEEE Signal Processing Magazine**, v. 37, n. 3, p. 50–60, maio 2020. https://doi.org/10.1109/MSP.2020.2975749.

LIEVEVROUW, Elisa; MARELLI, Luca; VAN HOYWEGHEN, Ine. The FDA's standard-making process for medical digital health technologies: co-producing technological and organizational innovation. **BioSocieties**, v. 17, n. 3, p. 549–576, set. 2022. https://doi.org/10.1057/s41292-021-00232-w.

LIPTON, Zachary C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. **Queue**, v. 16, n. 3, p. 31–57, jun. 2018. https://doi.org/10.1145/3236386.3241340.

LITJENS, Geert; KOOI, Thijs; BEJNORDI, Babak Ehteshami; SETIO, Arnaud Arindra Adiyoso; CIOMPI, Francesco; GHAFOORIAN, Mohsen; VAN DER LAAK, Jeroen A. W. M.; VAN GINNEKEN, Bram; SÁNCHEZ, Clara I. A survey on deep learning in medical image analysis. **Medical Image Analysis**, v. 42, p. 60–88, dez. 2017. https://doi.org/10.1016/j.media.2017.07.005.

LIU, Hongfang; ESTIRI, Hossein; WIENS, Jenna; GOLDENBERG, Anna; SARIA, Suchi; SHAH, Nigam. Artificial Intelligence Model Development and Validation. *In*: MATHENY, Michael E.; ISRANI, Sonoo Thadaney; AHMED, Mahnoor; WHICHER, Danielle (orgs.). **Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril**. Washington, DC: National Academy of Medicine, 2022. p. 131–158.

LIU, Xiaoxuan; GLOCKER, Ben; MCCRADDEN, Melissa M; GHASSEMI, Marzyeh; DENNISTON, Alastair K; OAKDEN-RAYNER, Lauren. The medical algorithmic audit. **The Lancet Digital Health**, v. 4, n. 5, e384–e397, maio 2022. https://doi.org/10.1016/S2589-7500(22)00003-6.

LONDON, Alex John. Artificial Intelligence and Black-Box Medical Decisions: *Accuracy versus Explainability*. **Hastings Center Report**, v. 49, n. 1, p. 15–21, jan. 2019. https://doi.org/10.1002/hast.973.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. 4 dez. 2017. **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 4 dez. 2017. p. 4768–4777. Disponível em: https://dl.acm.org/doi/10.5555/3295222.3295230.

MADIEGA, Tambiama. **Artificial intelligence act. Briefing: EU Legislation in Progress**. European Parliamentary Research Service, jun. 2023. Disponível em: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf.

MADIEGA, Tambiama; VAN DE POL, Anne Louise. **Artificial intelligence act and regulatory sandboxes**. European Parliamentary Research Service, jun. 2022. Disponível em: https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf.

MANN, Jonathan M.; GOSTIN, Lawrence; GRUSKIN, Sofia; BRENNAN, Troyen; LAZZARINI, Zita; FINEBERG, Harvey V. Health and human rights. **Health and human rights**, United States, v. 1, n. 1, p. 6–23, Fall 1994.

MARCUS, Gary Fred. **The algebraic mind: integrating connectionism and cognitive science**. Cambridge, Mass London: MIT, 2001
(Learning, development, and conceptual change).

MATHENY, Michael E.; ISRANI, Sonoo Thadaney; AHMED, Mahnoor; WHICHER, Danielle (Orgs.). **Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril**. Washington, DC: National Academy of Medicine, 2022.

MATTHEWS, Victoria; MURPHY, Matt. **White Paper: Addressing Bias in Artificial Intelligence: The Current Regulatory Landscape**. Thomson Reuters, ago. 2023. Disponível em: https://www.thomsonreuters.com/en-us/posts/wp-content/uploads/sites/20/2023/08/Addressing-Bias-in-AI-Report.pdf.

MATTHIAS, Andreas. The responsibility gap: Ascribing responsibility for the actions of learning automata. **Ethics and Information Technology**, v. 6, n. 3, p. 175–183, 2004. https://doi.org/10.1007/s10676-004-3422-1.

MAY, Thomas. Sociogenetic Risks - Ancestry DNA Testing, Third-Party Identity, and Protection of Privacy. **New England Journal of Medicine**, v. 379, n. 5, p. 410–412, 2 ago. 2018. https://doi.org/10.1056/NEJMp1805870.

MCCARTHY, John. Programs with Common Sense. Stanford University, 1959. Disponível em: http://jmc.stanford.edu/articles/mcc59.html.

MCCARTHY, John. What Is Artificial Intelligence? Stanford University, 2007. Disponível em: http://jmc.stanford.edu/articles/whatisai.html.

MCCARTHY, John; MINSKY, Marvin L.; ROCHESTER, Nathaniel; SHANNON, Claude E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. 1955. Disponível em: http://jmc.stanford.edu/articles/dartmouth.html.

MCNAIR, Douglas; PRICE, W. Nicholson. Health Care Artificial Intelligence: Law, Regulation, and Policy. *In*: MATHENY, Michael E.; ISRANI, Sonoo Thadaney; AHMED, Mahnoor; WHICHER, Danielle (orgs.). **Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril**. Washington, DC: National Academy of Medicine, 2022. p. 197–234.

