**MINISTÉRIO DA SAÚDE**
**INSTÂNCIA NACIONAL DE ÉTICA EM PESQUISA — INAEP**

# CADERNO REFERENCIAL

FUNDAMENTOS ÉTICOS, CIENTÍFICOS, NORMATIVOS E METODOLÓGICOS

## Inteligência Artificial em Pesquisa com Seres Humanos

*Documento de fundamentação do Guia de Uso Ético de Inteligência Artificial em Pesquisa com Seres Humanos e da MARIAH*

Preservação do trabalho técnico do Grupo de Trabalho — Versão 1.0

Brasília – DF

2026

**LISTA DE FIGURAS**



**LISTA DE SIGLAS**

IA — Inteligência artificial

AIA — Avaliação de Impacto Algorítmico (Algorithmic Impact Assessment, Canadá)

ANPD — Autoridade Nacional de Proteção de Dados

Anvisa — Agência Nacional de Vigilância Sanitária

APPI — Lei de Proteção de Informações Pessoais do Japão (Act on the Protection of Personal Information)

CDSCO — Organização Central de Controle de Padrões de Medicamentos da Índia (Central Drugs Standard Control Organization)

CEP — Comitê de Ética em Pesquisa

CFM — Conselho Federal de Medicina

CGI.br — Comitê Gestor da Internet no Brasil

CHS — Ciências Humanas e Sociais

CNPq — Conselho Nacional de Desenvolvimento Científico e Tecnológico

CNS — Conselho Nacional de Saúde

CNN — Rede neural convolucional (Convolutional Neural Network)

Conep — Comissão Nacional de Ética em Pesquisa

DNN — Rede neural profunda (Deep Neural Network)

DPDP — Lei de Proteção de Dados Pessoais Digitais da Índia (Digital Personal Data Protection Act)

EDPB — Comitê Europeu de Proteção de Dados (European Data Protection Board)

EHDS — Espaço Europeu de Dados de Saúde (European Health Data Space)

ELM — Máquina de aprendizado extremo (Extreme Learning Machine)

EMA — Agência Europeia de Medicamentos (European Medicines Agency)

FDA — Agência de Alimentos e Medicamentos dos Estados Unidos (Food and Drug Administration)

FHIR — Recursos Rápidos para Interoperabilidade em Saúde (Fast Healthcare Interoperability Resources, padrão HL7)

GAN — Rede adversária generativa (Generative Adversarial Network)

GDPR — Regulamento Geral sobre a Proteção de Dados da União Europeia (General Data Protection Regulation)

Grad-CAM — Mapeamento de ativação de classes ponderado por gradiente (Gradient-weighted Class Activation Mapping)

HL7 — Health Level Seven (organização internacional de padrões de interoperabilidade em saúde)

ICMJE — Comitê Internacional de Editores de Periódicos Médicos (International Committee of Medical Journal Editors)

ICMR — Conselho Indiano de Pesquisa Médica (Indian Council of Medical Research)

ICO — Escritório do Comissário de Informação do Reino Unido (Information Commissioner's Office)

IDATEN — Sistema de Avaliação Tempestiva de Inovação da PMDA (Improvement Design within Approval for Timely Evaluation and Notification)

INAEP — Instância Nacional de Ética em Pesquisa

LGPD — Lei Geral de Proteção de Dados Pessoais (Lei n.º 13.709, de 14 de agosto de 2018)

LIME — Explicações locais interpretáveis independentes de modelo (Local Interpretable Model-agnostic Explanations)

LLM — Grande modelo de linguagem (Large Language Model)

LNCC — Laboratório Nacional de Computação Científica

MARIAH — Matriz de Avaliação de Risco em Inteligência Artificial em Pesquisa com Seres Humanos

MCTI — Ministério da Ciência, Tecnologia e Inovação

MFDS — Ministério da Segurança Alimentar e de Medicamentos da Coreia do Sul (Ministry of Food and Drug Safety)

MHLW — Ministério da Saúde, Trabalho e Bem-Estar do Japão (Ministry of Health, Labour and Welfare)

MEXT — Ministério da Educação, Cultura, Esportes, Ciência e Tecnologia do Japão (Ministry of Education, Culture, Sports, Science and Technology)

MHRA — Agência Reguladora de Medicamentos e Produtos de Saúde do Reino Unido (Medicines and Healthcare products Regulatory Agency)

MRCT — Centro Multirregional de Ensaios Clínicos (Multi-Regional Clinical Trials Center)

NHC — Comissão Nacional de Saúde da China (National Health Commission)

NHMRC — Conselho Nacional de Saúde e Pesquisa Médica da Austrália (National Health and Medical Research Council)

NIC.br — Núcleo de Informação e Coordenação do Ponto BR

NMPA — Administração Nacional de Produtos Médicos da China (National Medical Products Administration)

OMS — Organização Mundial da Saúde

OPAS — Organização Pan-Americana da Saúde

OWASP — Fundação Open Worldwide Application Security Project

PIPL — Lei de Proteção de Informações Pessoais da China (Personal Information Protection Law)

PLN — Processamento de linguagem natural (Natural Language Processing)

PMDA — Agência de Dispositivos Médicos e Produtos Farmacêuticos do Japão (Pharmaceuticals and Medical Devices Agency)

RDC — Resolução da Diretoria Colegiada (Anvisa)

RIPD — Relatório de Impacto à Proteção de Dados Pessoais

RNDS — Rede Nacional de Dados em Saúde

RNN — Rede neural recorrente (Recurrent Neural Network)

SaMD — Programa de Computador como Dispositivo Médico (Software as a Medical Device)

SEIDIGI — Secretaria de Informação e Saúde Digital do Ministério da Saúde

SHAP — Explicações aditivas de Shapley (SHapley Additive exPlanations)

SINEP — Sistema Nacional de Ética em Pesquisa com Seres Humanos

SISA — Arquitetura de desaprendizado de máquina fragmentada, isolada, fatiada e agregada (Sharded, Isolated, Sliced, Aggregated)

SUS — Sistema Único de Saúde

TCLE — Termo de Consentimento Livre e Esclarecido

TCPS — Enunciado de Política das Três Agências do Canadá (Tri-Council Policy Statement)

UNESCO — Organização das Nações Unidas para a Educação, a Ciência e a Cultura

WCG — Organização norte-americana de serviços regulatórios e de ética em pesquisa clínica

XAI — Inteligência artificial explicável (Explainable Artificial Intelligence)

SUMÁRIO

## NOTA DE ENQUADRAMENTO

Este Caderno não estabelece requisitos adicionais para a apreciação ética dos protocolos. A finalidade é apresentar a fundamentação ética, científica, normativa e metodológica das orientações constantes do Guia INAEP e dos critérios operacionalizados pela MARIAH.

Este Caderno Referencial preserva e organiza a fundamentação produzida pelo Grupo de Trabalho que sustenta o Guia INAEP de Uso Ético de Inteligência Artificial em Pesquisa com Seres Humanos e a MARIAH — Matriz de Avaliação de Risco em Inteligência Artificial em Pesquisa com Seres Humanos.

Sua função é explicativa e referencial. O documento reúne a construção conceitual, ética, normativa, metodológica e comparada que fundamenta as orientações e os instrumentos operacionais, permitindo ao leitor compreender por que determinadas perguntas, dimensões de risco e salvaguardas foram incorporadas ao modelo brasileiro.

O Caderno Referencial não substitui o Guia INAEP, não constitui instrumento de classificação de risco e não deve ser utilizado isoladamente como checklist de avaliação. Para a aplicação prática, devem ser consultados o Guia INAEP e a MARIAH.

A opção por manter este documento autônomo reconhece o trabalho técnico acumulado pelo Grupo de Trabalho e preserva a memória metodológica da construção do referencial brasileiro, sem transferir para o Guia principal a extensão própria de um documento de fundamentação.

Este trabalho é destinado à sociedade brasileira, à comunidade científica que faz pesquisa com inteligência artificial em saúde no Brasil, às pesquisadoras e aos pesquisadores que propõem os protocolos, aos Comitês de Ética em Pesquisa que os avaliam, instituições e aos gestores de dados que os viabilizam.

# 1 AS PERGUNTAS ÉTICAS NA ERA DA INTELIGÊNCIA ARTIFICIAL

## 1.1 UM GUIA SOBRE FAZER PERGUNTAS

A ética em pesquisa com seres humanos sempre dependeu mais da capacidade de interrogar do que da capacidade de dominar. Pesquisadores conhecem detalhadamente os protocolos que elaboram. No contexto brasileiro, seus pares, integrantes de Comitês de Ética em Pesquisa (CEPs), conhecem os princípios que o protocolo precisa satisfazer. Entre esses dois conhecimentos, a pergunta certa é o instrumento que permite o diálogo.

Este guia foi construído sobre essa premissa. Seu propósito não é transformar seus leitores em especialistas em inteligência artificial (IA). É oferecer as perguntas que profissionais da ética em pesquisa precisam fazer a quem desenvolve e utiliza tecnologia. Essas perguntas não pertencem apenas a quem avalia. Quem propõe o estudo deve respondê-las no próprio protocolo. Esse é o diálogo sobre o qual este guia foi construído.

Essa distinção tem consequência prática imediata. A formação especializada em arquiteturas de redes neurais, métricas de desempenho de modelos ou protocolos de criptografia exige anos de estudo, o que não é exigido ou presumido pela composição multiprofissional dos CEPs.

A capacidade de interrogação ética exige precisão conceitual sobre o que está em jogo para o participante de pesquisa, transparência sobre quais garantias o protocolo deve oferecer e firmeza para exigi-las mesmo diante do entusiasmo científico e da assimetria de conhecimento técnico. Essa capacidade este guia pode construir, e é para construí-la que foi escrito (MRCT Center; WCG, 2025; OMS, 2021).

A pergunta certa, feita no momento certo, é o instrumento mais poderoso de proteção ética que um comitê possui. Não porque substitui o julgamento técnico do pesquisador, mas porque o obriga a explicitá-lo, documentá-lo e submetê-lo ao escrutínio de quem representa os interesses das pessoas que não estão na sala quando o protocolo é aprovado. Perguntar sobre a composição demográfica da base de dados de treino, o mecanismo de retirada do consentimento após o treinamento do modelo e o destino do sistema após o encerramento da pesquisa demonstra compreensão dos riscos e impede a aprovação daquilo que não pode ser adequadamente avaliado. A velocidade de desenvolvimento da inteligência artificial

coloca qualquer conhecimento técnico numa condição peculiar, uma vez que ele pode se tornar parcialmente (ou totalmente) obsoleto antes de ser completamente aprendido.

A arquitetura de redes neurais que domina o desempenho clínico hoje pode ser substituída por abordagem distinta em dois ou três anos. O modelo de linguagem que impressiona especialistas neste semestre apresentará limitações documentadas no próximo.

Para um guia que pretende orientar a escrita e a avaliação ética de protocolos de pesquisa por uma década, essa velocidade representa um desafio que não se resolve apenas por meios técnicos, mas também exige uma resposta ética.

As perguntas que este guia oferece não envelhecem na mesma velocidade que as arquiteturas. A questão da representatividade demográfica dos dados de treino permanece válida para qualquer sistema, de qualquer geração, treinado sobre qualquer modalidade de dado clínico. A verificação da supervisão humana real, distinta da supervisão apenas nominal, atravessa todas as arquiteturas e todos os contextos de uso. A reflexão sobre o que acontecerá com o modelo após o encerramento da pesquisa deve preceder qualquer discussão sobre seu desempenho técnico e sobrevive a qualquer atualização de protocolo.

São perguntas éticas com roupagem técnica. Este guia propõe-se torná-las acessíveis, mas sem exigir que o leitor domine essa roupagem, porque a roupagem muda e a ética permanece (Youssef *et al.*, 2024; Barucci *et al.*, 2026).

Há uma dimensão adicional nessa escolha. Quem aprende a fazer as perguntas certas, no comitê ou na bancada de pesquisa, desenvolve uma capacidade transferível. Pode aplicá-la a protocolos com sistemas que este guia não antecipou, em contextos que ainda não existem, sobre tecnologias que ainda não têm nome. A pessoa com especialização técnica que aprende uma arquitetura específica precisa reaprender quando a arquitetura muda. Quem aprende a perguntar não precisa recomeçar. Pergunta de novo.

O Guia de Uso Ético de Inteligência Artificial em Pesquisa com Seres Humanos da Inaep oferece um repertório de perguntas. Não um vocabulário técnico para impressionar pesquisadores, nem um manual de conformidade para satisfazer exigências formais. O Guia propõe um conjunto de perguntas precisas, fundamentadas e operacionais para proteger as pessoas que participam de pesquisas com inteligência artificial em saúde no Brasil.

A diferença entre um guia de conformidade e um guia de ética está justamente no tipo de pergunta que cada um propõe. O primeiro verifica se o protocolo atende aos requisitos

estabelecidos, o segundo questiona se as escolhas feitas na pesquisa são eticamente justificáveis.

A forma pode ser satisfeita com um Termo de Consentimento Livre e Esclarecido (TCLE) que menciona IA sem explicar o que isso significa para o participante. A substância exige que o TCLE descreva, em linguagem acessível, o que o sistema fará com os dados da pessoa, quem terá acesso ao modelo resultante e como o direito de retirada será exercido num sistema que já aprendeu com aqueles dados.

A forma pode ser satisfeita com a declaração de que um médico supervisionará as decisões do algoritmo. A substância exige demonstrar que esse profissional compreende a base da decisão algorítmica, dispõe de tempo para contestá-la, além de comprovar que o sistema foi desenhado para registrar quando discordou.

Perguntar bem é também um ato político. A pergunta formulada com precisão sobre temas como soberania dos dados, sobre colonialismo digital na pesquisa em saúde e sobre o destino dos dados do SUS em plataformas estrangeiras de IA não cumpre apenas função regulatória, mas função civilizatória.

É nesse sentido que o parecer pode contribuir para que a pesquisa com inteligência artificial no Brasil responde ao interesse das pessoas que geraram os dados, não apenas ao interesse de quem desenvolveu o sistema. O Sistema Nacional de Ética em Pesquisa com Seres Humanos (SINEP), instituído pela Lei n.º 14.874/2024, é a evolução do Sistema CEP/Conep, ao qual sucede na estrutura desenhada pela Lei e por sua regulamentação. As competências atribuídas pelas resoluções do Conselho Nacional de Saúde anteriores à Lei — entre elas a Resolução CNS n.º 738/2024, largamente citada neste guia — leem-se, aqui, como competências do SINEP (Brasil, 2024a). O sistema funda-se sobre a compreensão de que a proteção dos participantes exige um sistema independente com mandato para fazer as perguntas que o entusiasmo científico e a pressão de mercado tendem a silenciar. Este guia é o instrumento que torna esse mandato exercível no campo da IA em saúde.

## 1.2 O TEMPO ÉTICO DESTE GUIA

A inteligência artificial já integra protocolos de pesquisa em saúde submetidos aos CEPs brasileiros. Entre os exemplos, estão os sistemas de apoio ao diagnóstico por imagem treinados sobre bases de dados de hospitais universitários, modelos preditivos de risco

epidemiológico aplicados a populações do SUS, ferramentas de processamento de linguagem natural sobre prontuários eletrônicos e algoritmos generativos que produzem dados sintéticos para contornar restrições de privacidade. Esses protocolos são avaliados com instrumentos construídos para outro mundo, porque a regulação específica ainda não chegou com a profundidade necessária para cobrir o que a tecnologia já faz. Este guia nasce desse intervalo.

Seria reconfortante prometer que o intervalo é provisório, que a regulação virá, que a lacuna será preenchida e que este guia se tornará desnecessário. Essa promessa seria imprecisa. O intervalo entre a velocidade da inovação tecnológica e a velocidade da produção normativa é condição estrutural de campos em aceleração, não falha administrativa corrigível. A bioética não esperou pela conclusão do debate filosófico antes de precisar responder à realidade dos ensaios clínicos com novos fármacos.

A ética em pesquisa com dados não esperou pelo amadurecimento da ciência de dados antes de precisar proteger participantes. A ética em pesquisa com IA não aguardará regulação completa para exercer seu mandato. O CEP decide agora, com os princípios que existem, aplicados a tecnologias que os princípios não anteciparam nominalmente, mas que os princípios alcançam com precisão quando as perguntas certas são feitas (Soares; Chiavegatto Filho, 2026; Wilton Park, 2026).

Este guia é a resposta institucional do SINEP, por meio da Instância Nacional de Ética em Pesquisa (INAEP), vinculada ao Ministério da Saúde do Brasil, a esse momento. Não é documento de espera nem exercício acadêmico. É tomada de posição sobre o que o sistema de ética em pesquisa brasileiro exige da pesquisa com IA hoje, com base nos princípios estabelecidos e na experiência acumulada pelo Sistema CEP/Conep ao longo de sua trajetória, iniciada em 1988, com a publicação da Resolução n.º 01, de 1988, e considerando o marco legal vigente, a Lei n.º 14.874/2024 e o Decreto n.º 12.651/2025 que a regulamenta (Brasil, 2024a; Brasil, 2025a). O Brasil tem um sistema de ética em pesquisa com legitimidade histórica, capilaridade territorial e composição multiprofissional que poucos países no mundo construíram com a mesma solidez. Esse sistema reúne condições para avançar na liderança do campo da ética em pesquisa com IA. Este guia é um passo nessa direção.

## 1.3 PARA QUEM ESTE GUIA FOI ESCRITO

Este guia destina-se à sociedade brasileira, à comunidade científica que faz pesquisa com IA em saúde no Brasil: às pesquisadoras e aos pesquisadores que a propõem, às instituições e aos gestores de dados que a viabilizam, e aos CEPs que a avaliam. Os Comitês brasileiros reúnem especialistas como médicos, enfermeiros, odontólogos, assistentes sociais, filósofos, juristas, teólogos e representantes de usuários, e essa composição multiprofissional é uma força deliberada do sistema, não uma limitação. Nenhum de seus leitores precisa entender redes neurais para cumprir seu papel ético: cada conceito técnico que aparece aqui existe porque é necessário para formular uma pergunta ética precisa. Nenhum conceito técnico aparece por si mesmo.

As pesquisadoras e os pesquisadores que desenvolvem ou propõem pesquisa com IA em saúde no Brasil são o público primário deste guia. Para eles, o guia oferece clareza sobre o que o CEP vai perguntar e por que — não para preparar respostas convenientes a perguntas previsíveis, mas para construir protocolos que mereçam respostas afirmativas. A transparência sobre as exigências éticas serve ao pesquisador honesto tanto quanto protege o participante. O protocolo que antecipa as perguntas deste guia e as responde com honestidade é o protocolo que a pesquisa com IA em saúde no Brasil precisa produzir.

A estrutura do Guia permite tanto a leitura integral quanto a consulta por tema ou necessidade.

Quem propõe ou avalia um protocolo com aprendizagem federada pode ir diretamente ao Capítulo 7. Quem precisa entender o que é deriva de dados antes de avaliar um protocolo longitudinal encontra essa resposta no Capítulo 2. Quem quer o roteiro completo de avaliação ética passa pelos Capítulos 6 e 7 (integridade científica e governança de dados). Quem precisa do instrumento proporcional de avaliação de risco vai ao Capítulo 8 (Matriz de Avaliação de Risco em Inteligência Artificial). Quem quer a especificação por natureza da pesquisa, Programa de Computador como Dispositivo Médico (Software as a Medical Device — SaMD) e Ciências Humanas e Sociais, vai ao Capítulo 9. A leitura integral forma quem propõe e quem avalia. A consulta pontual instrumentaliza o ato de avaliação.

## 1.4 COMO USAR ESTE GUIA

O guia que o leitor tem nas mãos foi construído com arquitetura deliberada. O Capítulo 2 forma: oferece o vocabulário mínimo para que as perguntas éticas sobre IA façam sentido sem exigir especialização técnica.

Os Capítulos 4 e 5 situam: o contexto brasileiro e o panorama internacional que determinam o território onde essas perguntas serão feitas e as assimetrias de poder que as tornam necessárias.

O Capítulo 6 responsabiliza: trata da integridade científica e a autoria na pesquisa com IA, que definem a cadeia de responsabilidade pelo protocolo que chega ao CEP.

O Capítulo 7 governa: traduz a Lei Geral de Proteção de Dados Pessoais (LGPD) e a Resolução CNS n.º 738/2024 em exigências operacionais sobre os dados que o protocolo utiliza (Brasil, 2024a). O Capítulo 8 instrumentaliza: apresenta a Matriz de Avaliação de Risco em Inteligência Artificial em Pesquisa com Seres Humanos (MARIAH) como ferramenta de avaliação proporcional de risco.

O Capítulo 9 especifica: trata das especificidades de protocolos com Programa de Computador como Dispositivo Médico (SaMD) e com pesquisa em Ciências Humanas e Sociais. As Partes V e VI da MARIAH oferecem sustentação à capacitação que torna o exercício do mandato ético progressivamente mais qualificado ao longo do tempo.

As Partes III (Versão A) e IV (Versão B) da MARIAH são instrumentos operacionais de apoio à avaliação. A Matriz de Avaliação de Risco em Inteligência Artificial, disponível nas versões qualitativa e quantitativa, converte os princípios do guia em critérios concretos para aplicação nos protocolos, com análise individual de cada dimensão e questionamento. O guia forma quem pergunta. A matriz instrumentaliza o ato de avaliação. Os dois são necessários e os dois foram concebidos para funcionar juntos.

Este Caderno não é um manual jurídico, uma diretriz clínica ou um documento de certificação técnica de sistemas de IA. Sua função é oferecer fundamentação para a avaliação ética. Não substitui o julgamento do comitê, não dispensa a deliberação colegiada e não oferece respostas prontas para dilemas que exigem ponderação. Oferece as perguntas que tornam essa ponderação possível com a precisão que o campo exige.

A ética em pesquisa sempre foi, em última instância, o exercício coletivo e fundamentado do julgamento humano sobre o que se pode fazer a outros seres humanos em nome do conhecimento científico.

A inteligência artificial não altera esse fundamento. Exige que ele seja exercido com mais rigor, mais informação e mais consciência das consequências que um parecer mal fundamentado pode produzir sobre populações que não estavam na sala quando o protocolo foi aprovado.

Há também o que este guia não muda. Boa parte dos protocolos com algum uso de IA que chegam hoje a um CEP envolve estudos com dados tabulares e modelos preditivos consolidados. São usos que já fazem parte do serviço cotidiano dos comitês, e para eles este guia não cria barreira nova. A profundidade da avaliação permanece proporcional ao que o sistema faz no protocolo. O filtro de entrada da MARIAH classifica como rotineiro o uso que não influencia as decisões do estudo nem produz efeitos sobre a pessoa participante, e o protocolo segue o fluxo ordinário do CEP. Os usos que passam pelo filtro encontram nos níveis iniciais da matriz um percurso compatível com o risco que oferecem. O Nível I, o mais leve, já contempla protocolos como os de apoio à análise exploratória de dados não identificados. Para esses estudos, este guia pede apenas o que o comitê já faz bem.

## 1.5 O SUS COMO TERRITÓRIO E HORIZONTE

O Sistema Único de Saúde (SUS) do Brasil responde pelo Direito à Saúde de mais de 214 milhões de pessoas (Brasil, 2026e) em território continental, com diversidade epidemiológica, étnica, linguística e regional que nenhuma base de dados privada global reproduz. Essa diversidade é simultaneamente o maior ativo científico do sistema público de saúde brasileiro e sua maior responsabilidade ética. Os dados que o SUS produz diariamente, ao longo de décadas de atendimento universal, representam uma riqueza informacional sobre a saúde humana em contextos de desigualdade que o mundo não tem em outro lugar. Pesquisar com esses dados é um privilégio que carrega obrigações proporcionais.

Este guia foi escrito para esse território. Não para um sistema de saúde abstrato nem para o contexto clínico de países de alta renda onde a maioria dos sistemas de IA em saúde foi desenvolvida e validada. Escrevê-lo a partir do SUS não restringe o seu alcance: as mesmas diretrizes aplicam-se à pesquisa conduzida na saúde suplementar e em instituições privadas, sujeitas ao mesmo arcabouço ético e normativo da pesquisa com seres humanos no Brasil. As perguntas que oferece foram construídas com a especificidade do SUS em mente: a diversidade regional que torna o viés algorítmico um risco de saúde pública, a fragilidade de

infraestrutura que limita o que é possível exigir e o que é possível supervisionar, a presença de povos originários, quilombolas e povos das águas em contextos que nenhum modelo importado conhece. Ignorar essa especificidade seria produzir um guia sobre ética em pesquisa com IA que não serve ao Brasil. Seria mais um documento importado sem tradução ao Sul.

O horizonte deste guia é o SUS como projeto civilizatório, não apenas como infraestrutura de atendimento. Um sistema construído sobre o princípio de que a saúde é direito universal, que a pesquisa que o alimenta deve respeitar as pessoas que o sustentam e que a inteligência artificial que nele operar deve ser governada pelo interesse público. Esse horizonte determina o que as perguntas deste guia protegem: não apenas os participantes individuais de cada protocolo, mas o acervo coletivo de dados, de cuidado e de confiança que o SUS representa. A pesquisa com IA no SUS que não responder às perguntas deste guia não é apenas pesquisa eticamente inadequada. É pesquisa que compromete o patrimônio público que a torna possível.

Os capítulos que seguem desenvolvem essas perguntas. Começam pelo vocabulário necessário para formulá-las, no Capítulo 2.

## 1.6 NOTA METODOLÓGICA: O USO DE INTELIGÊNCIA ARTIFICIAL NA CONSTRUÇÃO DESTE GUIA

Este guia foi produzido no exercício da ética que recomenda. A Portaria CNPq n.º 2.664/2026, que instituiu a Política de Integridade na Atividade Científica, exige a declaração explícita do uso de ferramentas de IA generativa em qualquer fase da produção intelectual de pesquisa ou fomento (CNPq, 2026). O Capítulo 6 deste guia traduz essa exigência em parâmetros operacionais para protocolos submetidos aos Comitês de Ética em Pesquisa. A coerência editorial exige que o próprio guia seja transparente sobre o uso de IA em sua construção, e aplica sobre si o mesmo padrão que propõe aos demais.A construção deste guia envolveu o uso coordenado de cinco sistemas de IA generativa, entre março e abril de 2026, sob direção editorial humana explícita em todas as fases. O Claude Opus 4.7 (Anthropic) atuou na revisão editorial integral dos capítulos em quatro camadas (ortografia, estilo, conteúdo e estrutura), na reescrita de trechos sob direção do organizador, na harmonização de siglas e de referências bibliográficas, na elaboração dos quadros de navegação e dos mapas de

alterações, na composição do boneco do livro, na construção do índice remissivo e na articulação entre capítulos.

O ChatGPT 5.4 (OpenAI) foi utilizado em discussões exploratórias de estrutura dos capítulos, levantamento bibliográfico inicial e sugestão de alternativas conceituais em temas específicos. O Gemini 3.1 Pro (Google) foi utilizado em pesquisa comparativa sobre regulação internacional de IA em saúde, com ênfase em Brasil, Ásia (China, Japão, Coreia do Sul e Índia) e Rússia, que serviu de subsídio inicial ao Capítulo 5. O Microsoft Copilot foi empregado na assistência à elaboração de tabelas, quadros sinóticos e na verificação de consistência terminológica ao longo da redação. O DeepSeek foi utilizado na validação cruzada de conteúdo técnico e na identificação de lacunas em temas específicos de aprendizado de máquina e governança de dados.

Nenhum capítulo foi produzido autonomamente por qualquer sistema. A estrutura do livro, a seleção de argumentos, as decisões editoriais e a validação de cada afirmação couberam aos autores humanos, cujos nomes constam na ficha de identificação.

Todos os conteúdos gerados ou sugeridos por sistemas de IA foram revisados integralmente pelos autores antes de comporem o texto final. Referências bibliográficas foram confirmadas nas fontes originais; citações foram cotejadas com os documentos citados; dispositivos normativos foram conferidos no texto oficial das leis, decretos e resoluções. A verificação seguiu o padrão que o Capítulo 6 estabelece como mínimo: identificação da ferramenta, descrição da função exercida e atestação do responsável humano pela validação.

Os responsáveis humanos (supervisora e editor) nominados na ficha de identificação declaram que nenhum conteúdo gerado por IA foi submetido como de autoria humana sem revisão substantiva, que projetos de terceiros não foram inseridos em ferramentas de IA para elaboração de pareceres, e que a responsabilidade integral pelo conteúdo deste guia lhes pertence. A IA não figura como coautora, posição convergente da Organização Mundial da Saúde, do International Committee of Medical Journal Editors e da Portaria CNPq n.º 2.664/2026, retomada no Capítulo 6, §6.2.

O uso de IA acelera a produção textual, mas introduz riscos específicos que este guia reconhece em seus próprios capítulos: alucinação de referências e fatos (Capítulo 2, seção sobre alucinação algorítmica; Capítulo 6, §6.3); viés herdado dos dados de treinamento (Capítulo 3, §3.2; Capítulo 4, §4.5); tendência a confirmação retórica em grandes modelos de

linguagem; e imprecisão terminológica em dispositivos normativos brasileiros nos sistemas treinados predominantemente em inglês.

Três protocolos de mitigação foram adotados: (a) verificação de toda citação bibliográfica em fonte primária; (b) cotejo de cada afirmação normativa com o texto oficial da lei, decreto ou resolução referida; (c) leitura integral, por pesquisador humano responsável, de cada capítulo antes da inclusão no livro.

A declaração aqui prestada cumpre a Portaria CNPq n.º 2.664/2026 (CNPq, 2026), articula-se à Lei n.º 14.874/2024, que institui o SINEP e fixa a responsabilidade integral do pesquisador por todo o protocolo (Brasil, 2024a), e é coerente com a Resolução CFM n.º 2.454/2026 no que concerne ao uso de IA como ferramenta de apoio à decisão humana, jamais substituta (CFM, 2026).

A transparência aqui prestada serve ao duplo propósito de honrar o padrão ético proposto pelo guia e de oferecer, a autores e pesquisadores, um exemplo prático do que significa cumprir a exigência de declaração de uso de IA generativa na produção científica.

# 2 CONCEITOS PARA ENTENDER INTELIGÊNCIA ARTIFICIAL

## 2.1 INTELIGÊNCIA ARTIFICIAL: RAÍZES, HISTÓRIA E FUNDAMENTOS

Este capítulo apresenta os fundamentos necessários para compreender o que é um modelo de inteligência artificial, como ele aprende a partir de dados e quais famílias de algoritmos aparecem com maior frequência nos protocolos de pesquisa submetidos ao Sistema Nacional de Ética em Pesquisa (SINEP). As propriedades que transformam essas escolhas técnicas em questões éticas — como viés, explicabilidade, supervisão humana e ciclo de vida — serão examinadas no Capítulo 3.

A inteligência artificial já está presente em pesquisas avaliadas pelos Comitês de Ética em Pesquisa (CEPs), inclusive em sistemas de apoio ao diagnóstico, modelos preditivos, ferramentas de processamento de linguagem natural e arquiteturas generativas. Esses sistemas frequentemente chegam aos CEPs acompanhados de descrições técnicas que pressupõem conhecimentos especializados dos avaliadores.

O propósito deste capítulo não é formar especialistas em aprendizado de máquina, mas oferecer aos membros dos CEPs o vocabulário mínimo necessário para compreender o uso proposto, identificar informações ausentes e formular perguntas eticamente relevantes antes da aprovação do protocolo. Cada seção funciona como unidade autônoma de leitura. Quem avalia, por exemplo, um protocolo que utiliza um sistema generativo pode consultar diretamente a Seção 2.3.

Um modelo é uma função matemática (Figura 2.1). Recebe dados de entrada e produz uma saída: uma classificação, uma probabilidade, uma pontuação de risco, uma recomendação. O modelo opera por otimização de uma função matemática previamente definida durante o treinamento, sem realizar raciocínio clínico no sentido humano do termo.

**Figura 2.1— Modelo de IA como função matemática: entrada, transformação e saída**


Fonte: adaptado de Salih *et al.* (2023) e Santana, Moreno e Santos (2025), com exemplos documentados em pesquisas de saúde no Brasil (Barbosa *et al.*, 2021a, 2021b; Santos *et al.*, 2026).

A diferenciação entre otimização e raciocínio delimita o âmbito em que se concentram os principais riscos éticos associados. Um modelo otimizado para maximizar acurácia em um conjunto de dados históricos pode produzir resultados tecnicamente corretos naquele contexto e clinicamente inadequados em outro. A otimização matemática não tem acesso ao significado dos dados que processa (Santana; Moreno; Santos, 2025).

Em um breve contexto histórico, os conceitos fundadores da IA emergem de contribuições acumuladas ao longo do século XIX e da primeira metade do século XX. Figueiredo e Aguiar (2025) documentam esse percurso a partir de Charles Babbage, que concebeu a Máquina Analítica, e de Ada Lovelace, colaboradora direta de Babbage e reconhecida como a primeira programadora da história. Lovelace foi a primeira a compreender que uma máquina de calcular poderia ir além dos números e operar sobre símbolos de qualquer natureza. Essa intuição, formulada em 1843, antecipou em um século o que Alan Turing formalizaria com a máquina de Turing e o conceito de computabilidade.

Importante destacar que a contribuição das mulheres para a história da IA, não se limita a Ada Lovelace. Muitas mulheres tiveram papel essencial na história da IA, mas muitas não foram reconhecidas. As seis programadoras do Computador e Integrador Numérico Eletrônico (Electronic Numerical Integrator and Computer — ENIAC) desenvolveram o primeiro computador eletrônico dos EUA em 1946 e só foram creditadas publicamente décadas depois (Light, 1999).

Grace Hopper desenvolveu o primeiro compilador em 1952 e cunhou o conceito de linguagem de programação legível por humanos (Beyer, 2009). Hedy Lamarr patenteou, com George Antheil, o espalhamento espectral, fundamento das comunicações sem fio modernas (Figueiredo; Aguiar, 2025). Essas contribuições não constam dos livros canônicos da área. Essa ausência reflete o mesmo padrão de exclusão que os algoritmos treinados sobre dados históricos tendem a reproduzir e amplificar.

O apagamento não se restringe às mulheres. Shun'ichi Amari publicou, em 1967, a primeira rede neural com autoajuste de pesos por gradiente, fundamento matemático do aprendizado de máquina moderno (Amari, 1967). Kunihiko Fukushima desenvolveu, em 1980, o Neocognitron, a primeira rede convolucional profunda, precursora direta das redes neurais convolucionais (Convolutional Neural Networks — CNNs) que dominam o diagnóstico por imagem hoje (Fukushima, 1980).

O Prêmio Nobel de Física de 2024 foi concedido a John Hopfield e Geoffrey Hinton por contribuições que Amari e Fukushima haviam antecipado e, em aspectos centrais, fundamentado (Figueiredo; Aguiar, 2025). A Universidade de Durham documentou esse apagamento sistemático: pioneiros japoneses da IA foram escritos para fora da história canônica do campo (Durham University, 2024). O campo que se constrói esquecendo seus fundadores produz instrumentos que tendem a esquecer as populações que não dominam a narrativa.

Em 1956, pesquisadores reunidos em Dartmouth cunharam o termo inteligência artificial (IA) e inauguraram o campo como disciplina. Frank Rosenblatt criou o Perceptron em 1958, a primeira rede neural capaz de aprender padrões por meio de aprendizado supervisionado. A promessa era transformadora.

Entretanto, o desencanto veio cedo: as limitações matemáticas do Perceptron, documentadas por Minsky e Papert, geraram os chamados invernos da IA (*AI winters)*, períodos de corte de financiamento e descrédito acadêmico. O degelo chegou em 1986, quando o algoritmo de retropropagação (*backpropagation*) permitiu o treinamento de redes capazes de representar relações não linearmente separáveis e reabriu o campo para uma nova geração de pesquisadores (Figueiredo; Aguiar, 2025).

A ruptura seguinte foi estrutural. Em 2017, pesquisadores do Google publicaram o artigo "*Attention Is All You Need*" (Vaswani *et al.*, 2017), introduzindo a arquitetura dos transformadores (*transformers*). Esse trabalho tornou possível o desenvolvimento dos

grandes modelos de linguagem (Large Language Models — LLMs), sistemas treinados sobre vastos volumes de texto capazes de gerar linguagem, resumir documentos e apoiar decisões clínicas com sofisticação crescente (Figueiredo; Aguiar, 2025).

Em paralelo, o aprendizado profundo (*deep learning*) passou a processar imagens médicas, sinais fisiológicos e sequências genômicas com desempenho comparável ao de especialistas humanos em tarefas específicas.

Para os CEPs, essa trajetória oferece uma perspectiva que ultrapassa sua dimensão histórica. A evolução da inteligência artificial também revela processos de inclusão e exclusão presentes no desenvolvimento dos sistemas e na seleção dos dados utilizados para seu treinamento. Algoritmos aprendem padrões registrados em dados históricos, inclusive ausências, sub-representações e formas de apagamento neles presentes. A própria história desse campo, marcada pelo reconhecimento tardio ou insuficiente da contribuição de algumas de suas precursoras, demonstra que a invisibilidade não é apenas um problema dos dados, mas também das instituições e práticas que orientam a produção do conhecimento. Na pesquisa com seres humanos, esse legado exige atenção aos grupos que estão ausentes ou sub-representados nas bases de dados e aos efeitos discriminatórios que os sistemas podem reproduzir ou ampliar.

Essa dimensão histórica e ética se materializa, de forma concreta, nas etapas técnicas que estruturam o próprio processo de construção dos modelos.

## 2.2 COMO OS SISTEMAS DE IA APRENDEM: DADOS, TREINO E MODELO

Um sistema de inteligência artificial em saúde começa com dados clínicos produzidos em contextos reais, por pessoas reais e sob condições que o modelo não conhece diretamente. Antes do treinamento, esses dados já foram submetidos a sucessivas decisões humanas: quais informações coletar, quais populações incluir, quais instrumentos utilizar e em quais condições de infraestrutura realizar a coleta. Essas decisões influenciam a qualidade, a representatividade e as limitações do sistema resultante.

Youssef *et al.* (2024), em artigo publicado na *European Cardiology Review*, destacam que o protocolo de pesquisa com IA deve descrever o fluxo completo dos dados, incluindo coleta, pré-processamento, divisão dos conjuntos de dados, treinamento, validação, teste e eventual pós-processamento das saídas do modelo. A estratégia de divisão dos dados

acompanha essa sequência e organiza o processo de aprendizagem do modelo em etapas distintas.

O conjunto de treino reúne os exemplos rotulados com os quais o modelo ajusta seus parâmetros internos para associar padrões de entrada às saídas corretas. O conjunto de validação acompanha o desempenho durante o treinamento e orienta o ajuste dos hiperparâmetros, sem alterar diretamente os pesos do modelo. O conjunto de teste permanece isolado até o final do processo e permite avaliar, de forma imparcial, a capacidade de generalização para dados novos (Youssef *et al.*, 2024). Essa separação permite estimar, com maior rigor, a capacidade de generalização do modelo, sem assegurar que o desempenho observado será reproduzido em condições reais de uso.

Por conseguinte, os CEPs devem ter informações adequadas sobre os dados, incluindo rastreabilidade, vinculabilidade e auditabilidade, para assegurar que os dados de treinamento sejam compatíveis com o uso previsto do sistema de IA. Para o CEP, isso significa que o protocolo deve informar não apenas quais dados foram usados, mas de qual população foram extraídos, com qual critério de seleção, se passaram por pré-processamento como normalização ou harmonização, se alguma parte foi gerada sinteticamente e, se sim, com qual justificativa.

O *Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard* (MRCT Center) e o *WIRB-Copernicus Group* (WCG) recomendam ainda que os protocolos especifiquem, de forma clara, os níveis e privilégios de acesso aos dados e ao sistema de IA durante o treinamento, incluindo se as atividades ocorrem em infraestrutura local dedicada ou em plataformas de nuvem e recursos de terceiros (MRCT Center; WCG, 2025).

A transparência sobre a origem, a qualidade e o tratamento dos dados são indispensáveis para avaliar se o desempenho observado durante o desenvolvimento poderá ser mantido nas condições reais de uso. Resultados satisfatórios nos dados de treinamento ou em ambientes controlados não garantem que o sistema será igualmente útil, seguro ou preciso em outros contextos.

Um sistema de rastreamento de retinopatia diabética, por exemplo, pode apresentar bom desempenho quando validado em um centro de referência que dispõe de equipamentos de alta resolução e profissionais especializados. O mesmo sistema pode produzir resultados diferentes quando utilizado em unidades básicas de saúde com equipamentos de menor qualidade, condições distintas de coleta e profissionais sem treinamento específico. Isso

ocorre porque o modelo aprende relações presentes nos dados e no contexto em que eles foram produzidos. Quando esse contexto se modifica, o modelo não adapta automaticamente os parâmetros aprendidos: continua produzindo resultados com base nos padrões do ambiente original.

Por essa razão, o protocolo deve justificar a correspondência entre o contexto de desenvolvimento e o contexto de uso pretendido, além de prever validação local quando diferenças de população, infraestrutura, equipamentos, fluxos de trabalho ou qualificação dos usuários puderem afetar o desempenho do sistema. Na avaliação do protocolo, essa propriedade dos modelos tem implicação direta.

O MRCT Center e o WCG (2025) recomendam que os Comitês examinem o contexto de uso pretendido desde a avaliação inicial do protocolo. Entre as questões relevantes estão: para qual população o modelo foi treinado e validado? Essa população é suficientemente representativa das pessoas e dos grupos que poderão ser afetados pelo sistema em condições reais de uso? O modelo não identifica automaticamente as diferenças entre a população representada nos dados de desenvolvimento e aquela na qual será efetivamente aplicado. Cabe ao pesquisador reconhecer essas diferenças, avaliar suas possíveis consequências e propor medidas adequadas de validação e monitoramento.

Essa correspondência também pode se modificar ao longo do tempo. Modelos de inteligência artificial não acompanham automaticamente as mudanças ocorridas na população, na prática assistencial ou no ambiente em que são utilizados. Quando um sistema é treinado com dados clínicos de determinado período, ele aprende relações associadas às condições então existentes, como tecnologias disponíveis, normas vigentes, perfil epidemiológico da população, fluxos assistenciais e critérios diagnósticos.

Se essas condições se alteram e o sistema não é reavaliado ou atualizado, ele continua produzindo resultados com base nos padrões aprendidos anteriormente. A divergência progressiva entre os dados utilizados no desenvolvimento e os dados encontrados durante o uso é denominada **deriva de dados** (*data drift*). Youssef *et al.* (2024) destacam essa dinâmica como uma dimensão ainda insuficientemente examinada nos protocolos de pesquisa com IA, embora possa comprometer o desempenho do sistema, a segurança dos participantes e a validade científica dos resultados.

A deriva de dados assume formas distintas. A deriva de covariável (*covariate drift*) ocorre quando a distribuição dos dados de entrada muda sem que a relação entre entrada e

saída se altere. Um modelo de triagem treinado numa população urbana de renda média pode degradar quando aplicado a uma população rural com padrões distintos de comorbidade, mesmo que o critério diagnóstico permaneça o mesmo.

Já a deriva de conceito (*concept drift*) é mais profunda: ocorre quando a própria relação entre as variáveis de entrada e o desfecho se altera. A introdução de um novo protocolo terapêutico, a mudança nos critérios de uma classificação diagnóstica ou o surgimento de uma variante de agente infeccioso podem tornar obsoleta a lógica que o modelo aprendeu, sem que o sistema sinalize essa obsolescência.