MDCG. **Guidance on Clinical Evaluation (MDR) / Performance Evaluation (IVDR) of Medical Device Software**. Medical Device Coordination Group Document, mar. 2020. Disponível em: https://ec.europa.eu/docsroom/documents/40323.

MELLO, Michelle M.; GUHA, Neel. ChatGPT and Physicians' Malpractice Risk. **JAMA Health Forum**, v. 4, n. 5, e231938, 18 maio 2023. https://doi.org/10.1001/jamahealthforum.2023.1938.

MENDOZA, Isak; BYGRAVE, Lee A. The Right Not to be Subject to Automated Decisions Based on Profiling. *In*: SYNODINOU, Tatiana-Eleni; JOUGLEUX, Philippe; MARKOU, Christiana; PRASTITOU, Thalia (orgs.). **EU Internet Law**. Cham: Springer International Publishing, 2017. p. 77–98. Disponível em: https://doi.org/10.1007/978-3-319-64955-9_4.

MILLER, Randolph A. Medical Diagnostic Decision Support Systems — Past, Present, And Future: A Threaded Bibliography and Brief Commentary. **Journal of the American Medical Informatics Association**, v. 1, n. 1, p. 8–27, 1994. https://doi.org/10.1136/jamia.1994.95236141.

MILLER, Tim. Explanation in artificial intelligence: Insights from the social sciences. **Artificial Intelligence**, v. 267, p. 1–38, fev. 2019. https://doi.org/10.1016/j.artint.2018.07.007.

MINSKY, Marvin L. Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy. **AI Magazine**, v. 12, n. 2, p. 34, 15 jun. 1991. https://doi.org/10.1609/aimag.v12i2.894.

MINSSEN, Timo; GERKE, Sara; ABOY, Mateo; PRICE, Nicholson; COHEN, Glenn. Regulatory responses to medical machine learning. **Journal of Law and the Biosciences**, v. 7, n. 1, lsaa002, 25 jul. 2020. https://doi.org/10.1093/jlb/lsaa002.

MINSSEN, Timo; MIMLER, Marc; MAK, Vivian. When Does Stand-Alone Software Qualify as a Medical Device in the European Union?—The Cjeu's Decision in Snitem and What it Implies for the Next Generation of Medical Devices. **Medical Law Review**, v. 28, n. 3, p. 615–624, ago. 2020. https://doi.org/10.1093/medlaw/fwaa012.

MITCHELL, Tom M. **Machine Learning**. New York: McGraw-Hill, 1997 (McGraw-Hill series in computer science).

MITTELSTADT, Brent. Interpretability and Transparency in Artificial Intelligence. *In*: VÉLIZ, Carissa (org.). **The Oxford Handbook of Digital Ethics**. 1. ed. [*S. l.*]: Oxford University Press, 2022. DOI 10.1093/oxfordhb/9780198857815.013.20. Disponível em: https://academic.oup.com/edited-volume/37078/chapter/378567795.

MITTELSTADT, Brent; ALLO, Patrick; TADDEO, Mariarosaria; WACHTER, Sandra; FLORIDI, Luciano. The ethics of algorithms: Mapping the debate. **Big Data & Society**, v. 3, n. 2, p. 1–21, dez. 2016. https://doi.org/10.1177/2053951716679679.

MOLNAR, Christoph. **Interpretable Machine Learning: A Guide for Making Black Box Models Explainable**. 2. ed. 2022.
Disponível em: https://christophm.github.io/interpretable-ml-book/.

MONTEIRO, Renato; MACHADO, Caio V.; SILVA, Leoncio. The Right to Explanation in Brazilian Data Protection Law. **International Journal of Digital and Data Law**, v. 7, n. 1, p. 119–136, 5 ago. 2021.
https://ojs.imodev.org/?journal=RIDDN&page=article&op=view&path[]=406.

MOOR, Michael; BANERJEE, Oishi; ABAD, Zahra Shakeri Hossein; KRUMHOLZ, Harlan M.; LESKOVEC, Jure; TOPOL, Eric J.; RAJPURKAR, Pranav. Foundation models for generalist medical artificial intelligence. **Nature**, v. 616, n. 7956, p. 259–265, abr. 2023. https://doi.org/10.1038/s41586-023-05881-4.

MORLEY, Jessica; MACHADO, Caio C. V.; BURR, Christopher; COWLS, Josh; JOSHI, Indra; TADDEO, Mariarosaria; FLORIDI, Luciano. The ethics of AI in health care: A mapping review. **Social Science & Medicine**, v. 260, p. 113172, 1 set. 2020. https://doi.org/10.1016/j.socscimed.2020.113172.

MURDOCH, Blake. Privacy and artificial intelligence: challenges for protecting health information in a new era. **BMC Medical Ethics**, v. 22, n. 1, p. 1–5, 15 set. 2021. https://doi.org/10.1186/s12910-021-00687-3.

MURPHY, Kevin P. **Machine learning: a probabilistic perspective**. Cambridge, MA: MIT Press, 2012 (Adaptive computation and machine learning series).

NATURE, Editorial. AI's potential to accelerate drug discovery needs a reality check. **Nature**, v. 622, n. 7982, p. 217–217, 12 out. 2023a. https://doi.org/10.1038/d41586-023-03172-6.

NATURE, Editorial. ChatGPT is a black box: how AI research can break it open. **Nature**, v. 619, n. 7971, p. 671–672, 25 jul. 2023b. https://doi.org/10.1038/d41586-023-02366-2.

NDPC. Nigeria Data Protection Commission Resources. 2023. Disponível em: https://ndpc.gov.ng/Files/Nigeria_Data_Protection_Act_2023.pdf. Acesso em: 1 out. 2023.

NEWELL, Allen; SIMON, Herbert A. Computer science as empirical inquiry: symbols and search. **Communications of the ACM**, v. 19, n. 3, p. 113–126, mar. 1976. https://doi.org/10.1145/360018.360022.

NIEMIEC, Emilia. Will the EU Medical Device Regulation help to improve the safety and performance of medical AI devices? **DIGITAL HEALTH**, v. 8, p. 205520762210890, jan. 2022. https://doi.org/10.1177/20552076221089079.

NOVELLI, Claudio; TADDEO, Mariarosaria; FLORIDI, Luciano. Accountability in artificial intelligence: what it is and how it works. **AI & SOCIETY**, 7 fev. 2023. Disponível em: https://doi.org/10.1007/s00146-023-01635-y.

NTOUTSI, Eirini; FAFALIOS, Pavlos; GADIRAJU, Ujwal; IOSIFIDIS, Vasileios; NEJDL, Wolfgang; VIDAL, Maria-Esther; RUGGIERI, Salvatore; TURINI, Franco; PAPADOPOULOS, Symeon; KRASANAKIS, Emmanouil; KOMPATSIARIS, Ioannis; KINDER-KURLANDA, Katharina; WAGNER, Claudia; KARIMI, Fariba; FERNANDEZ, Miriam; ALANI, Harith; BERENDT, Bettina; KRUEGEL, Tina; HEINZE, Christian; … STAAB, Steffen. Bias in data-driven artificial intelligence systems—An introductory survey. **WIREs Data Mining and Knowledge Discovery**, v. 10, n. 3, e1356, 2020. https://doi.org/10.1002/widm.1356.

OBERMEYER, Ziad; EMANUEL, Ezekiel J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. **New England Journal of Medicine**, v. 375, n. 13, p. 1216–1219, set. 2016. https://doi.org/10.1056/NEJMp1606181.

OBERMEYER, Ziad; POWERS, Brian; VOGELI, Christine; MULLAINATHAN, Sendhil. Dissecting racial bias in an algorithm used to manage the health of populations. **Science**, v. 366, n. 6464, p. 447–453, 25 out. 2019. https://doi.org/10.1126/science.aax2342.

ODAIBO, Stephen G. Risk Management of AI/ML Software as a Medical Device (SaMD): On ISO 14971 and Related Standards and Guidances. 2021. Disponível em: https://arxiv.org/abs/2109.07905.

OECD. **Artificial Intelligence in Society**. Paris: Organisation for Economic Co-operation and Development, 2019. Disponível em: https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en.

OFFICE FOR CIVIL RIGHTS. Health Information Privacy. 9 jun. 2021. HHS.gov. Disponível em: https://www.hhs.gov/hipaa/index.html. Acesso em: 26 set. 2023.

OHCHR; WHO. **The Right to Health**. Geneva: UN, Office of the High Commissioner for Human Rights : World Health Organization, 2008 (Human rights fact sheet, 31). Disponível em: https://www.ohchr.org/en/publications/fact-sheets/fact-sheet-no-31-right-health.

O'NEIL, Cathy. **Weapons of math destruction: how big data increases inequality and threatens democracy**. New York: Crown, 2016.

OWENS, John; CRIBB, Alan. 'My Fitbit Thinks I Can Do Better!' Do Health Promoting Wearable Technologies Support Personal Autonomy? **Philosophy & Technology**, v. 32, n. 1, p. 23–38, mar. 2019. https://doi.org/10.1007/s13347-017-0266-2.

PARIKH, Ravi B.; OBERMEYER, Ziad; NAVATHE, Amol S. Regulation of predictive analytics in medicine. **Science**, v. 363, n. 6429, p. 810–812, 22 fev. 2019. https://doi.org/10.1126/science.aaw0029.

PARISI, German I.; KEMKER, Ronald; PART, Jose L.; KANAN, Christopher; WERMTER, Stefan. Continual lifelong learning with neural networks: A review. **Neural Networks**, v. 113, p. 54–71, maio 2019. https://doi.org/10.1016/j.neunet.2019.01.012.

PFOHL, Stephen R.; XU, Yizhe; FORYCIARZ, Agata; IGNATIADIS, Nikolaos; GENKINS, Julian; SHAH, Nigam H. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. 3 fev. 2022. Disponível em: http://arxiv.org/abs/2202.01906.