Para a pesquisa em saúde no Brasil, esse risco tem dimensão adicional. A diversidade regional do Sistema Único de Saúde (SUS) significa que um modelo validado num contexto pode degradar rapidamente quando transportado para outro. Soares e Chiavegatto Filho (2026) demonstraram que algoritmos desenvolvidos sem validação local produzem resultados inadequados quando aplicados a populações com perfis epidemiológicos distintos dos dados de treino.

A deriva de dados não é fenômeno futuro a ser gerenciado após a aprovação do protocolo. Começa no momento em que o modelo é desenvolvido num contexto e destinado a operar em outro. O CEP que avalia um protocolo multicêntrico com IA em múltiplas regiões do Brasil avalia, simultaneamente, o risco de deriva entre os contextos participantes.

O protocolo de pesquisa com IA deve prever mecanismos de monitoramento contínuo do desempenho do modelo ao longo de toda a duração da pesquisa. Youssef *et al.* (2024) descrevem três fases relevantes do ciclo de vida de um sistema de IA: treinamento, teste em condições reais antes da entrada em uso e monitoramento após a entrada em uso.

A ausência de qualquer dessas fases indica que o pesquisador não mapeou o comportamento esperado do sistema além da fase de desenvolvimento. Na avaliação do CEP, um protocolo sem plano de monitoramento contínuo descreve um sistema que será abandonado à própria deriva logo após a aprovação.

Em vista disso, o MRCT Center e o WCG (2025) orientam os comitês de ética a verificarem a definição de critérios nítidos de desempenho mínimo aceitável. O protocolo também precisa estabelecer limites de degradação que acionem revisão do sistema e prever mecanismos de comunicação aos participantes diante de alterações relevantes no comportamento do modelo. Esse conjunto de exigências sustenta a integridade ética e a confiança pública na pesquisa em saúde.

Esses critérios constituem condições essenciais para a validade contínua do consentimento informado. O consentimento ancora-se no desempenho declarado do sistema no momento da decisão do participante. Alterações decorrentes de deriva de dados modificam esse referencial e exigem resposta ética proporcional. Os comitês de ética em pesquisa devem exigir planejamento prévio para esses cenários, com salvaguardas que preservem a legitimidade do consentimento ao longo do tempo.

## 2.3 TIPOS DE SISTEMAS DE IA EM PESQUISA EM SAÚDE

A escolha do algoritmo é a primeira decisão técnica de um protocolo de pesquisa com IA. É também uma decisão ética. Cada família de algoritmos ocupa uma posição distinta no espectro que tensiona desempenho preditivo e explicabilidade, e essa posição determina quais riscos o sistema carrega e quais perguntas o CEP deve formular. Esta seção examina as principais famílias presentes em protocolos de pesquisa com IA no SUS, com base em estudos realizados no Brasil em diagnóstico de doenças infecciosas, classificação de lesões mamárias e previsão de arboviroses (Barbosa *et al.*, 2021a, 2021b; Santos *et al.*, 2026). Para cada família, o quadro indica o nível de explicabilidade e a pergunta ética central que o CEP deve saber formular. Essa pergunta não exige que o avaliador domine a matemática do algoritmo. Exige que saiba o que perguntar ao pesquisador sobre o sistema avaliado.

O nível de explicabilidade indica como o raciocínio do modelo pode ser facilmente entendido por profissionais de saúde e participantes. A alta explicabilidade não implica maior desempenho preditivo; o equilíbrio entre ambos deve ser avaliado conforme a finalidade e o contexto de uso. Os exemplos de uso em saúde no Brasil referem-se a pesquisas realizadas no SUS e em hospitais universitários brasileiros, documentadas nas referências ao final deste capítulo.

Os sistemas de IA empregados em pesquisa clínica distribuem-se, em sua forma mais elementar, em duas grandes famílias, segundo o eixo temporal da pergunta que respondem.

Os sistemas de apoio diagnóstico respondem a uma pergunta sobre o presente: o que este dado revela sobre o estado clínico atual deste paciente? Os sistemas de predição de risco respondem a uma pergunta sobre o futuro: qual a probabilidade de que este paciente evolua para determinado desfecho?

Essa distinção define o tipo de risco ético que o protocolo carrega e o tipo de pergunta que o CEP deve formular antes de qualquer análise de mérito científico.

Os sistemas de apoio diagnóstico processam dados de entrada como imagens médicas, perfis laboratoriais e registros clínicos, produzindo classificações ou recomendações probabilísticas. O Conselho Federal de Medicina (CFM) estabeleceu, por meio da Resolução n.º 2.454, de 11 de fevereiro de 2026, que ferramentas algorítmicas de apoio diagnóstico devem atuar exclusivamente como suporte à decisão médica (CFM, 2026). A resolução veda a comunicação direta de diagnósticos aos pacientes e mantém a responsabilidade integral do médico sobre o ato clínico final.

Os sistemas de predição de risco operam sobre variáveis multimodais como histórico clínico, determinantes sociais, dados epidemiológicos e fatores ambientais, produzindo escores de probabilidade de desfechos futuros.

Modelos preditivos de risco ampliam significativamente a capacidade de vigilância epidemiológica do SUS. Contudo, tais sistemas apresentam riscos éticos próprios, notadamente o fenômeno da profecia autorrealizável (*self-fulfilling prophecy*).

A profecia autorrealizável caracteriza-se por uma dinâmica na qual uma previsão, ao influenciar decisões e comportamentos, contribui para que o resultado previsto se concretize: a expectativa se converte em realidade.

Por exemplo, um paciente classificado por um algoritmo como de alto risco de abandono terapêutico recebe menos acompanhamento e, por isso, abandona o tratamento, confirmando a própria previsão. O desfecho previsto passa a ocorrer em razão da decisão orientada pelo modelo, e não por evolução clínica inevitável. Portanto, o sistema contribui para produzir o resultado que anunciou.

Apoio diagnóstico e sistemas preditivos compartilham um risco estrutural que o CEP deve verificar em qualquer protocolo: o desempenho diferencial entre subgrupos populacionais. Soares e Chiavegatto Filho (2026) demonstraram que sistemas treinados com dados predominantemente de determinados grupos étnicos, de gênero ou socioeconômicos produzem resultados piores para populações sub-representadas na base de treino. Um sistema validado em hospital universitário de referência pode degradar ao ser aplicado em unidade básica de saúde do interior do Nordeste ou da Amazônia, com equipamentos distintos, perfis epidemiológicos diferentes e populações ausentes dos dados de treino (Fiske *et al.*, 2025).

Esse desafio de desempenho e representatividade também atravessa os sistemas que operam sobre linguagem clínica não estruturada. O processamento de linguagem natural (Natural Language Processing — PLN) parte de um dado paradoxo da medicina: até 80% das informações clinicamente relevantes residem em texto livre, não em campos estruturados de banco de dados. Evoluções de enfermagem, anamneses narrativas, laudos anatomopatológicos e históricos familiares carregam conhecimento que nenhuma planilha captura. O PLN converte essas narrativas em dados analisáveis por algoritmos, mediante técnicas de tokenização, lematização e representações vetoriais de palavras.

O principal desafio ético do processamento de linguagem natural não está relacionado à tecnologia. O texto livre dos registros médicos contém muitos identificadores indiretos, como profissões incomuns, datas de acidentes singulares e estruturas familiares pouco comuns. Por inferência reversa e triangulação com bases abertas, esses fragmentos podem reidentificar o participante à sua revelia, mesmo após supressão de nome e CPF.

As LLMs ampliam o poder do PLN e ampliam seus riscos na mesma proporção. Diferentemente dos modelos convencionais, que classificam ou extraem informação dirigida, os LLMs geram respostas, produzem laudos, formulam hipóteses e sintetizam literatura com fluência que simula autoridade técnica. Dois riscos específicos emergem com força na literatura.

O primeiro risco é a memorização latente: se pesquisadores introduzem textos médicos identificáveis nas janelas de interação com LLMs em nuvem, a arquitetura pode absorver esse dado confidencial e reproduzi-lo a outros usuários de forma não intencional (Figueiredo; Aguiar, 2025).

O segundo é a alucinação (*hallucination*): o modelo fabrica citações científicas inexistentes, recomenda posologias clinicamente inadequadas ou descreve achados laboratoriais fictícios com aparência de evidência.

A Portaria CNPq n.º 2.664, de 6 de março de 2026, fixou que a responsabilidade por imprecisões clínicas, plágios ou dados forjados gerados por LLMs em pesquisa recai integralmente sobre os pesquisadores humanos, vedando a autoria oculta por sistemas computacionais (CNPq, 2026).

Para um CEP, essa portaria tem implicação direta: protocolos que utilizam LLMs em qualquer fase da pesquisa devem declarar essa utilização e descrever os mecanismos de verificação humana dos resultados gerados. Os sistemas generativos constituem uma

categoria distinta dentro da IA em saúde. Redes adversárias generativas (Generative Adversarial Networks — GANs), modelos de difusão e LLMs podem criar imagens médicas sintéticas, texto clínico evolutivo, hipóteses farmacológicas e dados tabulares inteiramente artificiais.

A geração de dados sintéticos representa a promessa de escapar das restrições severas de privacidade, oferecendo material volumoso sem conexão com indivíduos reais. A Agência Nacional de Proteção de Dados (ANPD) alertou, por meio de nota técnica, que a unicidade individual não pode ser subestimada: modelos generativos podem replicar fragmentos idiossincráticos dos sujeitos reais que serviram de base, burlando a anonimização teórica e violando a Lei Geral de Proteção de Dados Pessoais (LGPD) caso engenharia reversa revele os pacientes originais (Brasil, 2024d).

O risco adicional é a substituição epistêmica: dados sintéticos usados como substituto de dados reais sem validação rigorosa da fidelidade estatística funcionam como câmara de eco, amplificando as desigualdades da base original e criando o risco de fabricação de evidências com aparência matemática de legitimidade.

Um fator de risco comum entre os sistemas PLN, LLMs e generativos consiste no viés linguístico decorrente da assimetria global. A maioria das bibliotecas e modelos disponíveis foi treinada predominantemente em inglês, capturando nuances culturais anglo-saxônicas. Quando aplicados a dados em português do Brasil, esses modelos distorcem significados, falham no processamento de jargões médicos regionais e cometem erros sistemáticos na leitura de abreviações padronizadas do SUS (Soares; Chiavegatto Filho, 2026).

O Projeto SoberanIA, lançado pelo Ministério da Ciência, Tecnologia e Inovação (MCTI) em parceria com a Secretaria de Inteligência Artificial do Piauí em 2025, representa a resposta estrutural a esse problema: um modelo fundacional treinado em 150 bilhões de tokens de escopo público em língua portuguesa, com sensibilidade a sotaques regionais e nuances socioculturais do território brasileiro (Santos *et al.*, 2025).

Para o CEP, a pergunta sobre o idioma de treinamento do modelo constitui questão de validade científica e equidade para a população estudada.

A resposta brasileira a esse cenário não elimina, por si, os desafios associados à reutilização de modelos previamente treinados em outros contextos. A transferência de aprendizagem (*transfer learning*) resolve um problema prático que afeta a maioria das instituições públicas de saúde brasileiras: o custo proibitivo de treinar um modelo de IA do

zero. Um modelo pré-treinado acumula conhecimento geral processando terabytes de dados em supercomputadores de grande porte.

O pesquisador utiliza a arquitetura já pronta e a treina rapidamente com seus próprios dados, ajustando o modelo ao contexto da pesquisa em poucos dias e a um custo reduzido. Para laboratórios de universidades públicas periféricas com orçamentos subdimensionados, essa abordagem equaliza o acesso às arquiteturas de ponta mundial.

O risco está exatamente nesse aparente atalho: o modelo importado carrega os vieses das populações com que foi originalmente treinado, predominantemente europeias e estadunidenses. Um pesquisador do SUS que realiza transferência de aprendizagem para diagnóstico de lesões cutâneas herda a insensibilidade do modelo original a manifestações em pacientes pretos e pardos, cristalizando iniquidades raciais silenciosas sem que o protocolo as declare (Soares; Chiavegatto Filho, 2026). Na análise do protocolo, a pergunta essencial sobre modelos pré-treinados é sobre a procedência demográfica da base de treino original e sobre o poder estatístico da fase de ajuste fino para corrigir essas deficiências no contexto brasileiro.

Nada obstante, a transferência de aprendizagem resolve parcialmente o problema ambiental que o treinamento de grandes modelos impõe. O ajuste fino de um modelo pré-treinado consome fração mínima da energia exigida pelo treinamento original. Isso reforça, do ponto de vista da sustentabilidade, a escolha por arquiteturas reutilizáveis em vez de modelos construídos do zero.

O custo ambiental que persiste é o da inferência contínua: o uso do modelo após o treinamento acumula consumo proporcional ao volume de consultas, e sistemas operando em infraestrutura de nuvem internacional com matriz energética intensiva em carbono transferem esse impacto para territórios distantes dos beneficiários da pesquisa. O Guia do Usuário Brasileiro de Inteligência Artificial reconhece que o impacto ambiental dos sistemas de IA constitui dimensão ética relevante (Brasil, 2026c). Para um CEP, a pergunta não é apenas sobre o que o sistema faz, mas onde opera, com qual infraestrutura e qual sua matriz energética. Essa dimensão raramente aparece nos protocolos. É parte da análise de risco de um protocolo comprometido com a equidade entre as gerações que habitam o mesmo território que o SUS serve.

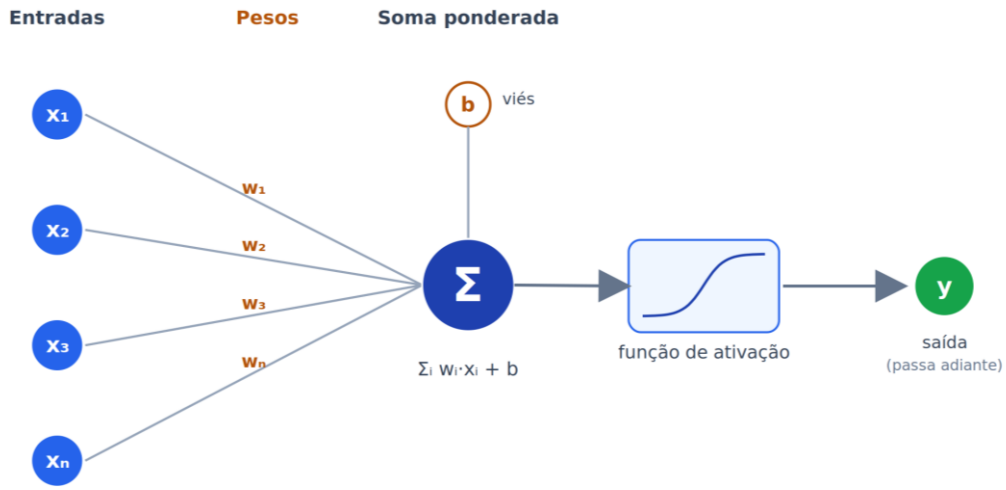
As redes neurais profundas (Deep Neural Networks — DNNs) estruturam-se em dezenas ou centenas de camadas consecutivas de neurônios artificiais. Cada neurônio recebe entradas, aplica uma função matemática e passa o resultado adiante (Figura 2.2). Cada

camada aprende representações progressivamente mais abstratas dos dados de entrada (Figura 2.3).

**Figura 2.2 — Arquitetura de uma rede neural profunda: camadas de neurônios artificiais**


Fonte: elaboração própria.

**Figura 2.3 — Funcionamento interno de um neurônio artificial em uma rede neural profunda**




Fonte: elaboração própria.

Três arquiteturas dominam a pesquisa em saúde: as redes convolucionais (Convolutional Neural Networks — CNNs) para classificação de imagens médicas, as redes recorrentes (Recurrent Neural Networks — RNNs) para séries temporais como eletrocardiogramas contínuos, e os transformadores (Transformers) para mineração profunda em prontuários extensos (Santana; Moreno; Santos, 2025).

A profundidade das DNNs garante alta precisão, mas as torna difíceis de explicar, funcionando como caixas-pretas. Outro risco é o sobreajuste: redes treinadas com poucos dados tendem a memorizar ruídos, exibindo alta acurácia no desenvolvimento, mas falhando diante da diversidade real (Youssef *et al.*, 2024). O CEP que recebe um protocolo com DNN deve verificar se há métricas documentadas de validação cruzada em conjuntos separados e se o protocolo prevê técnicas de IA explicável (Explainable Artificial Intelligence — XAI) para tornar as decisões do modelo contestáveis pelos profissionais de saúde envolvidos.

Os modelos leves (*lightweight models*) respondem a uma necessidade real do SUS: rodar sistemas de IA em dispositivos com capacidade computacional limitada, sem conexão estável com servidores de nuvem. Técnicas de compressão de modelos, destilação de conhecimento (*knowledge distillation*) e quantização reduzem o tamanho e o custo computacional das DNNs sem sacrifício crítico de desempenho.

O risco ético específico dos modelos leves é a degradação seletiva: os processos de compressão podem reduzir desempenho de forma desigual entre subgrupos populacionais, prejudicando exatamente as populações atípicas que o SUS mais precisa alcançar. O protocolo deve documentar que a compressão foi auditada para preservar equidade de desempenho entre grupos.

As máquinas de aprendizado extremo (Extreme Learning Machines — ELMs) representam uma abordagem radicalmente distinta das DNNs. Ao invés do processo iterativo de ajuste por retropropagação, a ELM sorteia aleatoriamente os pesos da camada oculta, trava esses pesos e resolve o resultado por inversão matricial direta. O ganho é expressivo: velocidade de treinamento incomparavelmente superior à das DNNs, com custo computacional mínimo.

O risco ético central da ELM é a instabilidade estocástica: por depender de pesos aleatórios, o modelo pode produzir resultados oscilantes quando treinado repetidas vezes sobre os mesmos dados, ferindo o princípio da reprodutibilidade científica. Em conjuntos de dados pequenos ou desbalanceados, essa instabilidade se agrava. O CEP deve verificar se o protocolo documenta testes de repetição com atestado de consistência das classificações e se o volume da base de dados é suficiente para sustentar o poder decisório da arquitetura proposta.

O tipo de sistema de IA determina o tipo de risco ético e o tipo de pergunta que o CEP deve formular. Essa afirmação, que atravessa todos os movimentos desta seção, exige uma

operacionalização prática: como o comitê identifica o tipo de sistema ao ler um protocolo? O MRCT Center e o WCG (2025) propõem que os comitês classifiquem os sistemas não pela doença estudada, mas pelo estágio de desenvolvimento da ferramenta no momento da pesquisa. Na fase de descoberta, o sistema ainda está em construção e os dados são usados para treinar a arquitetura. Na fase de tradução, o sistema é testado em condições reais pela primeira vez. Na fase de entrada em uso, o sistema opera de forma intervencionista sobre rotinas clínicas. Cada estágio impõe obrigações éticas distintas e exige vigilância do CEP em dimensões diferentes.

Barucci *et al.* (2026), na perspectiva europeia sobre o Regulamento da União Europeia sobre (Artificial Intelligence Act — AI Act) para comitês de ética, acrescentam uma dimensão regulatória relevante para o Brasil. O Regulamento (UE) 2024/1689 classifica como alto risco todos os sistemas de IA usados em contexto clínico como dispositivos médicos ou integrados a dispositivos médicos existentes.

No Brasil, a RDC n.º 657, de 24 de março de 2022, enquadra nessa categoria os programas de computador com finalidade diagnóstica, de prevenção, monitoramento ou tratamento de doença (Anvisa, 2022). Para o CEP brasileiro, essa articulação tem consequência prática: antes de avaliar o mérito científico do protocolo, o comitê deve determinar se o sistema proposto se enquadra como programa de computador como dispositivo médico (Software as a Medical Device — SaMD) e qual a classe de risco correspondente. Essa classificação antecede qualquer análise de desempenho.

O percurso deste capítulo termina aqui. O leitor que o completou sabe identificar o tipo de sistema diante de si, conhece a base de dados que o alimenta e a família de algoritmos que o move. O que essas escolhas técnicas produzem quando o sistema encontra participantes de pesquisa — o viés, a opacidade, a supervisão que falha, o ciclo de vida que não termina — é o objeto do próximo capítulo.

# 3 PROPRIEDADES CRÍTICAS DOS SISTEMAS DE IA: VIÉS, EXPLICABILIDADE, SUPERVISÃO E CICLO DE VIDA

Os conceitos do capítulo anterior descrevem o que os sistemas de inteligência artificial são e como funcionam. Este capítulo trata do que eles produzem quando entram numa pesquisa com seres humanos. Quatro propriedades concentram as questões éticas. O viés, que não é defeito de programação e se instala em três momentos distintos. A explicabilidade, que precisa funcionar para públicos diferentes. A supervisão humana, que falha por razões previsíveis. E o ciclo de vida, que não termina quando a pesquisa termina. Antes delas, a seção 3.1 apresenta as cinco dimensões que graduam a exigência sobre o protocolo e atravessam as quatro.

## 3.1 AS CINCO DIMENSÕES DO RISCO PROPORCIONAL

A Lei n.º 14.874, de 28 de maio de 2024, e o Decreto n.º 12.651, de 7 de outubro de 2025, consolidaram no Brasil a gradação das exigências e dos trâmites em proporção ao risco da pesquisa (Brasil, 2024a; 2025a). Os princípios éticos da avaliação não variam com o risco. O que varia é o rigor dos requisitos exigidos do protocolo e a instância designada para analisá-lo. Aplicada aos sistemas de IA, essa gradação organiza-se em cinco dimensões que o CEP deve percorrer ao ler qualquer protocolo.

A primeira é a natureza e sensibilidade dos dados processados: sistemas que operam sobre texto livre de prontuários ou imagens biométricas exigem escrutínio mais rigoroso do que sistemas que processam variáveis tabulares não identificáveis.

A segunda é o grau de automação e a consequência clínica da decisão: um sistema que produz recomendações sujeitas a revisão humana obrigatória carrega risco distinto de um sistema cujas saídas são incorporadas diretamente a protocolos de conduta.

A terceira é o perfil generativo ou classificatório da arquitetura, dado que sistemas capazes de criar conteúdo novo, como dados sintéticos ou laudos gerados, apresentam riscos de alucinação e fabricação de evidências ausentes nos classificadores convencionais.

A quarta dimensão é a plasticidade dos parâmetros: o sistema opera com pesos congelados após o treino, ou continua aprendendo durante a pesquisa? Sistemas com aprendizado contínuo ou adaptativo introduzem instabilidade que pode alterar o comportamento do modelo ao longo do estudo, comprometendo o consentimento informado

dado com base em características do sistema que podem não ser as mesmas ao final da pesquisa.

O MRCT Center e o WCG (2025) recomendam que protocolos com sistemas adaptativos definam limiares predeterminados de mudança que acionem revisão ética obrigatória. A quinta dimensão é o nível de explicabilidade da arquitetura: sistemas com capacidade de IA explicável permitem que profissionais de saúde identifiquem a base da decisão algorítmica, contestem resultados e documentem o raciocínio clínico de forma auditável (Santana; Moreno; Santos, 2025; Youssef *et al.*, 2024).

Essas cinco dimensões exigem precisão na leitura do protocolo e disposição para formular as perguntas que o pesquisador, imerso no entusiasmo científico, pode não ter se feito. A classificação do sistema antecede qualquer avaliação de desempenho. O CEP que avalia o quão bem o sistema funciona antes de perguntar o que ele é e como decide aprova uma promessa, não uma pesquisa.

## 3.2 VIÉS ALGORÍTMICO: O QUE É, COMO SURGE E POR QUE PERSISTE

O viés algorítmico resulta da estrutura dos dados usados no treinamento do modelo, não de erros técnicos. Um algoritmo de aprendizado de máquina não tem acesso ao mundo; tem acesso aos registros que o mundo produziu.

Esses registros carregam décadas de decisões humanas sobre quem foi atendido, quem foi documentado, quem recebeu diagnóstico preciso e quem foi subnotificado. Ao aprender padrões, o modelo também aprende ausências e distorções. Segundo Barucci *et al.* (2026), os principais fatores de opacidade em IA são a dependência de dados históricos com vieses implícitos, sutis, sistêmicos e difíceis de detectar.

A opacidade se instala em três momentos distintos do ciclo de desenvolvimento. Sendo eles: 1) No momento da coleta, quando os dados disponíveis para treino não representam a diversidade da população que será afetada pelo sistema. 2) No momento do desenho, quando as escolhas sobre quais variáveis incluir, quais desfechos prever e quais métricas de desempenho adotar incorporam pressupostos não declarados sobre quem importa e o que conta como resultado relevante. 3) No momento da avaliação, quando o desempenho do modelo é reportado como média geral, ocultando disparidades graves em subgrupos específicos.

A ausência de análise de desempenho diferencial por subgrupo populacional é uma das lacunas éticas mais frequentes nos protocolos de ensaios clínicos com IA, precisamente porque o bom desempenho médio mascara falhas concentradas nas populações mais vulneráveis (Youssef *et al.*, 2024).

A origem geográfica dos dados de treino é o vetor primário de viés em sistemas de IA aplicados à saúde global. A Organização Mundial da Saúde (OMS) documentou que a maioria dos grandes modelos disponíveis para uso clínico foi treinada predominantemente com dados de países de alta renda, com populações majoritariamente brancas e europeias (OMS, 2024).

O viés algorítmico tem, portanto, três camadas sobrepostas que o CEP precisa distinguir. A primeira é técnica: o modelo falha de forma mensurável num subgrupo porque esse subgrupo estava sub-representado nos dados de treino. A segunda é estrutural: o modelo reproduz hierarquias de raça, classe e território como se fossem regularidades naturais, porque os dados históricos as codificam dessa forma. A terceira é política: a ausência de exigência de análise de equidade nos protocolos de pesquisa não é omissão neutra. Essa escolha determina quem recebe proteção ética e quem sofre com falhas do sistema não reconhecidas.

Viés técnico e injustiça estrutural são conceitos distintos. O CEP deve diferenciá-los para agir de maneira precisa. O viés técnico é mensurável: o modelo produz erro estatisticamente superior num subgrupo em relação a outro, e esse erro pode ser identificado, quantificado e reduzido por intervenção metodológica.

A injustiça estrutural opera em outra escala. Quando o modelo codifica como padrão normal uma realidade construída sobre exclusão histórica, ele não apenas erra mais para determinados grupos. Reproduz como fato objetivo a hierarquia que gerou a exclusão.

Rajkomar *et al.* (2018) demonstraram empiricamente o mecanismo central: modelos preditivos com alto desempenho global podem ocultar disparidades graves em subgrupos específicos. Um modelo com 90% de acurácia geral pode ter 70% de acurácia para o grupo sub-representado no treino. O relatório de desempenho global não revela esse dado. Somente a análise estratificada por subgrupo o torna visível.

Fiske *et al.* (2025) avançam nessa direção ao demonstrar que o problema não se resolve pela simples inclusão de variáveis de raça e etnia nos modelos. Quando essas variáveis são incorporadas sem contextualização histórica e social adequada, o modelo pode reforçar

associações espúrias entre marcadores identitários e desfechos clínicos, produzindo discriminação algorítmica com aparência de evidência científica.

O MRCT Center e o WCG (2025) postulam que comitês de ética devem verificar se os protocolos documentam análises de desempenho diferencial por subgrupo antes de qualquer aprovação, porque o viés não verificado não é viés inexistente.

No Brasil, essa distinção entre viés técnico e injustiça estrutural tem consequência prática direta. O protocolo que apresenta apenas métricas globais de desempenho, sem análise estratificada por raça, gênero, faixa etária, região geográfica e condição socioeconômica, não oferece evidência de equidade. Oferece ausência de evidência de iniquidade.

No SUS, que atende populações com diversidade epidemiológica, étnica e regional que nenhuma base de dados privada reproduz, aprovar um protocolo sem essa análise é aprovar um sistema cujo comportamento real para parte da população estudada permanece desconhecido. O Projeto de Lei n.º 2.338/2023 (Brasil, 2023), cujo objeto é a regulamentação da IA no Brasil, reconhece explicitamente o risco de discriminação algorítmica em serviços públicos e demandou auditorias independentes para sistemas de alto risco.

A exigência de análise estratificada de desempenho é a operacionalização ética desse princípio na avaliação de protocolos de pesquisa. Soares e Chiavegatto Filho (2026) afirmam que a incorporação acrítica de IA no SUS, sem validação local robusta, corre o risco de reforçar um modelo tecnocrático e excludente que aprofunda iniquidades preexistentes.

A Organização Pan-Americana da Saúde (OPS/OMS, 2024) destaca que países de renda média enfrentam um desafio específico: os modelos disponíveis para adoção foram desenvolvidos em contextos de infraestrutura superior, e seu desempenho real em equipamentos mais antigos e com profissionais com formação distinta permanece frequentemente desconhecido.

Os dados do SUS permitem desenvolver modelos genuinamente representativos da população brasileira. O uso desses dados sem salvaguardas adequadas, porém, transforma a diversidade do SUS em matéria-prima para sistemas que nunca devolverão equidade às populações que os geraram. Santos *et al.* (2025) identificam esse risco como colonialismo digital interno: a extração de dados das populações mais vulneráveis para alimentar sistemas que servem prioritariamente quem já tem acesso.

Esse conjunto de riscos se traduz em exigências concretas sobre a forma como os dados são descritos e justificados no protocolo.

A primeira exigência é a documentação da composição demográfica da base de treino. O documento deve descrever a origem dos dados utilizados para treinar o modelo, incluindo a distribuição por raça, gênero, faixa etária, região geográfica e condição socioeconômica, e comparar essa distribuição com o perfil da população-alvo da pesquisa. Quando a base de treino provém de modelos pré-treinados estrangeiros, a documentação deve identificar a origem geográfica e étnica dos dados originais e descrever como a fase de ajuste fino endereçou as diferenças em relação à população brasileira. Protocolos que descrevem apenas o tamanho da base de treino sem informar sua composição demográfica fornecem ao CEP informação insuficiente para avaliar o risco de viés (Youssef *et al.*, 2024).

A segunda exigência é a análise de desempenho diferencial por subgrupo. Espera-se que o protocolo apresente, ou se comprometa a produzir durante a pesquisa, métricas de desempenho estratificadas pelos grupos populacionais relevantes para o contexto clínico estudado. Métricas globais de acurácia, sensibilidade e especificidade são insuficientes quando o modelo será aplicado a populações heterogêneas.

O Ministério da Segurança Alimentar e de Medicamentos da Coreia do Sul publicou, em 2022, diretrizes específicas para o desenho de ensaios clínicos com IA que instruem pesquisadores e comitês de ética a verificar três dimensões ausentes nos protocolos de medicamentos convencionais: prevenção de vazamento de dados (*data leakage*), auditoria para detecção de vieses populacionais ocultos (*hidden bias*) e metodologias para mitigar variações de desempenho entre unidades hospitalares distintas (MFDS, 2022). O CEP brasileiro pode e deve adotar exigência equivalente, proporcional ao risco clínico da aplicação proposta.

A terceira exigência é o plano de monitoramento de equidade durante e após a pesquisa. O viés não é propriedade estática do modelo. Pode surgir ou ampliar-se ao longo do tempo à medida que a população atendida diverge do perfil de treino. O MRCT Center e o WCG (2025) recomendam que protocolos com IA definam critérios predeterminados de degradação de desempenho que acionem revisão ética obrigatória, incluindo critérios específicos para disparidades entre subgrupos.

## 3.3 EXPLICABILIDADE E TRANSPARÊNCIA

Explicabilidade corresponde à capacidade de produzir, para cada decisão relevante, uma narrativa compreensível sobre os fatores que a determinaram, sem exigir acesso integral ao código-fonte do sistema de IA.

Santana, Moreno e Santos (2025) definem a IA explicável como a capacidade desses sistemas de fornecer explicações inteligíveis sobre suas decisões e processos internos, tornando o funcionamento dos algoritmos transparente para usuários e criadores. Essa definição carrega uma distinção operacional que o CEP precisa internalizar: explicar para um engenheiro é diferente de explicar para um médico, e explicar para um médico é diferente de explicar para um paciente. Cada nível tem requisitos distintos e serve a propósitos éticos distintos.

Três níveis de explicabilidade estruturam esse espectro. O primeiro é a explicabilidade técnica: a descrição dos pesos, gradientes e ativações internas do modelo para engenheiros de aprendizado de máquina. Esse nível permite auditar a arquitetura, identificar camadas problemáticas e propor correções no modelo.

O segundo é a explicabilidade funcional: a identificação das variáveis mais influentes na decisão para cientistas de dados e pesquisadores clínicos. Ferramentas como SHapley Additive exPlanations (SHAP) e Local Interpretable Model-agnostic Explanations (LIME) operam nesse nível, produzindo mapas que indicam quais características dos dados de entrada mais contribuíram para cada saída específica.

Já o terceiro é a explicabilidade clínica: a tradução da lógica da decisão em linguagem compreensível para o profissional de saúde e contestável pelo participante da pesquisa. Santana, Moreno e Santos (2025) documentam aplicações de IA explicável no diagnóstico de COVID-19, hanseníase e retinoblastoma precisamente porque a adoção clínica dessas ferramentas dependia da capacidade do médico de entender, verificar e questionar o raciocínio do algoritmo.

O CEP concentra sua avaliação no terceiro nível, voltado à garantia de explicabilidade no protocolo. A análise não recai sobre o código ou o funcionamento interno do modelo. A pergunta central dirige-se à capacidade de justificar, de forma compreensível, as decisões produzidas pelo sistema.

Na avaliação da explicabilidade, o CEP parte de três verificações concretas. A primeira: o protocolo descreve qual técnica de explicabilidade será adotada e para qual público? Uma técnica adequada para engenheiros pode ser inútil para médicos. A segunda: a explicabilidade está prevista como componente obrigatório do sistema ou como recurso opcional disponível para quem solicitar? Sistemas em que a explicabilidade é opcional reproduzem na prática o problema da caixa-preta, porque a pressão de produtividade no ambiente clínico tende a suprimir o uso de recursos opcionais.

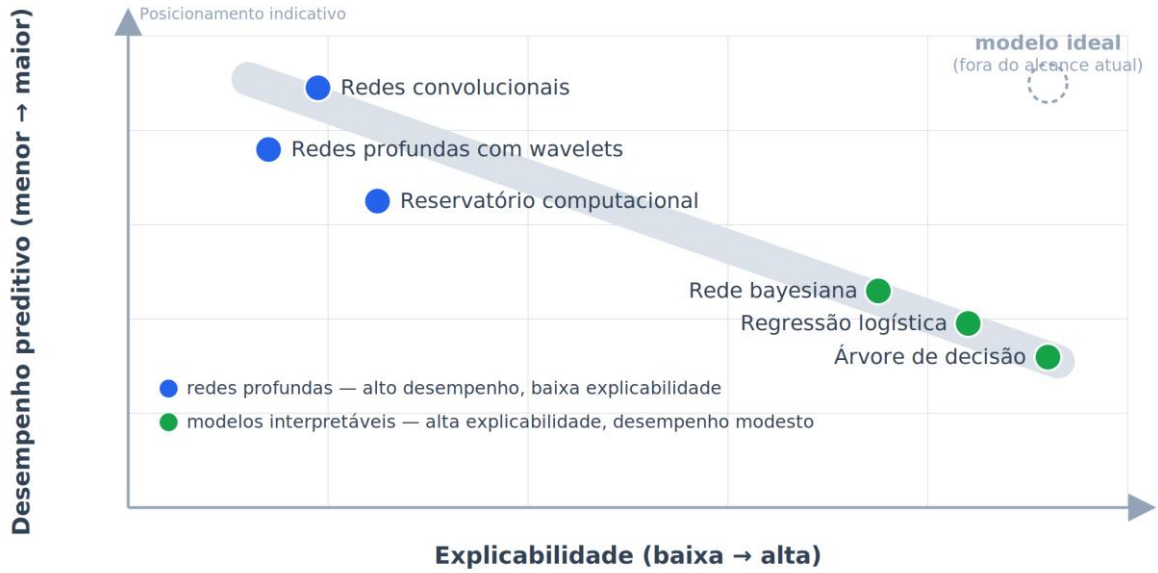
A terceira: as explicações produzidas pelo sistema foram validadas com profissionais de saúde do contexto onde o sistema será usado? Uma explicação válida para um especialista de oncologia num hospital terciário pode ser incompreensível para um médico generalista numa unidade básica de saúde no interior da Amazônia Legal. O MRCT Center e o WCG (2025) estabelecem que os comitês de ética devem verificar se as limitações do sistema e sua variabilidade potencial foram comunicadas de forma adequada a médicos, participantes e reguladores.

O conflito entre desempenho e explicabilidade é resultado das estruturas do aprendizado profundo, não de falhas de engenharia. Santana, Moreno e Santos (2025) documentam empiricamente que redes neurais profundas com alto desempenho clínico tendem a operar com menor explicabilidade do que modelos mais simples, como árvores de decisão e regressões logísticas.

Modelos mais precisos tendem a operar como caixas de difícil interpretação (Salih *et al.*, 2023). As DNNs alcançam alto desempenho ao identificar padrões complexos que escapam à leitura direta do clínico. Em sentido oposto, modelos como árvores de decisão permitem rastrear com maior transparência o caminho lógico da predição, ainda que com menor acurácia. Essa tensão não surge por acaso. Ela resulta da capacidade das redes neurais de construir representações abstratas em múltiplas dimensões, distantes dos conceitos clínicos tradicionais.

A Figura 3.1 representa esse espectro com os algoritmos mais usados em pesquisa de saúde no Brasil, posicionados segundo seu desempenho preditivo e seu nível de explicabilidade. No polo esquerdo, arquiteturas como redes neurais profundas com wavelets, redes convolucionais e redes de reservatório computacional concentram alto desempenho e baixa explicabilidade.

**Figura 3.1 — Relação indicativa entre desempenho e explicabilidade dos principais algoritmos em pesquisa em saúde**




Fonte: elaboração própria, com base em Barbosa *et al.* (2021a, 2021b), Salih *et al.* (2023), Santana, Moreno e Santos (2025) e Lima et al. (2026).

No polo direito, modelos como árvore de decisão, rede bayesiana e regressão logística oferecem explicabilidade alta com desempenho comparativamente mais modesto. O modelo ideal, que combina desempenho elevado com explicabilidade, segue sendo um objetivo fora do alcance das arquiteturas atuais. O posicionamento é indicativo e pode variar conforme o volume de dados, o contexto clínico e as escolhas de configuração (Barbosa *et al.*, 2021a, 2021b; Salih *et al.*, 2023; Santana; Moreno; Santos, 2025; Santos *et al.*, 2026).

Para um CEP, essa tensão tem implicação concreta na avaliação do protocolo. O pesquisador que propõe uma DNN para diagnóstico clínico em pesquisa com seres humanos está implicitamente aceitando um grau de opacidade. Essa aceitação precisa ser declarada e justificada.

A pergunta que o comitê deve formular é precisa: o ganho de desempenho em relação a modelos mais explicáveis justifica eticamente a redução de transparência para o participante e para o médico assistente?

Essa pergunta exige que o protocolo apresente comparação de desempenho entre a arquitetura proposta e alternativas mais interpretáveis, com análise explícita da troca entre acurácia e explicabilidade (Barucci *et al.*, 2026). Protocolos que adotam as redes neurais profundas sem essa comparação presumem que o desempenho superior justifica a opacidade.

A literatura documenta técnicas que mitigam, sem eliminar, a opacidade das DNNs. Mapas de calor e saliência (*heatmaps e saliency maps*) identificam as regiões de uma imagem médica que mais influenciaram a classificação do modelo, permitindo que o médico verifique se o sistema focou nas estruturas anatomicamente relevantes ou em artefatos irrelevantes. O *Gradient-weighted Class Activation Mapping* (Grad-CAM) e o SHAP operam nesse registro, traduzindo gradientes internos em representações visuais interpretáveis.

Essas técnicas também apresentam limites que precisam ser explicitados. Mapas de saliência indicam as regiões de maior atenção do modelo. A interpretação das razões desse foco e de sua relevância causal exige análise adicional.

Um mapa que destaca a região pulmonar numa radiografia de tórax não demonstra que o modelo aprendeu a anatomia pulmonar. Demonstra que essa região foi estatisticamente correlacionada com o desfecho nos dados de treino. A diferença entre correlação estatística e relevância causal é o espaço em que erros clínicos graves se instalam sem que o sistema sinalize nenhuma anomalia. Na prática, a presença de técnicas de explicabilidade num protocolo não equivale à garantia de que o sistema é clinicamente interpretável. Equivale a um instrumento que torna a interpretação possível. A supervisão humana significativa sobre esse instrumento permanece responsabilidade do pesquisador e exigência do comitê.

Três garantias adicionais estruturam a exigência de explicabilidade na aprovação de qualquer protocolo com IA.

A primeira é que o sistema produza, para cada decisão com consequência clínica sobre o participante, uma justificativa compreensível pelo profissional de saúde responsável. Essa justificativa não precisa descrever os pesos internos do modelo. Precisa responder, em linguagem clínica, à pergunta que o médico faria a um colega: por que esse resultado, para esse paciente, com esses dados?

A segunda é que essa justificativa possa ser contestada e sobreposta por julgamento humano documentado. No contexto da pesquisa, é o protocolo, sob responsabilidade do pesquisador, que descreve o fluxo pelo qual a saída do sistema pode ser contestada e sobreposta. Quando o sistema se destinar a uso clínico, o ato clínico decorrente observa a responsabilidade integral do médico prevista na Resolução CFM n.º 2.454, de 11 de fevereiro de 2026 (CFM, 2026). Para que essa supervisão seja real, é preciso base para contestação. Um sistema que produz saídas sem justificativa oferece resultado para aceitação ou rejeição

arbitrária. O protocolo deve descrever como o profissional responsável documenta sua discordância e registra o julgamento alternativo. Esse fluxo é condição para que a supervisão humana seja auditável.

A terceira é que os participantes sejam informados, no TCLE, sobre o nível de explicabilidade do sistema e o que isso significa para o seu direito à revisão da decisão. O artigo 20 da LGPD garante ao titular o direito de solicitar revisão de decisões automatizadas.

Fornasier (2022) demonstrou que esse direito permanece abstrato quando o participante não é informado sobre o grau de opacidade do sistema que o afeta. O TCLE de um protocolo com DNN deve informar, em linguagem acessível, que o sistema produz recomendações cuja lógica interna não é completamente transparente, descrever como o médico utilizará essa recomendação e indicar o canal pelo qual o participante pode solicitar esclarecimentos ou revisão. Explicabilidade que não chega ao participante não protege seus direitos, somente cria aparência de proteção.

O Ministério da Segurança Alimentar e de Medicamentos da Coreia do Sul exige, nas diretrizes para ensaios clínicos com IA de 2022, que fornecedores forneçam dados transparentes sobre a origem da base de treino e lógica de predição. Relatórios devem documentar metodologias para mitigar variações de desempenho entre hospitais (MFDS, 2022). Essa exigência regulatória sul-coreana é mais precisa do que qualquer orientação brasileira vigente sobre explicabilidade em protocolos de pesquisa com IA. O CEP que adota exigência equivalente não ultrapassa seu mandato. Exerce com rigor a função de guardião da dignidade dos participantes num campo onde a regulação ainda não chegou com a especificidade necessária.