PHANSALKAR, Shobha; VAN DER SIJS, Heleen; TUCKER, Alisha D; DESAI, Amrita A; BELL, Douglas S; TEICH, Jonathan M; MIDDLETON, Blackford; BATES, David W. Drug—drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. **Journal of the American Medical Informatics Association**, v. 20, n. 3, p. 489–493, maio 2013. https://doi.org/10.1136/amiajnl-2012-001089.

PLOTKIN, David. **Data stewardship: an actionable guide to effective data management and data governance**. Second edition. London, United Kingdom ; San Diego, CA: Academic Press, 2021.

PRICE, W. Nicholson. Medical Malpractice and Black-Box Medicine. *In*: COHEN, I. Glenn; LYNCH, Holly Fernandez; VAYENA, Effy; GASSER, Urs (orgs.). **Big Data, Health Law, and Bioethics**. Cambridge University Press, 2018. p. 295–306. DOI 10.1017/9781108147972.027. Disponível em: https://www.cambridge.org/core/product/identifier/9781108147972%23CN-bp-20/type/book_part.

PRICE, W. Nicholson; COHEN, I. Glenn. Privacy in the age of medical big data. **Nature Medicine**, v. 25, n. 1, p. 37–43, jan. 2019. https://doi.org/10.1038/s41591-018-0272-7.

PRICE, W. Nicholson; GERKE, Sara; COHEN, I. Glenn. Liability for Use of Artificial Intelligence in Medicine. Rochester, NY, 20 maio 2022. DOI 10.2139/ssrn.4115538. Disponível em: https://papers.ssrn.com/abstract=4115538.

PRICE, W. Nicholson, II; GERKE, Sara; COHEN, I. Glenn. Potential Liability for Physicians Using Artificial Intelligence. **JAMA**, v. 322, n. 18, p. 1765–1766, 12 nov. 2019. https://doi.org/10.1001/jama.2019.15064.

PRIVACY INTERNATIONAL. **The Keys to Data Protection: A Guide for Policy Engagement on Data Protection**. London, ago. 2018. Disponível em: https://privacyinternational.org/report/2255/data-protection-guide-complete-guide-full.

PRS LEGISLATIVE RESEARCH. The Digital Personal Data Protection Bill. 2023. **PRS Legislative Research**. Disponível em: https://prsindia.org/billtrack/digital-personal-data-protection-bill-2023. Acesso em: 1 out. 2023.

RAJKOMAR, Alvin; DEAN, Jeffrey; KOHANE, Isaac. Machine Learning in Medicine. **New England Journal of Medicine**, v. 380, n. 14, p. 1347–1358, 4 abr. 2019. https://doi.org/10.1056/NEJMra1814259.

RAJKOMAR, Alvin; HARDT, Michaela; HOWELL, Michael D.; CORRADO, Greg; CHIN, Marshall H. Ensuring Fairness in Machine Learning to Advance Health Equity. **Annals of internal medicine**, v. 169, n. 12, p. 866–872, 18 dez. 2018. https://doi.org/10.7326/M18-1990.

RAJKOMAR, Alvin; OREN, Eyal; CHEN, Kai; DAI, Andrew M.; HAJAJ, Nissan; HARDT, Michaela; LIU, Peter J.; LIU, Xiaobing; MARCUS, Jake; SUN, Mimi; SUNDBERG, Patrik; YEE, Hector; ZHANG, Kun; ZHANG, Yi; FLORES, Gerardo; DUGGAN, Gavin E.; IRVINE, Jamie; LE, Quoc; LITSCH, Kurt; … DEAN, Jeffrey. Scalable and accurate deep learning with electronic health records. **npj Digital Medicine**, v. 1, n. 1, p. 18, 8 maio 2018. https://doi.org/10.1038/s41746-018-0029-1.

RAJPURKAR, Pranav; LUNGREN, Matthew P. The Current and Future State of AI Interpretation of Medical Images. **New England Journal of Medicine**, v. 388, n. 21, p. 1981–1990, 25 maio 2023. https://doi.org/10.1056/NEJMra2301725.

RASCHKA, Sebastian. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. 10 nov. 2020. Disponível em: http://arxiv.org/abs/1811.12808.

RAWLS, John. **A theory of justice**. Rev. ed. Cambridge, Mass: Belknap Press of Harvard University Press, 1999.

RENDA, Andrea; ENGLER, Alex. **What's in a name? Getting the definition of Artificial Intelligence right in the EU's AI Act**. CEPS – Centre for European Policy Studies, fev. 2023. Disponível em: https://www.ceps.eu/ceps-publications/whats-in-a-name/.

REPUBLIC OF INDONESIA, Constitutional Court. Govt: Law on Personal Data Protection Provides Legal Protection. 13 fev. 2023. en.mkri.id. Disponível em: https://en.mkri.id/news/details/2023-02-13/Govt:_Law_on_Personal_Data_Protection_Provides_Legal_Protection. Acesso em: 30 set. 2023.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 13 ago. 2016. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 13 ago. 2016. p. 1135–1144. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939778.

RICHMAN, Barak. Health Regulation for the Digital Age — Correcting the Mismatch. **New England Journal of Medicine**, v. 379, n. 18, p. 1694–1695, nov. 2018. https://doi.org/10.1056/NEJMp1806848.