A explicabilidade e a equidade são decisões de arquitetura em IA, definidas antes mesmo da coleta dos dados, e não acrescentadas após o sistema estar pronto. Essa distinção define o princípio da ética desde a concepção (*ethics by design*): as escolhas éticas devem ser incorporadas à estrutura do sistema como critérios de engenharia, não sobrepostas como camada de conformidade posterior. O princípio estende ao campo ético o que a privacidade desde a concepção (*privacy by design*) já estabelece para a proteção de dados.

O artigo 46 da LGPD determina que os agentes de tratamento devem adotar medidas de segurança, técnicas e administrativas, desde a fase de concepção do produto ou serviço (Brasil, 2018). Aplicado à IA em saúde, esse mandato legal significa que o protocolo de

pesquisa deve demonstrar que as decisões éticas fundamentais foram tomadas no momento do desenho do sistema, não identificadas como pendências após a construção do modelo.

A OMS documentou que a governança ética de sistemas de IA em saúde se articula em três fases da cadeia de valor: o desenho e desenvolvimento, a provisão do serviço ou produto, e a entrada em uso clínico (OMS, 2024). Em cada fase, atores distintos carregam responsabilidades distintas, e certas decisões só podem ser tomadas na fase anterior.

A escolha pela explicabilidade, por exemplo, precisa ocorrer antes do treino, porque algumas arquiteturas mais opacas são difíceis ou impossíveis de tornar interpretáveis retroativamente.

A decisão sobre a composição demográfica da base de treino precisa ocorrer antes da coleta, porque vieses estruturais de representatividade não se corrigem com ajustes posteriores ao modelo.

O protocolo que prevê análise de equidade apenas como etapa final de avaliação declara implicitamente que os dados já foram coletados, o modelo já foi treinado e as decisões arquiteturais irreversíveis já foram tomadas. O CEP que recebe esse protocolo recebe um sistema formado, não um projeto avaliável.

A posição da equidade no ciclo de desenvolvimento do sistema revela se as decisões éticas foram incorporadas desde a concepção ou tratadas como verificação tardia, já fora do alcance avaliativo do CEP.

Para o CEP, o princípio da ética desde a concepção tem implicação operacional direta na leitura de qualquer protocolo com IA. O comitê deve verificar se as decisões éticas estruturantes foram tomadas e documentadas antes do desenvolvimento. Barucci *et al.* (2026) propõem que os protocolos descrevam explicitamente as escolhas arquiteturais com justificativa ética, incluindo o racional para a seleção da arquitetura, as metodologias de mitigação de sobreajuste e a documentação da qualidade e composição da base de dados. Um protocolo que justifica essas escolhas apenas com critérios de desempenho, sem consideração explícita de equidade e explicabilidade, não satisfaz o padrão ético desde a concepção. A distinção entre um protocolo que trata a ética como projeto e um protocolo que trata a ética como checklist final é visível na sequência das seções do documento. O CEP que aprende a ler essa sequência desenvolve uma ferramenta de avaliação que nenhuma métrica de desempenho substitui.

## 3.4 SUPERVISÃO HUMANA E AUTOMAÇÃO DE DECISÕES

A supervisão humana em IA varia continuamente quanto à influência das decisões humanas sobre os resultados algorítmicos. Em um extremo, o sistema produz informação que o pesquisador considera livremente e pode ignorar sem justificativa. No extremo oposto, o sistema toma decisões que o pesquisador ratifica por hábito ou pressão institucional, ou que se executam de forma autônoma sem intervenção humana real. Entre esses extremos existe uma gradação que os protocolos de pesquisa raramente descrevem com precisão, mas que determina na prática quem decide sobre o participante (OMS, 2024).

A capacidade de interromper o sistema é a supervisão humana em sua forma mais concreta. O "botão de desliga" (*kill switch*) não é um recurso de segurança técnica, mas um requisito ético. Um sistema que não pode ser desligado pelo pesquisador responsável durante o estudo não está sob supervisão humana. Está sob supervisão nominal. O AI Act determina que profissionais responsáveis pela supervisão de sistemas de alto risco devem ser capazes de intervir e reverter decisões algorítmicas quando necessário (Barucci *et al.*, 2026). O protocolo CEP deve descrever como o pesquisador pode discordar da saída do sistema e como suspender o sistema quando o desempenho comprometer a segurança do participante, além de especificar quem tem autoridade para acionar essa suspensão. Essa capacidade de interromper, porém, só existe na prática quando o pesquisador reconhece que precisa intervir.

A OMS identificou o viés de automação (*automation bias*) como um dos riscos centrais do uso de grandes modelos de linguagem e sistemas de IA em diagnóstico e cuidado clínico. O fenômeno ocorre quando o profissional de saúde aceita passivamente a saída do sistema, sem escrutínio crítico, em virtude de sobrecarga de trabalho, excesso de confiança na tecnologia ou pressão institucional de produtividade (OMS, 2024). A documentação da OMS descreve esse risco como anterior aos LLMs: está presente em qualquer sistema de apoio à decisão clínica cuja arquitetura favoreça a aceitação sobre a contestação. Soares e Chiavegatto Filho (2026) apontam que esse fenômeno se consolida como risco ético específico na inserção de IA no SUS, onde a carga horária dos profissionais e a escassez de recursos criam condições propícias para que o apoio algorítmico se transforme em delegação efetiva.

O tipo de sistema e o contexto clínico determinam onde no espectro cada protocolo se situa na prática. Um sistema de triagem que produz um escore de risco numa tela de difícil

interpretação, em um ambiente de atendimento com filas longas e tempo reduzido por paciente, situa-se funcionalmente mais próximo da automação do que da recomendação, independentemente de como o protocolo o descreve. Quando o sistema se destinar a apoio à decisão clínica, aplica-se a exigência da Resolução CFM n.º 2.454/2026 de que a ferramenta apoie apenas a decisão do profissional, sem comunicar diagnósticos diretamente aos pacientes, e de que a responsabilidade sobre o ato clínico permaneça integral (CFM, 2026). Na pesquisa, a supervisão cabe ao protocolo e ao pesquisador. O protocolo deve comprovar que o sistema alcança esse objetivo no uso real, não apenas em teoria.

No contexto do CEP, a principal questão acerca da supervisão humana não se refere à posição declarada pelo pesquisador, mas sim àquela que o projeto do sistema, a elaboração das saídas, as condições institucionais de uso e as demandas de produtividade do ambiente clínico tendem a favorecer. Barucci *et al.* (2026) identificaram que a capacidade do médico de bloquear ou sobrepor a decisão algorítmica é um critério fundamental na avaliação regulatória sul-coreana de sistemas de IA clínica. O MRCT Center e o WCG (2025) recomendam que os protocolos descrevam os processos que estarão disponíveis para que pesquisadores identifiquem e respondam a eventos adversos relacionados à saída do sistema. Ambas as exigências partem da mesma premissa: supervisão humana que não foi desenhada para ser exercida não será exercida.

Supervisão humana significativa exige três condições simultâneas. A primeira: o pesquisador compreende a base da decisão algorítmica com profundidade suficiente para contestá-la. A segunda: dispõe de tempo e condições institucionais para exercer esse escrutínio antes de agir. A terceira: o sistema está desenhado para facilitar, não para inibir, a discordância. Quando qualquer dessas condições falha, a supervisão torna-se nominal. Fornasier (2022) demonstrou que a ausência de obrigatoriedade de intervenção humana real nas revisões de decisões automatizadas na LGPD abre espaço para que a supervisão se reduza a um gesto formal que não altera o resultado produzido pelo algoritmo.

A primeira condição depende da explicabilidade tratada na seção anterior. Um sistema que não produz justificativa compreensível ao pesquisador não oferece base para contestação. O pesquisador que discorda da saída algorítmica sem entender por que o sistema chegou àquele resultado não exerce supervisão, emite uma opinião concorrente sem base dialógica com o modelo. A OMS estabelece que a explicabilidade e a inteligibilidade são

condições para que a supervisão humana seja real, não apenas nominal, precisamente porque sem elas o profissional não tem acesso ao raciocínio que deveria supervisionar (OMS, 2024).

A segunda condição depende do contexto institucional, não do protocolo. Um médico com oito minutos por consulta numa clínica particular que atende planos de saúde populares não tem as mesmas condições de exercer escrutínio sobre uma saída algorítmica que um especialista numa consulta eletiva de trinta minutos. O protocolo de pesquisa não controla o tempo disponível por paciente no ambiente de uso. Pode, contudo, descrever para qual contexto clínico o sistema foi desenhado e validado, e explicitar que seu uso em contextos com menor capacidade de supervisão exige salvaguardas adicionais.

A terceira condição é a mais diretamente controlável pelo desenho do sistema. Sistemas que apresentam suas saídas como resultados definitivos em vez de recomendações graduadas inibem a discordância por formulação linguística. Sistemas que não registram as discordâncias do pesquisador em relação à saída algorítmica eliminam a trilha de auditoria que tornaria a supervisão verificável. Sistemas que exigem passos adicionais para registrar discordância em relação a passos simplificados para aceitação criam assimetria de esforço que favorece a aceitação passiva. Barucci *et al.* (2026) propõem que os protocolos descrevam explicitamente o mecanismo pelo qual o médico pode bloquear ou sobrepor a decisão algorítmica, e que esse mecanismo seja auditável.

A IA agêntica (agentic AI) representa o ponto mais avançado desse espectro. Diferentemente dos sistemas convencionais, que produzem uma única saída para uma única entrada, agentes de IA planejam sequências de ações, utilizam ferramentas externas, executam etapas encadeadas e tomam decisões intermediárias sem revisão humana entre cada passo. Um agente pode, num único ciclo autônomo, consultar prontuários, cruzar informações com literatura médica, formular hipóteses diagnósticas e propor condutas terapêuticas antes de qualquer pesquisador ter visto o caso. O ponto de supervisão humana desloca-se. Em vez de revisar cada resultado, o profissional aprova ou rejeita um plano completo produzido por um processo não acompanhado passo a passo. Quando o sistema agêntico se destinar a uso clínico, a responsabilidade sobre o ato clínico permanece integral do médico, conforme a Resolução CFM n.º 2.454/2026 (CFM, 2026). Em qualquer fase da pesquisa, exercer a supervisão exige que o protocolo, sob responsabilidade do pesquisador, descreva não apenas o que o sistema produz, mas como cada etapa autônoma foi delimitada, monitorada e tornada contestável. Um CEP que recebe protocolo com componentes agênticos

avalia riscos que aumentam com a autonomia e exige supervisão humana cuidadosamente projetada.

Esse conjunto de requisitos se traduz, na prática, em critérios objetivos de avaliação que orientam a atuação do CEP.

Por isso, o CEP deve formular quatro perguntas sobre qualquer protocolo que envolva decisões algorítmicas com consequência clínica sobre participantes. Cada pergunta endereça uma dimensão distinta da supervisão humana e não pode ser substituída pelas demais.

A primeira pergunta é sobre a formulação da saída do sistema. O protocolo descreve como a saída algorítmica é apresentada ao pesquisador? Um sistema que exibe "diagnóstico: pneumonia" opera diferente de um sistema que exibe "probabilidade de pneumonia: 78%, com base em infiltrado no lobo inferior direito". A formulação linguística da saída influencia diretamente o comportamento do pesquisador. A OMS documenta que a forma como as informações de saúde são apresentadas afeta as decisões clínicas de forma tão significativa quanto o conteúdo dessas informações (OMS, 2024). O protocolo que não descreve a interface de apresentação da saída do sistema não fornece ao CEP informação suficiente para avaliar o risco de viés de automação.

A segunda pergunta é sobre o tempo disponível para revisão. O protocolo descreve o contexto clínico de uso do sistema com suficiente precisão para que o CEP avalie se as condições institucionais permitem supervisão real? Um sistema validado num serviço de radiologia com tempo médio de laudo de vinte minutos pode produzir dinâmica de supervisão completamente distinta quando aplicado numa triagem de pronto-socorro com tempo médio de três minutos por paciente.

A terceira pergunta é sobre o registro de discordâncias. O sistema documenta automaticamente as saídas algorítmicas e as decisões clínicas finais, permitindo identificar divergências e construir trilha de auditoria? Sem esse registro, a supervisão humana declarada no protocolo torna-se inverificável.

O MRCT Center e o WCG (2025) orientam que os protocolos devem descrever os processos que estarão disponíveis para que pesquisadores identifiquem e respondam a eventos adversos relacionados ao desempenho do sistema. O registro de discordâncias é a condição mínima para que esses processos funcionem: sem ele, nenhum evento adverso vinculado à saída algorítmica pode ser rastreado, atribuído ou analisado.

A quarta pergunta é sobre o consentimento informado quanto ao papel do sistema na cadeia de decisão. Os participantes foram informados, no TCLE, de que decisões clínicas sobre seu cuidado durante a pesquisa contarão com apoio de sistema de IA, de qual o grau de influência desse sistema e de como a supervisão humana está garantida? A Lei n.º 14.874/2024 estabelece que o consentimento informado deve cobrir os procedimentos que serão adotados na pesquisa (Brasil, 2024a). Um sistema de IA que influencia decisões diagnósticas ou terapêuticas sobre o participante é procedimento relevante para o consentimento. O protocolo que não o menciona no TCLE produz consentimento incompleto. O CEP que aceita esse TCLE aprova uma pesquisa em que os participantes não sabem como as decisões sobre eles são tomadas.

Aos quatro pontos acima soma-se uma quinta pergunta, decorrente de uma ameaça contemporânea à supervisão humana: a injeção de prompt (*prompt injection*). Trata-se de instrução não autorizada que se infiltra nas entradas de um sistema baseado em grandes modelos de linguagem e subverte o comportamento esperado (OWASP Foundation, 2025). A forma direta vem do próprio usuário; a forma indireta vem embarcada em dados que o sistema processa, como prontuários, páginas consultadas em buscas automatizadas ou artigos integrantes de revisões de literatura assistidas por IA (Greshake *et al.*, 2023). Em pesquisa em saúde, esse vetor pode comprometer a confidencialidade de dados de participantes, falsear sumarizações e codificações qualitativas, e burlar a própria supervisão automatizada. O CEP deve perguntar, em protocolos que envolvam grandes modelos de linguagem com acesso a fontes externas ou a dados sensíveis, se o pesquisador descreve as salvaguardas em vigor contra injeção de prompt direta e indireta, se há separação clara entre instruções do sistema e conteúdo processado, e se há monitoramento de saídas anômalas com revisão humana periódica.

## 3.5 O CICLO DE VIDA DA IA EM PESQUISA: DA COLETA AO DESCARTE

O ciclo de vida de um sistema de IA em pesquisa em saúde não coincide com o ciclo convencional de uma pesquisa clínica. A pesquisa tem início, meio e fim definidos no protocolo. O sistema de IA percorre fases com obrigações éticas distintas, algumas anteriores ao recrutamento dos primeiros participantes e outras posteriores ao encerramento formal do estudo.

Seis fases estruturam esse ciclo. A primeira é a coleta e preparação dos dados: os dados clínicos são reunidos, organizados, pseudonimizados e divididos nos conjuntos de treino, validação e teste. As decisões sobre representatividade demográfica, critérios de inclusão e exclusão e métodos de pré-processamento são tomadas nessa fase e são irreversíveis nas fases subsequentes.

A segunda é o treino do modelo: o algoritmo processa os dados de treino e ajusta seus parâmetros internos. O participante cujos dados alimentam essa fase não tem presença no processo, mas tem direito a ser informado sobre ela no consentimento.

A terceira é a validação e o teste: o desempenho do modelo é avaliado nos conjuntos reservados, com análise de métricas gerais e estratificadas por subgrupo. Barucci *et al.* (2026) estabelecem que essa fase corresponde ao teste em condições reais antes da entrada em uso, com obrigações regulatórias e éticas distintas da fase de treino.

A quarta fase é a entrada em uso na pesquisa: o sistema passa a produzir saídas que influenciam decisões sobre participantes reais. A supervisão humana, a explicabilidade e a documentação de discordâncias tornam-se exigências operacionais, não apenas declarações de intenção.

A quinta fase é o monitoramento durante o estudo: o desempenho do sistema é acompanhado ao longo do tempo para detectar deriva de dados, degradação de acurácia ou surgimento de disparidades entre subgrupos que não eram visíveis na fase de validação.

O MRCT Center e o WCG (2025) recomendam que os protocolos definam limiares predeterminados de mudança no desempenho que acionem revisão ética obrigatória, incluindo a possibilidade de reavaliação pelo CEP.

A sexta fase é a descontinuação ou transferência do sistema: o modelo é retirado de uso ao fim da pesquisa, transferido a outra instituição, licenciado ou publicado em repositório aberto. Cada uma dessas trajetórias tem implicações éticas para os participantes cujos dados formaram o modelo e para as populações que o sistema continuará a afetar.

Conhecer essas seis fases constitui condição necessária para que o CEP avalie se o protocolo submetido cobre o ciclo de vida completo do sistema ou apenas a fase de uso na pesquisa. Um protocolo que descreve com precisão o treino e a validação, mas silencia sobre o destino do modelo após o encerramento, descreve metade do ciclo.

O encerramento formal de uma pesquisa não encerra o ciclo de vida do modelo que ela produziu. Quando o estudo termina, o sistema de IA permanece. Os dados que o

alimentaram permanecem incorporados nos seus parâmetros matemáticos, distribuídos de forma difusa e matematicamente indissociável pela arquitetura do modelo, como tratado na seção 7.1 deste guia. O participante que assinou o TCLE, consentiu com determinada pesquisa, num contexto específico, com finalidade declarada. Não consentiu com o que acontece ao modelo depois que a pesquisa termina.

Os destinos possíveis do modelo após o encerramento da pesquisa são numerosos e cada um carrega implicações éticas distintas. O modelo pode ser transferido a uma instituição parceira nacional ou estrangeira. Pode ser licenciado a uma empresa de tecnologia. Pode ser publicado em repositório aberto de modelos de IA, tornando-se acessível a qualquer pesquisador ou desenvolvedor no mundo. Pode ser incorporado a um produto comercial que será oferecido ao mercado. Pode ser usado como modelo base para transferência de aprendizagem em pesquisas futuras sobre populações que nunca participaram da pesquisa original. Nenhum desses destinos é eticamente neutro. Todos eles prolongam a vida dos dados dos participantes em sistemas que operam além do alcance do consentimento original (MRCT Center; WCG, 2025).

A literatura brasileira e internacional documenta casos em que dados de saúde coletados para finalidade de pesquisa foram subsequentemente usados para desenvolver produtos comerciais sem novo consentimento dos participantes. O caso DeepMind-Royal Free NHS, tratado na seção 7.4 deste guia, é o exemplo mais documentado desse padrão. A Lei n.º 14.874/2024 estabelece que qualquer uso secundário de dados coletados em pesquisa exige avaliação ética específica (Brasil, 2024a). A LGPD determina que o tratamento de dados pessoais para finalidade distinta da original requer nova base legal ou novo consentimento (Brasil, 2018). Ambas as normas se aplicam ao modelo tanto quanto aos dados brutos. O modelo não é produto neutro da pesquisa. É repositório dos dados que o formaram.

O risco não se limita ao modelo completo. Técnicas de extração de conhecimento (*knowledge distillation*) permitem que um modelo derivado aprenda os padrões do modelo original sem acesso direto aos dados brutos. Modelos publicados em repositórios abertos podem ser submetidos a ataques de inversão de gradiente que reconstituem fragmentos dos dados originais, como documentado por Geiping *et al.* (2020). O participante que consentiu com a coleta dos seus dados para uma pesquisa específica não tem mecanismo de controle sobre nenhum desses usos derivados. O protocolo que não endereça o destino do modelo

após o encerramento da pesquisa não apenas omite uma informação. Cria uma zona de opacidade deliberada sobre a trajetória mais duradoura dos dados dos participantes.

O protocolo de pesquisa com IA deve cobrir o ciclo de vida do sistema do primeiro dado coletado à descontinuação ou transferência do modelo. Três exigências concretas estruturam a avaliação do CEP sobre essa dimensão.

A primeira exigência é a descrição do destino do modelo após o encerramento da pesquisa. O protocolo deve indicar explicitamente o que acontecerá ao modelo resultante: se será descontinuado e deletado, transferido a instituição parceira com identificação precisa e condições contratuais descritas, publicado em repositório com especificação das condições de acesso, ou incorporado a produto ou serviço com identificação do destinatário. Quando o destino previsto for a transferência a terceiros ou a publicação em repositório aberto, o protocolo deve descrever as salvaguardas que impedirão usos secundários incompatíveis com o consentimento dos participantes. O TCLE deve informar os participantes sobre esse destino em linguagem acessível, antes da assinatura. A Lei n.º 14.874/2024 estabelece que o consentimento deve cobrir os procedimentos da pesquisa (Brasil, 2024a). O destino do modelo é procedimento com consequências que se estendem além do encerramento do estudo e deve constar do consentimento com a mesma clareza que qualquer intervenção direta sobre o participante.

A segunda exigência é a definição de critérios de descontinuação do sistema durante a pesquisa. O protocolo deve estabelecer quais condições acionam a suspensão do uso do sistema antes do encerramento previsto do estudo: um limiar de degradação de desempenho abaixo do qual o sistema não oferece mais benefício clínico justificado, a identificação de disparidade de desempenho entre subgrupos que configure dano diferencial a populações específicas, ou a ocorrência de eventos adversos vinculados à saída do sistema que superem os riscos previstos.

O MRCT Center e o WCG (2025) recomendam que os protocolos descrevam os processos disponíveis para que pesquisadores identifiquem e respondam a eventos adversos relacionados ao desempenho do sistema, incluindo critérios predeterminados de paralisação. Protocolos que não definem esses critérios operam sem freio de segurança: o sistema continua em uso até o encerramento programado independentemente do que acontece com seu desempenho real.

A terceira exigência é a cobertura do ciclo de vida completo no TCLE. O termo de consentimento de um protocolo com IA deve informar os participantes sobre seis pontos que os modelos convencionais não preveem: a finalidade específica do modelo que será treinado com seus dados; o destino do modelo após o encerramento da pesquisa; as condições de transferência a terceiros e eventual uso comercial; o mecanismo técnico pelo qual o participante pode exercer o direito de retirada após o treinamento do modelo; o período durante o qual o modelo permanecerá ativo; e o canal pelo qual o participante pode obter informações sobre o uso do modelo após o encerramento. Fornasier (2022) demonstrou que a ausência dessas informações nos TCLEs de pesquisa com IA cria zona de opacidade jurídica que compromete tanto os direitos dos participantes quanto a validade ética da pesquisa.

O parecer do CEP ao aprovar uma pesquisa com IA, aprova o ciclo de vida completo do sistema que a pesquisa vai construir, não apenas a pesquisa que o protocolo descreve. Esse sistema continuará a existir e a operar depois que o estudo for encerrado, depois que os pesquisadores passarem a outros projetos e depois que os participantes deixarem de ser acompanhados. A avaliação ética que não alcança esse horizonte aprova apenas a fração visível de uma cadeia de consequências que se estende muito além dela.

# 4 O CONTEXTO BRASILEIRO

Este capítulo posiciona este guia no cenário brasileiro contemporâneo e destaca a inserção da Inteligência Artificial (IA) na pesquisa em saúde. Nenhum dos modelos regulatórios estrangeiros analisados no Capítulo 5 foi concebido para a escala, a diversidade populacional e a arquitetura federativa do SUS. O percurso a seguir parte dessa singularidade (4.1), percorre a arquitetura normativa em camadas que o Brasil construiu em menos de uma década (4.2), examina a infraestrutura de dados sobre a qual uma IA soberana pode ser desenvolvida (4.3), descreve a política estatal que a agencia (4.4), identifica os riscos estruturais que ela reproduz quando mal concebida (4.5) e conclui com as lacunas normativas que o CEP precisa preencher hoje, na prática (4.6). Cada seção funciona como unidade autônoma de leitura.

## 4.1 UM SISTEMA DE SAÚDE QUE DESAFIA QUALQUER MODELO IMPORTADO

O Sistema Único de Saúde (SUS) atende o conjunto da população brasileira. Funciona em 5.571 municípios (Brasil, 2026e). Abrange desde grandes centros universitários até postos de atenção primária em comunidades em locais isolados no Brasil profundo. Nenhum outro sistema de saúde universal do mundo reúne, em escala comparável, essa combinação de diversidade étnica, desigualdade regional e compromisso constitucional com a universalidade. Qualquer tecnologia que pretenda operar nesse sistema precisa conhecê-lo antes de entrar nele.

A IA segue o mesmo princípio. Os modelos internacionais analisados neste guia foram construídos para contextos com infraestrutura tecnológica mais homogênea, populações mais demograficamente uniformes e sistemas de saúde organizados de forma diferente. O modelo canadense foi concebido para serviços públicos governamentais. O australiano, para um sistema de alta renda com cobertura digital razoavelmente distribuída. O indiano, o mais próximo do SUS em escala e complexidade, ainda carece da base normativa protetiva que o Brasil construiu com a Lei n.º 13.709, de 14 de agosto de 2018, Lei Geral de Proteção de Dados Pessoais LGPD (Brasil, 2018). Nenhum deles foi pensado para os povos originários e ribeirinhos da Amazônia, nem para os territórios quilombolas.

Trata-se de uma afirmação de especificidade, não de lamentação. Com o avanço rápido da IA na área da saúde, o Brasil já possui uma base inicial e não parte do zero. Parte de um lugar distinto, com desafios próprios e com ativos que muitos países não têm. O SUS é, ao mesmo tempo, o maior desafio e a maior potência para o desenvolvimento de uma abordagem brasileira de avaliação ética da IA em pesquisa em saúde.

## 4.2 O ARCABOUÇO NORMATIVO BRASILEIRO: UMA CONSTRUÇÃO EM CAMADAS

A regulação da IA na saúde no Brasil constrói-se em camadas. Cada instrumento respondeu a uma questão social concreta: a proteção de dados pessoais, a ética em pesquisa com seres humanos, a segurança de dispositivos médicos digitais, a integridade científica, a proteção de crianças e adolescentes em ambientes digitais. Articulados, esses instrumentos formam um dos marcos protetivos mais robustos entre os países que hoje enfrentam o avanço acelerado da IA em saúde.

A LGPD é o pilar dessa arquitetura. A lei classifica dados de saúde como dados pessoais sensíveis. Exige base legal específica, finalidade declarada e identificação do Encarregado de Dados para seu tratamento. Dourado e Aith (2022) demonstraram que a LGPD introduz no ordenamento jurídico brasileiro o direito à explicação e à revisão de decisões automatizadas (art. 20). Esse direito é a expressão normativa do princípio da transparência algorítmica, ponto central na regulação de sistemas de IA. A regulação da IA na saúde no Brasil começa com a LGPD. Matsushita e Almeida (2025) observam que a articulação entre a LGPD e as resoluções sanitárias específicas constitui o marco protetivo de dados de saúde mais avançado da América Latina.

A Lei n.º 14.874, de 28 de maio de 2024, e o Decreto n.º 12.651, de 7 de outubro de 2025, que a regulamenta, elevaram a ética em pesquisa com seres humanos ao patamar de lei federal e instituíram o Sistema Nacional de Ética em Pesquisa com Seres Humanos (SINEP) (Brasil, 2024a; 2025a). O texto da lei não utiliza os termos "algoritmo" ou "inteligência artificial". Suas disposições, contudo, aplicam-se diretamente ao processamento de grandes bases de dados e ao desenvolvimento de programas de computador com finalidade diagnóstica ou de monitoramento.

A Resolução CNS n.º 738, de 1.º de fevereiro de 2024, complementa esse marco. A resolução orientou os CEPs sobre as obrigações éticas na constituição e gestão de repositórios

de dados para pesquisa científica e condicionou a transferência de dados identificadores a terceiros — inclusive a desenvolvedores de sistemas de IA — à previsão no protocolo, à justificativa e à aprovação do comitê, cabendo sua execução ao Controlador (art. 12 e parágrafos; Brasil, 2024b).

A Agência Nacional de Vigilância Sanitária (Anvisa) integra-se a esse arcabouço com a Resolução de Diretoria Colegiada n.º 657, de 24 de março de 2022. A norma estabeleceu o marco regulatório para Programas de Computador como Dispositivos Médicos (Software as a Medical Device — SaMD). Um sistema de IA enquadra-se nessa categoria quando sua finalidade abrange diagnóstico, prevenção, monitoramento, tratamento ou compensação de doenças (Anvisa, 2022). O ciclo de vida do sistema passa a ter dois momentos obrigatórios e sequenciais. O CEP resguarda os direitos dos participantes durante a pesquisa. A Anvisa atesta que o produto derivado é seguro para uso em escala. Um não substitui o outro.

A Política de Integridade na Atividade Científica do CNPq, instituída em março de 2026, determina que todo uso de IA generativa seja declarado, com especificação da ferramenta e da finalidade, em qualquer fase da pesquisa (CNPq, 2026).

A Lei n.º 15.211, de 17 de setembro de 2025, denominada Estatuto Digital da Criança e do Adolescente (ECA Digital), acrescenta ao arcabouço uma camada de proteção específica para pesquisas que envolvam crianças e adolescentes em ambientes digitais (Brasil, 2025b). A lei proíbe o perfilamento comportamental de menores para qualquer finalidade que não seja estritamente necessária à operação do serviço. Exige configurações de privacidade no nível mais elevado de proteção desde a concepção do produto. Determina a revisão regular das ferramentas de IA, com participação de especialistas e órgãos competentes, para garantir sua segurança e adequação ao uso por esse público. Para protocolos de pesquisa com IA que envolvam crianças e adolescentes, o protocolo deve demonstrar conformidade com esse estatuto, e o CEP a verifica antes de qualquer outra análise. A proteção integral não é princípio abstrato. É obrigação legal com vigência desde março de 2026.

O Plano Brasileiro de Inteligência Artificial 2024–2028 projeta investimentos da ordem de R$ 23 bilhões até 2028 e posiciona a saúde como setor prioritário (Brasil, 2025c). A velocidade desse avanço torna ainda mais urgente a consolidação do marco ético. O Projeto de Lei n.º 2.338, de 2023, de autoria do senador Rodrigo Pacheco, propõe o Marco Regulatório da Inteligência Artificial no Brasil e prevê obrigações específicas para sistemas de alto risco, categoria que abrange aplicações clínicas e diagnósticas (Brasil, 2023a). O texto do PL também

prevê mecanismos de *sandbox* regulatório para o teste controlado de sistemas de alto risco. Enquanto tramita no Congresso Nacional, o arcabouço vigente já oferece base normativa suficiente para a avaliação ética de protocolos com IA. O CEP não aguarda nova lei para agir. Age com o que o direito brasileiro já construiu.

## 4.3 A REDE NACIONAL DE DADOS EM SAÚDE E A INFRAESTRUTURA PARA A IA

A Rede Nacional de Dados em Saúde (RNDS) é uma plataforma nacional de interoperabilidade criada para conectar sistemas de informação em saúde de diferentes origens, públicos e privados, em todo o território brasileiro (Brasil, 2020a). Sua lógica está na padronização da linguagem com que sistemas distintos se comunicam. Uma informação produzida em um posto de atenção primária no Maranhão torna-se legível por um sistema hospitalar em São Paulo. O padrão técnico adotado é o HL7 FHIR (Health Level Seven — Fast Healthcare Interoperability Resources, ou Recursos Rápidos para Interoperabilidade em Saúde do Nível Sete), versão 4.0.1, referência internacional para interoperabilidade em saúde digital.

Esse padrão estrutura os dados clínicos em recursos discretos e legíveis por máquina. Cada registro torna-se um objeto computacional com estrutura previsível, consultável e integrável a outros sistemas. Uma consulta, um exame, uma dispensação de medicamento: todos seguem a mesma lógica de organização. Para a IA, essa estrutura é condição de entrada. Um sistema de IA que opere com dados do SUS depende de dados organizados, consistentes e padronizados. A RNDS oferece essa infraestrutura em escala nacional. Sem ela, a soberania brasileira no desenvolvimento de IA em saúde permaneceria promessa sem sustentação técnica.

Os dados que trafegam pela rede abrangem registros de vacinação, resultados de exames laboratoriais, dispensação de medicamentos, sumários de alta hospitalar e registros de atendimentos clínicos (Brasil, 2020a). O acesso ocorre por canais distintos conforme o perfil do usuário. Profissionais de saúde acessam pelo SUS Digital Profissional. Cidadãos acessam pelo Meu SUS Digital. Pesquisadores acessam por meio de extração de dados agregados e desidentificados, com salvaguardas ancoradas na LGPD.

O acesso de pesquisadores a dados da RNDS, seja por extração agregada ou por projeto específico, é disciplinado pela Resolução CNS n.º 738/2024, que estabelece as normas

na constituição e no uso de bancos de dados com finalidade de pesquisa científica envolvendo seres humanos. A Resolução define as figuras do Controlador e do Operador, exige a apresentação de Termos institucionais conforme a origem do banco — em especial o Termo de Anuência Institucional quando o banco é constituído fora do âmbito da pesquisa —, e regula, nos arts. 20 e 25, o consentimento para uso futuro de dados e as hipóteses de dispensa de Termo de Consentimento Livre e Esclarecido (TCLE) (Brasil, 2024b). O Capítulo 7 desenvolve essas figuras e os Termos institucionais em detalhe; o Capítulo 8 os integra como critérios operacionais do Eixo 3 da Matriz de Avaliação de Risco em Inteligência Artificial.

A distinção entre uso assistencial e uso em pesquisa tem consequências éticas diretas. No uso assistencial, o dado circula sob as bases legais do cuidado em saúde, com finalidade clínica imediata e sob o controle do profissional responsável. No uso em pesquisa, a base legal muda, o consentimento ganha papel central e a finalidade precisa ser declarada com precisão. A anonimização, nesse contexto, é salvaguarda necessária, mas não suficiente. Rocher, Hendrickx e Montjoye (2019) demonstraram que bases aparentemente anonimizadas podem permitir a reidentificação de indivíduos a partir do cruzamento de poucos atributos quase-identificadores. Em um acervo do tamanho da RNDS, esse risco se multiplica. Um protocolo com dados da RNDS precisa demonstrar não apenas que os dados foram anonimizados, mas em que condições essa anonimização resiste ao cruzamento com outras bases disponíveis.

Para a pesquisa com IA, esse acervo tem valor estratégico. Reúne dados produzidos pelo conjunto da população brasileira, em contextos clínicos, epidemiológicos e sociais de enorme diversidade. Nenhum consórcio hospitalar privado, nacional ou estrangeiro, reproduz essa amplitude. Um modelo de IA treinado sobre dados da RNDS aprende com o Brasil real, não com um recorte dele. Essa distinção define a diferença entre uma ferramenta útil ao SUS e uma ferramenta transferida para o SUS.

Soares e Chiavegatto Filho (2026) alertam que modelos desenvolvidos com dados estrangeiros ou de hospitais privados do Sudeste podem ser ineficazes ou prejudiciais ao serem aplicados em populações com diferentes perfis epidemiológicos.

O problema é estrutural: os dados que treinam a IA refletem desigualdades históricas. Quando o modelo aprende com dados de biobancos britânicos ou de hospitais privados paulistanos, aprende também os padrões de exclusão embutidos nesses acervos.

Essa lógica já produziu consequências verificáveis no Brasil: algoritmos de predição de dengue, de vigilância da COVID-19 e de análise de declarações de óbito por processamento de

linguagem natural funcionam com desempenho aceitável quando validados nas populações que geraram seus dados de treino, e a validade dos modelos colapsa quando são aplicados a contextos epidemiológicos e sociais distintos, sem revalidação local (Soares; Chiavegatto Filho, 2026).

A RNDS responde a esse risco com a única contramedida estruturalmente viável: dados produzidos pelo Brasil real, em escala nacional, sobre a totalidade das pessoas que o SUS atende. Um modelo treinado sobre esse acervo tem chance de aprender o que nenhum dado importado ensinaria. A disputa pela soberania dos dados é, portanto, também uma disputa pela utilidade e pela justiça da IA na saúde pública brasileira.

## 4.4 A SEIDIGI E A CONSTRUÇÃO DA SAÚDE DIGITAL BRASILEIRA

A Secretaria de Informação e Saúde Digital, conhecida pela sigla SEIDIGI, foi criada pelo Decreto n.º 11.358, de 1.º de janeiro de 2023, como braço executor da política de saúde digital do Ministério da Saúde. A missão institucional da Secretaria abrange o planejamento, a incorporação e a supervisão de produtos e serviços de tecnologia da informação no SUS, incluindo a gestão da RNDS, o desenvolvimento de sistemas de interoperabilidade e a proteção de dados em saúde (Brasil, 2023b). Para o campo da IA, a SEIDIGI ocupa posição de protagonista na definição dos padrões técnicos que qualquer sistema de IA precisará respeitar para operar com dados do SUS. O Decreto n.º 11.358/2023 foi revogado pelo Decreto n.º 11.798, de 28 de novembro de 2023, que aprovou a Estrutura Regimental vigente do Ministério da Saúde, na qual a SEIDIGI foi mantida (Brasil, 2023c).

Haddad *et al.* (2025) documentaram a trajetória que conduziu à criação da secretaria. O Brasil acumulou, ao longo de quase duas décadas, experiência pioneira em telessaúde pública. O Programa Nacional Telessaúde Brasil, iniciado em 2007, antecedeu a maioria dos países de renda média na organização de redes de teleconsultoria e telediagnóstico no sistema público de saúde. A pandemia de COVID-19 acelerou o marco regulatório, com a Lei n.º 14.510, de 27 de dezembro de 2022 (Brasil, 2022), que autorizou e disciplinou a telessaúde em todo o território nacional. Essa trajetória não foi linear. Enfrentou resistências corporativas, lacunas de financiamento e fragmentação sistêmica. A SEIDIGI nasce justamente para superar essa fragmentação.

O que distingue a SEIDIGI de secretarias anteriores com funções similares é a escala e a integração das atribuições. A SEIDIGI incorpora o DATASUS, historicamente responsável pelos sistemas de informação do SUS, e articula as ações por meio do Comitê Gestor de Saúde Digital, que reúne a Agência Nacional de Vigilância Sanitária, a Agência Nacional de Saúde Suplementar, o Conselho Nacional de Secretários de Saúde e o Conselho Nacional de Secretarias Municipais de Saúde (Haddad *et al.*, 2025). Essa arquitetura de governança transforma a saúde digital de iniciativa isolada em política de Estado com instâncias formais de coordenação federativa.

O Programa SUS Digital, instituído pela Portaria GM/MS n.º 3.232, de 1º de março de 2024, é a expressão operacional da agenda da SEIDIGI. O programa organiza a saúde digital em ações estratégicas que abrangem telessaúde, interoperabilidade, prontuário eletrônico, identidade digital do cidadão e inteligência artificial aplicada à gestão e à assistência (Brasil, 2024c). Cada ação tem metas, responsáveis e critérios de avaliação definidos. A saúde digital deixa de ser projeto e passa a ser programa estruturado, com lógica de ciclo de vida e prestação de contas.

O Chamamento Público n.º 01/2026 do Inova SUS (Brasil, 2026a), coordenado pela SEIDIGI, avança nessa direção com uma inovação de método. O Estado brasileiro deixa de ser apenas regulador de tecnologias desenvolvidas pelo setor privado e passa a contratá-las com critérios éticos e técnicos previamente definidos. Os editais especificam requisitos de transparência algorítmica, de representatividade dos dados de treino e de compatibilidade com a RNDS. Um sistema que não respeita esses critérios não passa da seleção. A ética deixa de ser exigência posterior ao desenvolvimento e passa a ser condição de entrada.

Segundo Santos *et al.* (2025), iniciativas como o programa federal "Agora Tem Especialistas", a gradual federalização da RNDS e o Inova SUS constituem uma agenda chamada soberania sanitária digital. Essa expressão refere-se à aptidão de um país do Sul Global para criar e controlar tecnologias em saúde baseadas nas próprias populações, sob gestão pública e com foco na equidade. Essa direção tem tradução técnica concreta. Uma IA desenvolvida especificamente para a realidade do SUS, considerando as condições epidemiológicas e desigualdades regionais brasileiras, difere técnica e eticamente de uma IA importada e adaptada.

Esse movimento tem consequência direta para a pesquisa com IA. Sistemas desenvolvidos em contexto de pesquisa e destinados a operar no SUS precisarão, em algum

momento, passar pelo crivo da SEIDIGI. O protocolo de pesquisa submetido ao CEP não é o produto final. É a etapa que precede a avaliação regulatória e tecnológica. Quanto mais rigorosa a avaliação ética do protocolo, maior a probabilidade de que o sistema resultante seja aprovado nas etapas subsequentes. CEP e SEIDIGI não competem. Integram momentos distintos de uma mesma cadeia de responsabilidade.

## 4.5 VIÉS ALGORÍTMICO, DESIGUALDADE REGIONAL E O RISCO DO MODELO IMPORTADO

O viés algorítmico é herança, não defeito de programação. Os dados que treinam sistemas de IA refletem os padrões de quem os produziu, de quem usa o sistema de saúde e daquilo que foi considerado suficientemente qualificado para ser registrado. Quando esses padrões refletem exclusão histórica, o algoritmo aprende a exclusão. Soares e Chiavegatto Filho (2026) documentam que mesmo modelos de última geração reproduzem vieses de suas bases de treino, incluindo o chamado viés do voluntário saudável (*healthy volunteer bias*) e padrões sistemáticos de ausência de dados associados a processos de coleta com alcance limitado.

O Brasil reúne condições que tornam esse risco mais agudo do que em países com infraestrutura sanitária homogênea. A desigualdade regional não é variação demográfica. Perfis epidemiológicos distintos habitam o mesmo país: a carga de arboviroses no Norte e Nordeste, a mortalidade materna entre mulheres negras, a tuberculose em populações carcerárias e em situação de rua, as doenças negligenciadas que não aparecem nos grandes consórcios hospitalares privados do Sudeste. Um modelo treinado sobre dados de hospitais universitários de São Paulo aprende um Brasil que não existe para a maioria das pessoas atendidas pelo SUS.

O racismo algorítmico é o nome preciso para o que acontece quando esse processo se concentra sobre a população afro-brasileira. Dados mal coletados, ausência sistêmica de determinadas populações nos conjuntos de treino e ajustes técnicos cegos à raça/cor convergem para produzir sistemas que, sob aparência de neutralidade matemática, reproduzem padrões de exclusão construídos ao longo de séculos. Não se trata de falha acidental. É o resultado previsível de ensinar um algoritmo sobre registros que nunca foram pensados para captar a experiência das populações excluídas.

A literatura internacional confirma essa assimetria com precisão crescente. Rajkomar *et al.* (2018) demonstraram que modelos preditivos com desempenho aceitável em

populações majoritárias produzem resultados substancialmente piores em grupos historicamente sub-representados nos dados de treino. O problema não se resolve com mais dados. Resolve-se com os dados certos, das populações certas, coletados com os protocolos certos. Essa é a razão pela qual a representatividade dos dados de treino deixou de ser questão metodológica e passou a ser questão ética com respaldo normativo expresso nas principais diretrizes internacionais analisadas neste guia.

A questão racial acrescenta uma camada de complexidade que nenhum ajuste técnico resolve sozinho. Dados de raça e etnia são indispensáveis para revelar iniquidades em saúde. A mortalidade materna entre mulheres negras, a subnotificação de doenças falciformes em regiões com baixa cobertura laboratorial e o acesso diferencial a exames diagnósticos compõem padrões que só aparecem quando a variável racial está presente e bem coletada. Fiske *et al.* (2025) demonstraram que o uso dessas variáveis sem contextualização adequada pode reproduzir estigmas e reforçar discriminações estruturais no próprio algoritmo que deveria corrigi-las. O dado racial mal usado transforma a ferramenta de equidade em instrumento de exclusão.