RIEKE, Nicola; HANCOX, Jonny; LI, Wenqi; MILLETARÌ, Fausto; ROTH, Holger R.; ALBARQOUNI, Shadi; BAKAS, Spyridon; GALTIER, Mathieu N.; LANDMAN, Bennett A.; MAIER-HEIN, Klaus; OURSELIN, Sébastien; SHELLER, Micah; SUMMERS, Ronald M.; TRASK, Andrew; XU, Daguang; BAUST, Maximilian; CARDOSO, M. Jorge. The future of digital health with federated learning. **npj Digital Medicine**, v. 3, n. 1, p. 1–7, 14 set. 2020. https://doi.org/10.1038/s41746-020-00323-1.

ROBERTS, Marc J.; HSIAO, William; BERMAN, Peter; REICH, Michael R. Regulation. *In*: ROBERTS, Marc; HSIAO, William; BERMAN, Peter; REICH, Michael (orgs.). **Getting Health Reform Right: A Guide to Improving Performance and Equity**. New York: Oxford University Press, 2008. Disponível em: https://doi.org/10.1093/acprof:oso/9780195371505.003.0011.

ROCHER, Luc; HENDRICKX, Julien M.; DE MONTJOYE, Yves-Alexandre. Estimating the success of re-identifications in incomplete datasets using generative models. **Nature Communications**, v. 10, n. 1, p. 3069, 23 jul. 2019. https://doi.org/10.1038/s41467-019-10933-3.

ROSCHER, Ribana; BOHN, Bastian; DUARTE, Marco F.; GARCKE, Jochen. Explainable Machine Learning for Scientific Insights and Discoveries. **IEEE Access**, v. 8, p. 42200–42216, 2020. https://doi.org/10.1109/ACCESS.2020.2976199.

ROSKI, Joachim; CHAPMAN, Wendy; HEFFNER, Jaimee; TRIVEDI, Ranak; DEL FIOL, Guilherme; KUKAFKA, Rita; BLEICHER, Paul; ESTIRI, Hossein; KLANN, Jeffrey; PIERCE, Joni. How Artificial Intelligence Is Changing Health and Health Care. *In*: MATHENY, Michael E.; ISRANI, Sonoo Thadaney; AHMED, Mahnoor; WHICHER, Danielle (orgs.). **Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril**. Washington, DC: National Academy of Medicine, 2022. p. 65–98.

RUDIN, Cynthia. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence**, v. 1, n. 5, p. 206–215, maio 2019. https://doi.org/10.1038/s42256-019-0048-x.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533–536, out. 1986. https://doi.org/10.1038/323533a0.

RUMELHART, David E.; MCCLELLAND, James L. **Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations**. Cambridge, MA: The MIT Press, 1986.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial intelligence: a modern approach**. 3rd ed. Upper Saddle River: Prentice Hall, 2010 (Prentice Hall series in artificial intelligence).

SAHNI, Nikhil; CARRUS, Brandon. Artificial Intelligence in U.S. Health Care Delivery. **New England Journal of Medicine**, v. 389, n. 4, p. 348–358, 27 jul. 2023. https://doi.org/10.1056/NEJMra2204673.

SAHNI, Nikhil; STEIN, George; ZEMMEL, Rodney; CUTLER, David M. The Potential Impact of Artificial Intelligence on Healthcare Spending. Working Paper Series. jan. 2023. DOI 10.3386/w30857. Disponível em: https://www.nber.org/papers/w30857.

SAITO, Takaya; REHMSMEIER, Marc. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. **PLOS ONE**, v. 10, n. 3, e0118432, 4 mar. 2015. https://doi.org/10.1371/journal.pone.0118432.

SANTORO, Adam; LAMPINEN, Andrew; MATHEWSON, Kory; LILLICRAP, Timothy; RAPOSO, David. Symbolic Behaviour in Artificial Intelligence. 2021. Disponível em: https://arxiv.org/abs/2102.03406.

SCHÖNBERGER, Daniel. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. **International Journal of Law and Information Technology**, v. 27, n. 2, p. 171–203, 1 jun. 2019. https://doi.org/10.1093/ijlit/eaz004.

SCHWALBE, Nina; WAHL, Brian. Artificial intelligence and the future of global health. **The Lancet**, v. 395, n. 10236, p. 1579–1586, 16 maio 2020. https://doi.org/10.1016/S0140-6736(20)30226-9.

SEARLE, John R. Minds, brains, and programs. **Behavioral and Brain Sciences**, v. 3, n. 3, p. 417–424, set. 1980. https://doi.org/10.1017/S0140525X00005756.

SELBST, Andrew D; POWLES, Julia. Meaningful information and the right to explanation. **International Data Privacy Law**, v. 7, n. 4, p. 233–242, nov. 2017. https://doi.org/10.1093/idpl/ipx022.

SENDAK, Mark P.; D'ARCY, Joshua; KASHYAP, Sehj; GAO, Michael; NICHOLS, Marshall; COREY, Kristin; RATLIFF, William; BALU, Suresh. A Path for Translation of Machine Learning Products into Healthcare Delivery. **EMJ Innovations**, jan. 2020. Disponível em: https://doi.org/10.33590/emjinnov/19-00172.