A explicabilidade é o outro lado desse problema. Modelos com baixa transparência, frequentemente chamados de caixas-pretas (*black boxes*), impedem que profissionais e usuários compreendam como uma decisão foi gerada. Essa opacidade fragiliza o controle social, dificulta a responsabilização institucional em caso de erro e torna impossível identificar em qual etapa do processo o viés foi introduzido (Soares; Chiavegatto Filho, 2026). No contexto do SUS, onde a decisão algorítmica pode afetar o acesso a um procedimento, a regulação de uma fila ou a triagem de um paciente, a opacidade tem consequências que vão além da técnica e afetam direitos.

A desigualdade no letramento algorítmico entre os próprios profissionais de saúde amplifica esse risco. Soares e Chiavegatto Filho (2026) identificaram que a maioria das categorias profissionais ainda carece de preparo para interpretar algoritmos preditivos, compreender a probabilidade de desfechos e reconhecer quando um sistema opera fora dos parâmetros para os quais foi validado. Um profissional sem letramento algorítmico não questiona o modelo. Aceita o resultado como dado. A automação do erro, nesse cenário, é mais rápida e mais abrangente do que qualquer falha humana individual poderia ser.

A avaliação da representatividade dos dados de treino é obrigação ética da pesquisa, não detalhe metodológico a ser verificado apenas pelos revisores de uma revista científica. O

protocolo submetido ao comitê precisa responder a perguntas concretas: para qual população o modelo foi treinado? Essa população compartilha o perfil epidemiológico, social e demográfico das pessoas que serão afetadas pela pesquisa? Os dados de treino incluem grupos historicamente sub-representados nos sistemas de informação em saúde brasileiros? O MRCT Center e o WCG (2025) posicionam essas perguntas como critérios de entrada na avaliação ética de qualquer protocolo com IA. O guia que o leitor tem nas mãos adota o mesmo princípio.

A ferramenta desenvolvida pela Organização Pan-Americana da Saúde (Opas) para avaliação da prontidão institucional de países para a adoção de IA em saúde pública incorpora dimensões técnicas, jurídicas, éticas e organizacionais (OPS/OMS, 2024). A dimensão ética inclui, explicitamente, a avaliação de equidade e justiça algorítmica como condição para a adoção responsável de qualquer sistema. Não é preciso dominar os fundamentos matemáticos de um modelo de aprendizado de máquina. O essencial é fazer as perguntas certas. A Matriz de Avaliação de Risco em Inteligência Artificial apresentada no Capítulo 8 deste guia organiza essas perguntas em cinco eixos, dos quais o Eixo 4, dedicado à representatividade dos dados e à transferibilidade do modelo, responde diretamente ao risco do viés herdado e do racismo algorítmico.

A equidade algorítmica não é aspiração futura. É exigência normativa presente. Rajkomar *et al.* (2018) demonstraram que garantir equidade nos modelos preditivos em saúde é condição para o avanço da equidade em saúde como um todo. No Brasil, onde o SUS carrega o compromisso constitucional de universalidade e equidade, essa afirmação tem peso normativo. Um sistema de IA que reproduz desigualdades no acesso ao diagnóstico ou no resultado do tratamento contraria os fundamentos do sistema público de saúde. Avaliar esse risco antes que o sistema entre em uso é a contribuição mais concreta que o CEP pode dar ao projeto de saúde que o Brasil escolheu construir.

O material operacional correspondente está nas Partes V e VI da MARIAH.

# 5 PANORAMA INTERNACIONAL: EXPERIÊNCIAS DE AVALIAÇÃO ÉTICA E REGULAÇÃO DE IA EM PESQUISA EM SAÚDE

## 5.1 INTRODUÇÃO: POR QUE OLHAR PARA FORA

A experiência internacional em regulação de inteligência artificial na saúde avançou com rapidez nos últimos anos. A OMS, juntamente com países europeus, Canadá, Austrália, Estados Unidos e nações do leste asiático, estabeleceram estruturas regulatórias, instrumentos de avaliação ética e diretrizes para comitês de pesquisa que proporcionam contribuições concretas e verificáveis. O Brasil aprende com cada uma dessas experiências. Não as importa, mas aprende com elas para construir a sua.

O percurso que levou à construção da Matriz de Avaliação de Risco em Inteligência Artificial em Pesquisa com Seres Humanos (MARIAH) do Sistema Nacional de Ética em Pesquisa (SINEP), descrita no Capítulo 8, passou pelo quadro ético da OMS e por nove modelos regulatórios em três continentes. Nenhum deles foi concebido para o contexto epidemiológico e social brasileiro, para a diversidade étnica e regional da população atendida pelo SUS ou para o marco protetivo da LGPD. Essas especificidades não são lacunas a lamentar. São o espaço onde o modelo brasileiro se diferencia e se justifica.

O panorama a seguir parte do quadro ético multilateral da OMS e examina nove experiências regulatórias nacionais: Canadá, Austrália, Estados Unidos, Europa, China, Japão, Coreia do Sul, Índia e Rússia. Para cada experiência, o texto identifica o marco regulatório relevante para comitês de ética em pesquisa, o papel atribuído à supervisão ética de sistemas algorítmicos e a contribuição específica que cada modelo ofereceu à MARIAH. A seção 5.9 sintetiza as lições e explicita o que o contexto brasileiro acrescenta ao que nenhuma dessas experiências previu.

## 5.2 O QUADRO ÉTICO MULTILATERAL: A ORGANIZAÇÃO MUNDIAL DA SAÚDE COMO REFERÊNCIA FUNDACIONAL

O primeiro instrumento normativo de alcance global sobre o tema é a Recomendação sobre a Ética da Inteligência Artificial, adotada pelos Estados-membros da UNESCO em 23 de novembro de 2021, que estabelece o respeito aos direitos humanos e à dignidade humana como fundamento e princípios como transparência, equidade, sustentabilidade e supervisão

humana dos sistemas de IA (UNESCO, 2022). A governança ética da IA em saúde tem na OMS sua referência multilateral mais consolidada. Em 2021, a OMS publicou o documento Ethics and Governance of Artificial Intelligence for Health, resultado de consulta com vinte especialistas internacionais. O documento estabeleceu seis princípios éticos consensuais para orientar governos, desenvolvedores, pesquisadores e prestadores de serviços de saúde no desenvolvimento e na adoção de sistemas de IA (OMS, 2021).

Os seis princípios orientam o debate ético internacional desde então. 1)O primeiro é a proteção da autonomia: os seres humanos devem permanecer no controle das decisões clínicas, com acesso às informações necessárias para usar os sistemas de IA com segurança e eficácia. 2)O segundo é a promoção do bem-estar humano, da segurança e do interesse público: os sistemas devem satisfazer requisitos regulatórios de segurança e eficácia para usos bem definidos. 3)O terceiro é a garantia de transparência, explicabilidade e inteligibilidade: os sistemas de IA devem ser compreensíveis para desenvolvedores, profissionais de saúde, pacientes e reguladores. 4) O quarto é o fomento à responsabilidade e à prestação de contas: o uso de IA deve ocorrer sob condições apropriadas, com pontos de supervisão humana estabelecidos tanto antes quanto depois do algoritmo. 5) O quinto é a garantia de inclusão e equidade: os sistemas devem ser projetados para uso amplo e equitativo, sem codificar vieses em detrimento de grupos identificáveis. 6) O sexto é a promoção de IA responsiva e sustentável: os sistemas devem ser consistentes com a sustentabilidade dos sistemas de saúde, do ambiente e dos locais de trabalho (OMS, 2021).

Esses seis princípios não nasceram da especulação. A OMS os formulou a partir de evidências concretas sobre os riscos dos sistemas de apoio diagnóstico, dos modelos preditivos e dos sistemas de triagem que já operavam em contextos clínicos reais em 2021. Países de renda média e baixa enfrentam desafios específicos na adoção de IA em saúde, como a dependência de modelos desenvolvidos em contextos com infraestrutura superior e a falta de dados representativos de suas populações. Essa dimensão de equidade global ressoa diretamente no contexto do SUS e fundamenta as exigências de validação local que este guia estabelece.

Em 2024, a OMS publicou uma atualização específica para os grandes modelos multimodais (Large Multi-Modal Models — LMMs): Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. O documento reconhece que os princípios de 2021 permanecem válidos, mas que o surgimento dos LMMs introduziu

desafios novos que exigem atenção específica (OMS, 2024). Os modelos generativos foram adotados mais rapidamente do que qualquer outra aplicação de consumo na história, sem que indivíduos, profissionais de saúde ou governos estivessem preparados para seus riscos.

A atualização de 2024 identifica três dimensões de risco ausentes no debate de 2021. A primeira é a fabricação de conteúdo: os LMMs produzem respostas que parecem autoritativas, mas podem ser factualmente incorretas, incluindo citações científicas inexistentes e recomendações clínicas inadequadas. A segunda é a opacidade das bases de treinamento: os dados usados para treinar os principais LMMs disponíveis não foram divulgados, tornando impossível verificar se são enviesados, se foram obtidos legalmente e se o desempenho do modelo reflete capacidade genuína ou memorização dos dados de treinamento. A terceira é o controle exclusivo por grandes corporações: a concentração do desenvolvimento de LMMs em poucas empresas privadas cria dependência estrutural e assimetria de poder que os sistemas de saúde públicos precisam reconhecer e enfrentar.

Os documentos da OMS constituem, neste guia, o padrão ético mínimo para a aprovação de protocolos com IA, independentemente da sofisticação técnica ou prestígio da instituição. Os seis princípios de 2021 e as recomendações específicas para LMMs de 2024 são parâmetros operacionais. Este guia os operacionaliza no contexto do SINEP por meio da MARIAH e das diretrizes desenvolvidas nos capítulos seguintes.

## 5.3 CANADÁ

O modelo canadense é o mais diretamente influente na construção da MARIAH. O Secretariado do Conselho do Tesouro do Canadá desenvolveu e mantém a Avaliação de Impacto Algorítmico (Algorithmic Impact Assessment — AIA), instrumento de uso obrigatório para sistemas de IA implantados por órgãos do governo federal canadense (Canadá, 2024). O AIA opera com perguntas de resposta binária distribuídas em quatro dimensões de risco: impacto, contexto do projeto, dados e tecnologia. Classifica os sistemas em quatro níveis, cada um associado a requisitos mínimos obrigatórios. A regra de que mitigações não alteram o nível de risco, apenas os requisitos exigidos, tem origem direta nesse modelo.

Dois elementos do AIA foram incorporados à MARIAH com adaptações. O primeiro é a lógica de pontuação ponderada da Versão B: blocos temáticos com pesos diferenciados, cuja soma determina o nível de risco em escala numérica. O segundo é a separação entre

classificação de risco e requisitos exigidos, que preserva a integridade do instrumento e impede que declarações de intenção substituam evidências de controle efetivo do risco. O AIA foi concebido para serviços públicos governamentais, não para protocolos submetidos a comitês de ética. Essa transposição exigiu reformulação dos eixos, das perguntas e da base normativa do instrumento.

O Canadá é referência também pelo Enunciado de Política das Três Agências: Ética da Pesquisa com Seres Humanos (TCPS 2), documento que orienta os comitês de ética em pesquisa canadenses. Em sua edição mais recente, o TCPS 2 trata do uso de IA em pesquisa e estabelece que a supervisão ética deve acompanhar o ciclo de vida completo da pesquisa, da concepção à disseminação dos resultados (Canadá, 2022). Esse princípio orienta a estrutura de passos da MARIAH.

## 5.4 AUSTRÁLIA

O modelo australiano é a principal referência para a lógica qualitativa da Versão A (qualitativa) da MARIAH. O Conselho Nacional de Saúde e Pesquisa Médica da Austrália (National Health and Medical Research Council — NHMRC) publicou em 2023 o Guia para Avaliação de Pesquisa Envolvendo IA (*Guide for Assessing Research Involving AI*), documento concebido especificamente para orientar comitês de ética em pesquisa na avaliação de protocolos com sistemas algorítmicos (Austrália, 2023). O guia australiano parte de um princípio que orienta todo o Capítulo 8 deste guia: o comitê de ética não precisa compreender o código do algoritmo. Precisa compreender o impacto que ele pode produzir sobre a pessoa participante.

O NHMRC organiza a avaliação em torno do ciclo de vida do sistema de IA e distingue três estágios com perfis de risco distintos: descoberta exploratória (*Discovery*), tradução clínica (*Translation*) e Execução (*Deployment*). Um sistema em fase de descoberta exploratória opera com dados retrospectivos e hipóteses abertas. Um sistema em fase de implantação influencia decisões clínicas em tempo real. Os riscos éticos de cada estágio são fundamentalmente distintos e exigem perguntas distintas do comitê. Essa distinção foi incorporada ao Bloco 1 da Versão B da MARIAH, que exige que o protocolo declare em qual estágio o sistema se encontra, e ao Passo 0 de ambas as versões, que orienta o filtro de entrada com base no papel efetivo do sistema no protocolo.

A contribuição mais estruturante do modelo australiano para a Versão A é a lógica do fluxograma sequencial. O NHMRC demonstrou que instrumentos de avaliação ética são mais eficazes quando percorrem dimensões em ordem progressiva, do mais simples ao mais complexo e guiam o julgamento do avaliador sem substituí-lo.

Essa lógica foi adotada integralmente na Versão A: percorrem-se os cinco eixos em sequência, e o nível de risco só sobe, nunca desce. O guia australiano é também a fonte principal do conceito de supervisão humana continuada, que fundamenta o Eixo 5 de ambas as versões. A presença de um pesquisador habilitado com condições reais de revisar e contestar os resultados do sistema não é apenas recomendável. É condição para que o protocolo não alcance automaticamente os níveis mais elevados de risco.

O modelo australiano contribuiu ainda para o Eixo 4, dedicado à representatividade e transferibilidade do modelo. O NHMRC relatou que sistemas desenvolvidos para certas populações tendem a ter desempenho inferior em grupos com diferentes características étnicas, socioeconômicas ou geográficas. A exigência de que o protocolo demonstre validação do modelo em contexto comparável ao da pesquisa tem fundamento direto nesse princípio.

Uma limitação relevante do modelo australiano para a adaptação brasileira é a ausência de critérios específicos para populações vulneráveis e para contextos de alta desigualdade. O guia foi elaborado para um sistema de saúde de alta renda, com infraestrutura tecnológica relativamente homogênea. A MARIAH incorpora perguntas específicas sobre diversidade populacional e validação em contexto comparável ao SUS para cobrir essa lacuna.

## 5.5 ESTADOS UNIDOS

A principal contribuição estadunidense para a MARIAH vem da diretriz da Administração de Alimentos e Medicamentos dos Estados Unidos (Food and Drug Administration — FDA), publicada em janeiro de 2025. O documento trata do uso de IA no ciclo de vida de produtos farmacêuticos e biológicos e introduz dois conceitos que informaram a estrutura da MARIAH: o contexto de uso e a consequência da decisão (FDA, 2025).

A principal contribuição dos EUA foi no contexto de uso que define o papel específico do sistema e o escopo da pergunta que ele foi desenvolvido para responder. A consequência da decisão avalia a gravidade do evento adverso potencial quando o modelo opera com alta

influência e baixa supervisão humana. Esses dois conceitos fundamentam, respectivamente, o Passo 1 e o Bloco 4 da Versão B.

A FDA também contribuiu com o conceito de manutenção do ciclo de vida: a exigência de que sistemas adaptativos sejam monitorados continuamente após o início do uso, com protocolos definidos para deriva de dados (*data drift*) e retreinamento do modelo. Esse conceito está incorporado ao Eixo 4 da Versão A e ao de medidas de redução de risco do Bloco 7 da Versão B. A diretriz da FDA demonstrou que a avaliação de desempenho de um sistema de IA utilizado em contexto com supervisão humana deve considerar o resultado da equipe humano-IA, não apenas o desempenho isolado do modelo. Esse princípio fundamenta as perguntas sobre mecanismo de supervisão efetiva em ambas as versões da MARIAH.

O contexto regulatório estadunidense conta ainda com o referencial do Centro de Ensaios Clínicos Multirregionais do Brigham and Women's Hospital e da Universidade de Harvard MRCT Center e WCG (Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard) que publicaram em 2025 um conjunto de ferramentas para o uso responsável de IA em pesquisa clínica (MRCT Center; WCG, 2025). Esse referencial foi a fonte da distinção entre uso rotineiro e uso não rotineiro que fundamenta o filtro de entrada de ambas as versões da MARIAH, e da preocupação com a capacidade dos grandes modelos de linguagem de inferir identidades e atributos sensíveis a partir de dados aparentemente anonimizados, incorporada ao Eixo 3 e ao Bloco 6.

## 5.6 EUROPA

A contribuição europeia para a MARIAH vem de duas fontes complementares. A primeira é o Regulamento Geral de Proteção de Dados (General Data Protection Regulation — GDPR), vigente desde 2018 (União Europeia, 2016), que estabelece o padrão mais exigente de proteção de dados pessoais em vigor no mundo e serviu de referência para a construção da Lei Geral de Proteção de Dados Pessoais brasileira (LGPD). O GDPR classifica dados de saúde como categoria especial de dados pessoais, com requisitos reforçados de base legal, finalidade declarada e minimização do tratamento. Essa arquitetura normativa fundamenta o Eixo 3 da Versão A e o Bloco 6 da Versão B, os de maior peso regulatório em ambos os instrumentos.

A segunda fonte é o Regulamento Europeu de Inteligência Artificial (EU AI Act), aprovado em 2024 e em processo de efetivação progressiva. O regulamento classifica sistemas

de IA em quatro categorias de risco: inaceitável, alto, limitado e mínimo. Sistemas de IA utilizados em contextos de saúde e pesquisa clínica são enquadrados na categoria de alto risco, com exigências obrigatórias de transparência, supervisão humana, robustez técnica e gestão de dados (União Europeia, 2024).

Essa lógica de classificação proporcional influenciou a estrutura de quatro níveis da MARIAH, embora a adaptação brasileira tenha substituído os critérios europeus por critérios orientados ao contexto da pesquisa com seres humanos e ao marco normativo nacional.

A contribuição da OMS para a MARIAH, incluindo os seis princípios de 2021 e as recomendações específicas para grandes modelos multimodais de 2024, está descrita na seção 5.2. O princípio da equidade é especialmente relevante para o contexto europeu e brasileiro: a OMS documentou que modelos treinados predominantemente com dados de populações de alta renda apresentam desempenho sistematicamente inferior em populações com diferentes características étnicas e socioeconômicas (OMS, 2024). Essa constatação fundamenta o Eixo 4 de ambas as versões da MARIAH e a ênfase nas perguntas sobre representatividade e validação em contexto comparável ao SUS.

## 5.7 ÁSIA

A experiência asiática em regulação de IA na saúde é marcada por abordagens distintas entre si, mas convergentes em um ponto central: em todos os três países examinados, o comitê de ética em pesquisa é reconhecido como primeira linha de defesa técnica contra vieses algorítmicos, opacidade de sistemas e violações de proteção de dados. Essa convergência legitima e reforça o propósito deste guia.

### 5.7.1 China

A China construiu o marco regulatório mais rigoroso da Ásia para sistemas de IA em saúde. A Administração Nacional de Produtos Médicos (National Medical Products Administration — NMPA) adota um sistema de três classes de risco para dispositivos médicos com IA. Sistemas de diagnóstico autônomo e suporte à decisão clínica em condições críticas são classificados na Classe III, com validação clínica obrigatória em território chinês e escrutínio ético de duplo nível (China, 2021). Esse princípio de precaução máxima influenciou

a lógica da MARIAH: protocolos que envolvem sistemas autônomos sem supervisão humana efetiva alcançam automaticamente o Nível IV, independentemente do desempenho nos demais eixos.

O instrumento regulatório mais relevante para comitês de ética é o conjunto de Medidas para a Revisão Ética de Pesquisas em Ciências da Vida e Médicas Envolvendo Seres Humanos, em vigor desde fevereiro de 2023. Esse documento expandiu formalmente o escopo da supervisão ética para abranger algoritmos, uso de grandes bases de dados e interfaces cérebro-computador (China, 2023).

Ele introduziu também o mecanismo de dupla revisão ética para protocolos com intervenções algorítmicas diretas: submissão a comitê institucional e, frequentemente, a instância superior. O artigo 32 dessas medidas estabelece isenção de revisão ética adicional para estudos retrospectivos com dados previamente anonimizados e de risco mínimo. Essa solução dialoga com o debate brasileiro sobre o uso secundário de dados em pesquisa e com as perguntas do Eixo 3 da MARIAH sobre consentimento e anonimização.

A principal lacuna do modelo chinês para fins de comparação com o Brasil é a ausência de programa nacional de capacitação de membros de comitês de ética para avaliação de sistemas algorítmicos. A supervisão recai sobre metodologias desenvolvidas autonomamente por hospitais de grande porte, sem padronização federal. A Lei de Proteção de Informações Pessoais da China (PIPL, do inglês Personal Information Protection Law), em vigor desde 2021, exige que os dados sejam mantidos localmente e que o consentimento explícito seja a principal base legal. Isso cria um sistema de soberania de dados que, na prática, impede a realização de estudos internacionais multicêntricos com compartilhamento de dados brutos. Essa restrição é o contraste mais direto com a LGPD brasileira, que permite transferência internacional de dados com base em garantias contratuais e consentimento específico.

### 5.7.2 Japão

O Japão adotou uma das abordagens regulatórias mais pragmáticas da Ásia para sistemas de IA adaptativa. A Agência de Dispositivos Médicos e Produtos Farmacêuticos (Pharmaceuticals and Medical Devices Agency — PMDA) desenvolveu o sistema de Avaliação Tempestiva de Inovação (IDATEN, do inglês *Improvement Design within Approval for Timely Evaluation Notice*), mecanismo que permite atualizações iterativas em sistemas de IA já

aprovados sem exigir novo processo regulatório completo, desde que as modificações ocorram dentro de parâmetros previamente pactuados (PMDA, 2023). Esse conceito foi incorporado ao Eixo 5 da Versão A e ao Bloco 7 da Versão B da MARIAH, nas perguntas sobre protocolo de retreinamento do modelo e monitoramento de deriva de dados.

As Diretrizes Éticas para Pesquisa Médica e Biológica Envolvendo Seres Humanos, publicadas em 2021 pelos Ministérios da Saúde e da Educação japoneses, trouxeram três inovações relevantes para comitês de ética: a revisão centralizada de protocolos multicêntricos por um único comitê; a validação do consentimento por via digital; e o mecanismo de exclusão (*opt-out*) para uso secundário de dados em pesquisa com inteligência artificial (MHLW; MEXT, 2021).

O mecanismo de exclusão permite que pesquisadores utilizem vastos bancos de dados clínicos sem reobter consentimento explícito, desde que publiquem os detalhes do estudo e garantam ao paciente uma via acessível para retirar seus dados. Esse princípio foi incorporado à MARIAH de forma mais protetiva: a pergunta 2.8 do Eixo 2 exige que o protocolo preveja mecanismo para que o participante retire o consentimento, a qualquer tempo, com exclusão de seus dados das bases do estudo e garantia de não utilização em treinamentos ou retreinamentos futuros.

### 5.7.3 Coreia do Sul

A Coreia do Sul oferece as duas contribuições mais inovadoras da experiência asiática para a MARIAH. A primeira vem do Ministério da Segurança Alimentar e de Medicamentos (Ministry of Food and Drug Safety — MFDS), que desenvolveu dois critérios para classificação de risco de sistemas de IA em saúde: o potencial de falha (*Malfunction Potential*) e a oportunidade de contestação (*Opportunity to Override*) (MFDS, 2022).

O potencial de falha avalia a probabilidade de que erros algorítmicos causem dano clínico ao paciente. A oportunidade de contestação avalia se o pesquisador tem condições reais de questionar o resultado do sistema antes de sua execução. Esses dois critérios foram incorporados ao Eixo 1 de ambas as versões da MARIAH e ao Bloco 2 da Versão B, nas perguntas sobre autonomia do sistema e capacidade efetiva de supervisão humana.

A segunda contribuição vem da Lei-Quadro para o Desenvolvimento da Inteligência Artificial e Estabelecimento da Confiança (*Framework Act on the Development of Artificial*

*Intelligence and Establishment of Trust*), aprovada em janeiro de 2025 e em vigor desde janeiro de 2026 (Coreia do Sul, 2025).

A lei impõe requisitos éticos para sistemas de alto impacto setorial, incluindo saúde, e torna a Coreia do Sul a segunda jurisdição do mundo, após a União Europeia, a promulgar lei abrangente de IA. As diretrizes do MFDS para ensaios clínicos com IA, publicadas em 2022, instruem os comitês de ética a auditarem três fenômenos ausentes da avaliação de medicamentos convencionais: vazamento de dados entre conjuntos de treinamento e teste, vieses populacionais ocultos e variação de desempenho entre unidades hospitalares com diferentes equipamentos (MFDS, 2022). Essas três dimensões estão incorporadas aos Eixos 3 e 4 de ambas as versões da MARIAH.

### 5.7.4 Índia

A Índia é o país que mais se aproxima do propósito deste guia. O Conselho Indiano de Pesquisa Médica (Indian Council of Medical Research — ICMR) publicou em 2023 as Diretrizes Éticas para a Aplicação de Inteligência Artificial em Pesquisa Biomédica e Saúde (Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare), o único documento no mundo concebido especificamente para capacitar comitês de ética na avaliação de protocolos com sistemas algorítmicos (ICMR, 2023).

As diretrizes estruturam a avaliação com dez princípios bioéticos e destacam que a IA é destinada à amplificação diagnóstica. Não substitui o julgamento clínico. O profissional responsável detém soberania irrevogável sobre os resultados do sistema. Esse princípio é o fundamento central do Eixo 5 de ambas as versões da MARIAH.

A contribuição indiana mais inovadora para a MARIAH é o mecanismo de exclusão tardia: o participante de pesquisa tem o direito de solicitar a supressão de seus dados do repositório utilizado pela IA a qualquer tempo, mesmo após o início do estudo. Nenhum dos modelos ocidentais analisados prevê esse direito. A MARIAH o incorpora como exigência ética na pergunta 2.8 do Eixo 2. A Organização Central de Controle de Padrões de Medicamentos (CDSCO, do inglês Central Drugs Standard Control Organization) contribuiu com o Protocolo de Mudança de Algoritmo (Algorithm Change Protocol), que exige que fabricantes submetam antecipadamente um plano que descreva como o modelo será retreinado dentro de

parâmetros técnicos previamente pactuados (CDSCO, 2025). Esse princípio fundamenta as perguntas sobre retreinamento do modelo no Eixo 5 e no Bloco 7.

A principal advertência do modelo indiano é legislativa. A Lei de Proteção de Dados Pessoais Digitais (Digital Personal Data Protection Act — DPDP Act), aprovada em 2023 (Índia, 2023), suprimiu inteiramente a categoria de dados sensíveis de seu texto e equiparou prontuários de saúde, registros genéticos e dados psiquiátricos a quaisquer outros dados pessoais (Burman, 2023). Essa escolha transfere para os Comitês de Ética a responsabilidade de proteger o participante que a lei deixou desguarnecido. A LGPD brasileira faz escolha oposta: dados de saúde constituem categoria sensível com proteção reforçada. Essa distinção é um dos pilares do Bloco 6 da Versão B e do Eixo 3 da Versão A.

## 5.8 RÚSSIA

A experiência russa oferece ao guia duas contribuições de naturezas distintas: um modelo regulatório de precaução máxima e um caso de advertência sobre os riscos de subordinar a supervisão ética à lógica da inovação acelerada.

A Agência Federal de Vigilância em Saúde (Roszdravnadzor) adota uma classificação automática: qualquer sistema de saúde baseado em inteligência artificial é enquadrado na Classe 3, o nível de maior risco, independentemente de sua aplicação (Rússia, 2020). Essa decisão obriga todos os protocolos com IA a percorrer o caminho mais rigoroso de validação clínica e aprovação ética. Consiste numa solução administrativa simples e abrangente. A MARIAH gradua as exigências em proporção ao risco, mas admite a necessidade de máxima precaução quando decisões autônomas podem causar impacto irreversível ao participante.

A Lei Federal n.º 152-FZ sobre Dados Pessoais da Rússia (Rússia, 2006), com emendas recentes, classifica dados de saúde como categoria especial e exige consentimento expresso para seu tratamento. Impõe também localização compulsória de dados em servidores russos, o que inviabiliza, na prática, estudos multicêntricos internacionais com compartilhamento de dados brutos.

## 5.9 SÍNTESE COMPARATIVA E LIÇÕES PARA O BRASIL

O quadro ético da OMS e nove experiências regulatórias nacionais mostram que avaliar eticamente protocolos com inteligência artificial vai além da simples verificação de conformidade documental. A avaliação exige um instrumento que traduza complexidade técnica em impacto sobre a pessoa participante. É exatamente isso que a MARIAH se propõe a fazer.

O Brasil herda de cada modelo uma contribuição específica. Da OMS, os seis princípios éticos fundacionais e o alerta sobre riscos específicos dos grandes modelos multimodais. Do Canadá, a lógica de pontuação ponderada e a regra crítica de que mitigações não alteram o nível de risco. Da Austrália, o fluxograma sequencial e a distinção entre estágios do ciclo de vida do sistema. Dos Estados Unidos, o conceito de contexto de uso como etapa prévia obrigatória e o princípio da manutenção do ciclo de vida com monitoramento de deriva de dados. Da Europa, a classificação proporcional de risco e o padrão de proteção reforçada para dados sensíveis. Da Coreia do Sul, os critérios de potencial de falha e oportunidade de contestação. Do Japão, o protocolo de retreinamento do modelo dentro de parâmetros previamente pactuados. Da Índia, o direito do participante de solicitar a exclusão de seus dados do treinamento a qualquer tempo e o princípio da autonomia médica irrevogável. Da Rússia, o princípio de que a precaução máxima tem fundamento quando sistemas autônomos podem causar impacto irreversível ao participante.

O que nenhum desses modelos previu é o contexto brasileiro. Nenhum considera a pluralidade étnica e regional da população atendida pelo SUS. Nenhum está ancorado na LGPD, que classifica dados de saúde como categoria sensível com proteção reforçada e oferece ao participante de pesquisa um conjunto de direitos que supera o de todas as legislações asiáticas analisadas. Nenhum foi concebido para um sistema de saúde universal que atende toda a população brasileira em condições de acesso radicalmente desiguais. Essas especificidades não são adaptações à margem do instrumento, são o núcleo do que torna a MARIAH necessária e distinta.

Um diferencial adicional do arcabouço brasileiro escapa a qualquer dos modelos internacionais comparados. A Resolução CNS n.º 738/2024 operacionaliza a LGPD no contexto específico da pesquisa com bancos de dados envolvendo seres humanos, com detalhamento que nenhuma das nove experiências nacionais examinadas apresenta: definição das figuras jurídicas do Controlador e do Operador, distinção operacional entre bancos constituídos no âmbito da pesquisa e bancos já constituídos fora desse âmbito, instituição dos quatro Termos

institucionais — Acordo, Anuência, Compromisso e Transferência — que estruturam a cadeia de custódia do dado, e enumeração exaustiva das hipóteses de dispensa de TCLE uso futuro, distintas conforme a origem do banco. Esse nível de operacionalização da proteção de dados em pesquisa com seres humanos é particularidade do sistema brasileiro e compõe a base sobre a qual a MARIAH se ancora nos Capítulos 7 e 8.

# 6 INTEGRIDADE CIENTÍFICA E AUTORIA NA PESQUISA COM IA

A inteligência artificial não entrou na pesquisa científica apenas como ferramenta de análise. Entrou como agente capaz de redigir protocolos, gerar referências bibliográficas, produzir dados sintéticos e formular hipóteses com fluência que simula autoridade técnica. Essa entrada cria um conjunto de responsabilidades novas que o ordenamento brasileiro começou a tratar com precisão.

Cinco instrumentos estruturam hoje a integridade científica em pesquisa com IA no Brasil. A Portaria CNPq n.º 2.664/2026, que institui a Política de Integridade na Atividade Científica, exige declaração obrigatória de uso de IA generativa. A Lei n.º 14.874/2024 e o Decreto n.º 12.651/2025 fixam a responsabilidade integral do pesquisador por todo o protocolo. A Resolução CNS n.º 738/2024 disciplina a constituição e a gestão de bancos de dados para pesquisa, e consolida deveres de integridade e governança na cadeia de tratamento. A Resolução CFM n.º 2.454/2026 preserva a responsabilidade médica sobre decisões apoiadas por IA. A LGPD fundamenta a responsabilidade sobre os dados que o sistema processa.

Este capítulo traduz esse marco normativo nas exigências que um protocolo deve satisfazer antes da avaliação de mérito, quando usa IA em qualquer fase da pesquisa.

## 6.1 O QUE MUDA QUANDO A IA ENTRA NA PESQUISA

A IA entra na pesquisa de formas radicalmente distintas, e essa distinção determina o conjunto de responsabilidades éticas que se aplica a cada caso. Há dois modos principais de uso da IA no contexto da pesquisa.

No primeiro uso, a IA opera como ferramenta de processamento: analisa imagens médicas, classifica exames laboratoriais, identifica padrões em bases de dados epidemiológicas, processa séries temporais de sinais fisiológicos. O pesquisador define a pergunta, seleciona os dados, interpreta os resultados e responde por todas essas escolhas. A IA é instrumento, como foram o microscópio e o software estatístico.

No segundo uso, a IA opera como agente gerador de conteúdo: escreve seções do protocolo, formula hipóteses, produz revisões de literatura, redige o Termo de Consentimento Livre e Esclarecido (TCLE), gera dados sintéticos, sugere análises. Nesse uso, a fronteira entre

o que o pesquisador criou e o que a máquina produziu torna-se a questão ética central. Este capítulo aborda o segundo uso, reconhecendo que ambos podem ocorrer no mesmo protocolo, o que atualmente é frequente.

O ordenamento brasileiro respondeu a esse desafio com instrumentos normativos precisos, embora permaneçam lacunas na cadeia algorítmica estendida que são retomadas em 6.4 e no Capítulo 7. A Portaria CNPq n.º 2.664, de 6 de março de 2026, estabeleceu a obrigação de declarar o uso de qualquer ferramenta de inteligência artificial generativa em atividades de fomento e pesquisa, com descrição da ferramenta utilizada, da função exercida e da forma de verificação humana dos resultados (CNPq, 2026).

O Decreto n.º 12.651/2025, que regulamenta a Lei n.º 14.874/2024, consolidou o princípio de que nenhuma etapa da pesquisa com seres humanos escapa à responsabilidade humana identificada e verificável (Brasil, 2025a). A responsabilidade por imprecisões, plágios ou dados gerados incorretamente por sistemas de IA recai integralmente sobre os pesquisadores humanos. Esse princípio não admite exceção por complexidade técnica, por limitação do sistema ou por desconhecimento do pesquisador sobre o funcionamento do algoritmo que utilizou.

A integridade científica em pesquisa com IA não é questão acadêmica sobre autoria e crédito intelectual. Tem consequência direta e verificável sobre as pessoas que participam da pesquisa. Um protocolo que usa sistema de linguagem de grande escala para redigir seções do TCLE pode produzir linguagem fluente e juridicamente imprecisa, com omissões sobre riscos específicos que o sistema não foi capaz de identificar ou que alucinou.

Um protocolo que usa IA generativa para formular hipóteses sem verificação humana rigorosa pode expor participantes a intervenções mal fundamentadas com aparência de evidência científica. Um protocolo que apresenta dados sintéticos como dados reais compromete a validade científica que justifica o risco imposto aos participantes. O protocolo que não explicita o uso de IA generativa em cada fase é um documento incompleto. A avaliação recai sobre o que foi descrito e aceita implicitamente o que foi omitido.

Sete instrumentos normativos estruturam as obrigações do pesquisador e o mandato do CEP na avaliação de protocolos com IA. A triangulação entre a Lei n.º 14.874/2024, o Decreto n.º 12.651/2025 e a Portaria CNPq n.º 2.664/2026 define o núcleo da responsabilidade do pesquisador pela integridade de cada etapa da pesquisa.

A Resolução CNS n.º 738/2024 soma-se a esse núcleo ao oferecer, no seu art. 3.º, XII, a definição normativa brasileira de integridade de pesquisa como "compromisso com valores éticos e princípios de boas práticas na execução de pesquisas", e ao impor, no art. 5.º, VII, o dever de respeitar esse princípio a pesquisadores, patrocinadores e instituições. É a primeira vez que o ordenamento brasileiro fixa essa definição em norma de alcance nacional.

A Resolução CFM n.º 2.454/2026 acrescenta a dimensão da responsabilidade médica sobre decisões apoiadas por IA. A LGPD fundamenta a responsabilidade sobre os dados pessoais produzidos e processados ao longo do ciclo de vida do sistema. As Resoluções CNS n.º 466/2012 (Brasil, 2013) e n.º 510/2016 (Brasil, 2016) preservam os princípios bioéticos que atravessam todo o SINEP e que a chegada da IA torna mais exigentes, não menos.

## 6.2 AUTORIA HUMANA: PRINCÍPIO, FUNDAMENTO E LIMITE

A discussão sobre autoria científica gira, quase sempre, em torno do crédito: quem assina o artigo, quem aparece nos agradecimentos, quem recebe o reconhecimento pelo trabalho. Na pesquisa com IA, a questão central é  quem responde pelo dano.

Sistemas de IA não comparecem perante CEPs, não respondem a processos de apuração de má conduta científica, não podem ser responsabilizados por danos causados a participantes de pesquisa, não têm registro profissional, não têm vínculo institucional e não têm patrimônio passível de reparação civil. A autoria humana é o mecanismo que garante que alguém responde. Retirar a autoria humana de qualquer etapa da pesquisa não é questão de crédito intelectual. Remove a proteção que garante a responsabilização perante os participantes (Brasil, 2024a; 2025a).

A comunidade científica internacional convergiu nos últimos anos para uma posição clara sobre IA e autoria. O *International Committee of Medical Journal Editors* (ICMJE, 2026) define autoria científica em saúde por quatro critérios cumulativos: (1) contribuição substancial para a concepção ou o desenho do estudo, ou para a aquisição, análise ou interpretação dos dados; (2) redação ou revisão crítica substantiva do manuscrito; (3) aprovação final da versão a ser publicada; (4) concordância em responder por todos os aspectos do trabalho, garantindo que questões relacionadas à exatidão ou integridade de qualquer parte possam ser investigadas e resolvidas. Um sistema de IA não cumpre nenhum dos quatro: não aprova, não responde e não se responsabiliza. As principais publicações

médicas internacionais, incluindo as da OMS, já estabeleceram que a IA não deve ser listada como coautora, e é permitida apenas como ferramenta reconhecida (OMS, 2021). O ordenamento brasileiro, pela Portaria CNPq n.º 2.664/2026, alinhou-se a essa posição.

A Lei n.º 14.874/2024 e o Decreto n.º 12.651/2025 consolidam esse princípio sem ambiguidade: o pesquisador principal responde por todo o protocolo, por todas as suas etapas e por todos os seus resultados, inclusive pelas etapas em que delegou produção de conteúdo ou processamento de dados a sistemas algorítmicos.

A Portaria CNPq n.º 2.664/2026 operacionaliza essa responsabilidade com três exigências específicas: identificação da ferramenta de IA utilizada, descrição da função exercida por essa ferramenta em cada etapa da pesquisa, e descrição do mecanismo de verificação humana dos resultados gerados (CNPq, 2026). O terceiro elemento é o mais exigente e o menos cumprido. Declarar que se usou um sistema de linguagem de grande escala para redigir o protocolo não satisfaz a portaria. O pesquisador deve descrever como verificou a correção, a completude e a adequação ética do conteúdo produzido pelo sistema antes da submissão ao CEP, e identificar quem realizou essa verificação.

Todo protocolo deve deixar claros dois pontos específicos, cuja presença será verificada pelo CEP. O primeiro: o protocolo declara explicitamente se a IA foi ou não utilizada em cada fase da pesquisa? A ausência de declaração num protocolo em que o uso de IA é provável ou esperado deve acionar pedido de elucidação antes de qualquer avaliação de mérito. O silêncio do protocolo sobre o uso de IA não é evidência de que a IA não foi usada. É informação ausente que impede a avaliação ética completa.

O segundo: o protocolo descreve o mecanismo concreto de verificação humana dos resultados gerados por IA? A declaração de uso sem descrição da verificação é declaração incompleta. O pesquisador que usa IA sem verificar o que ela produziu transfere ao CEP, e eventualmente ao participante, o risco de uma etapa do protocolo que nenhum ser humano revisou com responsabilidade identificada. A autoria humana em pesquisa com IA não é princípio nostálgico, nem resistência ao progresso tecnológico. É a única garantia disponível de que alguém responde pelo que acontece às pessoas que participam da pesquisa.

## 6.3 PLÁGIO ALGORÍTMICO, MÁ CONDUTA POR NEGLIGÊNCIA E FABRICAÇÃO DE EVIDÊNCIAS

Plágio algorítmico ocorre quando um pesquisador apresenta conteúdo gerado por IA como fruto de trabalho próprio, sem declarar o uso da ferramenta e sem controle crítico rigoroso. Diferentemente do plágio tradicional, nesse caso não há a cópia identificável de um autor específico, mas a diluição da autoria por meio de uma mediação algorítmica que mascara a origem do conteúdo. O risco central não está apenas na apropriação indevida, mas na transferência silenciosa de autoridade científica para um sistema que não compreende o significado, a veracidade ou as implicações éticas do que produz.

Os grandes modelos de linguagem (Large Language Models — LLMs) produzem texto com fluência que simula autoridade técnica. Essa fluência é exatamente o fator de risco abordado neste capítulo. Um sistema de linguagem pode gerar citações bibliográficas com formatação impecável referenciando estudos que nunca existiram. Pode descrever com precisão aparente os resultados de ensaios clínicos que não foram realizados. Pode produzir revisões de literatura internamente coerentes que omitem sistematicamente evidências contrárias sem sinalizar essa omissão. Pode redigir seções metodológicas com imprecisões que nenhum revisor sem formação técnica específica identificará.

O problema não é a intenção do pesquisador que usa essas ferramentas. É a propriedade estrutural dos sistemas: modelos de linguagem de grande escala são otimizados para produzir texto plausível, não texto verdadeiro. Plausibilidade e veracidade não são sinônimos. A distância entre os dois pode ser invisível ao revisor e consequente para o participante (OMS, 2021; Figueiredo; Aguiar, 2025). A Agência Nacional de Proteção de Dados (Brasil, 2024d), em seu Radar Tecnológico sobre Inteligência Artificial Generativa, documentou essa propriedade estrutural e recomendou parâmetros de transparência e supervisão humana como condição de uso responsável.