SHAH, Nigam H.; MILSTEIN, Arnold; BAGLEY, PhD, Steven C. Making Machine Learning Models Clinically Useful. **JAMA**, v. 322, n. 14, p. 1351–1352, 8 out. 2019. https://doi.org/10.1001/jama.2019.10306.

SHAH, Pratik; KENDALL, Francis; KHOZIN, Sean; GOOSEN, Ryan; HU, Jianying; LARAMIE, Jason; RINGEL, Michael; SCHORK, Nicholas. Artificial intelligence and machine learning in clinical development: a translational perspective. **npj Digital Medicine**, v. 2, n. 1, p. 69, 26 jul. 2019. https://doi.org/10.1038/s41746-019-0148-3.

SHERKOW, Jacob S. Regulatory Sandboxes and the Public Health. Rochester, NY, 1 fev. 2022. DOI 10.2139/ssrn.3792217. Disponível em: https://papers.ssrn.com/abstract=3792217.

SMOLENSKY, Paul. On the proper treatment of connectionism. **Behavioral and Brain Sciences**, v. 11, n. 1, p. 1–23, mar. 1988. https://doi.org/10.1017/S0140525X00052432.

SPECTOR-BAGDADY, Kayte; HUTCHINSON, Raymond; O'BRIEN KALEBA, Erin; KHETERPAL, Sachin. Sharing Health Data and Biospecimens with Industry — A Principle-Driven, Practical Approach. **New England Journal of Medicine**, v. 382, n. 22, p. 2072–2075, 28 maio 2020. https://doi.org/10.1056/NEJMp1915298.

SULLIVAN, Hannah R.; SCHWEIKART, Scott J. Are Current Tort Liability Doctrines Adequate for Addressing Injury Caused by AI? **AMA Journal of Ethics**, v. 21, n. 2, p. 160–166, fev. 2019. https://doi.org/10.1001/amajethics.2019.160.

SURESH, Harini; GUTTAG, John. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. 4 nov. 2021. **Equity and Access in Algorithms, Mechanisms, and Optimization**. New York, NY, USA: Association for Computing Machinery, 4 nov. 2021. p. 1–9.
Disponível em: https://dl.acm.org/doi/10.1145/3465416.3483305.

SUTTON, Reed T.; PINCOCK, David; BAUMGART, Daniel C.; SADOWSKI, Daniel C.; FEDORAK, Richard N.; KROEKER, Karen I. An overview of clinical decision support systems: benefits, risks, and strategies for success. **npj Digital Medicine**, v. 3, n. 1, p. 1–10, 6 fev. 2020. https://doi.org/10.1038/s41746-020-0221-y.

SUTTON, Richard S.; BARTO, Andrew G. **Reinforcement learning: an introduction**. Second edition. Cambridge, Massachusetts: The MIT Press, 2018
(Adaptive computation and machine learning series).

SZOLOVITS, P.; PATIL, R. S.; SCHWARTZ, W. B. Artificial intelligence in medical diagnosis. **Annals of internal medicine**, United States, v. 108, n. 1, p. 80–87, jan. 1988. https://doi.org/10.7326/0003-4819-108-1-80.

THE WHITE HOUSE. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 30 out. 2023. The White House. Disponível em: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. Acesso em: 2 nov. 2023.

THEODOS, Kim; SITTIG, Scott. Health Information Privacy Laws in the Digital Age: HIPAA Doesn't Apply. **Perspectives in Health Information Management**, v. 18, n. Winter, p. 1l, 7 dez. 2020. .

TOPOL, Eric J. As artificial intelligence goes multimodal, medical applications multiply. **Science**, v. 381, n. 6663, p. eadk6139, 15 set. 2023. https://doi.org/10.1126/science.adk6139.

TOPOL, Eric J. **Deep medicine: how artificial intelligence can make healthcare human again**. New York: Basic Books, 2019a.

TOPOL, Eric J. High-performance medicine: the convergence of human and artificial intelligence. **Nature Medicine**, v. 25, n. 1, p. 44–56, jan. 2019b. https://doi.org/10.1038/s41591-018-0300-7.

TOWNEND, David. Privacy. *In*: GANGULI-MITRA, Agomoni; SORBIE, Annie; MCMILLAN, Catriona; DOVE, Edward; POSTAN, Emily; LAURIE, Graeme; SETHI, Nayha (orgs.). **The Cambridge Handbook of Health Research Regulation**. Cambridge Law Handbooks. Cambridge: Cambridge University Press, 2021. p. 73–80. Disponível em: https://www.cambridge.org/core/books/cambridge-handbook-of-health-research-regulation/privacy/E584F9EF39E9597F2F8A0F83ECBE34D6.

TURING, A. M. Computing Machinery and Intelligence. **Mind**, v. 59, n. 236, p. 433–460, out. 1950. https://doi.org/10.1093/mind/LIX.236.433.

UN PRESS. Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight, in First Debate on Artificial Intelligence | UN Press. 18 jul. 2023. United Nations. Disponível em: https://press.un.org/en/2023/sgsm21880.doc.htm. Acesso em: 17 ago. 2023.