Nem todo uso inadequado de IA generativa em pesquisa configura má conduta científica, e a distinção entre três tipos de falha importa para orientar a resposta institucional. A primeira é o erro metodológico: o pesquisador usa sistema de linguagem para redigir uma seção do protocolo e não verifica o resultado com o rigor necessário. Não há intenção de fraude, e a resposta é corrigir e capacitar. A segunda é a má conduta por negligência: o pesquisador sabe que deveria verificar, dispõe de meios e tempo para verificar, e ainda assim não verifica — seja por pressa, por confiança ingênua no sistema ou por suposição de que o problema é de outra instância. Este é o tipo mais recorrente atualmente e é o que a Portaria CNPq n.º 2.664/2026 mira com mais força, por meio da exigência de descrição explícita do

mecanismo de verificação. A terceira é a má conduta intencional: o pesquisador apresenta dados sintéticos gerados por IA como dados coletados de participantes reais, ou assina como autor um protocolo gerado integralmente por IA sem revisão substantiva. Aqui há fabricação de evidências, apuração formal e responsabilização jurídica.

O CEP não é instância de apuração de má conduta científica, mas é a primeira instância capaz de identificar protocolos que apresentam riscos estruturais de produção de evidências comprometidas pela ausência de verificação humana dos resultados gerados por IA.

Na prática, a avaliação ética de protocolos com IA generativa exige que o comitê formule quatro perguntas concretas diante de qualquer proposta que recorra a essas ferramentas em qualquer fase.

A primeira: a IA foi usada na redação do protocolo, do TCLE ou de qualquer documento submetido ao comitê? A ausência de declaração não é evidência de ausência de uso. É informação que o comitê deve solicitar ativamente.

A segunda: os resultados gerados pela IA foram verificados por especialista humano com competência para identificar erros, alucinações e omissões específicos ao campo científico da pesquisa? Verificação genérica por pesquisador sem formação na área não satisfaz essa exigência.

A terceira: o protocolo usa dados sintéticos gerados por IA? Se sim, como foi validada a fidelidade estatística desses dados em relação à população real, e essa validação está documentada no protocolo?

A quarta: o TCLE informa os participantes, em linguagem acessível, sobre o papel da IA em qualquer fase da pesquisa que os afete diretamente?

Plágio algorítmico, má conduta por negligência e fabricação de evidências por IA generativa não são riscos de um futuro distante. Estão nos protocolos que chegam hoje às plataformas de submissão do SINEP.

## 6.4 RESPONSABILIDADE CIVIL E A CADEIA ESTENDIDA

Quando a pesquisa convencional produz dano a um participante, a cadeia de responsabilidade é relativamente legível: pesquisador principal, instituição executora e, quando aplicável, patrocinador. A pesquisa com IA introduz um quarto elo que essa cadeia tradicional não contempla: o desenvolvedor ou fornecedor do sistema algorítmico —

empresas como OpenAI, Google, Anthropic, Microsoft, ou desenvolvedores nacionais de software médico.

Quando o dano decorre de uma falha do modelo, de um viés estrutural da arquitetura ou de uma alucinação do sistema generativo, a questão de quem responde não tem resposta consolidada no ordenamento brasileiro vigente. O que o ordenamento estabelece com nitidez é a posição do pesquisador: a responsabilidade do pesquisador principal permanece, mesmo quando há um sistema algorítmico intermediando sua decisão e o dano. Escolher usar uma ferramenta é escolha humana. Escolha humana carrega responsabilidade humana. O pesquisador que delega etapas da pesquisa a sistemas de IA sem verificação adequada não transfere sua responsabilidade ao sistema. Acumula a responsabilidade pela escolha de delegar sem verificar.

Essa lógica desenha quatro elos concretos na cadeia estendida de responsabilidade. O pesquisador principal responde pela escolha da ferramenta, pela verificação humana dos resultados e pela integridade final do protocolo. A instituição executora responde pela estrutura de governança, pelo acesso a sistemas seguros e pela capacitação dos pesquisadores. O patrocinador, quando existe, responde pelos termos contratuais que viabilizam ou limitam o uso de IA na pesquisa que financia. O fornecedor do sistema é responsável pelos defeitos do produto. No entanto, ainda não existe critério definido para responsabilização civil por danos decorrentes de decisões automatizadas em ambientes clínicos. Enquanto essa lacuna persiste, o pesquisador tende a absorver, individualmente, um risco que deveria estar distribuído.

Quando o protocolo envolve bancos de dados, a Resolução CNS n.º 738/2024 nomeia dois dos elos da cadeia como figuras jurídicas específicas em pesquisa. O Controlador, conforme o art. 3.º, IV, é a pessoa natural ou jurídica responsável pelas decisões sobre o tratamento dos dados, podendo ser patrocinador, pesquisador do protocolo, responsável pelo biobanco ou seu designado. O Operador, definido no art. 3.º, XIII, realiza o tratamento em nome do Controlador e responde por uso conforme à finalidade aprovada pelo CEP. O art. 11 estabelece que as responsabilidades do Controlador são irrenunciáveis, admitindo-se a controladoria conjunta. Essa nomeação contribui para reduzir a indefinição das responsabilidades no âmbito da governança de dados, sendo detalhada no Capítulo 7, o qual aborda agentes, funções e direitos na pesquisa operacional com dados. A lacuna que subsiste,

portanto, diz respeito especificamente ao elo do fornecedor externo do sistema algorítmico, não aos elos internos já disciplinados pela Res. CNS n.º 738/2024 (Brasil, 2024b).

Três dispositivos atenuam essa concentração na prática atual. A Lei n.º 14.874/2024 e o Decreto n.º 12.651/2025 estabelecem que o pesquisador principal responde por todo o protocolo em todas as suas fases (Brasil, 2024a; 2025a), mas admitem a responsabilidade solidária da instituição e do patrocinador.

Quando a pesquisa envolver uso clínico da IA, a Resolução CFM n.º 2.454/2026 determina que a responsabilidade médica sobre decisões apoiadas por IA é integral e inalienável, veda a comunicação direta de diagnósticos ao paciente por sistemas algorítmicos e exige registro em prontuário (CFM, 2026). Na cadeia da pesquisa, o direito do participante à revisão humana decorre do artigo 20 da LGPD, que garante ao titular a revisão, por pessoa natural, de decisões tomadas unicamente com base em tratamento automatizado; o protocolo deve prever essa informação no TCLE, e o CEP verifica.

A LGPD fundamenta, ainda, nos arts. 42 a 45, a responsabilidade civil sobre dados pessoais de participantes processados por sistemas de IA durante a pesquisa, inclusive quando esse processamento ocorre em plataformas de terceiros (Brasil, 2018). As obrigações de governança, base legal, minimização e segurança que decorrem da LGPD são desenvolvidas no Capítulo 7, que trata a governança de dados como dimensão operacional da responsabilidade civil exposta nesta seção. As Resoluções CNS n.º 466/2012 (Brasil, 2013) e n.º 510/2016 (Brasil, 2016), que preservam sua relevância ética, preservam os princípios que atravessam toda a cadeia: beneficência, não maleficência, autonomia e justiça. Esses princípios não envelheceram com a adoção da IA; tornaram-se mais exigentes

## 6.5 O QUE O PROTOCOLO DEVE CONTER — E O QUE O CEP VERIFICA

Todo protocolo com IA deve conter cinco requisitos de integridade, e um sexto se houver uso de bancos de dados próprios ou de terceiros; o CEP verifica a presença de cada um. A ausência desses elementos determina se o documento submetido é avaliável em sua integralidade. Dos cinco requisitos iniciais, três possuem caráter eliminatório. A falta de qualquer um deles inviabiliza a análise de mérito e demanda providências obrigatórias. Os outros dois são sanáveis por complementação documental. O sexto, quando aplicável, é eliminatório: sem ele, a cadeia de custódia do dado não está formalizada.

O primeiro (eliminatório): há declaração explícita sobre o uso ou não uso de IA em cada fase da pesquisa? A ausência de declaração não é evidência de ausência de uso. É informação que o comitê deve solicitar ativamente antes de qualquer avaliação de mérito.

O segundo (eliminatório): a declaração identifica a ferramenta utilizada, a função exercida em cada etapa e o responsável humano pela verificação dos resultados? Declaração genérica que menciona IA sem especificar uso, função e verificação não satisfaz as exigências da Portaria CNPq n.º 2.664/2026 (CNPq, 2026).

O terceiro (eliminatório): o protocolo descreve mecanismo concreto de verificação humana dos resultados gerados por IA, incluindo quem realizou a verificação e qual a sua competência técnica para tanto?

O quarto (sanável): o TCLE foi produzido ou revisado por humano com capacidade de identificar imprecisões, omissões e alucinações específicas ao contexto clínico da pesquisa, e informa os participantes sobre o direito à revisão de decisões automatizadas (art. 20, LGPD)?
O quinto (sanável): o pesquisador principal assina o protocolo com declaração expressa de responsabilidade integral pelo conteúdo, incluindo o conteúdo gerado ou processado por sistemas de IA?

O sexto: se o protocolo utiliza banco de dados, o Controlador e o Operador (se houver) são identificados, e os Termos institucionais exigidos pela Resolução CNS n.º 738/2024 são apresentados? Para bancos constituídos no âmbito da pesquisa, a exigência mínima é o Termo de Compromisso de Uso de Dados; em protocolos multicêntricos ou com participação de mais de uma instituição, acresce o Termo de Acordo Institucional.

Para bancos já existentes fora do contexto de pesquisa, a Resolução exige também o Termo de Anuência Institucional assinado pelo responsável da instituição de origem dos dados. A ausência dos Termos aplicáveis significa que a cadeia de custódia do dado não foi formalizada, e o protocolo não é avaliável no mérito científico. O Capítulo 7 descreve os quatro Termos institucionais (Acordo, Anuência, Compromisso, Transferência) em detalhe.

A ausência de qualquer dos elementos eliminatórios não configura pendência menor sanável por complementação formal. Indica que o protocolo foi submetido sem o nível de responsabilidade humana que a Lei n.º 14.874/2024, o Decreto n.º 12.651/2025 e a Resolução CNS n.º 738/2024 exigem de toda pesquisa com seres humanos (Brasil, 2024a; 2024b; 2025a).

Quatro dimensões atravessam o protocolo inteiro e merecem atenção específica durante a avaliação:

- Nas referências bibliográficas: há indicadores de citações geradas por IA sem verificação, como referências com dados inconsistentes, identificadores digitais de objetos inexistentes ou descrições de estudos sem correspondência verificável em bases indexadas?
- Na metodologia: o uso de IA em alguma fase da análise está descrito com precisão suficiente para permitir a avaliação dos riscos específicos daquele sistema naquele contexto clínico e populacional?
- Nos dados: se o protocolo usa dados sintéticos gerados por IA, a fidelidade estatística foi validada, documentada e comparada com a população real que os dados pretendem representar?
- No TCLE: a linguagem é acessível ao público-alvo da pesquisa, as informações sobre o papel da IA em decisões que afetam os participantes estão presentes e o texto foi verificado por humano responsável identificado no protocolo?

Essas quatro dimensões não exigem conhecimento técnico sobre os sistemas de IA utilizados. Exigem precisão na leitura do protocolo e disposição para formular as perguntas que a assimetria de conhecimento técnico tende a suprimir.

Este capítulo fecha com uma distinção operacional que merece atenção cuidadosa: a diferença entre protocolo incompleto e protocolo inadequado. O protocolo incompleto apresenta lacunas sanáveis: declaração de uso de IA ausente, mas solicitável; mecanismo de verificação não descrito, mas implementável; TCLE com linguagem inadequada, mas corrigível. Esses casos admitem solicitação de informações adicionais com prazo definido e critérios explícitos para a complementação.

O protocolo inadequado apresenta falhas estruturais que não se resolvem por complementação documental: ausência de responsabilidade humana identificada em etapas críticas da pesquisa, uso de dados sintéticos sem validação apresentado como dado real coletado de participantes, TCLE que omite o papel da IA em decisões que afetam diretamente as pessoas participantes. Esses casos exigem recusa fundamentada nos princípios da Lei n.º 14.874/2024, do Decreto n.º 12.651/2025, da Resolução CNS n.º 738/2024 e nas exigências da Portaria CNPq n.º 2.664/2026 (Brasil, 2024a; 2024b; 2025a; CNPq, 2026).

Distinguir entre os dois com precisão e fundamentação não cria obstáculo à pesquisa com o Brasil. Longe disso, contribui com o único momento em que a governança ainda pode

ser plena. Os capítulos 7 e 8 estendem essa lógica para a governança de dados e para a avaliação proporcional de risco do sistema descrito no protocolo.

# 7 GOVERNANÇA DE DADOS E PROTEÇÃO DE PARTICIPANTES

A inteligência artificial transforma dados em matéria-prima de modelos que passam a operar em escala sobre a mesma população que os gerou. Essa transformação exige governança de dados desenhada para a natureza específica do tratamento algorítmico: dados que se dissolvem em pesos do modelo, modelos que memorizam fragmentos do que viram, consentimentos cuja retirada não exclui o dado já treinado, bases clínicas que atravessam fronteiras sob promessa de inovação.

O ordenamento brasileiro construiu, em menos de uma década, o arcabouço mais detalhado da América Latina para esse contexto. A LGPD dá a arquitetura geral da proteção de dados pessoais. A Resolução CNS n.º 738, de 1.º de fevereiro de 2024, operacionaliza essa arquitetura para pesquisa com bancos de dados envolvendo seres humanos. A Lei n.º 14.874/2024 e o Decreto n.º 12.651/2025 fixam a responsabilidade institucional sobre os dados tratados em pesquisa. A Resolução CFM n.º 2.454/2026 disciplina o uso de IA em decisões médicas. Este capítulo traduz esse marco em perguntas concretas que o protocolo de pesquisa com IA em saúde deve responder.

## 7.1 O DADO DE SAÚDE EM SISTEMAS DE IA: UMA NATUREZA DIFERENTE

Um banco de dados relacional armazena informações em campos discretos e endereçáveis. O registro de um paciente ocupa um espaço definido, pode ser localizado, alterado ou excluído com precisão cirúrgica. Essa lógica orienta a maior parte da legislação sobre proteção de dados pessoais, incluindo a LGPD. Quando um sistema de IA entra em cena, essa lógica colapsa.

O treinamento de uma rede neural funciona por um mecanismo radicalmente distinto. O algoritmo processa os dados de treino e extrai padrões estatísticos que se distribuem por milhões de parâmetros matemáticos, os chamados pesos do modelo (Cota et al., 2024). O dado original não permanece armazenado como arquivo. Dissolve-se na arquitetura do modelo, como a tinta se dissolve na água. O registro clínico de um paciente deixa de ocupar um endereço localizável e passa a habitar, de forma difusa e matematicamente indissociável, a estrutura inteira do sistema.

Essa distinção tem consequência direta para o desenho do protocolo. Protocolos de pesquisa que tratam dados de saúde como matéria-prima para treinamento de modelos de IA operam numa categoria jurídica e ética que as normativas convencionais não anteciparam. A promessa de exclusão do dado após o término da pesquisa, suficiente num banco de dados relacional, torna-se tecnicamente vazia num modelo de aprendizado profundo. O pesquisador que submete um protocolo desse tipo precisa reconhecer essa diferença antes de qualquer outra análise. Grandes modelos de linguagem (*Large Language Models* — LLMs) e redes neurais profundas têm uma propriedade que os distingue de qualquer sistema de informação anterior: a capacidade de memorizar fragmentos dos dados com que foram treinados e reproduzi-los em condições específicas de consulta. Essa memorização não ocorre por falha de projeto. Resulta do próprio mecanismo de otimização que torna esses modelos eficazes. Quanto mais um modelo aprende, maior o risco de que dados raros ou únicos, como registros clínicos de condições pouco prevalentes, fiquem retidos com precisão nos seus parâmetros.

A pseudonimização clássica não protege contra esse risco. Substituir o nome e o número de identificação de um paciente por um código alfanumérico é medida suficiente num banco de dados relacional. Num modelo de aprendizado profundo treinado com dados clínicos ricos, combinações de variáveis como diagnóstico, data de internação, município de residência e resultado de exame permitem reidentificação com alta probabilidade, mesmo sem qualquer identificador direto. O Comitê Europeu de Proteção de Dados (European Data Protection Board — EDPB), em sua Opinion 28/2024 (EDPB, 2024), afirmou que a anonimização verdadeira de dados clínicos de alta dimensionalidade em modelos de IA avançados é virtualmente impossível com as tecnologias atuais.

O risco de memorização tem consequência normativa direta no Brasil. A Agência Nacional de Proteção de Dados (ANPD) publicou o Radar Tecnológico sobre IA generativa em que reconhece esse risco e recomenda que qualquer sistema de IA treinado com dados pessoais sensíveis seja submetido a avaliação específica de risco de reidentificação antes da entrada em uso (Brasil, 2024d). Essa orientação tem endereço preciso: o protocolo de pesquisa deve demonstrar que o risco de memorização foi avaliado, não presumido como inexistente.

O Termo de Consentimento Livre e Esclarecido (TCLE) foi concebido para um mundo de registros localizáveis. Seus modelos tradicionais informam o participante sobre a finalidade da pesquisa, os riscos do procedimento e o direito de retirada. Nenhuma dessas cláusulas

responde às perguntas que o uso de dados em treinamento de modelos de IA exige. A Organização Mundial da Saúde (OMS) reconheceu, em suas diretrizes de ética e governança de IA para a saúde, que o consentimento informado convencional não captura adequadamente os riscos específicos do uso de dados pessoais em sistemas de aprendizado de máquina (OMS, 2021).

O protocolo submetido ao CEP precisa responder a perguntas que os modelos convencionais de consentimento nunca formularam. O dado será usado para treinar um modelo ou apenas para análise estatística? O modelo resultante poderá ser transferido a terceiros? Se o participante retirar o consentimento após o treinamento, qual mecanismo técnico garante a exclusão dos seus dados do modelo?

Fornasier (2022) demonstrou que a ausência de respostas a essas perguntas nos protocolos de pesquisa brasileiros cria uma zona de opacidade jurídica que compromete tanto os direitos dos participantes quanto a validade ética da pesquisa.

O CEP avalia se o pesquisador compreende o que está fazendo com os dados das pessoas que confiaram ao sistema de saúde as informações mais íntimas sobre seus corpos. Essa avaliação exige um novo repertório de perguntas. O Capítulo 8 deste guia, com a Matriz de Avaliação de Risco em Inteligência Artificial em Pesquisa com Seres Humanos (MARIAH), e as seções seguintes deste capítulo oferecem esse repertório.

O ponto de partida, contudo, é o reconhecimento de que o dado de saúde num modelo de IA ocupa uma categoria ética própria. Tratá-lo como qualquer outro dado é a primeira e mais grave falha que um protocolo pode cometer.

## 7.2 AGENTES, PAPÉIS E DIREITOS NA PESQUISA COM DADOS

A governança de dados em pesquisa com IA começa pela identificação precisa de quem é quem. A Resolução CNS n.º 738/2024, em seus artigos 3.º, 9.º a 15 e 16 a 18, nomeia três figuras jurídicas que estruturam toda a cadeia de responsabilidade sobre o banco de dados: o Controlador, o Operador e o Titular (Brasil, 2024b). Essas figuras têm correspondência com a LGPD e ganham, na Resolução, operacionalização específica para pesquisa.

O Controlador é a pessoa natural ou jurídica a quem compete as decisões referentes ao tratamento de dados. No contexto da pesquisa, pode ser o patrocinador, o pesquisador responsável pelo protocolo, o responsável pelo biobanco ou pessoa por ele designada. As

responsabilidades atribuídas são inalienáveis, embora passíveis de compartilhamentos mediante controladoria conjunta, porém não permitem sua transferência.

O Operador realiza o tratamento em nome do Controlador e responde por uso conforme à finalidade aprovada pelo CEP. O Titular é o indivíduo relacionado aos dados, seja como participante de pesquisa ou cidadão cujas informações foram coletadas em contexto assistencial e agora usadas em estudos com bancos de dados existentes.

Essa tripartição tem consequência prática direta para o protocolo. O protocolo deve identificar expressamente quem é o Controlador e, se houver, o Operador. A ausência dessa identificação não é formalidade: indica que o pesquisador não mapeou a cadeia de responsabilidade pelo dado que pretende usar. Em protocolos multicêntricos ou com parceiros externos, a controladoria conjunta deve ser formalizada em Termo de Acordo Institucional, conforme o art. 3.º, XVI e o § 2.º do art. 12 da Resolução CNS n.º 738/2024.

O participante de pesquisa é, também, Titular de dados na acepção da LGPD e da Resolução CNS n.º 738/2024. Essa sobreposição de categorias amplia o conjunto de direitos que lhe são reconhecidos. Os direitos do Titular estabelecidos nos artigos 16 a 18 da Resolução convergem para um princípio fundador, consagrado no artigo 4.º: a autodeterminação informativa, entendida como o direito de cada indivíduo de controlar e proteger seus dados pessoais.

Desse princípio decorrem quatro garantias específicas: acesso às informações armazenadas a qualquer tempo; retificação ou atualização das informações erroneamente inseridas; retirada parcial ou total das suas informações, valendo a desistência a partir da manifestação expressa; e indenização em caso de dano decorrente de uso indevido ou quebra de segurança.

A Lei n.º 14.874/2024 acrescenta a esse conjunto o direito irrenunciável de retirada do consentimento em qualquer fase da pesquisa (Brasil, 2024a). Somam-se a ele os direitos de correção de dados incompletos, inexatos ou desatualizados, assegurados pelo art. 18, III, da Lei n.º 13.709/2018 (LGPD), com repercussão direta sobre as bases de treinamento do modelo. Integra ainda esse conjunto a devolutiva dos resultados. Ao término da pesquisa, o participante tem direito a receber informação sobre os resultados do estudo e sobre eventuais achados com impacto em sua condição de saúde ou no seu aconselhamento em saúde, na forma prevista pelo protocolo. Em estudos convencionais, esse direito se materializa pela exclusão do registro do participante das bases de dados e pela interrupção de qualquer uso

futuro das suas informações. Em pesquisas que utilizam dados de saúde para treinar modelos de IA, esse mecanismo não funciona. O dado já incorporado aos pesos do modelo não tem endereço localizável. Não pode ser excluído por comando administrativo.

A engenharia de software desenvolveu uma resposta técnica a esse problema: o desaprendizado de máquina (*machine unlearning*). A técnica consiste na remoção seletiva da influência de um conjunto específico de dados de um modelo já treinado, sem a necessidade de retreinar o sistema completo a partir do zero, procedimento computacionalmente proibitivo em modelos de grande escala (Cota et al., 2024).

Uma das arquiteturas mais promissoras para viabilizar esse processo em pesquisa em saúde é o modelo SISA (Sharded, Isolated, Sliced, Aggregated). Nessa arquitetura, os dados de treino são particionados de forma que a influência de cada participante fique contida num fragmento específico do modelo, permitindo que a retirada do consentimento seja tecnicamente executada e auditada pelo retreinamento apenas daquela fração.

A existência dessa tecnologia tem consequência direta para o desenho do protocolo. A promessa de exclusão do dado do servidor, suficiente em pesquisas convencionais, não satisfaz o direito de retirada em protocolos com treinamento de modelos de IA. O protocolo precisa descrever o mecanismo técnico pelo qual a influência do dado de um participante poderá ser removida do modelo, caso o consentimento seja retirado. O protocolo que traz promessa genérica de exclusão, sem descrever o mecanismo, oferece uma garantia que a arquitetura do sistema não pode cumprir.

O artigo 20 da LGPD garante ao titular o direito de solicitar a revisão de decisões tomadas com base exclusivamente em tratamento automatizado de dados pessoais (Brasil, 2018). No contexto de pesquisa em saúde com IA, esse direito adquire peso adicional. Algoritmos podem definir elegibilidade para tratamentos experimentais, estabelecer estratificações de risco ou orientar diagnósticos diferenciais. Quando uma decisão com essa magnitude resulta de um modelo opaco, o direito à revisão precisa se traduzir em algo mais do que a possibilidade formal de protocolar um pedido.

A lacuna está na redação. A LGPD brasileira não exige que a revisão de decisões automatizadas seja realizada obrigatoriamente por um ser humano. Mariotto (2024) demonstrou que essa flexibilização abre espaço para que uma máquina audite outra máquina, diferentemente do Regulamento Geral sobre a Proteção de Dados europeu (General Data Protection Regulation — GDPR), que determina expressamente o direito à intervenção

humana no processo de revisão. Para Fornasier (2022), essa diferença torna o marco brasileiro excessivamente permissivo no contexto de aplicações clínicas de alto impacto, e transfere ao CEP a responsabilidade de suprir, por exigência no protocolo, o que a lei não determinou.

O protocolo deve definir o mecanismo concreto de supervisão humana significativa (*human-in-the-loop*) para qualquer decisão com consequências clínicas sobre os participantes. Supervisão humana significativa não equivale à presença formal de um profissional de saúde na cadeia de aprovação. Equivale à capacidade real desse profissional de compreender os critérios da decisão algorítmica, contestá-los com base em evidência clínica e sobrepor seu julgamento ao resultado do modelo. O protocolo que prevê supervisão humana sem garantir essa capacidade real oferece proteção nominal, não proteção efetiva.

O direito de saber completa o conjunto. O participante tem direito de ser informado sobre quais dados foram coletados, para qual finalidade específica foram usados, se foram compartilhados com terceiros e em que condições esse compartilhamento ocorreu. Em pesquisas com IA, esse direito adquire dimensões que os modelos convencionais de consentimento não alcançam. O participante precisa saber não apenas que seus dados foram usados, mas que foram usados para treinar um modelo que poderá operar em escala, ser transferido a outras instituições ou servir de base para produtos comerciais (OMS, 2021).

Em outras palavras, para que esse direito de saber seja exercido de modo efetivo, a transparência precisa ser incorporada ao próprio instrumento que formaliza a participação: o termo de consentimento. É nele que o protocolo deve traduzir, em linguagem clara e verificável, o destino dos dados e do modelo, bem como as condições de uso, compartilhamento e retirada.

O TCLE de um protocolo com IA deve apresentar informações adicionais compatíveis com as características e os riscos do sistema de IA.

- descrição do tipo de modelo a ser treinado e da finalidade específica do sistema resultante.
- identificação de todos os destinatários do modelo, incluindo parceiros comerciais e instituições internacionais.
- prazo de uso dos dados e do modelo.
- mecanismo técnico de retirada do consentimento após o treinamento, com descrição do método de desaprendizado de máquina adotado.
- política de compartilhamento do modelo com terceiros e condições contratuais aplicáveis.

- canal acessível pelo qual o participante pode exercer seus direitos a qualquer momento durante e após a pesquisa (Fornasier, 2022; OMS, 2021).

O protocolo com IA sem essas seis informações no termo de consentimento é um protocolo incompleto. A ausência não é detalhe formal. Indica que o pesquisador não mapeou com precisão o que fará com os dados das pessoas que confiaram ao sistema de saúde as informações mais sensíveis sobre sua existência. O consentimento informado em pesquisa com IA precisa ser tão específico quanto a tecnologia que pretende governar.

## 7.3 BANCOS DE DADOS DE PESQUISA: CONSTITUIÇÃO, USO FUTURO E DISPENSA DE TCLE

A pesquisa em IA para saúde geralmente utiliza bancos de dados, com modelos treinados em conjuntos amplos de registros clínicos, próprios de estudos ou obtidos de bases existentes. A Resolução CNS n.º 738/2024 disciplina essa prática com precisão e estabelece uma dupla via que o protocolo deve observar antes de qualquer avaliação de mérito.

A primeira via, disciplinada no Capítulo VII da Resolução, trata da constituição de bancos de dados no âmbito da pesquisa. Quando um protocolo prevê a criação de um banco durante a execução da pesquisa, o comitê exige quatro descrições: identificação do Controlador ou Controladores; relação dos dados e informações a serem coletados; mecanismos que garantem a confidencialidade e a segurança das informações, incluindo estratégias de acesso e armazenamento; e critérios para o compartilhamento e a transferência dos dados. Qualquer intenção de compartilhar dados ou de anonimizá-los de forma irreversível exige justificativa específica, com ponderação documentada de riscos e benefícios aos participantes (Brasil, 2024b, art. 19).

A segunda via, disciplinada no Capítulo VIII, trata da utilização de bancos já constituídos. Essa é, na prática, a situação mais comum em pesquisa com IA em saúde. O protocolo que pretende usar dados de um banco já existente enfrenta um conjunto maior de exigências, que varia conforme a origem do banco. Caso o banco seja proveniente de outra pesquisa, o protocolo deve incluir autorização do Controlador, descrição dos dados, mecanismos de confidencialidade, identificação do protocolo original, Termo de Compromisso de Uso de Dados assinado e TCLE (Brasil, 2024b, art. 24). Caso o banco seja externo à pesquisa (como registros hospitalares, bases do SUS ou acervos privados), o

protocolo deve incluir a contextualização da origem dos dados e o Termo de Anuência Institucional assinado pelo responsável da instituição (Brasil, 2024b, art. 27).

Essa distinção tem implicação concreta para o protocolo. Um protocolo que pretende treinar um modelo sobre dados do SUS ou de um hospital privado sem apresentar Termo de Anuência Institucional está incompleto por definição, mesmo que todos os outros documentos estejam corretos. O Termo de Anuência não é formalidade: é o ato pelo qual a instituição que detém os dados reconhece o uso previsto no protocolo e assume corresponsabilidade pela cadeia de custódia.

Quatro instrumentos institucionais estruturam essa cadeia de custódia em pesquisa com bancos de dados. O Termo de Acordo Institucional formaliza a controladoria conjunta em bancos constituídos com participação de mais de uma instituição, com critérios de partilha e destinação dos dados em caso de dissolução. O Termo de Anuência Institucional atesta que a instituição detentora dos dados autoriza seu uso no protocolo. O Termo de Compromisso de Uso de Dados é a declaração formal em que o pesquisador responsável e sua equipe se comprometem com o sigilo, a confidencialidade e o uso dos dados para a finalidade aprovada. O Termo de Transferência de Informações formaliza a passagem de dados entre bancos, e assume responsabilidade pela guarda e pelo respeito ao sigilo. A ausência de qualquer desses termos, quando o protocolo os exige pela sua natureza, transforma a governança de dados em figura decorativa.

O artigo 20 da Resolução CNS n.º 738/2024 submete ao SINEP toda pesquisa que pretenda constituir banco de dados ou utilizar banco constituído para outras finalidades, e fixa, em seus parágrafos, o regime do consentimento. A regra do § 1.º é o consentimento prévio do participante. Duas são as situações em que o novo consentimento pode ser dispensado: quando o uso futuro já foi consentido no Registro ou no TCLE da pesquisa original (§ 2.º) e quando os dados disponibilizados são anonimizados pelo Controlador (§§ 3.º e 5.º). O § 5.º é decisivo para a pesquisa com IA: em bancos constituídos fora do âmbito da pesquisa — prontuários, registros e bases administrativas, origem mais frequente dos dados de treinamento —, a anonimização pelo Controlador é a única via de dispensa. Para os bancos já constituídos no âmbito da pesquisa, o artigo 25 detalha esse regime e enumera cinco situações de dispensa: dados coletados originalmente sem identificação do participante; consentimento prévio do participante para uso de dados em pesquisas futuras; compartilhamento de dados anonimizados pelo Controlador; anonimização irreversível de

dados no protocolo original (Brasil, 2024b). O rol do artigo 25 não alcança os bancos constituídos fora do âmbito da pesquisa, disciplinados pelo artigo 27, cuja dispensa segue a regra do artigo 20, § 5º. A distinção é operacional: o CEP classifica primeiro a origem do banco e só então verifica a hipótese normativa aplicável. Isso é decisivo para a pesquisa com IA, porque a maior parte dos modelos é treinada sobre dados coletados anteriormente.

As hipóteses de dispensa previstas nas normas aplicáveis exigem apreciação individualizada. O protocolo deve indicar a hipótese normativa invocada e apresentar justificativa ética e metodológica, com descrição da origem e da finalidade do banco, dos riscos envolvidos e das salvaguardas adotadas. A origem do banco orienta a identificação dos requisitos aplicáveis, mas não constitui, isoladamente, fundamento para a dispensa. A decisão cabe ao CEP e deve ser expressamente fundamentada.

Quando o consentimento não pode ser dispensado e não foi autorizado na pesquisa original, o protocolo deve apresentar novo TCLE, ou do responsável legal, para cada participante cujos dados serão reutilizados. Em bancos grandes, isso pode ser operacionalmente inviável. A resposta correta a essa inviabilidade não é desconsiderar a exigência, mas reformular a pergunta de pesquisa para trabalhar apenas com dados originalmente anonimizados ou que tenham consentimento prévio para uso futuro.

## 7.4 SOBERANIA DE DADOS E TRANSFERÊNCIA INTERNACIONAL

Soberania de dados em saúde expressa uma escolha política. Significa que o Estado e os cidadãos de um país determinam quem acessa, processa e se beneficia dos dados produzidos pela sua população. No campo da IA, essa escolha adquire dimensão estratégica adicional: os dados de saúde de uma população constituem a matéria-prima dos modelos que, uma vez treinados, passam a orientar diagnósticos, triagens e decisões clínicas sobre essa mesma população (Santos *et al.*, 2025). Ceder esses dados sem salvaguardas equivale a ceder a capacidade de moldar a inteligência que atuará sobre o próprio sistema de saúde.

Dados de saúde não operam como insumo neutro. A literatura recente identifica um padrão estrutural de extração de dados do Sul Global por plataformas concentradas no Norte global, com efeitos diretos sobre dependência tecnológica e assimetria de poder sanitário. Santos *et al.* (2025) descrevem esse processo como colonialismo digital: países periféricos fornecem dados brutos enquanto plataformas globais concentram infraestrutura, capacidade

de processamento e captura de valor. O modelo treinado com dados da população brasileira retorna como produto estrangeiro. O sistema público perde a capacidade de governar o conhecimento produzido a partir da sua própria base.

O Brasil construiu um arcabouço normativo que reconhece esse caráter estratégico. A LGPD classifica dados de saúde como dados pessoais sensíveis e exige base legal robusta para qualquer transferência internacional (Brasil, 2018). A Lei n.º 14.874/2024 e o Decreto n.º 12.651/2025 reforçam a responsabilidade institucional sobre dados usados em pesquisa, inclusive quando o tratamento ocorre fora do território nacional (Brasil, 2024a; Brasil, 2025a).

Matsushita e Almeida (2025) situam esse conjunto normativo num movimento internacional de convergência regulatória em torno de princípios comuns: finalidade específica; minimização de dados; transparência; segurança; responsabilização. A potência desse arcabouço depende de uma condição que nenhuma norma garante sozinha: protocolos de pesquisa que identifiquem com precisão para onde vão os dados, em nome de quem e com que garantias.

Dois episódios, um internacional e um nacional, ilustram o que acontece quando essa condição não é cumprida. Em 2015, o Google DeepMind firmou acordo com o Serviço Nacional de Saúde (National Health Service — NHS) britânico para acessar dados clínicos de 1,6 milhão de pacientes de três hospitais de Londres. Os dados incluíam registros de internação, resultados de exames, diagnósticos e históricos de atendimento.

Eles foram transferidos sem consentimento dos pacientes, sem base legal adequada e sem que os titulares soubessem que suas informações alimentariam o desenvolvimento de um sistema de IA de uma das maiores corporações tecnológicas do mundo (Hern, 2017). A autoridade britânica de proteção de dados declarou que o Royal Free NHS Trust violou a lei ao não informar adequadamente os pacientes sobre o uso das suas informações. Powles e Hodson (2017) descreveram o caso como advertência para a era algorítmica na saúde: a ausência de supervisão ética independente e a velocidade com que o acordo foi firmado criaram um precedente de transferência de dados públicos de saúde a interesses privados sem prestação de contas.

O segundo episódio é brasileiro. Em novembro de 2020, em plena pandemia, uma planilha com dados de cerca de 16 milhões de pessoas diagnosticadas ou suspeitas de COVID-19 foi acidentalmente disponibilizada em repositório público do GitHub por um funcionário terceirizado de hospital privado que prestava serviços ao Ministério da Saúde.

A planilha continha nomes completos, documentos de identificação, endereços, diagnósticos clínicos e classificações de risco, incluindo dados do Presidente da República, de ministros de Estado e de suas famílias. A exposição durou pelo menos seis meses antes de ser detectada por pesquisadores independentes e noticiada.

Para tanto, a ANPD instaurou procedimento investigatório, mas o caso foi encerrado sem sanção consubstanciada, em parte porque a Agência ainda operava em fase transitória sem aplicar sanções administrativas (Brasil, 2026d). O episódio mostrou que a exposição de dados clínicos brasileiros ocorre tanto em acordos internacionais quanto na rotina da cadeia de custódia, incluindo terceirizados, consultorias, repositórios e ambientes de desenvolvimento. A cadeia de responsabilidade, quando não é desenhada com explicitude, é preenchida pelo acaso.

Os dois casos, lidos em conjunto, apontam para a mesma lição. O padrão identificado é estrutural. Instituições públicas e privadas manuseiam bases clínicas extensas sob lógica de projeto, com cadeias de custódia frágeis, sem supervisão ética independente dedicada ao dado e sem instrumentos contratuais suficientes. O sistema público perde o controle sobre o conhecimento produzido a partir da sua própria base (Santos *et al.*, 2025) e sobre a segurança cotidiana dos dados que deveria proteger.

O SUS produz diariamente um dos maiores acervos de dados clínicos do mundo em desenvolvimento. Esse acervo representa quase quatro décadas de atendimento a toda a população brasileira, com diversidade epidemiológica, étnica e regional que nenhuma plataforma privada global reproduz. Protocolos de pesquisa que preveem o acesso a esses dados por parceiros internacionais ou corporações de tecnologia precisam demonstrar, com precisão, que nem o acordo DeepMind nem a exposição da planilha de 2020 se repetirão no Brasil.

O protocolo de pesquisa que prevê transferência internacional de dados de saúde precisa responder a um conjunto mínimo de perguntas antes de qualquer análise de mérito científico:

- qual a base legal que sustenta a transferência, nos termos da LGPD?
- o país ou organização destinatária oferece nível adequado de proteção de dados, reconhecido pela ANPD, ou o protocolo apresenta salvaguardas contratuais específicas?
- o modelo treinado com os dados brasileiros poderá ser reutilizado, transferido a terceiros ou incorporado a produtos comerciais?

- como a instituição brasileira exercerá sua responsabilidade de custódia sobre os dados após a transferência, dado que a Lei n.º 14.874/2024 atribui dever integral de proteção independentemente do local de processamento (Brasil, 2024a)?

O protocolo que não responde a essas perguntas está incompleto. Sem essas respostas, aprova-se o que não se avaliou.

Duas alternativas técnicas emergem na literatura como caminhos para reduzir a necessidade de transferência internacional de dados brutos de saúde. A primeira é o uso de dados sintéticos: conjuntos gerados artificialmente que reproduzem os padrões estatísticos de uma base real sem expor diretamente as informações identificáveis dos pacientes. A segunda é a transferência de aprendizagem (*transfer learning*): modelos treinados em ambiente nacional que exportam apenas os parâmetros aprendidos, não os dados originais, permitindo adaptação a outros contextos sem circulação da base clínica.

Santos *et al.* (2025) identificam essas estratégias como alternativas ao extrativismo de dados, ao priorizar o compartilhamento do aprendizado em lugar do dado bruto. Ambas se alinham aos princípios de finalidade, necessidade e prevenção da LGPD (Brasil, 2018).

Essas alternativas têm limites que o pesquisador precisa conhecer. Dados sintéticos reproduzem padrões populacionais, mas podem não capturar com fidelidade subgrupos raros, condições pouco prevalentes ou perfis epidemiológicos específicos de populações como povos originários e comunidades quilombolas. Um modelo treinado sobre dados sintéticos que sub-representam esses grupos reproduz, no sintético, o mesmo viés que o dado real já continha. A transferência de aprendizagem, por sua vez, pressupõe que o modelo de origem foi treinado sobre dados suficientemente representativos da população-alvo. Quando essa condição não se verifica, o modelo transferido carrega os vieses do contexto original para o novo ambiente, com a agravante de que esses vieses se tornam mais difíceis de identificar (Borges *et al.*, 2026). O uso dessas técnicas reduz riscos, mas não os elimina.

A soberania de dados em pesquisa com IA resolve-se por decisão ética e política anterior à técnica. A avaliação de um protocolo com transferência internacional de dados define, naquele ato, se os dados clínicos produzidos pelo SUS atravessarão fronteiras sob proteção equivalente à garantida no território nacional. Essa decisão antecede qualquer escolha sobre métodos de anonimização, arquiteturas de modelo ou instrumentos contratuais. A soberania de dados em pesquisa começa no parecer do comitê.

## 7.5 APRENDIZAGEM FEDERADA E ARQUITETURAS DE PRIVACIDADE

A aprendizagem federada (*federated learning*) responde a um problema concreto de governança: como treinar modelos de IA com dados de múltiplas instituições sem que nenhuma delas renuncie à custódia dos seus registros clínicos. A resposta técnica inverte a lógica convencional. Em vez de os dados migrarem para o modelo, o modelo visita os dados. Cada instituição participante treina localmente uma versão do algoritmo com seu próprio acervo. O que circula entre as instituições e o servidor central são apenas os parâmetros matemáticos aprendidos, não os registros dos pacientes. O dado clínico permanece onde foi produzido (Rieke *et al.*, 2020).

O modelo aprende com vários contextos sem acessar dados individuais diretamente. Para pesquisa em saúde no Brasil, essa arquitetura tem implicação direta. Um protocolo que utilize a aprendizagem federada pode reunir, num único modelo, o aprendizado clínico de hospitais universitários do Sul, unidades básicas do Nordeste e laboratórios de referência da Amazônia, sem que nenhum registro atravesse os limites da instituição que o gerou. A diversidade epidemiológica do SUS torna-se ativo técnico, não obstáculo logístico.

A realidade brasileira, contudo, revela uma assimetria preocupante entre o potencial da técnica e seu uso efetivo. Borges *et al.* (2026) conduziram uma revisão de escopo sobre aplicações de IA na saúde pública brasileira e identificaram, entre 349 estudos recuperados, apenas seis que utilizaram explicitamente aprendizagem federada ou transferência de aprendizagem. Os autores concluem que essas técnicas permanecem marginais na prática cotidiana, apesar de sua capacidade de viabilizar soluções de IA com recursos computacionais limitados e de preservar a soberania de dados. A adoção dessas abordagens ocorreu principalmente em consórcios multicêntricos internacionais, o que indica que a colaboração interinstitucional e a capacitação técnica são condições para sua difusão no sistema público brasileiro.

A aprendizagem federada representa um avanço ético relevante. Um protocolo que a adota reduz o risco de transferência indevida de dados sensíveis e fortalece a governança nacional sobre o acervo clínico do SUS. O comitê que recebe um protocolo com essa arquitetura deve reconhecer esse mérito e, ao mesmo tempo, verificar se a técnica foi descrita com precisão suficiente para que seus benefícios sejam reais e não apenas retóricos.

A aprendizagem federada reduz o risco de exposição de dados clínicos, mas não o elimina. Essa distinção precisa constar nos protocolos com nitidez. Os parâmetros matemáticos que circulam entre as instituições e o servidor central carregam informação sobre os dados que os geraram. Em condições específicas, pesquisadores demonstraram que esses parâmetros podem ser revertidos para reconstruir fragmentos dos dados originais, num procedimento denominado ataque de inversão de gradiente (*gradient inversion attack*) (Geiping *et al.*, 2020). O dado que não saiu do hospital pode, portanto, ser parcialmente reconstituído a partir do que saiu.

A gravidade prática desse risco depende de fatores que o protocolo deve descrever: o tamanho do lote de dados usado em cada rodada de treinamento local, a frequência das atualizações enviadas ao servidor central e a adoção ou não de técnicas complementares de proteção. A privacidade diferencial (*differential privacy*) introduz ruído calibrado nos parâmetros antes do envio, reduzindo a precisão de eventuais ataques de inversão. A agregação segura (*secure aggregation*) impede que o servidor central acesse os parâmetros individuais de cada instituição, processando apenas o resultado agregado. Rieke *et al.* (2020) apontam que a aprendizagem federada alcança seu potencial protetivo pleno apenas quando combinada com essas salvaguardas adicionais.

Existe também um risco de outra natureza, menos visível e igualmente relevante. A aprendizagem federada pressupõe que os dados de cada instituição participante são suficientemente representativos para contribuir com aprendizado útil ao modelo global. Quando uma instituição concentra pacientes de um perfil demográfico específico, o modelo aprende desproporcionalmente com esse perfil. O resultado agregado pode apresentar desempenho inferior exatamente nas populações sub-representadas nos dados locais de cada nó da rede (Borges *et al.*, 2026). O problema do viés algorítmico não desaparece com a arquitetura federada, migra para dentro dela.

A criptografia homomórfica (*homomorphic encryption*) responde ao ponto de vulnerabilidade adicional que a aprendizagem federada deixa em aberto: o servidor de agregação. A criptografia homomórfica é um conjunto de técnicas matemáticas que permitem realizar operações sobre dados cifrados sem que seja necessário decifrá-los em nenhum momento do processo. O servidor de agregação recebe parâmetros cifrados, opera sobre eles e devolve um resultado cifrado às instituições. Nunca acessa o conteúdo. O dado permanece protegido mesmo do sistema que o processa.

Mesmo que os dados clínicos permaneçam nas instituições, o servidor precisa processar os parâmetros recebidos para gerar o modelo global. Um servidor comprometido, controlado por ator mal-intencionado ou sujeito a ordem judicial em jurisdição estrangeira, pode explorar esses parâmetros.

A aplicação dessa técnica à saúde pública brasileira já conta com evidência nacional. Lunkes, Lunkes e Oliveira (2024) demonstraram a viabilidade da criptografia homomórfica na análise de dados epidemiológicos de dengue, usando a biblioteca TensorSeal com Python para executar modelos sobre dados cifrados sem decifração em nenhuma etapa do processamento.

O trabalho, desenvolvido no Laboratório Nacional de Computação Científica (LNCC), vinculado ao Ministério da Ciência, Tecnologia e Inovações, indica que o Brasil possui capacidade técnica instalada para avançar nessa direção em pesquisa com dados sensíveis de saúde. O custo computacional ainda limita aplicações de larga escala, e a técnica permanece em fronteira de pesquisa para uso clínico rotineiro. Esse horizonte tem implicação prática imediata: protocolos que combinam aprendizagem federada com criptografia homomórfica oferecem o nível mais alto de proteção atualmente disponível para dados clínicos em pesquisa colaborativa com IA, e merecem reconhecimento diferenciado na avaliação ética.

Cinco perguntas estruturam a avaliação de um protocolo com aprendizagem federada.

A primeira: quem controla o servidor de agregação e sob qual jurisdição? Um servidor controlado por parceiro estrangeiro recria, de forma parcial, o problema de soberania que a arquitetura federada pretende resolver. A segunda: quais técnicas complementares de proteção foram adotadas, como privacidade diferencial, agregação segura ou criptografia homomórfica, e como sua adoção foi documentada? A terceira: as instituições participantes do consórcio representam adequadamente a diversidade epidemiológica da população-alvo, incluindo regiões periféricas, povos originários e comunidades quilombolas? A quarta: o modelo agregado ao final da pesquisa poderá ser transferido a terceiros, licenciado ou incorporado a produtos comerciais? A quinta: o TCLE informa os participantes sobre essa arquitetura distribuída de treino, em linguagem acessível e sem omitir o destino do modelo resultante?

O protocolo que responde a essas cinco perguntas com precisão oferece base suficiente para uma avaliação fundamentada. O protocolo que as omite ou as responde de

forma genérica sinaliza que o pesquisador não mapeou os riscos com a profundidade que a técnica exige.

A assimetria identificada por Borges *et al.* (2026), ou seja, a aprendizagem federada marginal no Brasil e adotada principalmente em consórcios internacionais, reforça a responsabilidade de descrever a técnica com precisão: quanto menos difundida a técnica, menor a familiaridade dos pesquisadores com suas exigências éticas, e maior a probabilidade de protocolos que invocam a aprendizagem federada sem cumprir as condições que a tornam efetivamente protetiva.

## 7.6 O QUE O PROTOCOLO DEVE CONTER E O QUE O CEP VERIFICA

O Relatório de Impacto à Proteção de Dados Pessoais (RIPD) é o instrumento previsto no artigo 38 da LGPD para documentar os processos de tratamento de dados pessoais que podem gerar riscos às liberdades civis e aos direitos fundamentais dos titulares (Brasil, 2018). A ANPD publicou orientação específica sobre o RIPD, que estabelece que o documento deve conter, no mínimo, a descrição dos tipos de dados coletados; a metodologia usada para a coleta e para a garantia da segurança das informações; e a análise do controlador com relação a medidas, salvaguardas e mecanismos de mitigação de risco adotados (Brasil, 2026d). Em pesquisa com IA que utiliza dados de saúde, o RIPD deixa de ser recomendação e passa a ser obrigação. O tratamento de dados sensíveis em larga escala, com finalidade de treinamento algorítmico e potencial de reidentificação, configura exatamente o cenário de alto risco que a lei visa mapear antes que o dano ocorra.

O RIPD para pesquisa com IA tem estrutura e exigências distintas do RIPD convencional. A ANPD (Brasil, 2024d), em seu Radar Tecnológico sobre inteligência artificial generativa, reconhece que modelos de linguagem de grande escala apresentam riscos específicos de memorização e de reidentificação que não constam dos modelos tradicionais de avaliação de impacto.

Um RIPD adequado para pesquisa com IA deve responder a quatro dimensões que os modelos convencionais não contemplam: o risco de memorização de dados sensíveis pelo modelo durante o treinamento; o risco de reidentificação de participantes a partir de combinações de variáveis clínicas mesmo após pseudonimização; o risco de deriva de dados (*data drift*) quando o modelo for aplicado a populações distintas da base de treino; e o risco

de transferência indevida do modelo a terceiros após o término da pesquisa. A ausência de qualquer dessas dimensões no RIPD submetido ao CEP indica que o documento foi elaborado sem considerar as especificidades técnicas e éticas da pesquisa com IA.

O RIPD submetido com um protocolo de pesquisa com IA deve mapear o fluxo de dados do início ao fim. Esse mapeamento abrange, no mínimo, seis etapas: a coleta dos dados clínicos e sua base legal; a pseudonimização ou anonimização e o método técnico adotado; o armazenamento e as medidas de segurança aplicadas; o treinamento do modelo e os riscos de memorização avaliados; o compartilhamento com parceiros e as salvaguardas contratuais aplicáveis; e o destino dos dados e do modelo após o encerramento da pesquisa.

Um RIPD que descreve apenas as etapas iniciais de coleta e armazenamento, sem mapear o que acontece com o modelo resultante, cobre a fase menos arriscada do ciclo e omite a fase mais crítica. O documento deve acompanhar o dado até o fim de sua vida útil na pesquisa, e o CEP verifica se ele o faz.

Avaliar um RIPD de pesquisa em IA não exige conhecimento técnico aprofundado em IA. Bastam perguntas críticas, objetivas e vinculadas ao interesse da pessoa participante e da sociedade: o documento descreve o ciclo de vida completo dos dados? Identifica os riscos específicos de reidentificação para a população estudada? Avalia o risco de memorização pelo modelo? Descreve as salvaguardas técnicas adotadas e demonstra sua adequação aos riscos identificados? Quantifica os riscos residuais após as salvaguardas? Define responsabilidades claras em caso de incidente de segurança? Um RIPD que responde afirmativamente a essas perguntas com evidência documental atende ao padrão mínimo para pesquisa com IA em saúde. Um RIPD que as ignora ou as responde de forma genérica sinaliza que o pesquisador não compreendeu os riscos do que propõe.

Além do RIPD, o protocolo de pesquisa com IA que usa dados de saúde exige um conjunto documental específico que os modelos convencionais de submissão não contemplam. Esse conjunto articula quatro documentos técnicos com os quatro Termos institucionais da Resolução CNS n.º 738/2024. O primeiro documento técnico é o plano de gestão de dados, que descreve o ciclo de vida completo das informações coletadas: origem, formato, padrão de interoperabilidade, local de armazenamento, controles de acesso, prazo de retenção e destino após o encerramento da pesquisa. O plano deve cobrir não apenas os dados brutos, mas o modelo treinado e os parâmetros intermediários gerados durante o processo de aprendizado.

O segundo documento é o Termo de Acordo Institucional, exigível sempre que o protocolo envolver mais de uma instituição na constituição ou utilização conjunta do banco de dados. O acordo define as respectivas responsabilidades, os critérios de partilha e a destinação dos dados em caso de dissolução futura. Em protocolos com IA, o instrumento adquire dimensão adicional porque o compartilhamento pode incluir o próprio modelo resultante, não apenas os dados de origem. A Carta Circular CNS n.º 1/2021 já orientava os CEPs sobre a necessidade de verificar instrumentos de compartilhamento em pesquisas colaborativas (Brasil, 2021); a Resolução n.º 738/2024 consolidou essa orientação em exigência formal.

O terceiro documento é o Termo de Anuência Institucional, obrigatório em protocolos que usam bancos de dados constituídos fora do âmbito da pesquisa, conforme o art. 27 da Resolução n.º 738/2024. A anuência é emitida pelo dirigente da instituição de onde os dados são provenientes e atesta a autorização para o uso previsto, e identifica cargo, função e assinatura do responsável.

O quarto documento é o Termo de Compromisso de Uso de Dados, pela qual o pesquisador responsável e sua equipe se comprometem com o sigilo, a confidencialidade dos dados e seu uso para a finalidade prevista na pesquisa. Soma-se, quando aplicável, o Termo de Transferência de Informações, exigível quando há passagem de dados entre bancos ou instituições.

A abrangência desse conjunto documental é proporcional ao volume e à sensibilidade dos dados da pesquisa. Pesquisas que processam grandes volumes de dados de saúde, como bases populacionais, repositórios de imagens, prontuários em escala ou monitoramento contínuo por dispositivos, exigem o conjunto completo descrito acima. Nas pesquisas de menor volume ou com dados menos sensíveis, a exigência documental é proporcional, sem redução do rigor da análise. A proporcionalidade gradua os requisitos exigidos do protocolo, não os princípios éticos da avaliação, que permanecem os mesmos para todas as pesquisas.

O TCLE com as cláusulas específicas para pesquisa com IA descritas na seção 7.2 permanece obrigatório quando a dispensa não se configura em nenhuma das hipóteses previstas nos arts. 20, §§ 2.º, 3.º e 5.º, e 25 da Resolução n.º 738/2024. A OMS (2021) recomenda que protocolos com IA apresentem termos de consentimento adaptados que descrevam, em linguagem acessível, os riscos específicos do uso de dados pessoais em

sistemas de aprendizado de máquina. Num protocolo com IA, um TCLE genérico configura um consentimento que não informa.

Quando o protocolo prevê transferência internacional de dados ou de modelo, um instrumento jurídico de salvaguarda contratual acresce a esse conjunto. A LGPD admite três mecanismos para transferência internacional de dados pessoais sensíveis: a transferência para países com nível de proteção reconhecido como adequado pela ANPD; a adoção de cláusulas contratuais padrão aprovadas pela ANPD; ou a elaboração de normas corporativas globais aprovadas pela autoridade competente (Brasil, 2018). O protocolo deve indicar qual mecanismo se aplica, apresentar o instrumento correspondente e demonstrar que suas cláusulas cobrem o uso do modelo após o término da pesquisa. Um acordo que protege os dados brutos, mas silencia sobre o modelo treinado protege a matéria-prima e libera o produto.

A governança de um protocolo de IA aplicado à saúde requer mais do que documentos; demanda transparência sobre as atribuições decisórias em cada etapa do ciclo de vida da pesquisa. O relatório Wilton Park (2026), produzido em associação com a OMS e o Co-Develop, propõe um quadro prático de responsabilidades para infraestruturas digitais em saúde que se aplica diretamente à avaliação ética de protocolos com IA.

O quadro organiza quatro funções distintas: quem responde diretamente pela execução; quem presta contas pelos resultados; quem deve ser consultado antes das decisões; e quem deve ser informado sobre o andamento. Em protocolos com IA, esse mapeamento precisa cobrir ao menos seis dimensões: controle sobre os dados; responsabilização em caso de incidentes de segurança; validação da arquitetura técnica; supervisão dos resultados do modelo; autorização para integração com sistemas externos; e prestação de contas aos participantes e à instituição executora. O protocolo que não mapeia essas responsabilidades opera sem governança real, independentemente da sofisticação técnica do sistema proposto.

O mesmo relatório identifica o aprisionamento a fornecedores (*vendor lock-in*) como risco persistente em sistemas digitais de saúde. Plataformas proprietárias oferecem conveniência imediata, mas produzem dependência de longo prazo, limitam auditoria e reduzem a capacidade do Estado de corrigir rumos (Wilton Park, 2026). Esse risco tem endereço ético preciso. Um protocolo de pesquisa com IA que depende de plataforma estrangeira fechada, sem acesso ao código-fonte, sem documentação auditável da arquitetura

e sem cláusula de portabilidade dos dados e do modelo, aprova junto com a pesquisa uma limitação estrutural a qualquer supervisão ética futura. O protocolo deve prever arquitetura auditável, padrões abertos e mecanismo de saída que preserve a soberania institucional sobre os dados produzidos pela pesquisa, e o CEP verifica se essas garantias estão presentes.

A governança de dados em pesquisa com IA não começa nos contratos de transferência, nos servidores de armazenamento, nem nos mecanismos técnicos de proteção. Começa no momento em que se lê um protocolo e se pergunta: esse pesquisador sabe o que está pedindo às pessoas que vão participar desta pesquisa? Sabe o que fará com os dados delas? Sabe para onde irão? O capítulo que o leitor conclui aqui oferece as ferramentas conceituais, jurídicas e práticas para que essa pergunta tenha resposta fundamentada. O SUS atende a totalidade da população brasileira. Os dados que esse atendimento gera pertencem a essas pessoas. O CEP existe para garantir que a pesquisa com IA os trate com a dignidade que essa pertença exige. O Capítulo 8 apresenta a Matriz de Avaliação de Risco em Inteligência Artificial, instrumento prático que traduz essa governança em avaliação proporcional de risco.

# 8 MARIAH: MATRIZ DE AVALIAÇÃO DE RISCO EM INTELIGÊNCIA ARTIFICIAL (VERSÕES A E B)

## 8.1 O PERCURSO DE CONSTRUÇÃO E AS NOVAS QUESTÕES PARA OS COMITÊS DE ÉTICA EM PESQUISA

O Capítulo 8 pode ser consultado de forma sequencial ou por seção, conforme a necessidade do leitor.

A construção de um instrumento de avaliação ética para protocolos de pesquisa que utilizam inteligência artificial exige, antes de qualquer decisão metodológica, uma escolha de postura: o instrumento existe para apoiar o julgamento do CEP ou para substituí-lo?

Este guia responde a essa pergunta sem ambiguidade. As matrizes apresentadas nas Partes III (Versão A) e IV (Versão B) da MARIAH são ferramentas de APOIO à deliberação. Elas não são novas normas, não criam obrigações além das já estabelecidas no âmbito do Sistema Nacional de Ética em Pesquisa (SINEP), e não têm caráter vinculante. Sua adoção é uma escolha institucional de cada CEP. O que elas oferecem é precisão: um percurso que torna a avaliação mais sistemática, mais documentada e mais justificável perante os participantes da pesquisa e a sociedade.

Esse posicionamento não é uma concessão, mas uma consequência direta do que a experiência internacional demonstrou. Nenhum dos oito modelos regulatórios avaliados conseguiu criar um instrumento universal eficaz para todos os contextos institucionais previstos.

O modelo canadense de Avaliação de Impacto Algorítmico é robusto, mas foi desenhado para serviços públicos governamentais, não para pesquisa com seres humanos (Canadá, 2024). As diretrizes do Conselho Nacional de Saúde e Pesquisa Médica da Austrália (National Health and Medical Research Council — NHMRC) são precisas no tratamento do ciclo de vida, mas não contemplam as especificidades jurídicas da proteção de dados em países de renda média (Austrália, 2023). As diretrizes do Conselho Indiano de Pesquisa Médica (Indian Council of Medical Research — ICMR) são as únicas criadas para orientar CEPs na avaliação de protocolos com IA, mas foram desenvolvidas para um contexto diferente do brasileiro (ICMR, 2023). Cada modelo ofereceu contribuições reais e apresentou lacunas reais. A MARIAH é a síntese dessas contribuições, ancorada no direito brasileiro e orientada pelo contexto do SUS.

O percurso que levou à construção das duas versões que serão apresentadas nas próximas seções foi deliberadamente longo. A análise considerou cinco eixos de avaliação consagrados internacionalmente e, durante a pesquisa, incluiu contribuições como o conceito de Oportunidade de Contestar (Opportunity to Override), que expressa a condição real de o profissional contestar o resultado do sistema, da experiência coreana (MFDS, 2022). O direito do participante de solicitar a exclusão de seus dados do treinamento do modelo provém do modelo indiano (ICMR, 2023). O Protocolo de Mudança de Algoritmos (Algorithm Change Protocol), que trata da responsabilidade do pesquisador de prever como o modelo será retreinado, decorre da experiência regulatória indiana e japonesa (CDSCO, 2025; PMDA, 2023). O conceito de contexto de uso como etapa prévia obrigatória à classificação de risco deriva da diretriz da Food and Drug Administration estadunidense (FDA, 2025). Cada um desses elementos ampliou e qualificou o instrumento final.

O resultado desse percurso são duas versões de um mesmo instrumento, com a mesma base ética e o mesmo conjunto de perguntas essenciais.

O material operacional correspondente está na Parte II da MARIAH.

A avaliação ética de protocolos com IA não pode ser reduzida a uma formalidade. Nenhuma das duas versões da matriz admite esse risco.

A IA assimila os vieses dos dados de treinamento, utiliza lógicas pouco transparentes e gera impactos sobre participantes que nenhum outro método contemporâneo produz com igual escala e rapidez. Avaliar um protocolo com IA não é avaliar uma tecnologia, mas o impacto que essa tecnologia pode produzir sobre seres humanos concretos, em contextos concretos, com vulnerabilidades concretas. As matrizes deste guia existem para tornar essa avaliação mais rigorosa. A responsabilidade pela avaliação permanece, integralmente, com os CEPs.

## 8.2 FUNDAMENTOS, FONTES E DESCRIÇÃO DAS MATRIZES

Esta seção apresenta as bases teóricas, normativas e metodológicas que sustentam as duas versões da MARIAH. Cada escolha de desenho tem fundamento em experiências regulatórias documentadas e no marco jurídico brasileiro. Isso vale para os eixos, as perguntas, os pesos e a lógica de chegada ao nível de risco. A descrição das duas versões está organizada em subseções independentes. A seção 8.2.1 trata da Versão A Qualitativa. A seção

8.2.2 trata da Versão B Quantitativa. As duas versões compartilham a mesma base ética e o mesmo conjunto de perguntas essenciais. O que as distingue é a lógica de operação. Essa distinção determina o público e o contexto para os quais cada versão é mais adequada.

MARIAH é um instrumento de gradação proporcional. Opera sobre um espectro que vai do risco baixo ao risco crítico e reconhece que a maioria dos protocolos com IA ocupa posições intermediárias que admitem salvaguardas, condicionantes e mitigações. Essa gradação tem um limite que o instrumento não transpõe. Três dimensões não são graduáveis pela MARIAH: dano direto e potencialmente irreversível ao participante sem mecanismo documentado de interrupção e reparação; ausência de supervisão humana efetiva em contexto em que o sistema influencia decisões clínicas críticas; e incerteza técnica não documentada sobre o comportamento do sistema no contexto de uso declarado.

Diante dessas dimensões, o princípio de não maleficência e o princípio de precaução, estabelecidos pela Lei n.º 14.874/2024, prevalecem sobre qualquer resultado que a soma dos blocos produza (Brasil, 2024a). A matriz estrutura o julgamento do CEP e não esgota seus limites éticos. Esses limites estão nos princípios que a lei estabelece e que nenhum instrumento de avaliação tem competência para relativizar.

### 8.2.1 Versão A qualitativa

A avaliação ética de protocolos de pesquisa que utilizam IA exige instrumentos que traduzam a complexidade técnica dos sistemas algorítmicos de impacto ético mensurável sobre a pessoa participante. A MARIAH, Versão A qualitativa, foi construída a partir dessa premissa, com fundamento em oito experiências regulatórias internacionais e ancorada no marco jurídico brasileiro.

A escolha pela abordagem qualitativa

A Versão A da Matriz adota uma lógica qualitativa estruturada: perguntas de resposta binária, organizadas em cinco eixos temáticos, percorridos em sequência por meio de um fluxograma. Essa escolha se apoia na experiência do modelo australiano, que demonstrou que instrumentos de avaliação ética são mais eficazes quando orientam o julgamento do avaliador sem substituí-lo. A avaliação não exige domínio da arquitetura técnica do sistema, mas a

compreensão do impacto que ele pode produzir. A lógica qualitativa preserva essa centralidade do julgamento ético, ao mesmo tempo que garante sistematicidade e rastreabilidade ao processo de avaliação (Austrália, 2023).

O formato de perguntas binárias, em que qualquer resposta de risco aumenta a classificação no eixo, foi adaptado do modelo canadense de Avaliação de Impacto Algorítmico (Algorithmic Impact Assessment — AIA). O AIA canadense demonstrou, em uso consolidado desde 2019, que perguntas objetivas e bem formuladas são instrumentos eficazes para avaliadores sem formação técnica em ciência da computação. A versão brasileira adota essa lógica e a amplia para o cenário da pesquisa que envolve seres humanos. Ela direciona a atenção da prestação de serviços públicos governamentais para a proteção dos participantes dessas pesquisas (Canadá, 2024).

O filtro de entrada e a caracterização do contexto de uso

O Passo 0 (filtro de entrada) distingue usos rotineiros de usos que demandam avaliação aprofundada. Essa distinção é essencial. Sem um filtro inicial, os comitês analisam indiscriminadamente desde ferramentas para organização bibliográfica até sistemas autônomos de apoio diagnóstico. Isso compromete a eficiência e a proporcionalidade do processo. A diretriz da FDA dos Estados Unidos da América (EUA), publicada em janeiro de 2025, demonstrou que nenhuma avaliação de risco algorítmico começa sem a definição precisa do contexto de uso. Trata-se do papel específico do sistema e do escopo da pergunta que ele foi desenvolvido para responder (FDA, 2025). A Matriz incorpora essa exigência como etapa prévia obrigatória aos cinco eixos.

Os cinco eixos e suas fontes

O Eixo 1, dedicado à natureza e à autonomia do sistema, articula dois conceitos complementares. O primeiro é o de influência do modelo: o quanto a IA pesa na decisão final, em relação a outras fontes de evidência (FDA, 2025). O Ministério da Segurança Alimentar e de Medicamentos da Coreia do Sul (Ministry of Food and Drug Safety — MFDS) define dois critérios (MFDS, 2022): Potencialidade de Má Função, ou seja, a chance de falhas algorítmicas causarem dano clínico; e Oportunidade de Contestação, que é a possibilidade real do

profissional questionar o resultado do sistema. A combinação dessas duas dimensões permite avaliar não apenas o que o sistema faz, mas o que acontece quando ele erra. O mais importante, nesse momento, é saber quem detém a capacidade de intervir.

O Eixo 2 examina o impacto sobre a pessoa participante. Suas bases estão na Lei n.º 14.874, de 28 de maio de 2024, e no Decreto n.º 12.651, de 2025, que estabelecem os princípios éticos fundamentais da pesquisa com seres humanos no Brasil (Brasil, 2024a; 2025a). O eixo inclui pergunta específica sobre populações vulneráveis: crianças, gestantes, idosos, povos indígenas e pessoas em situação de vulnerabilidade socioeconômica. Essa inclusão responde a uma lacuna dos modelos internacionais analisados. Nenhum deles foi concebido para o contexto epidemiológico e social brasileiro, marcado por desigualdades regionais acentuadas e pela pluralidade étnica e cultural da população atendida pelo Sistema Único de Saúde (SUS). O eixo incorpora ainda pergunta sobre mecanismo de exclusão de dados do treinamento do modelo. Essa pergunta tem fundamento nas diretrizes do Indian Council of Medical Research (ICMR), que institui o direito de o participante solicitar a supressão de seus dados do repositório utilizado pela IA a qualquer tempo (ICMR, 2023). Nenhum dos modelos ocidentais analisados prevê esse direito.

O Eixo 3, sobre sensibilidade e governança dos dados, é o eixo mais diretamente ancorado no direito brasileiro. A LGPD classifica dados de saúde como dados pessoais, sensíveis e exige base legal específica, finalidade declarada e identificação do Encarregado de Dados para seu tratamento. Esse conjunto de exigências é mais protetivo do que as legislações de proteção de dados de todos os países analisados neste guia, à exceção do Regulamento Geral de Proteção de Dados da União Europeia (GDPR).

O contraste torna-se ainda mais evidente em relação à Lei de Proteção de Dados Pessoais Digitais da Índia (Digital Personal Data Protection Act), de 2023, que eliminou a categoria de dados sensíveis do texto legal e passou a tratar dados de saúde como qualquer outro dado pessoal (Burman, 2023). Equiparar prontuários, dados genéticos e registros psiquiátricos a dados comerciais correntes expõe os participantes de pesquisa a riscos de discriminação sem respaldo jurídico para proteção diferenciada.

A LGPD brasileira faz escolha oposta. Dados de saúde constituem categoria sensível com proteção reforçada. Essa distinção é um dos pilares do Eixo 3 da Matriz. O eixo inclui ainda pergunta específica sobre risco de reidentificação. O MRCT Center e WCG (2025)

documentaram que grandes modelos de linguagem conseguem inferir identidades a partir de dados aparentemente anonimizados. Esse risco não desaparece com a anonimização formal.

A Resolução CNS n.º 738, de 1.º de fevereiro de 2024, operacionaliza a LGPD no contexto específico da pesquisa com bancos de dados envolvendo seres humanos. Define as figuras jurídicas do Controlador, do Operador e do Titular (art. 3.º, incisos IV, XIII e XX) e estabelece os quatro Termos institucionais que estruturam a cadeia de custódia do dado: Acordo, Anuência, Compromisso de Uso e Transferência (art. 3.º, XVI a XIX). A Resolução distingue ainda duas situações de uso de bancos de dados — constituídos no âmbito da pesquisa (Capítulo VII) e já constituídos fora desse âmbito (Capítulo VIII) — com exigências documentais distintas, e regula o uso futuro de dados e as hipóteses de dispensa de TCLE (art. 20, §§ 2.º, 3.º e 5.º, e art. 25). A MARIAH reúne essas exigências em duas subseções dedicadas a bancos de dados — Eixo 3.b na Versão A qualitativa e Bloco 6.b na Versão B quantitativa — que incorporam os elementos em três grupos de perguntas operacionais: o protocolo identifica o Controlador do banco de dados? Apresenta os Termos institucionais exigidos pela natureza do banco (especialmente o Termo de Anuência Institucional, obrigatório em bancos fora do âmbito da pesquisa)? Apresenta, quando houver pedido de dispensa de TCLE, a hipótese normativa invocada e justificativa ética e metodológica suficiente (art. 20, § 5.º, ou art. 25)? (Brasil, 2024b).

O material operacional correspondente está na Parte II da MARIAH.

O Eixo 4, sobre representatividade e transferibilidade, responde a uma das principais lacunas éticas identificadas na literatura internacional: modelos treinados predominantemente com dados de populações de alta renda e do hemisfério norte apresentam desempenho sistematicamente inferior quando aplicados a populações de diferentes características étnicas, socioeconômicas e geográficas (OMS, 2024). O guia australiano e as diretrizes do MFDS coreano documentaram que a ausência de validação local é um dos principais fatores de risco ético em protocolos que importam sistemas desenvolvidos em outros contextos (Austrália, 2023; MFDS, 2022). O eixo incorpora ainda pergunta sobre *data drift*. Modelos de aprendizado de máquina degradam seu desempenho à medida que os dados do mundo real divergem dos dados de treinamento. Esse fenômeno exige monitoramento contínuo ao longo de todo o ciclo de vida do sistema. A pergunta tem fundamento no conceito de manutenção do ciclo de vida desenvolvido pela FDA (2025) e

operacionalizado pelo sistema IDATEN da Agência de Dispositivos Médicos e Produtos Farmacêuticos do Japão (Pharmaceuticals and Medical Devices Agency — PMDA, 2023).

O Eixo 5 examina a supervisão humana e a explicabilidade. Seu fundamento central é o princípio da autonomia médica irrevogável enunciado pelo ICMR (2023) indiano. A inteligência artificial serve para amplificação diagnóstica. Não substitui o julgamento clínico. O princípio é formulado para a prática clínica. Em pesquisa, ele incide sobre o desenho da supervisão descrito no protocolo: o que o CEP avalia não é o ato assistencial futuro, mas se o protocolo garante revisão e contestação humanas antes da execução de cada decisão apoiada pelo sistema. A soberania sobre a saída do sistema cabe, nesse plano, ao pesquisador responsável, sem prejuízo da responsabilidade do profissional assistente quando houver ato clínico na etapa seguinte de cuidado. O pesquisador responsável é o garantidor da verificação humana de todo resultado gerado pela IA. Esse princípio articula-se com o conceito de humano no circuito decisório (*human-in-the-loop*) presente no modelo australiano (Austrália, 2023).

Ele se articula com a exigência da FDA de que, quando o contexto de uso envolve um humano no processo decisório, a avaliação de desempenho do sistema considere o resultado da equipe humano-IA, não apenas o desempenho isolado do modelo (FDA, 2025). O eixo incorpora ainda pergunta sobre o protocolo de retreinamento do modelo. Essa pergunta tem fundamento no Protocolo de Mudança de Algoritmo (*A*lgorithm Change Protocol) indiano, que exige que os fabricantes submetam antecipadamente um plano que descreva como o modelo será atualizado. Esse requisito foi estabelecido pela Organização Central de Controle de Padrões de Medicamentos da Índia (CDSCO, do inglês Central Drugs Standard Control Organization) (Índia, 2025). No contexto deste guia, esse princípio se traduz na responsabilidade do pesquisador de informar ao CEP como e quando o modelo será modificado durante o estudo.

A convergência de alguns eixos em torno do impacto clínico sobre o participante é intencional, não é sobreposição inadvertida. Os eixos de natureza e autonomia do sistema, de impacto sobre a pessoa participante e de transferibilidade e desempenho avaliam dimensões distintas de uma mesma realidade: o que o sistema pode fazer à pessoa que participa da pesquisa. O eixo de natureza examina como o sistema opera. O eixo de impacto examina o que o sistema produz sobre o participante. O eixo de transferibilidade examina se o que o sistema produz num contexto continuará válido noutro. São perguntas diferentes sobre o

mesmo risco porque o risco tem múltiplas dimensões que um único eixo não captura. Um protocolo que apresenta alto risco nesses três eixos simultaneamente não é penalizado três vezes pelo mesmo problema. Demonstra que o risco é consistente, estrutural e não mitigável por uma única salvaguarda. A convergência entre eixos num resultado de alto risco é evidência de gravidade, não artefato metodológico.

A lógica do fluxograma sequencial e a regra crítica

O fluxograma sequencial foi adotado por dois motivos. O primeiro é pedagógico. A sequência conduz a avaliação de forma progressiva, do mais simples ao mais complexo, e evita saltos que comprometam a consistência da análise. O segundo é ético. O nível de risco só sobe, nunca desce. Essa lógica é coerente com o princípio da precaução que fundamenta a ética em pesquisa com seres humanos (Brasil, 2025a). Um único eixo em Nível IV é suficiente para elevar o protocolo inteiro ao nível mais alto de exigência, independentemente do desempenho nos demais eixos. A ausência de supervisão humana em um sistema que influencia decisões terapêuticas, por exemplo, não é compensada por boas respostas nos eixos de dados ou representatividade.

A regra crítica tem origem direta no modelo canadense de Avaliação de Impacto Algorítmico (Canadá, 2024). O modelo demonstrou que permitir que mitigações reduzam o nível de risco cria um incentivo perverso. Pesquisadores passam a declarar salvaguardas para obter classificações mais favoráveis. Deixam de demonstrar que o risco foi efetivamente controlado. A regra resolve esse problema com uma separação clara. O nível de risco é determinado pelas características do sistema e do protocolo. Os requisitos exigidos são ajustados em função das garantias demonstradas. Essa separação preserva a integridade do instrumento e a clareza da comunicação entre pesquisador e comitê.

### 8.2.2 Versão B quantitativa

A Versão B da MARIAH utiliza perguntas com pesos numéricos variados em sete blocos temáticos. A soma dos pontos determina o risco do protocolo em uma escala de I a IV. Essa abordagem torna suas decisões mais rastreáveis e auditáveis. A pontuação identifica áreas de concentração de risco e mostra como as escolhas do pesquisador afetam esse risco.

## A escolha pela abordagem quantitativa

A lógica quantitativa tem inspiração no modelo canadense de Avaliação de Impacto Algorítmico (Canadá, 2024). O modelo opera com 65 perguntas de risco, pontuação máxima de 169 pontos, e 41 perguntas de mitigação, pontuação máxima de 75 pontos, distribuídas em quatro níveis de risco.

A experiência canadense demonstrou duas vantagens concretas do modelo quantitativo. A primeira é a precisão: a pontuação identifica em quais dimensões o risco está concentrado e orienta as exigências do comitê de forma proporcional. A segunda é a consistência: protocolos avaliados em momentos diferentes e por membros diferentes do CEP tornam-se comparáveis, o que fortalece a coerência das decisões institucionais (Canadá, 2024). A Versão B incorpora essa lógica e a adapta ao contexto da pesquisa com seres humanos no Brasil. O foco deixa de ser serviços públicos governamentais e passa a ser a proteção dos participantes de pesquisa e a conformidade com a LGPD.

Este guia propõe uma estrutura em sete blocos: Projeto, Sistema, Algoritmo, Decisão, Impacto, Dados e Mitigação. Os blocos de maior peso ético e regulatório recebem perguntas com valor individual mais alto. Essa escolha torna explícita a hierarquia de prioridades do instrumento. Os blocos de Impacto e Dados concentram juntos 139 dos 275 pontos máximos da matriz. Representam 50,5% da pontuação total. Essa concentração expressa uma escolha deliberada. Em pesquisa com seres humanos, o que mais importa é o que o sistema pode fazer à pessoa participante e como seus dados são tratados.

## O filtro de entrada e a caracterização do contexto de uso

A Versão B mantém, da Versão A, o Passo 0 e a caracterização do contexto de uso como etapas obrigatórias anteriores à pontuação. A diretriz da FDA (EUA) demonstrou que nenhuma avaliação de risco algorítmico começa sem a definição precisa do contexto de uso: o papel específico do sistema e o escopo da pergunta que ele foi desenvolvido para responder (FDA, 2025). A caracterização não pontua. Delimita o objeto da avaliação e impede que o sistema seja avaliado em abstrato, desvinculado do protocolo concreto submetido à análise.

Os sete blocos e suas fontes

O Bloco 1 avalia a adequação do protocolo para o uso de IA. Examina a transparência na descrição do sistema, a qualificação da equipe, a clareza da justificativa metodológica e a adequação do Termo de Consentimento Livre e Esclarecido. A pontuação máxima é de 22 pontos. Cada pergunta vale três pontos, com exceção da pergunta sobre consulta técnica externa, que vale quatro. O bloco opera como verificação de completude documental.

O protocolo deve descrever o estágio de desenvolvimento do sistema: descoberta exploratória (*Discovery*), tradução clínica (*Translation*) ou implantação (*Deployment*). Essa exigência tem fundamento no referencial do MRCT Center e WCG (2025), que demonstrou que os riscos éticos variam conforme o momento do ciclo de vida em que a IA opera. Um sistema em fase de descoberta exploratória apresenta riscos distintos de um sistema já em implantação clínica. Essa distinção deve ser visível no protocolo submetido ao CEP.

O Bloco 2 avalia a natureza técnica do sistema: autonomia, adaptabilidade, explicabilidade e rastreabilidade. A pontuação máxima é de 17 pontos. Cada pergunta vale dois ou três pontos. O bloco articula dois conceitos desenvolvidos por referenciais distintos. O primeiro é o de influência do modelo: o quanto a IA pesa na decisão final em relação a outras fontes de evidência. Esse conceito foi desenvolvido pela FDA (2025). O segundo é o de potencial de falha (*Malfunction Potential)* e oportunidade de contestação (*Opportunity to Override*): a probabilidade de que falhas algorítmicas causem dano clínico e a existência de condições reais para que o profissional responsável questione o resultado do sistema.

Esse conceito foi desenvolvido pelo Ministério da Segurança Alimentar e de Medicamentos da Coreia do Sul (MFDS, 2022). As perguntas de risco direto valem três pontos cada. Tratam de situações sem intervenção humana obrigatória, de sistemas adaptativos e de sistemas opacos. Sua proximidade com o risco direto ao participante justifica o peso maior. As perguntas de ausência de garantia valem dois pontos. Tratam de déficits documentais que, embora graves, são mais facilmente corrigíveis.

O Bloco 3 avalia a qualidade, a origem e a representatividade dos dados de treinamento e de teste. A pontuação máxima é de 14 pontos. Cada pergunta vale dois pontos. O bloco tem fundamento direto no guia australiano e nas diretrizes do MFDS coreano, que documentaram que a ausência de validação local é um dos principais fatores de risco ético em protocolos que utilizam sistemas desenvolvidos em outros contextos (Austrália, 2023; MFDS,

2022). O bloco inclui pergunta sobre independência entre dados de treinamento e de teste. A diretriz da FDA estabelece que dados de teste devem ser rigorosamente independentes dos dados de desenvolvimento para que a avaliação de desempenho do modelo seja válida (FDA, 2025). Inclui também pergunta sobre representatividade por subgrupos populacionais. A Organização Mundial da Saúde identificou que modelos treinados predominantemente com dados de populações de alta renda e do hemisfério norte apresentam desempenho sistematicamente inferior quando aplicados a populações de diferentes características étnicas, socioeconômicas e geográficas (OMS, 2024). Essa lacuna é especialmente relevante para o contexto brasileiro.

O Bloco 4 avalia a consequência da decisão. A pontuação máxima é de oito pontos. As duas primeiras perguntas valem três pontos cada. Essa assimetria é deliberada. As perguntas desse bloco tratam das situações de maior gravidade potencial em todo o instrumento: o sistema como único determinante de uma decisão sem revisão humana e a irreversibilidade do dano em caso de erro. A concentração de pontos em poucas perguntas de alto peso reflete a lógica da consequência da decisão desenvolvida pela FDA (2025). Quando o modelo opera com alta influência e baixa supervisão humana, a severidade do evento adverso potencial deve ser o fator dominante na avaliação de risco. A pontuação máxima desse bloco, por si só, pode deslocar um protocolo do Nível II para o Nível III. Esse deslocamento expressa a gravidade das situações que essas perguntas descrevem.

O Bloco 5 avalia o impacto sobre a pessoa participante. A pontuação máxima é de 62 pontos. Cada pergunta vale cinco ou seis pontos. As perguntas de maior impacto direto valem seis pontos: influência em decisões clínicas, risco de dano físico, risco de dano psíquico ou social e envolvimento de populações vulneráveis. O bloco tem suas bases na Lei n.º 14.874, de 28 de maio de 2024, e no Decreto n.º 12.651, de 2025, que estabelecem os princípios éticos fundamentais da pesquisa com seres humanos no Brasil (Brasil, 2024a; 2025a). O bloco inclui pergunta sobre inferência de características sensíveis não coletadas explicitamente. O MRCT Center e WCG (2025) documentaram que grandes modelos de linguagem conseguem inferir identidades e atributos sensíveis a partir de dados aparentemente neutros. Esse risco não desaparece com a anonimização formal. O bloco inclui ainda pergunta sobre mecanismo de exclusão de dados do treinamento do modelo. As diretrizes do Indian Council of Medical Research instituem o direito do participante de solicitar a supressão de seus dados do

repositório utilizado pela IA a qualquer tempo (ICMR, 2023). Nenhum dos modelos ocidentais analisados prevê esse direito. Este guia o incorpora como exigência ética.

O Bloco 6 avalia a sensibilidade e a governança dos dados. A pontuação máxima é de 77 pontos. Cada pergunta vale entre cinco e sete pontos. As perguntas diretamente vinculadas à Lei Geral de Proteção de Dados Pessoais valem sete pontos cada. O bloco expressa uma escolha central do guia. A LGPD é o diferencial protetivo mais robusto do marco brasileiro em relação a todos os modelos internacionais analisados.

A lei classifica dados de saúde como dados pessoais sensíveis e exige base legal específica, finalidade declarada e identificação do Encarregado de Dados para seu tratamento. Esse nível de proteção supera o da Lei de Proteção de Informações Pessoais chinesa (PIPL), o da Lei de Proteção de Informações Pessoais japonesa (APPI) e, especialmente, o da Lei de Proteção de Dados Pessoais Digitais indiana (*Digital Personal Data Protection* Act), de 2023, que suprimiu inteiramente a categoria de dados sensíveis de sua redação e equiparou prontuários de saúde a quaisquer outros dados pessoais (Burman, 2023).

A pergunta sobre transferência de dados para servidores fora do Brasil vale sete pontos. Conforme o país receptor e as garantias contratuais estabelecidas, essa transferência pode comprometer integralmente a proteção conferida pela LGPD aos participantes da pesquisa. A pergunta sobre risco de reidentificação incorpora a preocupação documentada com a capacidade dos sistemas de IA de inferir identidades a partir de dados aparentemente anonimizados (MRCT Center; WCG, 2025).

A subseção Bloco 6.b da Versão B reúne cinco pontos específicos para protocolos que utilizam bancos de dados, com pesos calibrados pela centralidade operacional na Resolução CNS n.º 738/2024 (Brasil, 2024b): identificação nominal do Controlador do banco — P6.b.1 (sete pontos); apresentação do Termo de Anuência Institucional quando exigível pela origem do banco — P6.b.2 (sete pontos, ELIMINATÓRIA); apresentação do Termo de Compromisso de Uso de Dados — P6.b.3 (cinco pontos); enquadramento fundamentado de eventual dispensa de TCLE na hipótese cabível segundo a origem do banco — P6.b.4 (cinco pontos); e formalização da controladoria conjunta em pesquisa multicêntrica, por Termo de Acordo Institucional — P6.b.5 (cinco pontos).

O Bloco 6.b soma 29 pontos à pontuação total quando ativado pelo filtro do Passo 0. A ausência de resposta afirmativa em P6.b.2, quando aplicável, torna o protocolo não avaliável no mérito pela regra crítica (item "A regra crítica na Versão B", adiante), independentemente

da pontuação total. A subseção integra ainda três itens de diligência — P6.b.4.1 (necessidade e proporcionalidade do acesso a dados identificáveis), P6.b.6 (critérios cumulativos para dispensa de TCLE em pesquisa com IA) e P6.b.7 (declaração do Controlador) — que não pontuam nem se somam ao teto: operam como *gate* cumulativo, e a ausência de qualquer critério implica devolução do protocolo.