UNESCO. **Preliminary study on the technical and legal aspects relating to the desirability of a standard-setting instrument on the ethics of artificial intelligence**. Paris: United Nations Educational, Cultural and Scientific Organization, 2019. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000367422.

UNESCO. **Recommendation on the Ethics of Artificial Intelligence**. Paris: United Nations Educational, Scientific and Cultural Organization, 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137.

UNITED NATIONS. International Covenant on Civil and Political Rights. 1966a. OHCHR. Disponível em: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights. Acesso em: 25 set. 2023.

UNITED NATIONS. International Covenant on Economic, Social and Cultural Rights. 1966b. OHCHR. Disponível em: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights. Acesso em: 22 set. 2023.

UNITED NATIONS. Universal Declaration of Human Rights. 1948. United Nations. Disponível em: https://www.un.org/en/about-us/universal-declaration-of-human-rights. Acesso em: 22 set. 2023.

UNITED NATIONS, Economic and Social Council. **General Comment No. 14: The Right to the Highest Attainable Standard of Health (Art. 12 of the Covenant)**. Geneva: UN Committee on Economic, Social and Cultural Rights, 2000(E/C.12/2000/4). Disponível em: https://digitallibrary.un.org/record/425041.

UNITED NATIONS, Human Rights Committee. **CCPR General Comment No. 16: Article 17 (Right to Privacy), The Right to Respect of Privacy, Family, Home and Correspondence, and Protection of Honour and Reputation**. Geneva: UN Human Rights Committee, 1988. Disponível em: https://www.refworld.org/docid/453883f922.html.

UNIVERSITY OF CAMBRIDGE. “The best or worst thing to happen to humanity” - Stephen Hawking launches Centre for the Future of Intelligence. 19 out. 2016. University of Cambridge. Disponível em: https://www.cam.ac.uk/research/news/the-best-or-worst-thing-to-happen-to-humanity-stephen-hawking-launches-centre-for-the-future-of. Acesso em: 16 ago. 2023.

USACM. **Statement on Algorithmic Transparency and Accountability**. Association for Computing Machinery US Public Policy Council (USACM), 12 jan. 2017. Disponível em: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

VALVERDE-ALBACETE, Francisco J.; PELÁEZ-MORENO, Carmen. 100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox. **PLOS ONE**, v. 9, n. 1, e84217, 10 jan. 2014. https://doi.org/10.1371/journal.pone.0084217.

VAMATHEVAN, Jessica; CLARK, Dominic; CZODROWSKI, Paul; DUNHAM, Ian; FERRAN, Edgardo; LEE, George; LI, Bin; MADABHUSHI, Anant; SHAH, Parantu; SPITZER, Michaela; ZHAO, Shanrong. Applications of machine learning in drug discovery and development. **Nature Reviews Drug Discovery**, v. 18, n. 6, p. 463–477, jun. 2019. https://doi.org/10.1038/s41573-019-0024-5.

VAN DE VEN, Gido M.; TUYTELAARS, Tinne; TOLIAS, Andreas S. Three types of incremental learning. **Nature Machine Intelligence**, v. 4, n. 12, p. 1185–1197, dez. 2022. https://doi.org/10.1038/s42256-022-00568-3.

VAN LAERE, Sven; MUYLLE, Katoo M.; CORNU, Pieter. Clinical Decision Support and New Regulatory Frameworks for Medical Devices: Are We Ready for It? - A Viewpoint Paper. **International Journal of Health Policy and Management**, v. 11, n. 12, p. 3159–3163, 18 out. 2021. https://doi.org/10.34172/ijhpm.2021.144.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Lukasz; POLOSUKHIN, Illia. Attention Is All You Need. 2017. DOI 10.48550/ARXIV.1706.03762. Disponível em: https://arxiv.org/abs/1706.03762.

VAYENA, Effy; BLASIMME, Alessandro. Biomedical Big Data: New Models of Control Over Access, Use and Governance. **Journal of Bioethical Inquiry**, v. 14, n. 4, p. 501–513, dez. 2017. https://doi.org/10.1007/s11673-017-9809-6.

VAYENA, Effy; BLASIMME, Alessandro; COHEN, I. Glenn. Machine learning in medicine: Addressing ethical challenges. **PLOS Medicine**, v. 15, n. 11, e1002689, 6 nov. 2018. https://doi.org/10.1371/journal.pmed.1002689.

VEALE, Michael; BINNS, Reuben. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. **Big Data & Society**, v. 4, n. 2, p. 205395171774353, dez. 2017. https://doi.org/10.1177/2053951717743530.

VICKERS, Andrew J.; ELKIN, Elena B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. **Medical Decision Making**, v. 26, n. 6, p. 565–574, nov. 2006. https://doi.org/10.1177/0272989X06295361.

VLADECK, David. Machines Without Principals: Liability Rules and Artificial Intelligence. **Washington Law Review**, v. 89, n. 1, p. 117, mar. 2014. Disponível em: https://digitalcommons.law.uw.edu/wlr/vol89/iss1/6/

VOKINGER, Kerstin N.; GASSER, Urs. Regulating AI in medicine in the United States and Europe. **Nature machine intelligence**, v. 3, n. 9, p. 738–739, set. 2021. https://doi.org/10.1038/s42256-021-00386-z.