O Bloco 7 avalia as mitigações. É o único bloco com lógica bidirecional: perguntas de risco somam pontos ao total e perguntas de mitigação subtraem. A pontuação máxima é de 75 pontos. A capacidade de redução por mitigações chega a 65 pontos — às quais se somam até 23 pontos de evidências do sub-bloco 7C. O bloco operacionaliza uma distinção herdada do modelo canadense. Mitigações não alteram o nível de risco determinado pelos demais blocos quando avaliados isoladamente. Reduzem a pontuação total e, por isso, podem deslocar o protocolo para um nível inferior na escala consolidada (Canadá, 2024). Essa lógica é mais flexível do que a da Versão A, em que mitigações não afetam o nível em nenhuma hipótese. Reflete a compreensão de que, na Versão B, a pontuação já incorpora a qualidade das salvaguardas adotadas pelo pesquisador como parte integrante da avaliação de risco. O sub-bloco 7C, de evidências de transparência, opera em regime distinto, de só-abate: a presença documentada da evidência subtrai pontos, e a ausência não soma.

O bloco se divide em três sub-blocos. O Sub-bloco 7A (consultas) vale dez pontos. Avalia se o pesquisador buscou orientação técnica externa antes da submissão, inclusive consulta à Anvisa sobre o enquadramento do sistema como SaMD. O Sub-bloco 7B (medidas de redução de risco, *de-risking*) vale 65 pontos. Avalia a qualidade dos mecanismos de supervisão humana, explicabilidade, monitoramento do ciclo de vida e gestão de falhas. O sub-bloco inclui pergunta sobre monitoramento de deriva de dados (*data drift*), ou seja, a degradação do desempenho do modelo ao longo do tempo. Essa pergunta tem fundamento no conceito de manutenção do ciclo de vida desenvolvido pela FDA (2025), operacionalizado pelo sistema IDATEN da agência regulatória japonesa (PMDA, 2023) e pelo Protocolo de Mudança de Algoritmo (Algorithm Change Protocol) indiano (CDSCO, 2025). O Sub-bloco 7C (evidências de transparência) registra evidências documentais — documentação padronizada do modelo e da base de dados, avaliação de impacto algorítmico ou Relatório de Impacto à Proteção de Dados Pessoais e disponibilização de código-fonte — que abatem até 23 pontos quando presentes e documentadas, sem somar quando ausentes.

## As faixas de nível e a calibração da pontuação

As quatro faixas de nível foram calibradas a partir de dois princípios. O primeiro é a proporcionalidade. A faixa do Nível I é estreita, de zero a 58 pontos, porque protocolos com uso de IA verdadeiramente de baixo risco são raros e devem ser classificados com rigor. O segundo é a sensibilidade ao risco. A faixa do Nível IV cobre 67 pontos, de 209 a 275, porque protocolos nesse nível apresentam combinações de características graves em múltiplos blocos. A amplitude da faixa reflete a variação de gravidade dentro do próprio nível crítico. Quando o filtro do Passo 0 ativa o Bloco 6.b (banco de dados), o teto da matriz passa de 275 para 304 pontos e as faixas são recalibradas proporcionalmente: Nível I de 0 a 64 pontos, Nível II de 65 a 141 pontos, Nível III de 142 a 230 pontos e Nível IV de 231 a 304 pontos (teto avaliável 297 pontos, descontada a questão eliminatória P6.b.2). A recalibração preserva a distância relativa entre os níveis e evita que os 29 pontos adicionais do Bloco 6.b distorçam a classificação. As faixas serão testadas empiricamente durante o período de uso e teste da versão 1.0; eventuais ajustes serão considerados na versão 2.0, a ser submetida à Inaep.

O material operacional correspondente está na Parte II da MARIAH.

## A regra crítica na Versão B

Na Versão B, a regra crítica assume formulação diferente da Versão A. O nível determinado pela pontuação não pode ser alterado por decisão subjetiva do avaliador. A pontuação é o resultado sistemático da aplicação do instrumento. As mitigações já estão incorporadas na pontuação do Bloco 7. Constituem a única forma legítima de reduzir a pontuação final. Essa arquitetura preserva a integridade do instrumento. Impede que declarações de intenção não documentadas substituam evidências de controle efetivo do risco. O modelo canadense demonstrou que esse princípio é essencial para a credibilidade do instrumento perante pesquisadores, patrocinadores e a sociedade (Canadá, 2024).

A gradação descrita até aqui convive com quatro salvaguardas que não se graduam. São a regra especial do Eixo 3.b, a Cláusula de Prevalência Ética, a eliminatória de cadeia de custódia e a diligência impeditiva de novo consentimento, todas descritas no Suplemento à Nota Técnica de Premissas. Operam de modo automático, sem depender do julgamento de quem aplica a matriz. As duas primeiras determinam o nível de risco por si sós. As duas últimas

suspendem a análise de mérito e devolvem o protocolo ao pesquisador em diligência. Nenhuma delas admite compensação por mitigações declaradas, porque todas tratam de condições que precedem a gradação do risco. Os gatilhos e o procedimento de cada uma estão na Parte II da MARIAH.

O material operacional correspondente está na Parte II da MARIAH.

# 9 ESPECIFICIDADES POR NATUREZA DA PESQUISA

Os capítulos anteriores construíram os fundamentos: os conceitos, o contexto brasileiro, o panorama internacional, a integridade científica, a governança de dados e a Matriz de Avaliação de Risco em Inteligência Artificial. Este capítulo aplica esses fundamentos a dois territórios com perfis de risco específicos que exigem perguntas que os instrumentos gerais não contemplam. A pesquisa em saúde com sistemas classificados como dispositivos médicos de software envolve uma dupla regulação, do CEP e da Anvisa, com mandatos complementares que o protocolo deve satisfazer de forma distinta. A pesquisa em Ciências Humanas e Sociais com algoritmos de análise comportamental, mineração de dados e modelos preditivos produz riscos políticos, sociais e epistêmicos que a moldura biomédica não alcança. Identificar o território antes de aplicar o instrumento é o que permite avaliar com a precisão que o campo exige.

## 9.1 PESQUISA EM SAÚDE: IMPACTO ASSISTENCIAL, INTERFACE ANVISA E SAMD

### 9.1.1 Quando o sistema de IA é também dispositivo médico

A Resolução de Diretoria Colegiada n.º 657, de 24 de março de 2022, da Agência Nacional de Vigilância Sanitária (Anvisa), estabelece o critério que deve ser conhecido antes de qualquer outra análise de um protocolo com sistema de IA em saúde: a finalidade do sistema determina sua classificação regulatória, não sua arquitetura técnica (Anvisa, 2022).

Um algoritmo de aprendizado profundo que analisa imagens de tomografia para identificar nódulos pulmonares é um programa de computador como dispositivo médico. O mesmo algoritmo, com a mesma arquitetura, treinado para otimizar escalas de plantão hospitalar, não é. A distinção não está no código. Está no que o sistema faz com os dados clínicos e nas consequências dessa operação sobre decisões que afetam pacientes.

A RDC n.º 657/2022 classifica como dispositivo médico o programa de computador com finalidade de diagnóstico, prevenção, monitoramento, predição, prognóstico, tratamento ou compensação de doença, lesão ou deficiência. Ficam expressamente excluídos os programas com finalidade puramente administrativa, de faturamento, de comunicação geral ou de apoio à gestão sem relação com decisões clínicas individuais. Essa distinção produz

uma pergunta operacional que o protocolo deve responder antes da avaliação de mérito: o sistema de IA descrito neste protocolo se enquadra como SaMD segundo a RDC n.º 657/2022?

O pesquisador deve declarar explicitamente a classificação e justificá-la. A ausência dessa declaração em protocolo com sistema de apoio diagnóstico, preditivo ou de monitoramento é informação que impede a avaliação ética completa do risco imposto aos participantes.

A identificação não depende exclusivamente da declaração do pesquisador. Dois indicadores no protocolo permitem uma leitura independente.

O primeiro é a descrição da saída do sistema: se o sistema produz classificações, escores de risco, recomendações terapêuticas ou alertas clínicos que chegam ao profissional de saúde ou ao paciente, o enquadramento como SaMD é provável.

O segundo é a descrição do fluxo de uso: se a saída do sistema integra, influencia ou substitui etapas de um processo assistencial, o enquadramento é praticamente certo. O CEP que identifica esses indicadores e não encontra declaração correspondente no protocolo deve solicitar esclarecimento antes de qualquer outra análise.

### 9.1.2 A sequência CEP-Anvisa: mandatos complementares, não concorrentes

Um sistema classificado como SaMD em pesquisa ocupa simultaneamente dois territórios regulatórios com mandatos distintos. O CEP avalia se a pesquisa que desenvolve ou testa o sistema respeita os direitos e a segurança dos participantes. A Anvisa avalia se o produto resultante dessa pesquisa é seguro e eficaz para uso clínico generalizado. Os dois mandatos são complementares e sequenciais, não concorrentes. Para que um sistema de IA obtenha registro comercial na Anvisa, o desenvolvedor precisa apresentar evidência de acurácia e segurança clínica produzida em pesquisa com seres humanos. Essa pesquisa, antes de ser conduzida, precisa de aprovação do CEP, cujo mandato é proteger os participantes durante a geração da evidência. A Anvisa atesta que a evidência é suficiente para o uso amplo. Nenhuma das duas instâncias substitui a outra (Anvisa, 2022; Barucci *et al.*, 2026).

A fase em que o sistema se encontra no momento da pesquisa determina o que precisa ser avaliado com maior atenção. Barucci *et al.* (2026) e o MRCT Center e o WCG (2025) propõem uma taxonomia em três fases com obrigações regulatórias e éticas distintas: treino, testes em condições reais e monitoramento pós-entrada em uso.

Na fase de treino, o sistema está em desenvolvimento e os dados clínicos dos participantes alimentam o aprendizado do modelo. O impacto sobre o cuidado dos participantes é indireto, mas o risco de exposição de dados sensíveis é máximo. Nessa fase, o protocolo deve detalhar a governança dos dados, a composição demográfica da base de treino e as garantias de consentimento.

Na fase de teste em condições reais, o sistema já treinado é avaliado sobre dados de participantes em ambiente clínico real, ainda sem interferência direta nas decisões assistenciais. O risco de dados e o risco de automação da decisão coexistem.

Na fase de monitoramento pós-entrada em uso, o sistema já opera clinicamente e é acompanhado quanto à manutenção de seu desempenho. O risco de deriva de dados e de degradação silenciosa de acurácia é o mais relevante nessa fase.

Para o protocolo, essa taxonomia tem consequência direta. O pesquisador deve declarar em qual fase o sistema se encontra no momento da pesquisa e descrever as obrigações éticas específicas daquela fase.

Um protocolo que descreve um sistema como "ferramenta de apoio diagnóstico em fase de validação" sem especificar se isso corresponde à fase de treino, à fase de teste em condições reais ou ao monitoramento pós-comercialização fornece informação insuficiente para avaliar o risco imposto aos participantes.

### 9.1.3 O que o protocolo com SaMD deve conter — e o que o CEP verifica

O protocolo com SaMD deve conter duas declarações: a classificação regulatória e a declaração de fase; é o primeiro ponto que o CEP verifica.

A primeira é a declaração de classificação regulatória: o pesquisador afirma explicitamente se o sistema se enquadra ou não como SaMD segundo a RDC n.º 657/2022, com justificativa baseada na finalidade do sistema e não apenas na sua arquitetura técnica.

A segunda é a declaração de fase: o pesquisador identifica em qual das três fases do ciclo de vida o sistema se encontra no momento da pesquisa e descreve as obrigações éticas específicas daquela fase.

Se alguma dessas declarações estiver ausente em um protocolo que utiliza sistemas de apoio diagnóstico, preditivo ou de monitoramento, não é possível corrigir essa falta apenas

com uma complementação genérica. Isso ocorre porque tal informação é fundamental para definir quais questões éticas devem ser consideradas no protocolo.

O segundo requisito é a descrição do impacto assistencial potencial sobre os participantes durante a fase de pesquisa. Um sistema em fase de treino que processa dados clínicos históricos sem interferir em decisões sobre os participantes ativos da pesquisa tem perfil de risco concentrado na governança dos dados e no consentimento. Mas um sistema em fase de teste que já orienta ou influencia condutas clínicas sobre participantes ativos tem perfil de risco que inclui, adicionalmente, o viés de automação, a supervisão humana real e a possibilidade de dano decorrente de erro algorítmico.

O protocolo deve descrever com precisão o fluxo de uso do sistema durante a pesquisa: quem acessa a saída do sistema, em que momento, com qual nível de influência sobre a decisão clínica e com qual mecanismo de contestação disponível ao pesquisador. Sem essa descrição, o impacto no cuidado dos participantes permanece incerto.

O terceiro requisito é o plano de transição entre a aprovação ética da pesquisa e o eventual registro na Anvisa. Quando a pesquisa termina, o sistema permanece. Os dados dos participantes que alimentaram o treino ou a validação do modelo continuam incorporados nos seus parâmetros.

O modelo pode seguir para o processo de registro na Anvisa, ser transferido a parceiro comercial, publicado em repositório aberto ou descontinuado. Cada um desses destinos tem implicações éticas para os participantes que consentiram com a pesquisa, não com o destino posterior do sistema.

O encerramento gera também um dever de devolutiva. O protocolo deve prever como os participantes serão informados sobre o produto desenvolvido e sobre qualquer resultado com impacto em sua condição de saúde ou no seu aconselhamento em saúde. O protocolo deve descrever explicitamente o que acontecerá ao sistema após o encerramento da pesquisa, quais dados dos participantes permanecerão incorporados no modelo que seguirá para registro regulatório, e se o consentimento obtido cobre esse uso subsequente ou se novo consentimento será necessário. Sem essa descrição, aprova-se implicitamente um destino do sistema que não foi avaliado.

Os estudos que avaliam clinicamente um sistema de apoio à decisão contam com referência internacional de relato. A diretriz DECIDE-AI, publicada simultaneamente pela Nature Medicine e pelo BMJ em 2022, organiza o que um estudo de avaliação clínica em fase

inicial deve relatar sobre o sistema, os dados, os resultados e as falhas observadas (Vasey *et al.*, 2022). Um protocolo em fase de avaliação clínica inicial que organiza seu plano por essa referência chega ao comitê com a completude que a verificação exige, e o relato posterior dos resultados ganha padrão reconhecido pela comunidade científica. A adoção da diretriz é voluntária e não substitui as declarações exigidas nesta seção. Oferece a elas um formato consolidado internacionalmente.

Quando o protocolo com SaMD utiliza bancos de dados — próprio, de outra pesquisa ou de instituição externa, situação mais comum em sistemas de apoio diagnóstico treinados sobre registros hospitalares —, aplicam-se as exigências operacionais da Resolução CNS n.º 738/2024 (Brasil, 2024b): identificação do Controlador, apresentação dos Termos institucionais exigíveis conforme a origem do banco, e enquadramento de eventual dispensa de TCLE na hipótese cabível segundo a origem do banco (art. 20, § 5.º, ou art. 25).

O Capítulo 7 desenvolve essas figuras e o Eixo 3 da MARIAH (Capítulo 8) as traduz em perguntas operacionais. Para protocolos com SaMD, essas exigências compõem um segundo eixo de verificação, paralelo ao filtro regulatório da RDC n.º 657/2022, e não se substituem nem se anulam entre si.

## 9.2 PESQUISA EM CIÊNCIAS HUMANAS E SOCIAIS (CHS)

### 9.2.1 Um território com riscos próprios

A pesquisa em Ciências Humanas e Sociais (CHS) com inteligência artificial (IA) não produz diagnósticos clínicos nem orienta condutas terapêuticas. Produz algo diferente e igualmente consequente: inferências sobre comportamento, identidade, crenças, condição social e saúde mental de pessoas e grupos.

Um algoritmo de análise de sentimentos aplicado a postagens públicas de redes sociais pode inferir estados emocionais que o participante nunca declarou. Um modelo preditivo treinado sobre padrões de consumo de informação pode inferir filiação política, condição socioeconômica ou vulnerabilidade psicológica a partir de dados que o participante considera triviais. Um sistema de reconhecimento de padrões em dados de mobilidade urbana pode revelar rotinas, relações e condições de vida que o participante nunca autorizou tornar visíveis. A IA reúne informações ditas pelas pessoas e deduz também sobre o não dito.

Os riscos que essa capacidade produz não são clínicos. São políticos, sociais e epistêmicos.

O risco político é o da vigilância: sistemas de IA em CHS podem ser usados para monitorar dissidência, identificar ativistas ou mapear redes de organização social com consequências que vão além do âmbito acadêmico.

Já o risco social é o da discriminação e estigmatização: inferências sobre grupos vulneráveis produzidas em pesquisa podem migrar para contextos de aplicação onde fundamentam decisões sobre acesso a serviços, crédito, emprego ou moradia.

O risco epistêmico é o da produção de conhecimento que justifica exclusão: quando um modelo treinado sobre dados históricos de desigualdade confirma padrões de desigualdade como se fossem regularidades naturais, a pesquisa transforma injustiça estrutural em evidência científica (OMS, 2021; Soares; Chiavegatto Filho, 2026).

Essa mudança de escala, que vai do dano potencialmente individual ao efeito sobre comunidades e categorias sociais, exige que a avaliação ética também mude de pergunta, de lente e de instrumento.

Avaliar protocolos de CHS com IA utilizando apenas instrumentos destinados à pesquisa biomédica emprega ferramentas que não são apropriadas para esse contexto específico. Por exemplo, a pergunta sobre impacto em pesquisa clínica não tem correspondente direto em CHS.

Já uma pergunta sobre supervisão humana em decisões diagnósticas não se traduz automaticamente para contextos em que a decisão é sobre publicar uma análise de discurso ou treinar um modelo sobre dados de uma comunidade quilombola. O Capítulo 9 propõe que o comitê aprenda a formular esses princípios com as perguntas certas para esse território específico.

A Resolução n.º 510, de 7 de abril de 2016, do Conselho Nacional de Saúde (Brasil, 2016), foi construída para um mundo de questionários, entrevistas, grupos focais e observação participante.

Suas categorias foram formuladas com a pesquisa qualitativa e a etnografia em mente. O participante era alguém que o pesquisador encontrava, abordava e sobre quem tomava decisões de inclusão conscientes. O dado era algo que o participante produzia em interação com o pesquisador, ou que existia em registros institucionais com acesso regulado.

Naquele mundo, a distinção entre dado público e dado privado era relativamente estável, e a categoria de pesquisa que utiliza informações de acesso público dispensada de submissão ao SINEP fazia sentido operacional preciso.

A IA dissolveu essa estabilidade. Dados publicamente disponíveis em plataformas digitais, quando processados por algoritmos de aprendizado de máquina, produzem inferências que não são públicas e que o participante nunca autorizou.

Por exemplo: uma postagem em rede social é pública. A inferência sobre o estado de saúde mental do seu autor, produzida por um modelo treinado sobre milhões de postagens semelhantes, não é. Um comentário em fórum aberto é público. O perfil comportamental construído pelo cruzamento algorítmico desse comentário com centenas de outros rastros digitais do mesmo usuário não é.

A Resolução n.º 510/2016 tratava esses dados como de risco mínimo porque não antecipou o que a IA faz com eles (Brasil, 2016). Aplicar a dispensa de consentimento da resolução a protocolos com algoritmos de mineração de dados em redes sociais é aplicar a norma correta ao contexto errado.

A resolução permanece válida como referência subsidiária à Lei n.º 14.874/2024 e seus princípios continuam pertinentes. O que precisa de reinterpretação são suas categorias operacionais à luz do que a IA faz com dados que a norma tratava como de risco mínimo.

Três categorias exigem essa reinterpretação. A dispensa de consentimento para pesquisas com dados públicos: quando esses dados serão processados por algoritmos que produzem inferências sensíveis, a dispensa não se justifica automaticamente. A classificação de risco mínimo: pesquisas com dados de redes sociais que parecem de risco mínimo individualmente podem produzir risco coletivo significativo para os grupos representados nos dados. E a definição de participante: quando um algoritmo processa dados de milhões de pessoas sem nenhuma interação direta, quem é o participante e quais são seus direitos?

Essas perguntas não têm resposta consolidada no ordenamento brasileiro. Formulá-las com precisão contribui para construir a jurisprudência ética que o campo ainda não tem.

A pesquisa em CHS com IA apresenta um perfil de risco que se organiza em três dimensões a percorrer antes de avaliar o mérito científico do protocolo: a natureza das inferências, a escala e o risco coletivo.

As três dimensões são distintas das que estruturam a avaliação de protocolos biomédicos e exigem perguntas específicas que os instrumentos convencionais de avaliação ética não contemplam.

Na primeira dimensão, destaca-se a natureza das inferências geradas pelo sistema: os algoritmos de aprendizado de máquina utilizados em CHS tendem a produzir inferências sobre atributos considerados dados sensíveis pela legislação brasileira, tais como saúde mental, orientação sexual, filiação política, condição socioeconômica, bem como origem étnica ou racial. Essas inferências são produzidas a partir de dados que o participante nunca associou a esses atributos e que, isoladamente, não revelariam nenhuma informação sensível.

O cruzamento algorítmico de rastros digitais aparentemente triviais pode produzir perfis que violam a privacidade de forma mais profunda do que qualquer questionário direto sobre os mesmos temas.

A Lei Geral de Proteção de Dados (LGPD) classifica esses atributos como dados sensíveis e exige base legal específica para seu tratamento (Brasil, 2018). O protocolo deve identificar as inferências sensíveis que o sistema produzirá e a base legal que autoriza esse tratamento.

A segunda dimensão é a escala. A pesquisa em CHS com IA frequentemente opera sobre dados de milhares ou milhões de pessoas, a maioria das quais nunca soube que participava de uma pesquisa e nunca terá a oportunidade de exercer qualquer direito sobre os dados que o sistema processou.

Essa escala cria uma assimetria ética que os instrumentos convencionais de proteção de participantes não alcançam: o consentimento individual é operacionalmente impossível, a identificação dos participantes é incompleta e o monitoramento de danos é estruturalmente limitado. A avaliação desses protocolos não pode exigir o impossível, mas o protocolo deve demonstrar que as salvaguardas adotadas são proporcionais à escala do risco produzido (MRCT Center; WCG, 2025).

A terceira dimensão é o risco coletivo. As inferências produzidas por sistemas de IA em CHS frequentemente afetam não apenas os indivíduos cujos dados foram processados, mas os grupos sociais que esses indivíduos representam nos dados.

Um modelo treinado sobre dados de uma comunidade periférica pode produzir inferências que estigmatizam todos os moradores daquela comunidade, inclusive os que não participaram da pesquisa. Um algoritmo de análise de discurso treinado sobre mensagens de

um grupo religioso pode produzir representações que afetam a percepção pública de todos os membros daquele grupo.

O dano coletivo em CHS com IA ocorre sem a necessidade de identificar indivíduos. O protocolo deve indicar os grupos afetados, avaliar o risco de estigmatização coletiva e apresentar salvaguardas para esse risco, que vai além da proteção dos participantes individuais.

### 9.2.2 Consentimento em CHS com IA: o que muda

A Resolução n.º 510/2016 dispensava a submissão ao SINEP, instituído pela Lei n.º 14.874/2024, para pesquisas que utilizavam informações de acesso público.

Essa dispensa repousava sobre um pressuposto implícito: que o uso de dados públicos em pesquisa não produz riscos além dos que o participante já aceitou ao torná-los públicos. Quem publica uma postagem em rede social, concede uma entrevista a um veículo de comunicação ou registra uma manifestação em espaço público aceita que essa informação seja lida, citada e analisada.

A pesquisa qualitativa convencional operava dentro desse pressuposto com razoabilidade: o pesquisador lê o que foi publicado, interpreta o que foi dito e produz análise que permanece no registro do que o participante tornou acessível.

A IA rompe esse pressuposto de forma estrutural. O algoritmo não lê o que foi publicado. Cruza, combina e extrapola o que foi publicado com milhares de outros registros para produzir inferências que o participante nunca tornou públicas e que frequentemente ele próprio desconhece sobre si mesmo.

O consentimento em CHS com IA não pode ser resolvido pela categoria do dado de origem. Precisa ser resolvido pela categoria da inferência produzida. Um dado de origem público que alimenta um algoritmo capaz de inferir diagnóstico psiquiátrico, orientação sexual ou vulnerabilidade econômica não é tratado de forma eticamente equivalente a um dado público analisado por um pesquisador humano.

A inferência produzida pelo algoritmo é dado sensível segundo a LGPD, independentemente de o dado de origem ser público (Brasil, 2018). Dispensar o consentimento pela Resolução n.º 510/2016 sem considerar as possíveis inferências do sistema aplica a norma apenas ao dado original e desconsidera os dados realmente gerados

pela pesquisa (Brasil, 2016). Aprova-se a coleta e aceita-se implicitamente a produção de informação sensível sobre pessoas que nunca consentiram com essa produção.

Três situações concretas ilustram o problema do consentimento em CHS com IA e as perguntas que o protocolo deve responder em cada uma delas.

A primeira é a pesquisa que usa dados de redes sociais publicamente disponíveis para treinar modelo de análise de saúde mental.

Os dados de origem são públicos. As postagens foram feitas voluntariamente por seus autores em plataformas abertas. A Resolução n.º 510/2016 classificaria essa pesquisa como candidata à dispensa de submissão ao SINEP (Brasil, 2016).

O algoritmo, contudo, produzirá inferências sobre estado psiquiátrico, risco de suicídio ou uso de substâncias a partir dessas postagens. Essas inferências são dados de saúde e dados sensíveis segundo a LGPD. A dispensa de consentimento não se aplica automaticamente porque a inferência produzida pertence a categoria distinta da do dado de origem. O CEP deve perguntar: o protocolo identifica as inferências sensíveis que o modelo produzirá? Qual a base legal que autoriza o tratamento desses dados sensíveis? Os autores das postagens serão informados de que seus dados foram utilizados para produzir esse tipo de inferência?

A segunda é a pesquisa que usa gravações de entrevistas obtidas com consentimento prévio para treinar modelo de reconhecimento de padrões de discurso.

O consentimento original foi obtido para a realização da entrevista e para o uso de seu conteúdo em análise qualitativa. Não antecipou o uso da gravação para treinar algoritmo de aprendizado de máquina, que processará não apenas o conteúdo semântico da fala, mas seus padrões prosódicos, pausas, variações de tom e características vocais. Esses padrões podem revelar estado emocional, condição de saúde ou origem regional que o participante não pretendia divulgar. O CEP deve perguntar: o consentimento original cobre o uso da gravação para treinamento algorítmico? Se não cobre, novo consentimento foi obtido ou será obtido? O protocolo descreve quais atributos o modelo infere a partir das características vocais dos participantes?

A terceira é a pesquisa longitudinal que reutiliza dados coletados em estudos anteriores para treinar modelo preditivo sobre comportamento. Os dados foram coletados com consentimento obtido para finalidade específica e distinta. A reutilização para treinamento de modelo de IA configura uso secundário que, segundo a Lei n.º 14.874/2024, regulamentada pelo Decreto n.º 12.651/2025, exige avaliação ética específica (Brasil, 2024a;

2025a). O CEP deve perguntar: o consentimento original autorizava uso secundário para desenvolvimento de sistemas de IA? Se não autorizava, o protocolo demonstra que novo consentimento foi obtido ou justifica adequadamente sua dispensa com base nas hipóteses admitidas pelas resoluções do CNS, uma vez que a Lei n.º 14.874/2024 não prevê dispensa? O modelo treinado sobre esses dados produzirá inferências sobre os participantes originais ou apenas padrões populacionais gerais?

O consentimento informado em CHS com IA é consentimento sobre o que o sistema vai concluir, não apenas sobre o que o pesquisador vai coletar. Essa distinção reorganiza o que se deve exigir do protocolo e do TCLE em pesquisas nesse campo.

Para tanto, três exigências estruturam essa reorganização:

Que o protocolo descreva não apenas a origem dos dados, mas as inferências que o sistema produzirá, e avalie explicitamente se o consentimento existente ou a dispensa prevista cobre essas inferências. Um protocolo que descreve com precisão os dados que serão coletados, mas silencia sobre o que o algoritmo inferirá a partir desses dados fornece informação incompleta sobre o risco real da pesquisa. A pergunta pertinente não é apenas a procedência dos dados. É o que o sistema produzirá com eles e quem será afetado por essa produção.

Quando a inferência produzida pelo sistema é sobre atributo sensível, o consentimento deve ser obtido mesmo que os dados de origem sejam públicos. A publicidade do dado de origem não transfere ao participante a responsabilidade por inferências que ele não antecipou e não autorizou. Aceitar a dispensa de consentimento com base na publicidade do dado sem verificar a natureza das inferências produzidas pelo algoritmo aplica o critério correto ao elemento errado da pesquisa. A base legal para o tratamento de dados sensíveis exige justificativa específica que o protocolo deve apresentar e o comitê deve avaliar (Brasil, 2018).

TCLE em CHS com IA deve informar os participantes sobre o tipo de inferência que será produzida, não apenas sobre o tipo de dado que será coletado. Um termo que descreve a coleta de postagens públicas sem mencionar que essas postagens serão usadas para inferir estado de saúde mental não é termo completo. Um termo que descreve a gravação de entrevistas sem mencionar que os padrões vocais serão processados algoritmicamente para identificar características que o participante não declarou não é termo transparente.

A Lei n.º 14.874/2024 exige que o consentimento cubra os procedimentos da pesquisa (Brasil, 2024a). O processamento algorítmico de dados para produção de inferências sensíveis é procedimento da pesquisa com consequências que o participante tem direito de conhecer antes de decidir se consente.

### 9.2.3 Risco coletivo e grupos vulneráveis em CHS

O risco em pesquisa biomédica tem endereço individual preciso: o participante é exposto a uma intervenção que pode causar dano à sua saúde, e o sistema de proteção ética foi construído para proteger essa pessoa específica que consentiu, que pode retirar o consentimento e que tem direito a acompanhamento durante e após a pesquisa.

O risco de uma pesquisa em CHS que mobiliza IA opera em outra escala. Um modelo treinado sobre dados de moradores de uma comunidade periférica pode produzir inferências que estigmatizam todos os moradores daquela comunidade, inclusive os que não participaram da pesquisa, não sabem que ela foi realizada e nunca terão oportunidade de contestar as conclusões do algoritmo sobre si mesmos.

Um sistema de análise de discurso treinado sobre mensagens de um movimento social pode produzir representações que afetam a percepção pública de todos os seus membros, inclusive os que nunca postaram nenhuma mensagem. O dano coletivo em CHS com IA não exige identificação individual para ocorrer. Ocorre no nível do grupo, da comunidade ou da categoria social que os dados representam, e seus efeitos se distribuem por pessoas que o protocolo de pesquisa nunca nomeou.

Essa característica cria uma lacuna estrutural nos instrumentos convencionais de proteção ética em pesquisa. O SINEP foi desenhado para proteger participantes: pessoas que consentiram, que poderiam ser identificadas, que têm direitos exercíveis e que o pesquisador tem obrigação de acompanhar.

O participante coletivo de uma pesquisa em CHS com IA não tem essas características. É um grupo social cujos membros podem nunca saber que foram afetados, não têm mecanismo de consentimento coletivo reconhecido pelo ordenamento brasileiro e não estão cobertos pelos instrumentos de proteção previstos na Lei n.º 14.874/2024 com a mesma nitidez que os participantes individuais (Brasil, 2024a).

A avaliação de pesquisa em CHS com IA enfrenta, portanto, um desafio que vai além da aplicação dos instrumentos existentes: exige formular perguntas sobre risco coletivo para as quais o sistema ainda não construiu respostas consolidadas, e exige fazê-lo com base nos princípios que a ética em pesquisa acumulou, aplicados a um território que esses princípios não anteciparam explicitamente.

Populações que já sofrem discriminação estrutural são as mais expostas ao risco coletivo em pesquisa com IA em CHS. A razão é dupla e se reforça.

Primeiro, essas populações estão sistematicamente sub-representadas nos dados que treinam os modelos: as bases de dados disponíveis para pesquisa tendem a refletir as populações que tiveram mais acesso a serviços, a registros institucionais e a plataformas digitais, que não são as populações em situação de maior vulnerabilidade.

Segundo, as categorias sociais dessas populações estão sistematicamente sobre-representadas nos dados históricos de exclusão, criminalização e estigma que os modelos aprendem como padrões.

O algoritmo treinado sobre dados que carregam séculos de injustiça aprende essa injustiça como regularidade. Reproduz como achado científico o que é produto histórico de discriminação (Soares; Chiavegatto Filho, 2026; Fiske *et al.*, 2025).

Para populações pretas e pardas, indígenas, quilombolas, LGBTQIA+, população privada de liberdade, moradores de periferias e pessoas em situação de vulnerabilidade econômica, esse mecanismo tem consequências que extrapolam o âmbito acadêmico. Inferências produzidas por algoritmos sobre esses grupos podem fundamentar decisões sobre acesso a serviços públicos, elegibilidade a benefícios sociais, prioridade no atendimento de saúde ou risco criminal atribuído a territórios inteiros.

Quando a pesquisa em CHS com IA produz modelos que serão eventualmente aplicados em contextos de políticas públicas, aprová-la sem avaliar o risco coletivo sobre grupos vulneráveis é aprovar também a cadeia de consequências que o modelo carregará para além do ambiente acadêmico. O princípio da não maleficência, que a Resolução n.º 466/2012 (Brasil, 2013) estabeleceu como fundamento da ética em pesquisa e que a Lei n.º 14.874/2024 preserva, não se esgota na proteção do participante individual durante o estudo. Alcança os grupos que os dados representam e os contextos em que as inferências serão usadas (Brasil, 2012; Brasil, 2024a).

### 9.2.4 O que o protocolo de CHS com IA deve conter — e o que o CEP verifica

As seções anteriores percorreram o diagnóstico completo da pesquisa em CHS com IA: os riscos não são clínicos, mas são reais; a Resolução n.º 510/2016 precisa de reinterpretação à luz do que os algoritmos fazem com dados que a norma tratava como de baixo risco; o consentimento se organiza pela inferência produzida, não pelo dado de origem; e o risco coletivo afeta grupos que o protocolo nunca nomeou (Brasil, 2016).

Esta seção entrega o instrumento operacional que transforma esse diagnóstico em avaliação. Quatro blocos de perguntas estruturam a leitura de qualquer protocolo de CHS com IA.

O primeiro bloco é sobre a origem e a natureza dos dados. De onde vêm os dados que o sistema processará: coleta prospectiva com consentimento, dados institucionais com acesso regulado, dados publicamente disponíveis em plataformas digitais, ou reutilização de dados coletados em pesquisas anteriores? Cada origem tem implicações distintas para o consentimento e para a base legal do tratamento. O protocolo descreve com precisão a origem de cada tipo de dado utilizado e a base legal que autoriza seu processamento pelo sistema de IA?

O segundo bloco é sobre as inferências produzidas. O protocolo descreve o que o sistema inferirá a partir dos dados coletados? As inferências produzidas envolvem atributos sensíveis segundo a LGPD, como saúde mental, orientação sexual, filiação política, condição socioeconômica ou origem étnica? Se sim, qual a base legal que autoriza o tratamento desses dados sensíveis? O consentimento obtido ou a dispensa prevista cobre especificamente essas inferências, e não apenas os dados de origem? O TCLE informa os participantes sobre o tipo de inferência que o sistema produzirá em linguagem acessível e precisa?

O terceiro bloco é sobre o risco coletivo. O protocolo identifica os grupos sociais que serão afetados pelas inferências do sistema, inclusive os grupos não representados diretamente nos dados mas que compartilham as categorias sociais dos participantes? O protocolo avalia o risco de estigmatização coletiva e descreve salvaguardas específicas para esse risco, distintas das salvaguardas para proteção individual? O pesquisador demonstra que considerou o potencial de uso das inferências produzidas em contextos que amplificam discriminação estrutural sobre grupos vulneráveis?

O quarto bloco é sobre o destino do modelo. O protocolo descreve o que acontecerá ao modelo após o encerramento da pesquisa? Quando o destino previsto inclui publicação em repositório aberto ou transferência a parceiros, o protocolo avalia o risco de uso do modelo em contextos discriminatórios que os participantes não anteciparam e não autorizaram? O TCLE informa os participantes sobre o destino do modelo de forma honesta, incluindo os usos secundários possíveis e os riscos coletivos que esses usos podem produzir?

A distinção entre protocolo incompleto e protocolo inadequado aplica-se em CHS com IA com a mesma lógica da seção 6.5, ajustada ao território específico. O protocolo incompleto não descreve as inferências que o sistema produzirá, não avalia se a dispensa de consentimento da Resolução n.º 510/2016 cobre essas inferências ou não identifica os grupos afetados além dos participantes formais (Brasil, 2016).

Essas lacunas são sanáveis por solicitação de informações adicionais com critérios explícitos. O protocolo inadequado apresenta dados públicos processados por algoritmo que produz inferências sensíveis como pesquisa de risco mínimo dispensada de consentimento, sem qualquer análise das inferências produzidas nem da base legal para seu tratamento. Esse caso exige recusa fundamentada nos princípios da Lei n.º 14.874/2024 e nas exigências da LGPD, com argumentação que demonstre ao pesquisador por que a classificação de risco mínimo não se sustenta diante do que o sistema efetivamente produz (Brasil, 2024a; Brasil, 2018).

O presente capítulo encerra-se com uma afirmação que atravessa todo o guia: a especificidade não é exceção. É a regra. Todo protocolo com IA tem natureza específica que determina o conjunto de perguntas pertinentes. A pesquisa biomédica com sistema classificado como SaMD e a pesquisa em Ciências Humanas e Sociais com algoritmos de análise comportamental não são variações do mesmo problema ético com intensidades diferentes. São territórios distintos que exigem instrumentos distintos, perguntas distintas e critérios distintos de avaliação. Identificar o território antes de aplicar o instrumento não cria obstáculos à pesquisa com IA: exerce-se o mandato de avaliação com a precisão que a complexidade do campo exige e que as pessoas que participam dessas pesquisas têm o direito de esperar.

## 10 RELAÇÃO COM O GUIA E A MARIAH

O conjunto documental passa a ser composto por três peças complementares:

- Guia de Uso Ético de Inteligência Artificial em Pesquisa com Seres Humanos (Guia INAEP) — documento institucional de orientação para a avaliação ética, com linguagem recomendatória e foco nos critérios que devem ser considerados por pesquisadores e CEPs.
- MARIAH — matriz, checklists e instrumentos de aplicação para identificação, classificação, verificação e documentação dos riscos.
- Caderno Referencial — documento de fundamentação que preserva a construção teórica, normativa, metodológica e comparada desenvolvida pelo Grupo de Trabalho.

Em caso de diferença de formulação entre este Caderno Referencial e o Guia INAEP, a orientação institucional vigente deve ser consultada no Guia e nas normas aplicáveis. O Caderno Referencial explica a racionalidade e registra a trajetória técnica que sustenta essas orientações.

# 11 ATUALIZAÇÃO REGULATÓRIA INTERNACIONAL — 2024–2026

A rápida evolução dos marcos regulatórios internacionais exige que a fundamentação ética da pesquisa com inteligência artificial seja interpretada como um referencial dinâmico. Entre 2024 e 2026, a União Europeia e autoridades reguladoras de medicamentos aprofundaram a articulação entre regulação de IA, desenvolvimento de medicamentos, proteção de dados e uso secundário de dados de saúde. Os desenvolvimentos abaixo complementam o panorama internacional deste Caderno e não representam transposição automática de requisitos estrangeiros para o Sistema Nacional de Ética em Pesquisa com Seres Humanos.

## 11.1 REGULAMENTO EUROPEU DE INTELIGÊNCIA ARTIFICIAL (AI ACT)

O Regulamento (UE) 2024/1689 (União Europeia, 2024) estabeleceu uma arquitetura baseada em risco para sistemas de inteligência artificial. Para a pesquisa, sua contribuição metodológica está em exigir que o enquadramento considere a finalidade prevista, o contexto de uso e a influência do sistema sobre decisões e direitos. A simples presença de IA em um protocolo não determina, isoladamente, que o sistema seja de alto risco. O enquadramento depende das hipóteses e critérios previstos no Regulamento.

Implicação para a avaliação ética: recomenda-se partir do contexto de uso, da consequência potencial da decisão, do grau de autonomia, da supervisão humana e dos efeitos possíveis sobre participantes e grupos, evitando classificações automáticas baseadas apenas no tipo de tecnologia.

## 11.2 EMA E O CICLO DE VIDA DO MEDICAMENTO

Em setembro de 2024, a European Medicines Agency (EMA) adotou o *Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle* (EMA/CHMP/CVMP/83833/2023). O documento considera aplicações de IA e aprendizado de máquina ao longo do ciclo de vida do medicamento e orienta que seu uso seja compreendido em articulação com os requisitos europeus sobre medicamentos, inteligência artificial e proteção de dados.

Para este Caderno, o documento reforça uma abordagem de ciclo de vida: proveniência e adequação dos dados, desempenho no contexto pretendido, documentação, supervisão humana, controle de mudanças e monitoramento permanecem relevantes à medida que o sistema e o ambiente de uso evoluem.

## 11.3 PRINCÍPIOS CONJUNTOS EMA–FDA DE BOAS PRÁTICAS DE IA NO DESENVOLVIMENTO DE MEDICAMENTOS

Em janeiro de 2026, EMA e U.S. Food and Drug Administration (FDA) publicaram conjuntamente dez princípios orientadores de boas práticas de IA no desenvolvimento de medicamentos. O documento abrange IA utilizada para gerar ou analisar evidências nas fases não clínica, clínica, pós-comercialização e de fabricação e estabelece uma base comum para o desenvolvimento de futuras boas práticas.

A convergência reforça elementos incorporados à MARIAH: abordagem centrada no ser humano, contexto de uso claramente definido, governança e qualidade dos dados, desempenho adequado à finalidade, avaliação proporcional ao risco, transparência, documentação, gestão do ciclo de vida e monitoramento.

## 11.4 ESPAÇO EUROPEU DE DADOS DE SAÚDE (EHDS)

O Regulamento (UE) 2025/327, que institui o European Health Data Space (EHDS), acrescenta uma camada relevante à governança de dados de saúde na Europa. Entre seus objetivos está estabelecer condições harmonizadas para o uso secundário de determinadas categorias de dados eletrônicos de saúde, inclusive para pesquisa, inovação e atividades regulatórias, observadas as salvaguardas previstas no Regulamento.

A experiência do EHDS evidencia que governança de dados para pesquisa não se reduz à obtenção de consentimento. Envolve finalidade, fundamento jurídico, acesso controlado, minimização, segurança, responsabilidades institucionais, rastreabilidade e condições para reutilização. Trata-se de referência comparada, não de norma diretamente aplicável ao Brasil.

## 11.5 EDPB E PESQUISA CIENTÍFICA — GUIDELINES 1/2026

Em abril de 2026, o European Data Protection Board (EDPB) submeteu à consulta pública as *Guidelines 1/2026 on processing of personal data for scientific research purposes*. A consulta foi encerrada em 25 de junho de 2026. O documento aborda bases jurídicas, papéis de controlador e operador, direitos dos titulares e princípios aplicáveis ao tratamento de dados pessoais para pesquisa científica.

Por se tratar de documento submetido a consulta, sua utilização neste Caderno é referencial e preserva esse *status*. O material é especialmente relevante para distinguir consentimento para participação em pesquisa do fundamento jurídico para tratamento de dados pessoais, questões relacionadas, mas não necessariamente equivalentes.

## 11.6 EDPB E ANONIMIZAÇÃO — GUIDELINES 02/2026

Em julho de 2026, o EDPB publicou para consulta as *Guidelines 02/2026 on Anonymisation*. Na data desta atualização, a consulta permanece aberta até 30 de outubro de 2026. O documento, portanto, não é apresentado aqui como orientação europeia definitivamente consolidada.

Sua relevância para pesquisas com IA decorre da necessidade de distinguir anonimização de pseudonimização e de avaliar o risco de reidentificação no contexto do tratamento. Em sistemas treinados com dados de alta dimensionalidade, a simples declaração de que um conjunto é 'anônimo' não substitui a análise técnica do risco de reidentificação.

## 11.7 RECOMENDAÇÕES BRITÂNICAS PARA A REGULAÇÃO DA IA EM SAÚDE

Em setembro de 2026, a National Commission into the Regulation of AI in Healthcare publicou recomendações para um futuro marco regulatório britânico, em iniciativa apoiada pelo Department of Health and Social Care e pela Medicines and Healthcare products Regulatory Agency (MHRA) (Reino Unido, 2026). O documento constitui referência comparada e não produz efeitos normativos no Brasil.

Sua contribuição para a revisão ética está na articulação entre finalidade, contexto de uso, desempenho, equidade e responsabilidades ao longo do ciclo de vida. A avaliação não se encerra na validação inicial: atualizações, recalibrações, retreinamento, substituição do

modelo, mudanças nos dados ou expansão da população de uso podem alterar o perfil de risco e exigir nova apreciação.

Para o Guia INAEP e a MARIAH, as recomendações reforçam a necessidade de: identificar versão e configuração do sistema; justificar a adequação das evidências ao contexto brasileiro; avaliar desempenho e possíveis diferenças entre subgrupos; definir supervisão humana efetiva; documentar controle de mudanças; estabelecer monitoramento, critérios de interrupção e resposta a incidentes; e distribuir claramente as responsabilidades entre pesquisadores, instituições, desenvolvedores, fornecedores e demais agentes.

A equidade deve ser tratada também como dimensão de segurança. Um sistema pode apresentar desempenho agregado satisfatório e, ainda assim, produzir resultados inferiores ou danos desproporcionais em grupos sub-representados. Para as Ciências Humanas e Sociais, essa perspectiva amplia a análise para inferências sobre comunidades, estigmatização, discriminação, relações de poder e pessoas afetadas que não se enquadram formalmente como participantes.

## 11.8 CONSEQUÊNCIAS PARA O REFERENCIAL BRASILEIRO E PARA A MARIAH

Os desenvolvimentos de 2024–2026 reforçam cinco conclusões: (i) o risco deve ser avaliado a partir do contexto de uso e das consequências potenciais; (ii) governança de dados e governança algorítmica devem ser examinadas de forma integrada; (iii) validação, supervisão e monitoramento acompanham o ciclo de vida; (iv) uso secundário de dados exige análise própria de finalidade, fundamento jurídico, acesso, segurança e direitos; e (v) convergência internacional não elimina a necessidade de contextualização às normas brasileiras, à diversidade populacional e institucional e às características das Ciências da Saúde e das Ciências Humanas e Sociais.

Esses elementos complementam, sem substituir, os fundamentos do Guia INAEP e da MARIAH. Referenciais estrangeiros são fontes comparadas de aprendizagem regulatória; a avaliação ética no Brasil permanece orientada pelo marco jurídico e ético nacional aplicável.

## REFERÊNCIAS


AMARI, S. A theory of adaptive pattern classifiers. IEEE Transactions on Electronic Computers, v. EC-16, n. 3, p. 299-307, 1967. DOI: 10.1109/PGEC.1967.264666. Disponível em: https://doi.org/10.1109/PGEC.1967.264666. Acesso em: 24 jul. 2026.

AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA. Resolução da Diretoria Colegiada - RDC n.º 657, de 24 de março de 2022. Dispõe sobre a regularização de software como dispositivo médico (Software as a Medical Device — SaMD). Diário Oficial da União: seção 1, Brasília, DF, edição 61, 30 mar. 2022. Disponível em: https://anvisalegis.datalegis.net/action/ActionDatalegis.php?acao=abrirTextoAto&tipo=RDC&numeroAto=00000657&seqAto=000&valorAno=2022&orgao=RDC/DC/Anvisa/MS&codTipo=&desItem=&desItemFim=&cod_menu=9434&cod_modulo=310&pesquisa=true. Acesso em: 19 abr. 2026.

AUSTRÁLIA. National Health and Medical Research Council (NHMRC). Guide for assessing research involving AI. Canberra: NHMRC, 2023. Disponível em: https://www.nhmrc.gov.au/about-us/resources/guide-assessing-research-involving-ai. Acesso em: 10 mar. 2026.

BRASIL. Ministério da Justiça e Segurança Pública. Agência Nacional de Proteção de Dados. Relatório de Impacto à Proteção de Dados Pessoais (RIPD). Brasília, DF: ANPD, 23 set. 2026d. Disponível em: https://www.gov.br/anpd/pt-br/canais_atendimento/agente-de-tratamento/relatorio-de-impacto-a-protecao-de-dados-pessoais-ripd. Acesso em: 24 set. 2026.

BARBOSA, V. A. F. *et al.* Covid-19 rapid test by combining a Random Forest-based web system and blood tests. Journal of Biomolecular Structure and Dynamics, v. 40, n. 22, p. 11948-11967, 2021a. DOI: 10.1080/07391102.2021.1966509. Disponível em: https://doi.org/10.1080/07391102.2021.1966509. Acesso em: 24 maio 2026.

BARBOSA, V. A. F. *et al.* Heg.IA: an intelligent system to support diagnosis of covid-19 based on blood tests. **Research on Biomedical Engineering**, v. 38, n. 1, p. 99–116, 2021. DOI: 10.1007/s42600-020-00112-5. Disponível em: https://doi.org/10.1007/s42600-020-00112-5. Acesso em: 24 maio 2026.

BARUCCI, A. *et al.* Ethical and regulatory frameworks for artificial intelligence in clinical research: a european perspective on the artificial intelligence act for ethics committees and researchers. European Cardiology Review, v. 21, e01, Jan. 2026. DOI: 10.15420/ecr.2025.59. Disponível em: https://doi.org/10.15420/ecr.2025.59. Acesso em: 19 abr. 2026.

BEYER, K. W. Grace Hopper and the invention of the information age. Cambridge: MIT Press, 2009.

BORGES, F. T. *et al.* LLM-assisted scoping review of artificial intelligence in brazilian public health: lessons from transfer and federated learning for resource-constrained settings. International Journal of Environmental Research and Public Health, v. 23, n. 1, 81, 2026. DOI:10.3390/ijerph23010081. Disponível em: https://doi.org/10.3390/ijerph23010081. Acesso em: 24 maio 2026.

BRASIL. Autoridade Nacional de Proteção de Dados. Inteligência artificial generativa. Brasília, DF: ANPD, 2024d. (Radar tecnológico, 3). Disponível em: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/radar_tecnologico_ia_generativa_anpd.pdf. Acesso em: 23 mar. 2026.

BRASIL. Instituto Brasileiro de Geografia e Estatística. Estimativas da população residente para os municípios e para as unidades da federação com data de referência em 1.º de julho de 2026. Rio de Janeiro: IBGE, 2026e. Publicado no Diário Oficial da União em 28 ago. 2026. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html. Acesso em: 24 set. 2026.

BRASIL. Ministério da Ciência, Tecnologia e Inovação. IA para o bem de todos: Plano Brasileiro de Inteligência Artificial. Brasília, DF: Ministério da Ciência, Tecnologia e Inovação; Centro de Gestão e Estudos Estratégicos, 2025c. Disponível em: https://www.cgee.org.br/documents/37878/157505/000149+-+CGEE+Plano+Brasileiro+de+Intelige%CC%82ncia+Artificial+%28VN%29+8+Defeso+Eleitoral.pdf/0a05e153-87fb-0ac1-bfa3-af27471204dc?version=2.0&t=1784645088790. Acesso em: 24 set. 2026.

BRASIL. Ministério da Justiça e Segurança Pública. **Guia do usuário brasileiro de inteligência artificial**. Brasília, DF: MJSP, 2026c. Disponível em: https://www.gov.br/mjsp. Acesso em: abr. 2026.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. Carta Circular n.º 1, de 3 de março de 2021. Orienta os CEPs sobre compartilhamento de dados em pesquisas colaborativas. Brasília, DF: CNS, 2021.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. Resolução nº 466, de 12 de dezembro de 2012. Aprova as seguintes diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos; Revoga as (RES. 196/96); (RES. 303/00); (RES. 404/08). Diário Oficial da União: seção 1, Brasília, DF, n. 12, p. 59, 13 jun. 2013. Disponível em: https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2012/resolucao-no-466.pdf/view. Acesso em: 24 jun. 2026.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. Resolução n.º 510, de 7 de abril de 2016. Esta Resolução dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais cujos procedimentos metodológicos envolvam a utilização de dados diretamente obtidos com os participantes ou de informações identificáveis ou que possam acarretar riscos maiores do que os existentes na vida cotidiana, na forma definida nesta Resolução. **Diário Oficial da União**: seção 1, Brasília, DF, n. 98, p. 44-46, 24 maio 2016. Disponível em: https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2016/resolucao-no-510.pdf/view. Acesso em: 19 abr. 2026.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. Resolução n.º 738, de 1.º de fevereiro de 2024. Dispõe sobre o uso de bancos de dados com finalidade de pesquisa científica envolvendo seres humanos. Brasília, DF: CNS, 2024b. Disponível em: https://www.gov.br/conselho-nacional-de-saude/pt-br/camaras-tecnicas-e-comissoes/conep/legislacao/resolucoes/resolucao-no-738-de-01-de-fevereiro-de-2024. Acesso em: 19 abr. 2026.

BRASIL. Ministério da Saúde. Portaria GM/MS n.º 3.232, de 1º de março de 2024. Altera a Portaria de Consolidação GM/MS nº 5, de 28 de setembro de 2017, para instituir o Programa SUS Digital. **Diário Oficial da União**: seção 1, Brasília, DF, edição 43, p. 52, 4 mar. 2024c. Disponível em: https://www.in.gov.br/en/web/dou/-/portaria-gm/ms-n-3.232-de-1-de-marco-de-2024-546278935. Acesso em: 23 mar. 2026.

BRASIL. Ministério da Saúde. **Portaria n.º 1.434, de 28 de maio de 2020**. Institui a Rede Nacional de Dados em Saúde. Brasília, DF: Ministério da Saúde, 2020a. Disponível em: https://www.gov.br/saude/pt-br/composicao/seidigi/rnds. Acesso em: 23 mar. 2026.

BRASIL. Ministério da Saúde. Secretaria de Informação e Saúde Digital. Edital n.º 1/2026 — Chamamento Público do Laboratório Inova SUS Digital. Brasília, DF: Ministério da Saúde/Seidigi, 7 jan. 2026a.

BRASIL. Ministério da Saúde. Secretaria-Executiva. Departamento de Informática do SUS. Estratégia de Saúde Digital para o Brasil 2020–2028. Brasília: Ministério da Saúde Departamento de Informática do SUS, 2020b.

BRASIL. Presidência da República. Decreto n.º 11.358, de 1.º de janeiro de 2023. Aprova a Estrutura Regimental e o Quadro Demonstrativo dos Cargos em Comissão e das Funções de Confiança do Ministério da Saúde e remaneja cargos em comissão e funções de confiança. Diário Oficial da União, Brasília, DF, edição especial. 1º jan. 2023b. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/decreto/d11358.htm. Acesso em: 19 abr. 2026.

BRASIL. Presidência da República. Decreto n.º 11.798, de 28 de novembro de 2023. Aprova a Estrutura Regimental e o Quadro Demonstrativo dos Cargos em Comissão e das Funções de Confiança do Ministério da Saúde e remaneja e transforma cargos em comissão e funções de confiança. Diário Oficial da União: edição extra, Brasília, DF, 28 nov. 2023c. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/decreto/d11798.htm. Acesso em: 24 set. 2026.

BRASIL. Presidência da República. Decreto n.º 12.651, de 7 de outubro de 2025. Regulamenta a Lei nº 14.874, de 28 de maio de 2024, que dispõe sobre a pesquisa com seres humanos e institui o Sistema Nacional de Ética em Pesquisa com Seres Humanos. **Diário Oficial da União**: seção 1, Brasília, DF, edição 192, p. 3, 8 out. 2025a. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/decreto/d12651.htm. Acesso em: 19 jun. 2026.

BRASIL. Presidência da República. **Lei n.º 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União**: seção 1, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 19 abr. 2026.

BRASIL. Presidência da República. Lei n.º 14.510, de 27 de dezembro de 2022. Altera a Lei nº 8.080, de 19 de setembro de 1990, para autorizar e disciplinar a prática da telessaúde em todo o território nacional, e a Lei nº 13.146, de 6 de julho de 2015; e revoga a Lei nº 13.989, de 15 de abril de 2020. Diário Oficial da União: seção 1, Brasília, DF, 28 dez. 2022. Disponível

em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2022/lei/l14510.htm. Acesso em: 19 abr. 2026.

BRASIL. Presidência da República. Lei n.º 14.874, de 28 de maio de 2024. Dispõe sobre a pesquisa com seres humanos e institui o Sistema Nacional de Ética em Pesquisa com Seres Humanos. Diário Oficial da União: seção 1, Brasília, DF, n. 103, p. 6, 29 maio 2024a. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/l14874.htm. Acesso em: 23 mar. 2026.

BRASIL. Presidência da República. Lei n.º 15.211, de 17 de setembro de 2025. Dispõe sobre a proteção de crianças e adolescentes em ambientes digitais (Estatuto Digital da Criança e do Adolescente). Diário Oficial da União: seção 1, Brasília, DF, edição extra, 17 set. 2025b. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/l15211.htm. Acesso em: 29 jun. 2026.

BRASIL. Senado Federal. Projeto de Lei n.º 2.338, de 2023. Dispõe sobre o uso da Inteligência Artificial. Brasília, DF: Senado Federal, 2023a. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 12 abr. 2026.

BURMAN, A. Understanding India's new data protection law. Washington, DC: Carnegie Endowment for International Peace, Oct. 2023. Disponível em: https://carnegieendowment.org/research/2023/10/understanding-indias-new-data-protection-law. Acesso em: 11 mar. 2026.

CANADÁ. Canadian Institutes of Health Research; Natural Sciences and Engineering Research Council of Canada; Social Sciences and Humanities Research Council of Canada. Tri-Council Policy Statement: ethical conduct for research involving humans – TCPS 2 (2022). [S. l.]: Canadian Institutes of Health Research, 2022. Disponível em: https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html. Acesso em: 10 mar. 2026.

CANADÁ. Treasury Board of Canada Secretariat. Algorithmic Impact Assessment (AIA). Ottawa: Treasury Board of Canada Secretariat, 21 Nov. 2024. Disponível em: https://open.canada.ca/data/en/dataset/5423054a-093c-4239-85be-fa0b36ae0b2e. Acesso em: 24 set. 2026.

CHINA. National Medical Products Administration. NMPA announcement on guidance for the classification defining of AI-based medical software products. Beijing: National Medical Products Administration, 8 July 2021. Disponível em: https://english.nmpa.gov.cn/2021-07/08/c_660267.htm. Acesso em: 23 jun. 2026.

CHINA. Comissão Nacional de Saúde (NHC). *Medidas para a revisão ética de pesquisas em ciências da vida e médicas envolvendo seres humanos*. Pequim: NHC, fev. 2023.

CONSELHO FEDERAL DE MEDICINA. Resolução CFM n.º 2.454, de 11 de fevereiro de 2026. Normatiza o uso da inteligência artificial na medicina. Diário Oficial da União: seção 1, Brasília, DF, edição 39, p. 158, 27 fev. 2026. Disponível em: https://sistemas.cfm.org.br/normas/arquivos/resolucoes/BR/2026/2454_2026.pdf. Acesso em: 22 set. 2026.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. Portaria n.º 2.664, de 6 de março de 2026. Institui a Política de Integridade na Atividade Científica e estabelece obrigações sobre o uso de inteligência artificial generativa em pesquisa. Diário Oficial da União: seção 1, Brasília, DF, p. 4, 11 mar. 2026. Disponível em: http://memoria2.cnpq.br/web/guest/view/-/journal_content/56_INSTANCE_0oED/10157/23142775. Acesso em: 13 mar. 2026.

COREIA DO SUL. Ministry of Food and Drug Safety (MFDS). Guidance on Clinical Trials Design of Artificial Intelligence (AI)-based Medical Devices. Chungcheongbuk-do: MFDS, July 2022. Disponível em: https://www.mfds.go.kr/eng/brd/m_40/view.do?seq=72628&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=1. Acesso em: 19 jun. 2026.

COREIA DO SUL. Ministry of Government Legislation. Korean Law Information Center. Framework Act on the Development of Artificial Intelligence and Establishment of Trust. [Enforced January 22, 2026] [Law No. 20676, Enacted January 21, 2025]. Seoul: Korean Law Information Center, 9 July 2025. Disponível em: https://cset.georgetown.edu/publication/south-korea-ai-law-2025/. Acesso em: 11 mar. 2026.

COTA, D. O. C.; JESUS, D. C. S.; ROCHA, A. A. A. Desaprendizado de Maquinas e a LGPD: da privacidade ao direito ao esquecimento. In: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS, 42., 2024, Niterói, RJ. Minicursos […]. Porto Alegre: Sociedade Brasileira de Computação, 2024. Chapter 4, p. 143-185. Disponível em: https://books-sol.sbc.org.br/index.php/sbc/catalog/view/154/661/1202. Acesso em: 24 maio 2026.

DOURADO, D. A.; AITH, F. M. A. A regulação da inteligência artificial na saúde no Brasil começa com a Lei Geral de Proteção de Dados Pessoais. Revista de Saúde Pública, São Paulo, v. 56, 80, 2022. DOI: https://doi.org/10.11606/s1518-8787.2022056004461. Disponível em: https://doi.org/10.11606/s1518-8787.2022056004461. Acesso em: 24 maio 2026.

DURHAM UNIVERSITY. Japanese scientists were pioneers of AI, yet they're being written out of its history. Durham: Durham University, 27 Nov. 2024. Disponível em: https://www.durham.ac.uk/research/current/thought-leadership/2024/11/japanese-scientists-were-pioneers-of-ai-yet-theyre-being-written-out-of-its-history/. Acesso em: 2 abr. 2026.

EUROPEAN DATA PROTECTION BOARD. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. [*S. l.*]: EDPB, 17 Dec. 2024. Disponível em: https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en. Acesso em: 18 abr. 2026.

FIGUEIREDO, K.; AGUIAR, R. A. T. Artificial Intelligence: promises and liabilities. *In*: VIANNA SOBRINHO, L. *et al.* (org.). **Artificial Intelligence and bioethics: perspectives**. Boca Raton: CRC Press, 2025. Cap. 1.

FISKE, A. *et* al. Weighing the benefits and risks of collecting race and ethnicity data in clinical settings for medical artificial intelligence. Lancet Digital Health, v. 7, n. 4, p. e286-e294, Apr. 2025. DOI: 10.1016/j.landig.2025.01.003. Disponível em: https://doi.org/10.1016/j.landig.2025.01.003. Acesso em: 19 abr. 2026.

FORNASIER, M. O. The use of AI in digital health services and privacy regulation in GDPR and LGPD: between revolution and (dis)respect. Revista de Informação Legislativa, Brasília, DF, v. 59, n. 233, p. 201-220, jan./mar. 2022. Disponível em: https://www12.senado.leg.br/ril/edicoes/59/233/ril_v59_n233_p201. Acesso em: 19 jun. 2026.

FUKUSHIMA, K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, v. 36, n. 4, p. 193-202, 1980. Disponível em: https://doi.org/10.1007/BF00344251. Acesso em: 24 jul. 2026.

GEIPING, J. et al. Inverting gradients: how easy is it to break privacy in federated learning? In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 34., 2020, Vancouver, Canada. Proceedings [...]. Red Hook, NY, United States: Curran Associates Inc., 2020. v. 21, p. 16937-16947.

GOOGLE DEEPMIND. **Gemini Ultra**. Version 1.5. [Mountain View]: Google DeepMind, 2026. Ferramenta de inteligência artificial generativa utilizada para elaboração de figura científica. Disponível em: https://deepmind.google/technologies/gemini. Acesso em: abr. 2026.

GRESHAKE, K. *et al*. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. *In*: ACM WORKSHOP ON ARTIFICIAL INTELLIGENCE AND SECURITY (AISec '23), 16., 2023, Copenhagen, Denmark. Proceedings […]. New York: Association for Computing Machinery, 2023. p. 79-90. Disponível em: https://doi.org/10.1145/3605764.3623985. Acesso em: 24 maio 2026.

HADDAD, A. E. *et al.* Evolução da regulação brasileira de telessaúde: do Programa Nacional de Telessaúde ao SUS Digital. Revista de Direito Sanitário, São Paulo, v. 25, n. 1, e0008, 2025. DOI: 10.11606/issn.2316-9044.rdisan.2025.238280. Disponível em: https://doi.org/10.11606/issn.2316-9044.rdisan.2025.238280. Acesso em: 24 maio 2026.

HERN, A. Royal Free breached UK data law in 1.6m patient deal with Google's DeepMind. The Guardian, London, 3 July 2017. Disponível em: https://www.theguardian.com/technology/2017/jul/03/google-deepmind-16m-patient-royal-free-deal-data-protection-act. Acesso em: 19 jun. 2026.

ÍNDIA. Central Drugs Standard Control Organization (CDSCO). *Draft guidance on medical device software*. New Delhi: CDSCO, out. 2025. Disponível em: https://community.nasscom.in/communities/public-policy/cdsco-issues-draft-guidance-medical-device-software. Acesso em: 11 mar. 2026.

INDIAN COUNCIL OF MEDICAL RESEARCH. Ethical guidelines for application of artificial intelligence in biomedical research and healthcare. New Delhi: ICMR, 2023. Disponível em: https://www.icmr.gov.in/ethical-guidelines-for-application-of-artificial-intelligence-in-biomedical-research-and-healthcare. Acesso em: 11 mar. 2026.

ÍNDIA. Ministry of Law and Justice. The Digital Personal Data Protection Act, 2023. The Gazette of India Extraordinary, New Delhi, n. 22, 11 Aug. 2023. Disponível em: https://www.meity.gov.in/static/uploads/2024/06/2bf1f0e9f04e6fb4f8fef35e82c42aa5.pdf. Acesso em: 11 mar. 2026.

INTERNATIONAL COMMITTEE OF MEDICAL JOURNAL EDITORS. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. *[S. l.]:* International Committee of Medical Journal Editors, updated Jan. 2026. Disponível em: https://www.icmje.org/recommendations/. Acesso em: 18 abr. 2026.

JAPÃO. Ministry of Health, Labour and Welfare (MHLW); Ministry of Education, Culture, Sports, Science and Technology (MEXT). *Ethical guidelines for medical and biological research involving human subjects*. Tokyo: MHLW; MEXT, jun. 2021. Disponível em: https://www.hhs.gov/ohrp/international/compilation-human-research-standards/index.html. Acesso em: 11 mar. 2026.

JAPÃO. Pharmaceuticals and Medical Devices Agency (PMDA). Report on AI-based software as a medical device (SaMD). [Tokyo]: PMDA, 8 Aug. 2023. Disponível em: https://www.pmda.go.jp/files/000266100.pdf. Acesso em: 11 mar. 2026.

- LIGHT, J. S. When computers were women. Technology and Culture, v. 40, n. 3, p. 455-483, 1999. DOI: 10.1353/tech.1999.0128. Disponível em: https://doi.org/10.1353/tech.1999.0128. Acesso em: 24 jul. 2026.

- LIMA, C. L. et al. Leveraging Climate Data Through Intelligent Systems for the Prediction of Arbovirus Transmission by Aedes aegypti. International Journal of Environmental Research and Public Health, v. 23, n. 1, p. 12, 2026. DOI: 10.3390/ijerph23010012. Disponível em: https://doi.org/10.3390/ijerph23010012. Acesso em: 24 maio 2026.

LUNKES, A. L. Z.; LUNKES, A. F. Z.; OLIVEIRA, F. B. Aplicação da criptografia homomórfica na análise de dados da dengue. Caderno Pedagógico, [*S. l.*], v. 21, n. 8, p. e7263, 2024. DOI: 10.54033/cadpedv21n8-278. Disponível em: https://ojs.studiespublicacoes.com.br/ojs/index.php/cadped/article/view/7263. Acesso em: 24 maio 2026.

MARIOTTO, C. Brazilian legal framework on automated decision-making. International Bar Association, 9 June 2024. Disponível em: https://www.ibanet.org/Brazilian-legal-framework-on-automated-decision-making. Acesso em: 23 mar. 2026.

MATSUSHITA, P. M. B.; ALMEIDA, C. L. G. Regulation for data protection and privacy (in health). *In*: VIANNA SOBRINHO, L. *et al.* (org.). **Artificial Intelligence and bioethics: perspectives**. Boca Raton: CRC Press, 2025. Cap. 5.

MRCT CENTER; WCG ARTIFICIAL INTELLIGENCE TASK FORCE. Framework for review of clinical research involving artificial intelligence. [*S. l.*]: MRCT Center; WCG, June 2025. Disponível em: https://mrctcenter.org/resource/framework-for-review-of-clinical-research-involving-ai/. Acesso em: 23 mar. 2026.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Genebra: OMS, 2024. Disponível em: https://www.who.int/publications/i/item/9789240084759. Acesso em: 23 mar. 2026.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. Ethics and governance of artificial intelligence for health: WHO guidance. Genebra: World Health Organization, 2021. Disponível em: https://www.who.int/publications/i/item/9789240029200. Acesso em: 23 mar. 2026.

ORGANIZACIÓN PANAMERICANA DE LA SALUD; ORGANIZACIÓN MUNDIAL DE LA SALUD. Inteligencia Artificial en Salud Pública: Kit de herramientas de evaluación de la preparación. Washington, D.C.: OPS/OMS, 2024. Disponível em: https://www.paho.org/es/documentos/inteligencia-artificial-salud-publica-kit-herramientas-evaluacion-preparacion. Acesso em: 23 mar. 2026.

OWASP FOUNDATION. OWASP Top 10 for LLM Applications 2025. Version 2025. [*S. l.*]: OWASP Foundation, 18 Nov. 2024. Disponível em: https://genai.owasp.org/llm-top-10/. Acesso em: 24 maio 2026.

POWLES, J.; HODSON, H. Google DeepMind and healthcare in an age of algorithms. Health and Technology, v. 7, n. 4, p. 351-367, 2017. DOI: 10.1007/s12553-017-0179-1. Disponível em: https://doi.org/10.1007/s12553-017-0179-1. Acesso em: 24 maio 2026.

RAJKOMAR, A. *et al*. Ensuring fairness in machine learning to advance health equity. **Annals of Internal Medicine**, v. 169, n. 12, p. 866-872, 2018. DOI: 10.7326/M18-1990. Disponível em: https://doi.org/10.7326/M18-1990. Acesso em: 23 mar. 2026.

RIEKE, N. *et al.* The future of digital health with federated learning. NPJ Digital Medicine, v. 3, 119, 2020. DOI: https://doi.org/10.1038/s41746-020-00323-1. Disponível em: https://doi.org/10.1038/s41746-020-00323-1. Acesso em: 24 maio 2026.

ROCHER, L.; HENDRICKX, J. M.; MONTJOYE, Y. A. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications, v. 10, 3069, 2019. DOI: https://doi.org/10.1038/s41467-019-10933-2. Disponível em: https://doi.org/10.1038/s41467-019-10933-2. Acesso em: 24 maio 2026.

RÚSSIA. Lei Federal n.º 123-FZ, de 24 de abril de 2020. Regime legal experimental em inovação digital no município de Moscou. Moscou: Duma Federal, 2020. Disponível em: https://cis-legislation.com/document.fwx?rgn=124089. Acesso em: 11 mar. 2026.

RÚSSIA. Lei Federal n.º 152-FZ, de 27 de julho de 2006. Sobre dados pessoais. Moscou: Duma Federal, 2006.

SALIH, A. *et al.* Explainable artificial intelligence and cardiac imaging: toward more interpretable models. Circulation: Cardiovascular Imaging, v. 16, n. 4, e014519, Apr. 2023. DOI:10.1161/CIRCIMAGING.122.014519.

SANTANA, M. A.; MORENO, G. M. M.; SANTOS, W. P. Explainable artificial intelligence. *In*: VIANNA SOBRINHO, L. *et al.* (org.). **Artificial Intelligence and bioethics: perspectives**. Boca Raton: CRC Press, 2025. Cap. 8.

SANTOS, W. P. *et al*. Nations of the Global South: pioneering the future of health with artificial intelligence and digital and health sovereignty. *In*: VIANNA SOBRINHO, L. *et al.* (org.). **Artificial Intelligence and bioethics: perspectives**. Boca Raton: CRC Press, 2025. Cap. 3.

SOARES, M. Q.; CHIAVEGATTO FILHO, A. D. P. Inteligência artificial na saúde pública brasileira: potencialidades, desafios e implicações éticas para o Sistema Único de Saúde. Revista Brasileira de Epidemiologia, São Paulo, v. 29, e260006, 2026. DOI: https://doi.org/10.1590/1980-549720260006. Disponível em: https://doi.org/10.1590/1980-549720260006. Acesso em: 19 abr. 2026.

UNESCO. Recomendação sobre a Ética da Inteligência Artificial. Aprovada em 23 de novembro de 2021. Paris, FR: UNESCO, 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137_por. Acesso em: 12 jul. 2026.

UNIÃO EUROPEIA. Parlamento Europeu; Conselho da União Europeia. Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados e que revoga a Diretiva 95/46/CE (Regulamento Geral sobre a Proteção de Dados) (Texto relevante para efeitos do EEE). Jornal Oficial da União Europeia, L 119, p. 1–88, 4 maio 2016. Disponível em: https://eur-lex.europa.eu/eli/reg/2016/679/oj. Acesso em: 14 abr. 2026.

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que cria regras harmonizadas em matéria de inteligência artificial e que altera os Regulamentos (CE) n.° 300/2008, (UE) n.° 167/2013, (UE) n.° 168/2013, (UE) 2018/858, (UE) 2018/1139 e (UE) 2019/2144 e as Diretivas 2014/90/UE, (UE) 2016/797 e (UE) 2020/1828 (Regulamento da Inteligência Artificial) (Texto relevante para efeitos do EEE). **Jornal Oficial da União Europeia**, série L, 12 jul. 2024. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=OJ:L_202401689. Acesso em: 14 abr. 2026.

REINO UNIDO. Department of Health & Social Care. National Commission into the Regulation of AI in Healthcare. National Commission into the Regulation of AI in Healthcare: Recommendations for a future regulatory framework. [London]: ©Crown, 10 Sep. 2026. Disponível em: https://www.gov.uk/government/publications/national-commission-into-the-regulation-of-ai-in-healthcare-recommendations-for-a-future-regulatory-framework/national-commission-into-the-regulation-of-ai-in-healthcare-recommendations-for-a-future-regulatory-framework. Acesso em: 10 set. 2026.

ESTADOS UNIDOS. Food and Drug Administration (FDA). Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: draft guidance for industry and other interested parties. Rockville, MD: FDA, Jan. 2025. Disponível em: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological. Acesso em: 10 mar. 2026.

VASEY, B. *et al.* Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature Medicine, v. 28, n. 5, p. 924–933, May 2022. DOI: https://doi.org/10.1038/s41591-022-01772-9, Publicado simultaneamente

no BMJ, v. 377, e070904, 2022. DOI: 10.1038/s41591-022-01772-9. Disponível em: https://doi.org/10.1038/s41591-022-01772-9. Acesso em: 24 jul. 2026.

VASWANI, A. *et al*. Attention is all you need. arXiv:1706.03762 [cs.CL], 2017. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 2 abr. 2026.

WILTON PARK. Leveraging digital public infrastructure for sustainable health system transformation. [*S. l.*]: Wilton Park, Feb. 2026. Disponível em: https://www.wiltonpark.org.uk/reports/leveraging-digital-public-infrastructure-for-sustainable-health-system-transformation/. Acesso em: 25 mar. 2026.

YOUSSEF, A. *et al.* Ethical considerations in the design and conduct of clinical trials of artificial intelligence. JAMA Network Open, v. 7, n. 9, e2432482, Sep. 2024. DOI: 10.1001/jamanetworkopen.2024.32482. Disponível em: https://doi.org/10.1001/jamanetworkopen.2024.32482. Acesso em: 24 maio 2026.

## LEITURAS RECOMENDADAS


ALBUQUERQUE, A.; REGO, S. Parecer ético n.º 1/2026, Resolução CFM n.º 2.454, de 11 de fevereiro de 2026, normatiza o uso da inteligência artificial na medicina. [*S. l.*]: Sociedade Brasileira de Bioética, 2026. 18 p. Disponível em: https://static.galoa.com.br/file/Assoc-Private/attendee_dashboard_page/pdf/Parecer%20SBB%20Resolu%C3%A7%C3%A3o%20do%20CFM%20sobre%20Intelig%C3%AAncia%20Artificial%20Final.pdf. Acesso em: 11 set. 2026.

BOLLOS, R. H. IA na medicina e a resolução do CFM: avanços e zonas cinzentas. Nexo Políticas Públicas, São Paulo, 11 mar. 2026. Ponto de Vista. Disponível em: https://pp.nexojornal.com.br/ponto-de-vista/2026/03/11/ia-na-medicina-e-a-resolucao-do-cfm-avancos-e-zonas-cinzentas. Acesso em: 26 jul. 2026.

BONAWITZ, K. et al. Practical secure aggregation for privacy-preserving machine learning. In: ACM SIGSAC CONFERENCE ON COMPUTER AND COMMUNICATIONS SECURITY, 17., 2017, Dallas, Texas, USA. Proceedings […]. New York: Association for Computing Machinery, 2017. p. 1175–1191. DOI: https://doi.org/10.1145/3133956.3133982.

BRASIL. Conselho Nacional de Desenvolvimento Científico e Tecnológico. CNPq institui Política de Integridade na Atividade Cientifica, que estabelece normas e boas práticas de atuação. Brasília, DF: CNPq, 11 mar. 2026. Disponível em: https://www.gov.br/cnpq/pt-br/assuntos/noticias/cnpq-em-acao/cnpq-publica-portaria-que-institui-politica-de-integridade-na-atividade-cientifica. Acesso em: 13 mar. 2026.

BRASIL. Ministério da Justiça e Segurança Pública. Guia do uso ético de inteligência artificial: sociedade civil e poder público ampliam contribuições no MJSP. Brasília, DF: Ministério da Justiça e Segurança Pública, 10 abr. 2026b. Disponível em: https://www.gov.br/mj/pt-br/assuntos/noticias/guia-do-uso-etico-de-inteligencia-artificial-sociedade-civil-e-poder-publico-ampliam-contribuicoes-no-mjsp. Acesso em: 12 abr. 2026.

BRASIL. Presidência da República. Lei n.º 8.080, de 19 de setembro de 1990. Dispõe sobre as condições para a promoção, proteção e recuperação da saúde, a organização e o funcionamento dos serviços correspondentes e dá outras providências. Diário Oficial da União: seção 1, Brasília, DF, 20 set. 1990. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8080.htm. Acesso em: 29 jun. 2026.

BRASIL. Presidência da República. Lei n.º 15.378, de 6 de abril de 2026. Institui o Estatuto dos Direitos do Paciente. Diário Oficial da União, Brasília, DF, 7 abr. 2026. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2026/lei/l15378.htm. Acesso em: 26 jul. 2026.

COHEN, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, v. 20, n. 1, p. 37–46, 1960. DOI: https://doi.org/10.1177/001316446002000010.

COMITÊ GESTOR DA INTERNET NO BRASIL; NÚCLEO DE INFORMAÇÃO E COORDENAÇÃO DO PONTO BR. Estudos setoriais: inteligência artificial na saúde. São Paulo: CGI.br; NIC.br, 2024.

CORRÊA, N. K. *et al.* Worldwide AI ethics: a review of 200 guidelines and recommendations for AI governance. Patterns, New York, N. Y., v. 4, n. 10, art. 100857, 13 Oct. 2023. DOI: 10.1016/j.patter.2023.100857. Disponível em: https://doi.org/10.1016/j.patter.2023.100857. Acesso em: 18 jun. 2026.

CORRÊA, N. K. *et al.* Crossing the principle–practice gap in AI ethics with ethical problem-solving. AI and Ethics, v. 5, p. 1271-1288, 2025. DOI: https://doi.org/10.1007/s43681-024-00469-8. Disponível em: https://doi.org/10.1007/s43681-024-00469-8. Acesso em: 18 jun. 2026.

DWORK, C.; ROTH, A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, v. 9, n. 3-4, p. 211-407, 2014. DOI: https://doi.org/10.1561/0400000042.

ESCOLA NACIONAL DE SAÚDE PÚBLICA SÉRGIO AROUCA. Portaria n.º 69, de 1.º de junho de 2026. Disciplinar o uso de ferramentas de inteligência artificial e de inteligência artificial generativa e estabelece diretrizes para a verificação de similaridade e de integridade acadêmica no âmbito da ENSP/Fiocruz. Rio de Janeiro: ENSP/Fiocruz, 2026. 8 p. Processo n.º 25388.000033/2026-21, SEI n. 6277864. Disponível em: https://informe.ensp.fiocruz.br/assets/anexos/5ed838d9f860fa3f9f09e181d68d5069.PDF. Acesso em: 11 set. 2026.

HADDAD, A. E.; LIMA, N. T. Saúde Digital no Sistema Único de Saúde (SUS). Interface – Comunicação, Saúde, Educação, Botucatu, v. 28, e230597, 2024. DOI: https://doi.org/10.1590/interface.230597. Disponível em: https://doi.org/10.1590/interface.230597. Acesso em: 9 abr. 2026.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. Biometrics, v. 33, n. 1, p. 159–174, 1977.

LOBATO, A. V. C. A China e a zona de alcance. 2025. 151 f. Tese (Doutorado em Políticas Públicas, Estratégias e Desenvolvimento) – Instituto de Economia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2025. Disponível em: https://www.ie.ufrj.br/images/IE/PPED/Teses/2025/A%20China%20e%20a%20Zona%20de%20Alcance.pdf. Acesso em: 22 set. 2026.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. Nature, v. 323, n. 6088, p. 533–536, 9 Oct. 1986. Disponível em: https://doi.org/10.1038/323533a0. Acesso em: 2 abr. 2026.

RÚSSIA. Ministério da Saúde. Ordem n.º 686n: alteração das regras de classificação de risco para softwares médicos. Moscou: Ministério da Saúde da Federação Russa, 2020. Disponível em: https://medicaldevicesinrussia.com/2020/09/. Acesso em: 11 mar. 2026.

SILVA, A. B. et al. Diagnostic evaluation of institutions as a basis for designing the Brazilian maturity model of telehealth services. BMC Health Services Research, v. 24, n. 1, p. 372, 2024. DOI: 10.1186/s12913-024-10723-8. Disponível em: https://doi.org/10.1186/s12913-024-10723-8. Acesso em: 9 abr. 2026.

UNESCO. Recomendação sobre a Ética da Inteligência Artificial. Aprovada em 23 de novembro de 2021. Paris, FR: UNESCO, 2022. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000381137_por. Acesso em: 12 jul. 2026.

## REFERÊNCIAS DA ATUALIZAÇÃO REGULATÓRIA 2024–2026


EUROPEAN DATA PROTECTION BOARD. Guidelines 02/2026 on Anonymisation. Version 1.0, adopted on 07 July 2026, adoption of the guidelines for public consultation. [*S. l.*]: EDPB, 2026. Disponível em: https://www.edpb.europa.eu/public-consultations/guidelines-022026-on-anonymisation_en#no-back. Acesso em: 22 set. 2026.

EUROPEAN DATA PROTECTION BOARD. Guidelines 1/2026 on processing of personal data for scientific research purposes. Adopted on 15 April 2026. Version submitted for public consultation. [*S. l.*]: EDPB, 2026. Disponível em: https://www.edpb.europa.eu/public-consultations/guidelines-12026-on-processing-of-personal-data-for-scientific-research_en. Acesso em: 22 set. 2026.

EUROPEAN MEDICINES AGENCY. Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023. Amsterdam: EMA, 9 Sep. 2024. Disponível em: https://www.ema.europa.eu/system/files/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle-en.pdf. Acesso em: 22 set. 2026.

EUROPEAN MEDICINES AGENCY; U.S. FOOD AND DRUG ADMINISTRATION. Guiding principles of good AI practice in drug development. [*S. l.*]: EMA; FDA Jan. 2026. Disponível em: https://www.ema.europa.eu/en/documents/other/guiding-principles-good-ai-practice-drug-development_en.pdf. Acesso em: 22 set. 2026.

EUROPEAN PARLIAMENT; COUNCIL OF THE EUROPEAN UNION. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance. Official Journal of the European Union, L series, 12 July 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Acesso em: 22 set. 2026.

EUROPEAN PARLIAMENT; COUNCIL OF THE EUROPEAN UNION. Regulation (EU) 2025/327 of the European Parliament and of the Council of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847 (Text with EEA relevance). Official Journal of the European Union, L series, 5 Mar. 2025. Disponível em: https://eur-lex.europa.eu/eli/reg/2025/327/oj/eng. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/engAcesso em: 22 set. 2026.

MALTBY, J.; LIGUORI, L.; STEFANINI, E. Navigating Europe's AI regulatory labyrinth: A strategic guide for US pharma innovators. [S. l.]: Pharmaphorum, 31 Aug., 2026. [Fonte secundária utilizada para identificação de temas; os fundamentos normativos foram conferidos em fontes oficiais]. Disponível em: https://pharmaphorum.com/market-access/navigating-europes-ai-regulatory-labyrinth-strategic-guide-us-pharma-innovators. Acesso em: 22 set. 2026.

## APÊNDICE A — MAPA DE CORRESPONDÊNCIA COM O GUIA E A MARIAH

Este mapa orienta futuras referências cruzadas editoriais. A numeração definitiva poderá ser atualizada após o fechamento da edição.

- Fundamentos éticos e perguntas centrais — sustentam escopo, proporcionalidade e critérios gerais do Guia e os eixos iniciais da MARIAH.
- Conceitos e propriedades críticas — fundamentam contexto de uso, explicabilidade, supervisão, viés, ciclo de vida e validação.
- Contexto brasileiro e panorama internacional — fundamentam a adaptação do instrumento ao Brasil e a metodologia comparada.
- Integridade científica e autoria — fundamentam as recomendações sobre declaração de uso de IA, verificação humana e responsabilidade.
- Governança de dados — fundamenta origem, finalidade, compartilhamento, reidentificação, transferência e destino dos dados e modelos.
- Saúde e SaMD — fundamentam recomendações e fatores de risco específicos das Ciências da Saúde.
- Ciências Humanas e Sociais — fundamentam inferências sensíveis, consentimento, dados públicos, risco coletivo, estigmatização e usos secundários.