WACHTER, Sandra; MITTELSTADT, Brent; FLORIDI, Luciano. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. **International Data Privacy Law**, v. 7, n. 2, p. 76–99, maio 2017. https://doi.org/10.1093/idpl/ipx005.

WANG, Chao; ZHANG, Jieyu; LASSI, Nicholas; ZHANG, Xiaohan. Privacy Protection in Using Artificial Intelligence for Healthcare: Chinese Regulation in Comparative Perspective. **Healthcare**, v. 10, n. 10, p. 1878, 27 set. 2022. https://doi.org/10.3390/healthcare10101878.

WANG, Yichuan; KUNG, LeeAnn; BYRD, Terry Anthony. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. **Technological Forecasting and Social Change**, v. 126, p. 3–13, 2018. https://doi.org/10.1016/j.techfore.2015.12.019.

WARREN, Samuel D.; BRANDEIS, Louis D. The Right to Privacy. **Harvard Law Review**, v. 4, n. 5, p. 193, 15 dez. 1890. https://doi.org/10.2307/1321160.

WATSON, David S.; KRUTZINNA, Jenny; BRUCE, Ian N.; GRIFFITHS, Christopher EM; MCINNES, Iain B.; BARNES, Michael R.; FLORIDI, Luciano. Clinical applications of machine learning algorithms: beyond the black box. **BMJ**, v. 364, p. l886, 12 mar. 2019. https://doi.org/10.1136/bmj.l886.

WHO. **Delivering Quality-Assured medical Products for All 2019-2023: WHO's five-year plan to help build effective and efficient regulatory systems**. Geneva: World Health Organization, 2019(WHO/MVP/RHT/2019.01). Disponível em: https://apps.who.int/iris/handle/10665/332461.

WHO. **Digital health (A/71/A/CONF./1). Seventy-first World Health Assembly**. Geneva: World Health Organization, 21 maio 2018. Disponível em: https://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_ACONF1-en.pdf.

WHO. **Ethics and governance of artificial intelligence for health: WHO guidance**. Geneva: World Health Organization, 2021. Disponível em: https://apps.who.int/iris/handle/10665/341996.

WHO. **Regulatory considerations on artificial intelligence for health**. Geneva: World Health Organization, 2023. Disponível em: https://iris.who.int/handle/10665/373421.

WIENS, Jenna; SARIA, Suchi; SENDAK, Mark; GHASSEMI, Marzyeh; LIU, Vincent X.; DOSHI-VELEZ, Finale; JUNG, Kenneth; HELLER, Katherine; KALE, David; SAEED, Mohammed; OSSORIO, Pilar N.; THADANEY-ISRANI, Sonoo; GOLDENBERG, Anna. Do no harm: a roadmap for responsible machine learning for health care. **Nature Medicine**, v. 25, n. 9, p. 1337–1340, set. 2019. https://doi.org/10.1038/s41591-019-0548-6.

WOLFF, Robert F.; MOONS, Karel G.M.; RILEY, Richard D.; WHITING, Penny F.; WESTWOOD, Marie; COLLINS, Gary S.; REITSMA, Johannes B.; KLEIJNEN, Jos; MALLETT, Sue. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. **Annals of Internal Medicine**, v. 170, n. 1, p. 51–58, jan. 2019. https://doi.org/10.7326/M18-1376.

WOOD, Charlie. Microsoft president Brad Smith predicts AI will be as transformative to society as the combustion engine over the next 3 decades. 6 nov. 2019. Business Insider. Disponível em: https://www.businessinsider.com/brad-smith-ai-will-play-similar-role-to-combustion-engine-2019-11. Acesso em: 18 set. 2023.

WORLD BANK. **Global Experiences from Regulatory Sandboxes**. World Bank, Washington, DC, 2020. DOI 10.1596/34789. Disponível em: https://openknowledge.worldbank.org/handle/10986/34789.

WYNANTS, Laure; VAN SMEDEN, Maarten; MCLERNON, David J.; TIMMERMAN, Dirk; STEYERBERG, Ewout W.; VAN CALSTER, Ben. Three myths about risk thresholds for prediction models. **BMC Medicine**, v. 17, n. 1, p. 192, 25 out. 2019. https://doi.org/10.1186/s12916-019-1425-3.

YANG, Qiang; LIU, Yang; CHEN, Tianjian; TONG, Yongxin. Federated Machine Learning: Concept and Applications. **ACM Transactions on Intelligent Systems and Technology**, v. 10, n. 2, p. 1-19, 28 jan. 2019. https://doi.org/10.1145/3298981.

YU, Kun-Hsing; BEAM, Andrew L.; KOHANE, Isaac S. Artificial intelligence in healthcare. **Nature Biomedical Engineering**, v. 2, n. 10, p. 719–731, out. 2018. https://doi.org/10.1038/s41551-018-0305-z.

YU, Kun-Hsing; KOHANE, Isaac S. Framing the challenges of artificial intelligence in medicine. **BMJ Quality & Safety**, v. 28, n. 3, p. 238–241, mar. 2019. https://doi.org/10.1136/bmjqs-2018-008551.